Minimax Optimal Conditional Independence Testing

Matey Neykov   Sivaraman Balakrishnan   Larry Wasserman

Department of Statistics & Data Science
Carnegie Mellon University
Pittsburgh, PA 15213

{mneykov, siva, larry}@stat.cmu.edu

Abstract

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 [34], one cannot hope to design non-trivial 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 [34].

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 [14], to its applications in estimation and inference for graphical models [26, 28] and in causal inference and causal discovery [30, 36, 46], 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 [14], 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 [34] 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 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 [34], the work of Canonne et al. [13] 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. [13] 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 [23, 25]. 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, i.e. 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 non-trivial test which controls the type I error for *all* conditionally independent absolutely continuous triplets [34]. 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. [13] 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. [13] are instrumental in the development of the minimax rates in the present work. In particular [13] 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 [35] 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 [22] 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 [20], extend the independence testing results of Gretton and Györfi [19] 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 [34]. In particular, in the proof of Theorem 1 of Györfi and Walk [20], 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 et al. [29] 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. [32]. Fukumizu et al. [18] 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. [46] and Doran et al. [17] respectively. The recent work of Shah and Peters [34], 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 [37, 38, 39, 43]). Su and White [38] 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 [37] 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 [39] the authors propose an empirical likelihood based approach to CI testing. Wang and Hong [43] 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 [37].

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, 13, 31, 45] 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,

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

In addition, we remove distributions under H_0 and H_1 which are not smooth enough, i.e., $p_{X,Y|Z=z}$ 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 pre-specified 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.

4

	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	$n^{-2/5}$	$\frac{(\ell_1\ell_2)^{1/5}}{n^{2/5}}$	$n^{-2s/(5s+2)}$

Table 1: This is a summary of the minimax results obtained in the main text of our paper. In addition, in the appendix we provide extensions to the case when $Z \in [0,1]^{d_Z}$ and X, Y are either continuous or discrete with fixed many categories. For more details see Appendix G.

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 [34]. 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 [34]. 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 [34] 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

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

where the last identity assumes ν is a common dominating measure of p and q, i.e., $p \ll \nu \; q \ll \nu$ 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 co-authors [23, 25], and which has since then been considered by many authors (see for instance [2, 4, 5, 6, 11, 13, 16, 40]). Formally, consider the testing problem

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

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 pre-specified sets of distributions. We define the minimax risk of testing as

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

where the infimum is taken over all Borel measurable test functions $\psi : \mathrm{supp}(\mathcal{D}_n) \mapsto [0, 1]$ (which gives the probability of rejecting the null hypothesis), and $\mathrm{supp}(\mathcal{D}_n)$ is the support of the random

---

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

variables $\mathcal{D}_n = \{(X_1, Y_1, Z_1), \ldots (X_n, Y_n, Z_n)\}$. 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

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

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 [34], and our own results in Section 3, we must impose some restrictions on the distributions under consideration in order to obtain non-trivial 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:

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:

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

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]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 $[\ell_1] \times [\ell_2]$.

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:*

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

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:*

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

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*

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

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

In 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 smoothnes 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*

$$\sup_{k \leq \lfloor s \rfloor} \left| \frac{\partial^k}{\partial x^k} \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} f(x,y) - \frac{\partial^k}{\partial x^k} \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} f(x', y') \right| \leq L((x - x')^2 + (y - y')^2)^{\frac{s - \lfloor s \rfloor}{2}}, \qquad (2.4)$$

and in addition

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

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 [34]. 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, i.e., 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 pth 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, i.e., 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*
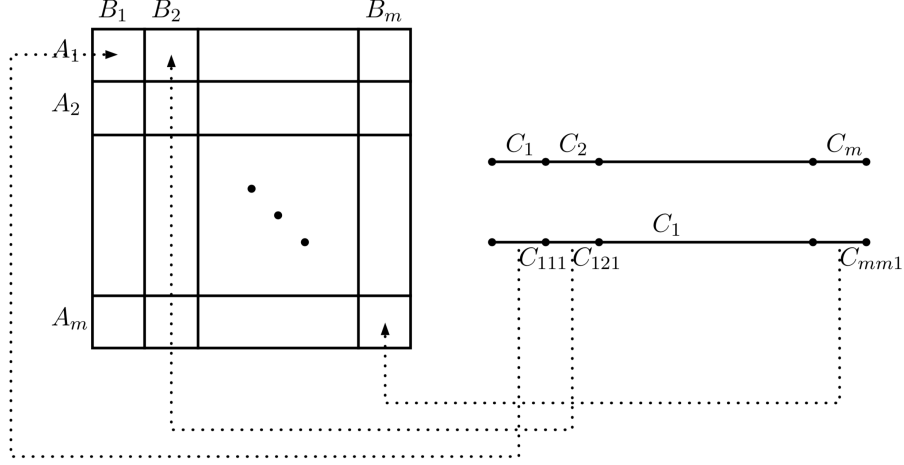
$$W_p(P, Q) \leq \varepsilon.$$

Figure 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 sub-divided into m^2 smaller sub-intervals. The interval C_1 is displayed along with its sub-divisions 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 equi-partition of $[-M, M]$ in intervals. Similarly let $\{B_1, \ldots, B_m\}$ and $\{C_1, \ldots, C_m\}$ be equi-partitions of $[-M, M]$. Divide each C_k further in m^2 sub-intervals 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. $\qquad \square$

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

---

[3]For a precise expression of the density of Q refer to Appendix A.

$\mathcal{E}_{0,M}$ 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 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 [34] 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 [34, see Theorem 2]. For convenience of the reader we restate the no-free-lunch theorem below, and give a complete proof in Appendix A. 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 [34].

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*

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

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, 23] (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 [23, 40], 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. We note that at a high-level we follow the strategy of Ingster [25] 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 4, 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.

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 re-introduce certain unbiased estimators from the work of Canonne et al. [13]. Our exposition and treatment of their estimators is novel, and builds on classical work on U-statistics [21, 33].

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

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

---

[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.

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

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

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 in equation (18) of Canonne et al. [13]. However, using Proposition 4.2 of [13] 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 [13]. 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*

$$\mathrm{Var}[U(\mathcal{D})] \leq C\left(\frac{\mathbb{E}[U(\mathcal{D})]\max(\|p_{X',Y'}\|_2, \|p_{X'}p_{Y'}\|_2)}{\sigma} \right.$$
$$\left. + \frac{\max(\|p_{X',Y'}\|_2^2, \|p_{X'}p_{Y'}\|_2^2)}{\sigma^2}\right).$$

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 $N \sim Poi(\frac{n}{2})$. If $N > n$ accept the null hypothesis. If $N \leq n$ 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 $\{C_1, \ldots, C_d\}$, so that $\cup_{i\in[d]} C_i = [0, 1]$, 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 $\sigma_m \geq 4$ 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. [13] and consider the following test statistic

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

14

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

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

where $\mathcal{D}_N = \{(X_1, Y_1, Z_1), \ldots, (X_N, Y_N, Z_N)\}$. 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 $\varepsilon \geq cn^{-2/5}$, for a sufficiently large constant c (depending on ζ, L, ℓ_1, ℓ_2). Then we have that*

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

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

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.

- 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, 16, 40]). 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

15

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 [41], 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, e.g., [13, 16]. 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 $h_{ijkl}^{\boldsymbol{a}}$, we construct the (weighted) U-statistic

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

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 the notation we now give a definition from [16].

Definition 5.3 (Split Distribution). *Given a discrete distribution p over $[d_1] \times [d_2]$ and a multi-set 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 multi-set A by adding a_{xy} occurrences of the pair (x,y) to A. Using this notation it now follows that

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

where $p_{X',Y',A}$ is the A-split distribution $p_{X',Y'}$, and $p_{X',Y',A}^\Pi$ is the A-split distribution $p_{X',Y'}^\Pi$ where $p_{X',Y'}^\Pi = p_{X'}p_{Y'}$. 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*

$$\mathrm{Var}[U_W(\mathcal{D})|\mathcal{D}_{X'}, \mathcal{D}_{Y'}] \leq C\bigg(\frac{\mathbb{E}[U_W(\mathcal{D})|\mathcal{D}_{X'}, \mathcal{D}_{Y'}]\|p_{X',Y',A}^\Pi\|_2}{\sigma}$$
$$+ \frac{\mathbb{E}[U_W(\mathcal{D})|\mathcal{D}_{X'}, \mathcal{D}_{Y'}]^{3/2}}{\sigma} + \frac{\|p_{X',Y',A}^\Pi\|_2^2}{\sigma^2} + \frac{\mathbb{E}[U_W(\mathcal{D})|\mathcal{D}_{X'}, \mathcal{D}_{Y'}]}{\sigma^2}\bigg).$$

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 $p_{X',Y',A}^{\Pi}$, instead of the L_2 norm of the original distribution $p_{X',Y'}$. As emphasized in [13, 16] 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 $N \sim Poi(\frac{n}{2})$ 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 $\cup_{i \in [d]} C_i = [0, 1]$, 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 $\sigma_m \geq 4$ 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

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

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

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

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*

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

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

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 $\left(\frac{n}{\ell_1}\right)^{2/5} \lesssim \frac{n}{\ell_1}$ (for our choice of d) which is implied by

18

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). We in turn show that this discretization error due to binning is appropriately small when the null distribution satisfies the χ^2 Lipschitzness condition, i.e. belongs to $\mathcal{P}'_{0,[0,1],\chi^2}(L)$.

- As we detail further in Appendix C, 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, 40]. 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 $N \sim Poi(\frac{n}{2})$, 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 $\cup_{i \in [d]} C_i = [0,1]$. 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 $\cup_{i \in [d']} C'_i = [0,1]$.

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*

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

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

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 (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).

- In addition, in Appendix G we derive another upper bound for the case where $Z \in [0,1]^{d_Z}$ and $s \geq 1$. It turns out that the minimax rate in this case is $n^{-2s/((4+d_Z)s+2)}$, which generalizes the above result (for the case $s \geq 1$). In addition we derive a matching lower bound.

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*

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

By the T2 Lemma it is simple to see that

$$d_{\chi^2}(p_{X|Z=z}, p_{X|Z=z'}) = \sum_x \frac{(p_{X|Z}(x|z) - p_{X|Z}(x|z'))^2}{p_{X|Z}(x|z')}$$

$$\geq \frac{(\sum_x |p_{X|Z}(x|z) - p_{X|Z}(x|z')|)^2}{\sum_x p_{X|Z}(x|z')} = \|p_{X|Z=z} - p_{X|Z=z'}\|_1^2.$$

Hence we have that

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

and therefore we obtain the inclusion in (6.1). To derive the second inclusion note that when $z, z' \in [0,1]$ we have $|z - z'| \leq \sqrt{|z - z'|}$, and therefore

$$\|p_{X|Z=z} - p_{X|Z=z'}\|_1 \leq \sqrt{L}|z - z'| \implies \|p_{X|Z=z} - p_{X|Z=z'}\|_1 \leq \sqrt{L|z - z'|}.$$

$\square$

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 $z, z' \in [0,1]$:*

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

Then

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

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

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

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 sub-additive on product distributions [40] so that

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

$$\square$$

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

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

Proof. We start by showing the first inclusion. Note that by the T2 Lemma

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

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 $p_{X,Y,Z} \in \mathcal{P}'_{0,[0,1],\chi^2}(L)$. $\square$

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 $|z - z'| \le 1$:

$$\frac{p_{X|Z}(x|z)}{p_{X|Z}(x|z')} - 1 \le \exp(L|z - z'|) - 1 = L|z - z'| + \sum_{k \ge 2} (L|z - z'|)^k/k!$$

$$\le L|z - z'| + L|z - z'| \sum_{k \ge 2} L^{k-1}/k! = L|z - z'| + L|z - z'|(e^L - 1 - L)/L$$

$$= (e^L - 1)|z - z'|.$$

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

$$\sum_x |p_{X|Z}(x|z) - p_{X|Z}(x|z')|$$

$$= \sum_x \left(\frac{\max(p_{X|Z}(x|z), p_{X|Z}(x|z'))}{\min(p_{X|Z}(x|z), p_{X|Z}(x|z'))} - 1 \right) \min(p_{X|Z}(x|z), p_{X|Z}(x|z'))$$

$$\le \sum_x \left(\frac{\max(p_{X|Z}(x|z), p_{X|Z}(x|z'))}{\min(p_{X|Z}(x|z), p_{X|Z}(x|z'))} - 1 \right) p_{X|Z}(x|z).$$

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, i.e.*

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

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 appendix. 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

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

Next we use Jensen's inequality and the fact that $-\log$ is a convex function to show that

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

Putting things together we get

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

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. $\square$

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, i.e.,* $|g(x,y,z) - g(x',y',z')| \le L(|x - x'| + |y - y'| + |z - z'|)$. *Take* $p_{X,Y,Z}(x,y,z) \propto \exp(g(x,y,z))$. *Then*

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

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

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

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

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

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

We conclude that

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

which is our desired result. $\qquad \square$

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 [46]). In more details, 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 $\{T_1, \ldots, T_M\}$ 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

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

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

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

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

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

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

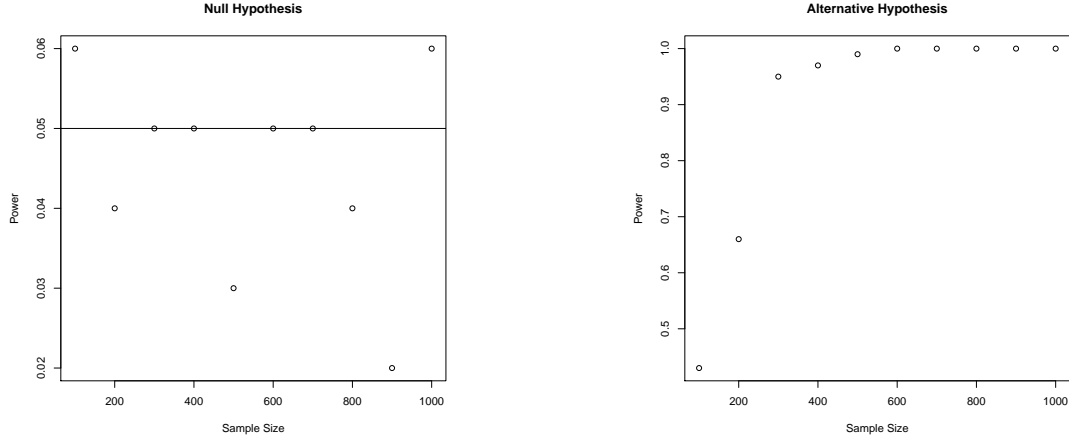$$X = \frac{U_1 + U + Z}{3} \text{ and } Y = \frac{U_2 + U + Z}{3},$$

Figure 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.

where $U, U_1, U_2, Z \sim U([0,1])$ are independent. A straightforward calculation (see Appendix F) 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.

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 [34]. 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
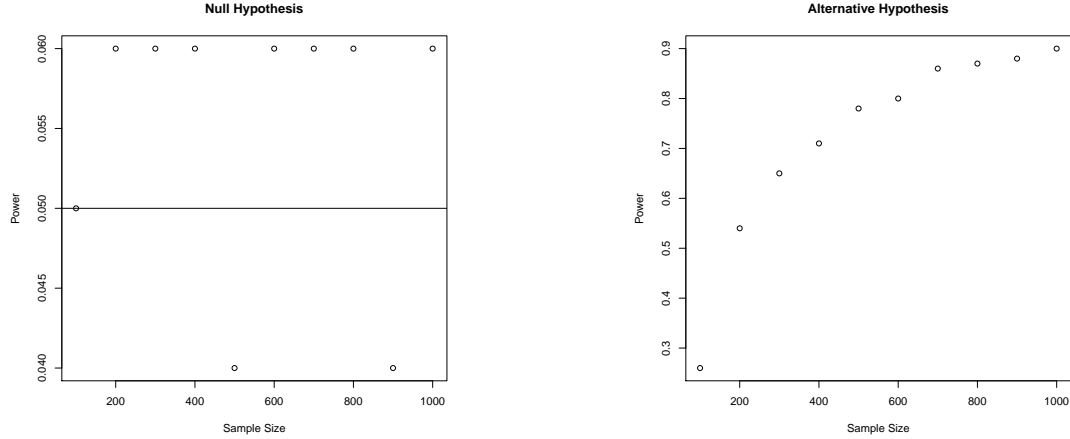
Figure 3: This figure displays the size and power of the test in the continuous X, Y, Z example. We see that under the null hypothesis the size is very slightly inflated at 0.06 for most of the simulations, which may be due to the limited number of replications of each simulation and also due to the 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.

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.

References

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

[2] Ery Arias-Castro, Bruno Pelletier, and Venkatesh Saligrama. Remember the curse of dimensionality: the case of goodness-of-fit testing in arbitrary dimension. *Journal of Nonparametric Statistics*, 30(2):448–471, 2018.

[3] Mona Azadkia and Sourav Chatterjee. A simple measure of conditional dependence. *arXiv preprint arXiv:1910.12327*, 2019.

[4] Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for high-dimensional multinomials: A selective review. *Ann. Appl. Stat.*, 12(2):727–749, 06 2018.

[5] Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *Ann. Statist.*, 47(4):1893–1927, 08 2019.

[6] Yannick Baraud. Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8 (5):577–606, 2002.

[7] Wicher Bergsma. Nonparametric testing of conditional independence by means of the partial copula. *Available at SSRN 1702981*, 2010.

[8] Wicher Pieter Bergsma. *Testing conditional independence for continuous random variables.* Eurandom, 2004.

[9] Thomas B Berrett, Yi Wang, Rina Foygel Barber, and Richard J Samworth. The conditional permutation test. *arXiv preprint arXiv:1807.05405*, 2018.

[10] Eric Blais, Clément L Canonne, and Tom Gur. Distribution testing lower bounds via reductions from communication complexity. *ACM Transactions on Computation Theory (TOCT)*, 11(2): 1–37, 2019.

[11] Clément L Canonne. A survey on distribution testing: your data is big. but is it blue? In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 22, pages 1–1, 2015.

[12] Clément L Canonne. A short note on poisson tail bounds, 2017.

[13] Clément L Canonne, Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Testing conditional independence of discrete distributions. In *Symposium on the Theory of Computing*, pages 1–57, 2018.

[14] Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15, 1979.

[15] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[16] Ilias Diakonikolas and Daniel M Kane. A new approach for testing properties of discrete distributions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 685–694. IEEE, 2016.

[17] Gary Doran, Krikamol Muandet, Kun Zhang, and Bernhard Schölkopf. A permutation-based kernel conditional independence test. In *UAI*, pages 132–141, 2014.

[18] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in neural information processing systems*, pages 489–496, 2008.

[19] Arthur Gretton and László Györfi. Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11(Apr):1391–1423, 2010.

[20] László Györfi and Harro Walk. Strongly consistent nonparametric tests of conditional independence. *Statistics & Probability Letters*, 82(6):1145–1150, 2012.

[21] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.*, 19(3):293–325, 09 1948.

[22] Tzee-Ming Huang. Testing conditional independence using maximal nonlinear conditional correlation. *The Annals of Statistics*, 38(4):2047–2091, 2010.

[23] J.I. Ingster and I.A. Suslina. *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*. Lecture Notes in Statistics. Springer, 2003.

[24] Yuri Ingster and Irina A Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169. Springer Science & Business Media, 2012.

[25] Yuri Izmailovich Ingster. On the minimax nonparametric detection of signals in white gaussian noise. *Problemy Peredachi Informatsii*, 18(2):61–73, 1982.

[26] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[27] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

[28] Dimitris Margaritis. Distribution-free learning of bayesian network structure in continuous domains. In *AAAI*, volume 5, pages 825–830, 2005.

[29] Rohit K Patra, Bodhisattva Sen, and Gábor J Székely. On a nonparametric notion of residual and its applications. *Statistics & Probability Letters*, 109:208–213, 2016.

[30] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference.* Elsevier, 2014.

[31] Paul R Rosenbaum. Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, 49(3):425–435, 1984.

[32] Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. In *Advances in Neural Information Processing Systems*, pages 2951–2961, 2017.

[33] Robert J. Serfling. *Approximation Theorems of Mathematical Statistics.* Wiley-Interscience, 2001.

[34] Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *To appear in the Annals of Statistics*, 2018.

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

[36] Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search.* MIT press, 2000.

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

[38] Liangjun Su and Halbert White. A nonparametric hellinger metric test for conditional independence. *Econometric Theory*, 24(4):829–864, 2008.

[39] Liangjun Su and Halbert White. Testing conditional independence via empirical likelihood. *Journal of Econometrics*, 182(1):27–44, 2014.

[40] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

[41] Gregory Valiant and Paul Valiant. Estimating the unseen: Improved estimators for entropy and other properties. *J. ACM*, 64(6):37:1–37:41, 2017.

[42] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[43] Xia Wang and Yongmiao Hong. Characteristic function based testing for conditional independence: a nonparametric regression approach. *Econometric Theory*, 34(4):815–849, 2018.

[44] Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

[45] Qing Yao and David Tritchler. An exact analysis of conditional independence in several 2 x 2 contingency tables. *Biometrics*, pages 233–236, 1993.

[46] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

A Proofs from Section 3

Proof of Theorem 3.3. For the sake of simplicity we only consider the case $d_X = d_Y = d_Z = 1$. The more general case follows the same strategy of proof, with some minor modifications. Following the proof of Shah and Peters [34], it suffices to re-prove the following key lemma in their argument.

Lemma A.1. *Suppose $(X, Y, Z) \in \mathbb{R}^3$ have a distribution supported either on $[-M, M]^3$ for some $M \in (0, \infty)$, or on $(-\infty, \infty)^3$. Let $(X_i, Y_i, Z_i)_{i \in [n]}$ be n i.i.d. copies of (X, Y, Z). Given $\delta > 0$ there exists $C := C(\delta)$ such that for all $\varepsilon > 0$ and all Borel sets $D \subseteq \mathbb{R}^{3n} \times [0, 1]$, it's possible to construct an i.i.d. sequence $(\widetilde{X}_i, \widetilde{Y}_i, \widetilde{Z}_i)_{i \in [n]}$ such that $\widetilde{X}_i \perp\!\!\!\perp \widetilde{Y}_i | \widetilde{Z}_i$ for all i and*

   i. $\mathbb{P}(\max_{i \in [n]} \|(\widetilde{X}_i, \widetilde{Y}_i, \widetilde{Z}_i) - (X_i, Y_i, Z_i)\|_\infty < \varepsilon) > 1 - \delta,$

   ii. If U is uniform on $[0, 1]$ independently of $(\widetilde{X}_i, \widetilde{Y}_i, \widetilde{Z}_i)_{i \in [n]}$ then

$$\mathbb{P}(((\widetilde{X}_i, \widetilde{Y}_i, \widetilde{Z}_i)_{i \in [n]}, U) \in D) \leq C\mu(D),$$

where μ is the Lebesgue measure.

Remark A.2. *The lemma is stated and proved assuming $(X, Y, Z) \in \mathbb{R}^3$, but the proof trivially extends to any $(X, Y, Z) \in \mathbb{R}^{d_X + d_Y + d_Z}$ for $d_X, d_Y, d_Z \in \mathbb{N}$.*

We prove this result below. With this the proof is complete.   □

Proof of Lemma A.1. **Step I (preparation).** First consider the case that the support is $(-\infty, \infty)^3$, i.e., $M = \infty$. We can always find an $M' := M'(\delta) < \infty$ such that $\mathbb{P}(\|(X, Y, Z)\|_\infty > M') < \delta/2n$. Construct $\bar{X}, \bar{Y}, \bar{Z}$ which coincide with (X, Y, Z) if $\|(X, Y, Z)\|_\infty \leq M'$ and are uniform on $[-M', M']^3$ otherwise. By the union bound $\mathbb{P}(\forall i \in [n] : (\bar{X}_i, \bar{Y}_i, \bar{Z}_i) = (X_i, Y_i, Z_i)) > 1 - \delta/2$. We henceforth work with $(\bar{X}_i, \bar{Y}_i, \bar{Z}_i)_{i \in [n]}$ (denoted with $(X_i, Y_i, Z_i)_{i \in [n]}$ for convenience), and will show that there exist $(\widetilde{X}_i, \widetilde{Y}_i, \widetilde{Z}_i)_{i \in [n]}$ satisfying $\mathbb{P}(\max_{i \in [n]} \|(\widetilde{X}_i, \widetilde{Y}_i, \widetilde{Z}_i) - (\bar{X}_i, \bar{Y}_i, \bar{Z}_i)\|_\infty < \varepsilon) = 1$ which implies i., and we will show that ii. is also satisfied. Hence we will assume $M < \infty$ from now on.

Second we note that (without loss of generality) we may assume that the density $p_{X,Y,Z}(x, y, z)$ is bounded by some constant $L := L(\delta)$. This is so since each distribution of (potentially) unbounded density can be well approximated by a distribution of bounded density with high probability. To see this note that the set $S_{\bar{L}} := \{(x, y, z) | p_{X,Y,Z}(x, y, z) > \bar{L}\} \downarrow \varnothing$ when $\bar{L} \to \infty$. Therefore for any δ we can take $\bar{L}(\delta)$ large enough so that $\mathbb{P}((X, Y, Z) \in S^c_{\bar{L}(\delta)}) > 1 - \delta/2n$. Thus we can construct $\bar{X}, \bar{Y}, \bar{Z}$ as X, Y, Z if $(X, Y, Z) \in S^c_{\bar{L}(\delta)}$ and $\bar{X}, \bar{Y}, \bar{Z}$ being uniform on $[-M, M]^3$ otherwise. This distribution has density bounded by $L(\delta) := \bar{L}(\delta) + \delta/(2n(2M)^3)$ and satisfies $\mathbb{P}((\bar{X}, \bar{Y}, \bar{Z}) = (X, Y, Z)) > 1 - \delta/2n$, and therefore by the union bound $\mathbb{P}(\forall i \in [n] : (\bar{X}_i, \bar{Y}_i, \bar{Z}_i) = (X_i, Y_i, Z_i)) > 1 - \delta/2$. As before, we will henceforth work with $(\bar{X}_i, \bar{Y}_i, \bar{Z}_i)_{i \in [n]}$ (denoted with $(X_i, Y_i, Z_i)_{i \in [n]}$ for convenience), and will show that there exist $(\widetilde{X}_i, \widetilde{Y}_i, \widetilde{Z}_i)_{i \in [n]}$ satisfying $\mathbb{P}(\max_{i \in [n]} \|(\widetilde{X}_i, \widetilde{Y}_i, \widetilde{Z}_i) - (\bar{X}_i, \bar{Y}_i, \bar{Z}_i)\|_\infty < \varepsilon) = 1$ which implies i., and we will show that ii. is also satisfied.

Step II (construction). Let $\{A_1, \ldots, A_m\}$ denote an equi-partition of $[-M, M]$ in intervals. Similarly let $\{B_1, \ldots, B_m\}$ and $\{C_1, \ldots, C_m\}$ be equi-partitions of $[-M, M]$. Divide each C_k further in m^2 sub-intervals of equal length denoted by C_{ijk}, so that each of these small intervals corresponds to a pair (A_i, B_j). 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) we construct $(\widetilde{X}, \widetilde{Y}, \widetilde{Z})$ 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 $\widetilde{X} \perp\!\!\!\perp \widetilde{Y} | \widetilde{Z}$. We refer to Figure 1 for a visualization of this construction. In addition it is clear that by construction $\mathbb{P}(\max_{i \in [n]} \|(\widetilde{X}_i, \widetilde{Y}_i, \widetilde{Z}_i) - (X_i, Y_i, Z_i)\|_\infty < \frac{2M}{m}) = 1$. Hence if we take m large enough so that $\frac{2M}{m} < \varepsilon$ we guarantee that i. is satisfied. What is more we may write out the density of $(\widetilde{X}, \widetilde{Y}, \widetilde{Z})$ as

$$p_{\widetilde{X}, \widetilde{Y}, \widetilde{Z}}(\widetilde{x}, \widetilde{y}, \widetilde{z}) = \sum_{i,j,k} \frac{m^5}{(2M)^3} \mathbb{1}(\widetilde{x} \in A_i, \widetilde{y} \in B_j, \widetilde{z} \in C_{ijk}) \mathbb{P}(X \in A_i, Y \in B_j, Z \in C_k)$$

Step III (showing part ii.). Recall that we are assuming that the distribution $p_{X,Y,Z}(x, y, z) \leq L$ for some constant $L > 0$. It is simple to see that the probability that $(X, Y) \in A_i \times B_j$ is bounded as

$$\int_{A_i \times B_j \times [-M,M]} p_{X,Y,Z}(x, y, z) dx dy dz \leq \frac{L(2M)^3}{m^2}.$$

It follows that if we have n observations $(X_i, Y_i, Z_i)_{i \in [n]}$ the probability to have at least two points (X_k, Y_k) and (X_l, Y_l) in one set $A_i \times B_j$ for some i and j is bounded by

$$\left(\frac{L(2M)^3}{m^2} \right)^n \left((m^2)^n - m^2(m^2 - 1) \ldots (m^2 - n + 1) \right) = O\left(\frac{(L(2M)^3)^n}{m^2} \right),$$

since the number of all possible arrangements with points belonging to different sets $A_i \times B_j$ is $m^2(m^2 - 1) \ldots (m^2 - n + 1)$ while the total number of possible arrangements for the n points is $(m^2)^n$. Denote by S the complement of this event. Note that when S happens all $(\widetilde{X}_i, \widetilde{Y}_i, \widetilde{Z}_i)_{i \in [n]}$ have points $(\widetilde{X}_i, \widetilde{Y}_i)$ in different rectangles and vice versa.

Next, suppose that D is an arbitrary fixed Borel set. We have

$$\mathbb{P}(((\widetilde{X}_i, \widetilde{Y}_i, \widetilde{Z}_i)_{i \in [n]}, U) \in D) \leq \mathbb{P}(((\widetilde{X}_i, \widetilde{Y}_i, \widetilde{Z}_i)_{i \in [n]}, U) \in D \cap (S \times [0,1]))$$
$$+ \mathbb{P}(((\widetilde{X}_i, \widetilde{Y}_i, \widetilde{Z}_i)_{i \in [n]}, U) \notin S \times [0,1])$$

We already have a bound on the second term on the RHS above:

$$\mathbb{P}(((\widetilde{X}_i, \widetilde{Y}_i, \widetilde{Z}_i)_{i \in [n]}, U) \notin S \times [0,1]) = O\left(\frac{(L(2M)^3)^n}{m^2} \right).$$

Suppose now that we randomize the assignment on the set C_{ijk}. In other words there is a permutation $\pi : [m^2] \mapsto [m^2]$[5] that assigns each pair A_i, B_j to an interval $C_{\pi_{ij}k}$. Denote by $(\widetilde{X}^\pi, \widetilde{Y}^\pi, \widetilde{Z}^\pi)$ the vectors generated in such manner. Clearly all properties described above hold for

---

[5]Here we use $\pi : [m^2] \mapsto [m^2]$ with a slight abuse of notation. We mean π permuting from all ordered pairs of indices (i, j) where $i, j \in [m]$ to all ordered pairs of indices (k, l) where $k, l \in [m]$.

$(\widetilde{X}_i^\pi, \widetilde{Y}_i^\pi, \widetilde{Z}_i^\pi)_{i\in[n]}$ for any permutation π. We have that

$$\frac{1}{(m^2)!} \sum_{\pi\in[(m^2)!]} \mathbb{P}(((\widetilde{X}_i^\pi, \widetilde{Y}_i^\pi, \widetilde{Z}_i^\pi)_{i\in[n]}, U) \in D \cap (S \times [0,1]))$$

$$= \frac{1}{(m^2)!} \sum_{\pi\in[(m^2)!]} \int_{D\cap(S\times[0,1])} \prod_{l\in[n]} \sum_{i,j,k} \frac{m^5}{(2M)^3} \mathbb{1}(\widetilde{x}_l \in A_i, \widetilde{y}_l \in B_j, \widetilde{z}_l \in C_{\pi_{ij}k}) \mathbb{P}(X\in A_i, Y\in B_j, Z\in C_k)$$

$$\leq \frac{(Lm^2)^n}{(m^2)!} \sum_{\pi\in[(m^2)!]} \int_{D\cap(S\times[0,1])} \prod_{l\in[n]} \sum_{i,j,k} \mathbb{1}(\widetilde{x}_l \in A_i, \widetilde{y}_l \in B_j, \widetilde{z}_l \in C_{\pi_{ij}k})$$

$$= \frac{(Lm^2)^n}{(m^2)!} \sum_{\{i_l\}_{l\in[n]}, \{j_l\}_{l\in[n]}, \{k_l\}_{l\in[n]}} \int_{D\cap(S\times[0,1])} \sum_{\pi\in[(m^2)!]} \mathbb{1}\left(\widetilde{\boldsymbol{x}} \in \prod_{l\in[n]} A_{i_l}, \widetilde{\boldsymbol{y}} \in \prod_{l\in[n]} B_{j_l}, \widetilde{\boldsymbol{z}} \in \prod_{l\in[n]} C_{\pi_{i_l j_l}k_l}\right),$$

where in the above summation we have $\{i_l\}_{l\in[n]}, \{j_l\}_{l\in[n]}, \{k_l\}_{l\in[n]}$ are sequences of n numbers from $[m]$. Since the integration is over the set $D\cap(S\times[0,1])$ all pairs of (i_l, j_l) need to be unique otherwise the integral is 0. Hence, the summation is over all $\{i_l\}_{l\in[n]}, \{j_l\}_{l\in[n]}, \{k_l\}_{l\in[n]}$ sequences of n numbers from $[m]$ for which no two pairs (i_l, j_l) and (i_k, j_k) are the same. Thus to fix $\pi_{i_l j_l}$ for all $(i_l, j_l)_{l\in[n]}$ and permute all others there are $(m^2-n)!$ permutations. Next note that $\prod_{l\in[n]} C_{k_l} = \prod_{l\in[n]} \sum_{ij} C_{ijk_l}$ contains all unique permutations of n elements (and more) and therefore the summation above is bounded as

$$\frac{(Lm^2)^n}{(m^2)!} \sum_{i_l,j_l,k_l} \int_{D\cap(S\times[0,1])} \sum_{\pi\in[(m^2)!]} \mathbb{1}\left(\widetilde{\boldsymbol{x}} \in \prod_{l\in[n]} A_{i_l}, \widetilde{\boldsymbol{y}} \in \prod_{l\in[n]} B_{j_l}, \widetilde{\boldsymbol{z}} \in \prod_{l\in[n]} C_{\pi_{i_l j_l}k_l}\right)$$

$$\leq \frac{(Lm^2)^n(m^2-n)!}{(m^2)!} \sum_{i_l,j_l,k_l} \int_{D\cap(S\times[0,1])} \mathbb{1}\left(\widetilde{\boldsymbol{x}} \in \prod_{l\in[n]} A_{i_l}, \widetilde{\boldsymbol{y}} \in \prod_{l\in[n]} B_{j_l}, \widetilde{\boldsymbol{z}} \in \prod_{l\in[n]} C_{k_l}\right)$$

$$\leq \frac{(Lm^2)^n(m^2-n)!}{(m^2)!} \mu(D\cap(S\times[0,1])) \leq \frac{(Lm^2)^n(m^2-n)!}{(m^2)!} \mu(D).$$

Therefore there exists a permutation π^* such that

$$\mathbb{P}(((\widetilde{X}_i^{\pi^*}, \widetilde{Y}_i^{\pi^*}, \widetilde{Z}_i^{\pi^*})_{i\in[n]}, U) \in D \cap (S\times[0,1])) \leq \frac{(Lm^2)^n(m^2-n)!}{(m^2)!} \mu(D).$$

This finishes the proof, by taking m sufficiently large so that

$$\mathbb{P}(((\widetilde{X}_i^{\pi^*}, \widetilde{Y}_i^{\pi^*}, \widetilde{Z}_i^{\pi^*})_{i\in[n]}, U) \in D) \leq \frac{(Lm^2)^n(m^2-n)!}{(m^2)!} \mu(D) + O\left(\frac{(L(2M)^3)^n}{m^2}\right) \leq C\mu(D).$$

$\square$

Proof of Corollary 3.4. The following proof was suggested to us by the AE. Let $d_Z = 1$ for simplicity (the proof obviously extends to the more general case). Suppose there exists a valid test ψ_n for the discrete case and an alternative distribution Q where it has nontrivial power. ψ_n can be converted into a test for the continuous case by applying it to binned versions of X and Y. Since independence is preserved under binning, this test continues to be valid (i.e. controls the Type I error for all absolutely continuous null distributions). Next one can create a continuous alternative related to

Q by first generating $W \sim Q$ and then reporting a random variable uniformly distributed on the (X, Y) cell associated with W. Clearly, such a distribution is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^3$. The modified test will have power against this alternative — a contradiction to Theorem 3.3.

<div style="text-align: right">□</div>

B Proofs from Section 4

B.1 Poissonization

In this section we will demonstrate that the results of the Section 4 remain valid under the assumption of Poissonization. Concretely, suppose that instead of a fixed sample size n, we are given a random sample of size $N \sim \text{Poi}(n)$. In order for us to redefine the minimax risk, suppose that we are given now a sequence of tests $\{\psi_k\}_{k=0}^{\infty}$ indexed by the sample size, where as before each ψ_k is a Borel measurable function such that $\psi_k : \text{supp}(\mathcal{D}_k) \mapsto [0, 1]$. We define the Poissonized minimax risk as

$$\overline{R}_n(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1, \varepsilon) = \inf_{\{\psi_k\}_{k=0}^{\infty}} \left\{ \sup_{p \in \mathcal{H}_0} \sum_{k=0}^{\infty} \mathbb{P}(N = k) \mathbb{E}_p[\psi_k(\mathcal{D}_k)] \right.$$

$$\left. + \sup_{p \in \{p \in \mathcal{H}_1 : \inf_{q \in \overline{\mathcal{H}}_0} \|p-q\|_1 \geq \varepsilon\}} \sum_{k=0}^{\infty} \mathbb{P}(N = k) \mathbb{E}_p[1 - \psi_k(\mathcal{D}_k)] \right\}, \quad \text{(B.1)}$$

As before we also define the corresponding critical radius

$$\overline{\varepsilon}_n(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1) = \inf \left\{ \varepsilon : \overline{R}_n(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1, \varepsilon) \leq \frac{1}{3} \right\}. \quad \text{(B.2)}$$

We will now state a lemma which relates the minimax risk (B.1) to the minimax risk (2.2). Our arguments are based on the proof of equation (11) in [44], but for completeness we provide them in Appendix B.

Lemma B.1. *Suppose that $\mathcal{H}_0, \mathcal{H}_1$ are dominated sets of measures by some common σ-finite measure. We have that*

$$\overline{R}_{2n}(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1, \varepsilon) - \exp(-(1 - \log 2)n) \leq R_n(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1, \varepsilon) \leq 2\overline{R}_{n/2}(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1, \varepsilon).$$

Using Theorems 4.1 and 4.2 and the right inequality of Lemma B.1 we arrive at the following corollaries which are stated without proof. These results show that the lower bounds in Theorems 4.1 and 4.2, which were developed for a fixed sample size n, continue to hold under Poissonization.

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

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

Corollary B.3. *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 $c_0 > 0$,*

$$\overline{\varepsilon}_n(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1) \geq \frac{c_0}{n^{2/7}}.$$

Proof of Lemma B.1. We first observe that $R_n(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1, \varepsilon) \leq 1$ and furthermore it is clear that $R_n(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1, \varepsilon)$ is a decreasing function in n. Hence for $N \sim \mathrm{Poi}(2n)$

$$\overline{R}_{2n}(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1, \varepsilon) \leq \sum_{0 \leq k \leq n} \mathbb{P}(N=k) R_k(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1, \varepsilon) + \sum_{k>n} \mathbb{P}(N=k) R_k(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1, \varepsilon)$$

$$\leq \mathbb{P}(\mathrm{Poi}(2n) \leq n) + R_n(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1, \varepsilon)$$

$$\leq \exp(-(1 - \log 2)n) + R_n(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1, \varepsilon),$$

where in the last inequality we used a Chernoff bound [44]. By Lemma 1 on page 476 of [27], we know that

$$R_n(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1, \varepsilon) = \sup_{\pi_0, \pi_1} \inf_{\psi} \left\{ \mathbb{E}_{p \sim \pi_0}[\psi(\mathcal{D}_n)] + \mathbb{E}_{q \sim \pi_1}[1 - \psi(\mathcal{D}_n)] \right\},$$

where π_0 and π_1 range over prior distributions over the sets $\mathcal{H}_0$ and $\{p \in \mathcal{H}_1 : \inf_{q \in \overline{\mathcal{H}}_0} \|p - q\|_1 \geq \varepsilon\}$. For the second inequality first fix two prior distributions π_0 and π_1, and take an arbitrary sequence of tests $\psi_0, \psi_1, \psi_2, \ldots$ indexed by the sample size.

It is unclear whether the sequence

$$\alpha_k = \mathbb{E}_{p \sim \pi_0}[\psi_k(\mathcal{D}_k)] + \mathbb{E}_{q \sim \pi_1}[1 - \psi_k(\mathcal{D}_k)],$$

is decreasing with k, but we can make this sequence monotone in the following way. Define $\{\widetilde{\alpha}_k\}$ recursively as $\widetilde{\alpha}_k = \widetilde{\alpha}_{k-1} \wedge \alpha_k$, and define the corresponding sequence of tests

$$\widetilde{\psi}_k(\mathcal{D}_k) = \begin{cases} \widetilde{\psi}_{k-1}(\mathcal{D}_{k-1}), & \text{if } \widetilde{\alpha}_k = \widetilde{\alpha}_{k-1} \\ \psi_k(\mathcal{D}_k), & \text{otherwise} \end{cases}$$

Take $N \sim \mathrm{Poi}(n/2)$. This sequence of estimates satisfies

$$\sum_{k=0}^{\infty} \mathbb{P}(N=k)\{\mathbb{E}_{p \sim \pi_0}[\psi_k(\mathcal{D}_k)] + \mathbb{E}_{q \sim \pi_1}[1 - \psi_k(\mathcal{D}_k)]\}$$

$$\geq \sum_{k=0}^{\infty} \mathbb{P}(N=k)\{\mathbb{E}_{p \sim \pi_0}[\widetilde{\psi}_k(\mathcal{D}_k)] + \mathbb{E}_{q \sim \pi_1}[1 - \widetilde{\psi}_k(\mathcal{D}_k)]\}$$

$$\geq \frac{1}{2}\left\{\mathbb{E}_{p \sim \pi_0}[\widetilde{\psi}_n(\mathcal{D}_n)] + \mathbb{E}_{q \sim \pi_1}[1 - \widetilde{\psi}_n(\mathcal{D}_n)]\right\}$$

$$\geq \frac{1}{2} \inf_{\widetilde{\psi}_n} \left\{\mathbb{E}_{p \sim \pi_0}[\widetilde{\psi}_n(\mathcal{D}_n)] + \mathbb{E}_{q \sim \pi_1}[1 - \widetilde{\psi}_n(\mathcal{D}_n)]\right\},$$

where we used that $\mathbb{P}(\mathrm{Poi}(n/2) \geq n) \leq \frac{1}{2}$ by Markov's inequality. Taking a supremum over π_0 and π_1 we obtain $R_n(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1, \varepsilon)$ on the right hand side. On the left hand side using that the Bayes

Figure 4: The construction of the matrix $\widetilde{\Delta}$ from Δ.

risk is upper bounded by the $\sup_p$ and $\sup_q$ and taking an infimum over all sequences $\psi_0, \psi_1, \psi_2, \ldots$ concludes that

$$\overline{R}_{n/2}(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1, \varepsilon) \geq \frac{1}{2} R_n(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1, \varepsilon).$$

$\square$

Proof of Theorem 4.1. To derive a lower bound we will first show how to obtain multiple distributions which are far from independent by perturbing the uniform discrete distribution. Suppose for simplicity that $\ell_1 = 2\ell_1'$ and $\ell_2 = 2\ell_2'$ for some integers ℓ_1' and ℓ_2' (although this simplifies our calculation we will remark how to fix the calculation for the odd case as well).

We first construct a single null hypothesis distribution. Suppose that $Z \sim U[0,1]$. Let the basic null distribution be given by the density $p_{X,Y|Z}(x,y|z) = \frac{1}{\ell_1 \ell_2}$ for each $x, y \in [\ell_1] \times [\ell_2]$ and $z \in [0,1]$. Clearly, this distribution belongs to all three sets of null distributions $\mathcal{P}'_{0,[0,1],\mathrm{TV}}(L)$, $\mathcal{P}'_{0,[0,1],\mathrm{TV}^2}(L)$ and $\mathcal{P}'_{0,[0,1],\chi^2}(L)$.

Next we will perturb this null distribution p to obtain alternative distributions. Let $\Delta = (\delta_{xy})_{x \in [\ell_1'], y \in [\ell_2']}$ be a matrix of ± 1 numbers δ_{xy}. We create the $\ell_1 \times \ell_2$ matrix $\widetilde{\Delta}$ so that

$$\widetilde{\delta}_{xy} = \delta_{xy} \text{ for } x, y \in [\ell_1'] \times [\ell_2'],$$
$$\widetilde{\delta}_{xy} = -\delta_{(x-\ell_1')y} \text{ for } x > \ell_1', y \in [\ell_2'],$$
$$\widetilde{\delta}_{xy} = -\delta_{x(y-\ell_2')} \text{ for } x \in [\ell_1'], y > \ell_2',$$
$$\widetilde{\delta}_{xy} = \delta_{(x-\ell_1')(y-\ell_2')} \text{ for } x > \ell_1', y > \ell_2'.$$

It is simple to check that all row sums and column sums of the matrix $\widetilde{\Delta}$ are 0 (see also Figure 4). We perturb the null distribution p using the following procedure: for $x, y \in [\ell_1] \times [\ell_2]$ we take $q_{X,Y|Z}(x,y|z) = \frac{1}{\ell_1 \ell_2} + \widetilde{\delta}_{xy}\eta_\nu(z)$, where

$$\eta_\nu(z) = \rho \sum_{j \in [d]} \nu_j h_{j,d}(z),$$

36

where $\rho > 0$ is a constant, $d \in \mathbb{N}$, $\nu_i \in \{-1, +1\}$, and $h_{j,d}(z) = \sqrt{d}h(dz - j + 1)$ for $z \in [(j - 1)/d, j/d]$, and h is an infinitely differentiable function supported on $[0, 1]$ such that $\int h(z)dz = 0$ and $\int h^2(z)dz = 1$. Since the row sums and column sums of $\widetilde{\Delta}$ are 0, it is simple to verify that the marginals of the distribution remain unchanged under this perturbation, i.e. $q_{X|Z}(x|z) = \sum_y q_{X,Y|Z}(x, y|z) = p_{X|Z}(x|z) = \frac{1}{\ell_1}$ and similarly $q_{Y|Z}(y|z) = \sum_x q_{X,Y|Z}(x, y|z) = p_{Y|Z}(y|z) = \frac{1}{\ell_2}$. We note that in the case that one of ℓ_1 or ℓ_2 or both is odd, the fix is to add one row and/or column to the matrix Δ to be fixed, and reason as in the even case.

When perturbing, in order to ensure that we create valid probability distributions, we need to satisfy the conditions that

$$\frac{1}{\ell_1 \ell_2} - \rho\sqrt{d}\|h\|_\infty \geq 0,$$

and

$$\frac{1}{\ell_1 \ell_2} + \rho\sqrt{d}\|h\|_\infty \leq 1.$$

We will ensure this by our choices of ρ and d. Next, we need to verify that $q_{X,Y,Z} \in \mathcal{Q}'_{0,[0,1],\text{TV}}(L)$. We start by showing that $\|q_{X,Y|Z=z} - q_{X,Y|Z=z'}\|_1 \leq L|z - z'|$. We have

$$\|q_{X,Y|Z=z} - q_{X,Y|Z=z'}\|_1 = \ell_1 \ell_2 |\eta_\nu(z) - \eta_\nu(z')|.$$

Now the derivative of $\eta_\nu(z)$, $|\frac{d}{dz}\eta_\nu(z)|$ is bounded by $d^{3/2}\rho\|h'\|_\infty$. Thus the above holds when

$$\|q_{X,Y|Z=z} - q_{X,Y|Z=z'}\|_1 \leq \ell_1 \ell_2 \rho d^{3/2}\|h'\|_\infty \leq L.$$

We let $Z \sim U([0, 1])$. Next we will show that the constructed distributions $q_{X,Y,Z}$ are ε far from being independent, that is we will show that

$$\inf_{p \in \mathcal{P}'_{0,[0,1]}} \|q_{X,Y,Z} - p\|_1 \geq \varepsilon,$$

for some $\varepsilon > 0$. To this end we need the following result which is essentially proved in [13] for the discrete case, here we prove it for continuous Z:

Lemma B.4. *Suppose that a distribution q satisfies*

$$\varepsilon = \inf_{p \in \mathcal{P}'_{0,[0,1]}} \|q_{X,Y,Z} - p\|_1.$$

Then

$$\|q_{X,Y,Z} - q_{X|Z}q_{Y|Z}q_Z\|_1 \leq 6\varepsilon.$$

The proof of Lemma B.4 can be found in Appendix B. Suppose that h satisfies $\int |h(z)|dz = c$ for some $0 < c < 1$. We then have that the L_1 distance between $q_{X,Y,Z}$ and $q_{X|Z}q_{Y|Z}q_Z$ satisfies

$$\varepsilon := \mathbb{E}_Z \ell_1 \ell_2 |\eta_\nu(Z)| = \ell_1 \ell_2 \sum_{j \in [d]} \int \rho|\nu_j h_{j,k}(z)|dz = \ell_1 \ell_2 \rho\sqrt{d}c,$$

where in the above we used that $h_{j,k}$ have disjoint support. By Lemma B.4 this shows that $q_{X,Y,Z}$ is at least $\varepsilon/6$ from any conditionally independent distribution in L_1 distance.

Next, we put uniform priors over ν and Δ, i.e., the random variables $(\nu_i)_{i\in[d]}$ and $(\delta_{xy})_{x,y\in[\ell_1]\times[\ell_2]}$ are taken as i.i.d. Rademachers. The likelihood ratio is

$$W = \mathbb{E}_{\nu,\Delta} \prod_{i=1}^{n}(1 + \widetilde{\delta}_{X_i,Y_i}\bar{\eta}_\nu(Z_i))$$

where $\bar{\eta}_\nu(z) = \ell_1\ell_2\eta_\nu(z)$ and the expectation is taken over all Rademacher sequences ν and Δ. By a standard argument [2, 5, 23] the risk of the likelihood ratio (which is the optimal test by Neyman-Pearson's Lemma) is bounded from below by $1 - \frac{1}{2}\sqrt{\mathrm{Var}_0 W}$. Hence it suffices to study $\mathbb{E}_0(W^2) - 1$ (here $\mathbb{E}_0$ is the expectation under the null hypothesis). We have

$$\mathbb{E}_0W^2 = \mathbb{E}_{\nu,\nu',\Delta,\Delta'} \prod_{i=1}^{n}\mathbb{E}_0(1 + \widetilde{\delta}_{X_i,Y_i}\bar{\eta}_\nu(Z_i))(1 + \widetilde{\delta}'_{X_i,Y_i}\bar{\eta}_{\nu'}(Z_i))$$

$$= \mathbb{E}_{\nu,\nu',\Delta,\Delta'} \prod_{i=1}^{n}\mathbb{E}_{Z_i} \sum_{(x,y)\in[\ell_1]\times[\ell_2]} \frac{(1 + \widetilde{\delta}_{xy}\bar{\eta}_\nu(Z_i))(1 + \widetilde{\delta}'_{xy}\bar{\eta}_{\nu'}(Z_i))}{\ell_1\ell_2}$$

$$= \mathbb{E}_{\nu,\nu',\Delta,\Delta'} \prod_{i=1}^{n}\left(1 + \frac{\mathbb{E}_{Z_i}\bar{\eta}_\nu(Z_i)\bar{\eta}_{\nu'}(Z_i)}{\ell_1\ell_2} \sum_{(x,y)\in[\ell_1]\times[\ell_2]} \widetilde{\delta}_{xy}\widetilde{\delta}'_{xy}\right)$$

$$= \mathbb{E}_{\nu,\nu',\Delta,\Delta'} \prod_{i=1}^{n}(1 + 4\ell_1\ell_2\rho^2\langle\nu,\nu'\rangle\langle\Delta,\Delta'\rangle)$$

$$\leq \mathbb{E}_{\nu,\nu',\Delta,\Delta'} \exp(4n\ell_1\ell_2\rho^2\langle\nu,\nu'\rangle\langle\Delta,\Delta'\rangle)$$

In the above $\langle\Delta,\Delta'\rangle = \mathrm{Tr}(\Delta^\top\Delta')$ is the standard matrix dot product, while $\langle\nu,\nu'\rangle$ is the standard vector dot product and $\mathbb{E}_{\nu,\nu',\Delta,\Delta'}$ is the expectation with respect to independent Rademacher draws of ν,ν',Δ,Δ'. Thus,

$$\mathbb{E}W^2 \leq \mathbb{E}_{\Delta,\Delta'}\mathbb{E}_{\nu,\nu'}[\exp(4n\ell_1\ell_2\rho^2\langle\nu,\nu'\rangle\langle\Delta,\Delta'\rangle)] = \mathbb{E}_{\Delta,\Delta'}\cosh(4n\ell_1\ell_2\rho^2\langle\Delta,\Delta'\rangle)^d$$

$$\leq \mathbb{E}_{\Delta,\Delta'}\exp((4n\ell_1\ell_2\rho^2\langle\Delta,\Delta'\rangle)^2d/2),$$

where we used the inequality $\cosh(x) \leq \exp(x^2/2)$, which can be verified by a Taylor expansion. Next, since when we condition on one value of Δ' all values of $\langle\Delta,\Delta'\rangle$ happen with the same probability as if we conditioned on any other value of Δ' we have the identity

$$\mathbb{E}_{\Delta,\Delta'}\exp((4n\ell_1\ell_2\rho^2\langle\Delta,\Delta'\rangle)^2d/2) = \mathbb{E}_\Delta \exp((4n\ell_1\ell_2\rho^2 \sum_{xy\in[\ell_1']\times[\ell_2']}\delta_{xy})^2d/2) \tag{B.3}$$

Note that if one has i.i.d. Rademacher random variables δ_{xy} and i.i.d. standard normal variables W_{xy} for any nonnegative integers a_{xy} for $x,y \in [\ell_1'] \times [\ell_2']$ one has

$$\mathbb{E} \prod_{xy\in[\ell_1']\times[\ell_2']}\delta_{xy}^{a_{xy}} \leq \mathbb{E} \prod_{xy\in[\ell_1']\times[\ell_2']}W_{xy}^{a_{xy}}$$

Expanding the exponential function in (B.3) one can control all moments of δ_{xy} using the inequality above with corresponding moments of W_{xy}. Thus we conclude

$$\mathbb{E}_\Delta \exp((4n\ell_1\ell_2\rho^2 \sum_{xy\in[\ell_1']\times[\ell_2']}\delta_{xy})^2d/2) \leq \mathbb{E}_W \exp((4n\ell_1\ell_2\rho^2 \sum_{xy\in[\ell_1']\times[\ell_2']}W_{xy})^2d/2) \tag{B.4}$$

The random variable $\sum_{xy \in [\ell_1'] \times [\ell_2']} W_{xy} \sim N(0, \ell_1' \ell_2')$ and therefore $(\sum_{xy \in [\ell_1'] \times [\ell_2']} W_{xy})^2 / \ell_1' \ell_2' := \chi^2$ has a $\chi^2(1)$ distribution. We have

$$\mathbb{E}_{\boldsymbol{W}} \exp\left((4n\ell_1\ell_2\rho^2 \sum_{xy \in [\ell_1'] \times [\ell_2']} W_{xy})^2 d/2\right) \leq \mathbb{E}_{\chi^2} \exp\left((4n\ell_1\ell_2\rho^2)^2 \ell_1' \ell_2' \chi^2 d/2\right). \tag{B.5}$$

Suppose now that $(4n\ell_1\ell_2\rho^2)^2 d\ell_1'\ell_2' < 1$. The above is the mgf of a chi-squared random variable and hence equals to

$$\sqrt{\frac{1}{1 - (4n\ell_1\ell_2\rho^2)^2 d\ell_1'\ell_2'}}.$$

This quantity can be made arbitrarily close to 1 provided that $(4n\ell_1\ell_2\rho^2)^2 d\ell_1'\ell_2'$ is small. Based on this select $\frac{1}{d} \asymp \frac{(\ell_1\ell_2)^{1/5}}{n^{2/5}} \wedge 1$, $\rho \asymp \frac{1}{\ell_1\ell_2 d^{3/2}}$ for some sufficiently small constants. This ensures $\frac{1}{\ell_1\ell_2} - \rho\sqrt{d}\|h\|_\infty \geq 0$ and $\ell_1\ell_2\rho d^{3/2}\|h'\|_\infty \leq L$. With these choices we obtain that the critical radius is bounded from below by a constant times $\frac{1}{d} \asymp \frac{(\ell_1\ell_2)^{1/5}}{n^{2/5}} \wedge 1$. $\qquad\square$

Proof of Lemma B.4. We first start by showing several bounds on the L_1 norm between two arbitrary distributions p and p' from $\mathcal{E}'_{0,[0,1]}$. We have that the L_1 norm between the distributions is

$$\|p - p'\|_1 = \int_{z \in [0,1]} \sum_{x,y \in [\ell_1] \times [\ell_2]} |p_Z(z)p_{X,Y|Z}(x,y|z) - p'_Z(z)p'_{X,Y|Z}(x,y|z)|dz.$$

Using the triangle inequality we conclude that

$$\|p - p'\|_1 \geq \int_{z \in [0,1]} \left| \sum_{x,y \in [\ell_1] \times [\ell_2]} p_Z(z)p_{X,Y|Z}(x,y|z) - p'_Z(z)p'_{X,Y|Z}(x,y|z) \right| dz$$

$$= \int_{z \in [0,1]} |p_Z(z) - p'_Z(z)|dz = \|p_Z - p'_Z\|_1.$$

Next we observe the following identities

$$\|p - p'\|_1$$
$$= \int_{z \in [0,1]} \sum_{x,y \in [\ell_1] \times [\ell_2]} |p_Z(z)(p_{X,Y|Z}(x,y|z) - p'_{X,Y|Z}(x,y|z)) + (p_Z(z) - p'_Z(z))p'_{X,Y|Z}(x,y|z)|dz$$

$$\geq \int_{z \in [0,1]} \sum_{x,y \in [\ell_1] \times [\ell_2]} p_Z(z)|p_{X,Y|Z}(x,y|z) - p'_{X,Y|Z}(x,y|z)| - |p_Z(z) - p'_Z(z)|p'_{X,Y|Z}(x,y|z)dz$$

$$= \mathbb{E}_Z\|p_{X,Y|Z} - p'_{X,Y|Z}\|_1 - \|p_Z - p'_Z\|_1.$$

Combining the last two identities we obtain that

$$\max(\mathbb{E}_Z\|p_{X|Z} - p'_{X|Z}\|_1, \mathbb{E}_Z\|p_{Y|Z} - p'_{Y|Z}\|_1) \leq \mathbb{E}_Z\|p_{X,Y|Z} - p'_{X,Y|Z}\|_1$$
$$\leq \|p_Z - p'_Z\|_1 + \|p - p'\|_1 \leq 2\|p - p'\|_1, \tag{B.6}$$

39

where the first inequality follows by the triangle inequality. Next, reversing the triangle inequality from before we obtain,

$$
\begin{aligned}
\|p - p'\|_1 \\
&= \int_{z \in [0,1]} \sum_{x,y \in [\ell_1] \times [\ell_2]} |p_Z(z)(p_{X,Y|Z}(x,y|z) - p'_{X,Y|Z}(x,y|z)) + (p_Z(z) - p'_Z(z))p'_{X,Y|Z}(x,y|z)|dz \\
&\leq \int_{z \in [0,1]} \sum_{x,y \in [\ell_1] \times [\ell_2]} p_Z(z)|p_{X,Y|Z}(x,y|z) - p'_{X,Y|Z}(x,y|z)| + |p_Z(z) - p'_Z(z)|p_{X,Y|Z}(x,y|z)dz \\
&= \mathbb{E}_Z \|p_{X,Y|Z} - p'_{X,Y|Z}\|_1 + \|p_Z - p'_Z\|_1.
\end{aligned}
$$

Next, suppose that q is a distribution which is ε far from being conditionally independent. We will denote the distribution $q_{X|Z} q_{Y|Z} q_Z$ with $\widetilde{q}$. Let p' be a distribution which is ε away from q in $\| \cdot \|_1$ and p' is conditionally independent (if such a distribution does not exist we can take a sequence that approximates the infimum). We have

$$
\|q - \widetilde{q}\|_1 \leq \|q - p'\|_1 + \|\widetilde{q} - p'\|_1 \leq \varepsilon + \|\widetilde{q} - p'\|_1
$$

We now handle the second term

$$
\begin{aligned}
\|\widetilde{q} - p'\|_1 &\leq \mathbb{E}_Z \|\widetilde{q}_{X,Y|Z} - p'_{X,Y|Z}\|_1 + \|q_Z - p'_Z\|_1 \\
&\leq \mathbb{E}_Z \|\widetilde{q}_{X,Y|Z} - p'_{X,Y|Z}\|_1 + \varepsilon
\end{aligned}
$$

Using that TV is sub-additive on product distributions we now have

$$
\begin{aligned}
\mathbb{E}_Z \|\widetilde{q}_{X,Y|Z} - p'_{X,Y|Z}\|_1 &\leq \mathbb{E}_Z \|\widetilde{q}_{X|Z} - p'_{X|Z}\|_1 + \mathbb{E}_Z \|\widetilde{q}_{Y|Z} - p'_{Y|Z}\|_1 \\
&= \mathbb{E}_Z \|q_{X|Z} - p'_{X|Z}\|_1 + \mathbb{E}_Z \|q_{Y|Z} - p'_{Y|Z}\|_1 \\
&\leq 4\varepsilon,
\end{aligned}
$$

where we used (B.6) in the last bound. We conclude that

$$
\|q - \widetilde{q}\|_1 \leq 6\varepsilon,
$$

which completes the proof. □

Proof of Theorem 4.2. Suppose $(X, Y, Z) \in [0,1]^3$ are three variables with a joint density with respect to the Lebesgue measure in $[0,1]^3$. Under the null hypothesis we specify the distribution as $p_{X,Y,Z}(x,y,z) = 1$ for all $(x,y,z) \in [0,1]^3$, or in other words the three variables have independent uniform distributions on $[0,1]$. Clearly this distribution belongs to the sets $\mathcal{P}_{0,[0,1]^3,\mathrm{TV}}(L)$ and $\mathcal{P}_{0,[0,1]^3,\chi^2}(L)$. Under the alternative hypothesis we specify the distribution as

$$
q_{X,Y|Z}(x,y|z) = 1 + \gamma_\Delta(x,y)\eta_\nu(z),
$$

where as in the proof of Theorem 4.1

$$
\eta_\nu(z) = \rho \sum_{j \in [d]} \nu_j h_{j,d}(z),
$$

where $\rho > 0$ is a constant, $d \in \mathbb{N}$, $\nu_i \in \{-1, +1\}$,and $h_{j,d}(z) = \sqrt{d}h(dz - j + 1)$ for $z \in [(j - 1)/d, j/d]$, and h is an infinitely differentiable function supported on $[0, 1]$ such that $\int h(z)dz = 0$ and $\int h^2(z)dz = 1$. Furthermore we take

$$\gamma_\Delta(x, y) = \rho^2 \sum_{j \in [d']} \sum_{i \in [d']} \delta_{ij} h_{i,d'}(x) h_{j,d'}(y), [6]$$

and we let the marginal distribution of Z be uniform on $[0, 1]$ (i.e. we let $q_Z = p_Z \equiv 1$). In order for this perturbation to be meaningful we need that $1 \geq \sqrt{(d')^2 d}\|h\|_\infty^3 \rho^3$. It is simple to check that $\int_{[0,1]^2} q_{X,Y|Z}(x, y|z)dxdy = 1$. Let us now check what are the marginals of such a distribution conditioned on z. We have

$$q_{X|Z}(x|z) = \int_{[0,1]} q_{X,Y|Z}(x, y|z)dy = 1 + \eta_\nu(z) \int_{[0,1]} \gamma_\Delta(x, y)dy = 1.$$

Similarly $\int_{[0,1]} q_{X,Y|Z}(x, y|z)dx = 1$. It is therefore clear that $q_{X|Z}(x|z)q_{Y|Z}(y|z) \in \mathcal{H}^{2,s}(L)$ for any L. We now check how far away is the distribution $q_{X,Y,Z}(x, y, z) = q_{X,Y|Z}(x, y|z)q_Z(z) = q_{X,Y|Z}(x, y|z)p_Z(z) = q_{X,Y|Z}(x, y|z)$ with respect to $p_{X,Y,Z}(x, y, z)$ (note that $p_{X,Y,Z} = q_{X|Z}q_{Y|Z}q_Z$) in total variation. We have

$$\|q_{X,Y|Z}q_Z - p_{X,Y,Z}\|_1 = \int_{[0,1]^3} |\gamma_\Delta(x, y)\eta_\nu(z)|dxdydz$$

$$= \int_{[0,1]} \rho \sum_{j \in [d']} |h_{j,d'}(x)|dx \int_{[0,1]} \rho \sum_{j \in [d']} |h_{j,d'}(y)|dy \int_{[0,1]} \rho \sum_{j \in [d]} |\eta_{j,d}(z)|dz.$$

Calculating each of the above integrals and multiplying them yields

$$\|q_{X,Y|Z}p_Z - p_{X,Y,Z}\|_1 = \|q_{X,Y,Z} - q_{X|Z}q_{Y|Z}q_Z\|_1 = \sqrt{d(d')^2}\rho^3 c^3,$$

where $c = \int_{[0,1]} |h(x)|dx$. Using Lemma B.4 (here we use this lemma with a slight abuse of notation since the lemma is only valid for discrete X, Y and continuous Z, but the same proof extends to the continuous case) we have that

$$\inf_{p \in \mathcal{P}_{0,[0,1]^3}} \|q - p\|_1 \geq \frac{\sqrt{d(d')^2}\rho^3 c^3}{6}.$$

Next we check that the TV between the distributions $(X, Y|Z = z)$ and $(X, Y|Z = z')$ is Lipschitz.

$$\int_{[0,1]^2} |\gamma_\Delta(x, y)||\eta_\nu(z) - \eta_\nu(z')|dxdy = \sqrt{(d')^2}\rho^2 c^2|\eta_\nu(z) - \eta_\nu(z')|.$$

We now observe that the derivative of $\eta_\nu(z)$ is bounded by $\sqrt{d}d\rho\|h'\|_\infty$, therefore the above is bounded by

$$\sqrt{(d')^2}\rho^2 c^2 \sqrt{d}d\rho\|h'\|_\infty |z - z'|.$$

---

[6]Here $\Delta = \{\delta_{ij}\}_{i,j \in [d]}$.

Next we check that $q(x,y|z)$ belongs to the Hölder class in x and y. We have that

$$\left| \frac{\partial^k}{\partial x^k} \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} \gamma_\Delta(x,y)\eta_\nu(z) - \frac{\partial^k}{\partial x^k} \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} \gamma_\Delta(x',y')\eta_\nu(z) \right|$$

$$\leq \sqrt{d}\rho \|h\|_\infty \left| \frac{\partial^k}{\partial x^k} \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} \gamma_\Delta(x,y) - \frac{\partial^k}{\partial x^k} \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} \gamma_\Delta(x',y') \right|$$

$$= \sqrt{d}\rho \|h\|_\infty \rho^2 (\sqrt{d'})^2 (d')^{\lfloor s \rfloor} \left| \sum_{i,j \in [d']} \delta_{ij} \Big[h^{(k)}(d'x - i + 1)h^{(\lfloor s \rfloor - k)}(d'y - j + 1) \right.$$

$$\left. - h^{(k)}(d'x' - i + 1)h^{(\lfloor s \rfloor - k)}(d'y' - j + 1) \Big] \right|.$$

Suppose now that $x \in \left[\frac{i_x - 1}{d'}, \frac{i_x}{d'}\right], y \in \left[\frac{j_y - 1}{d'}, \frac{j_y}{d'}\right]$, and $x' \in \left[\frac{i_{x'} - 1}{d'}, \frac{i_{x'}}{d'}\right], y' \in \left[\frac{j_{y'} - 1}{d'}, \frac{j_{y'}}{d'}\right]$. Therefore the above summation can be bounded as

$$\sqrt{d}\rho \|h\|_\infty \rho^2 (\sqrt{d'})^2 (d')^{\lfloor s \rfloor} \left[\left| h^{(k)}(d'x - i_x + 1)h^{(\lfloor s \rfloor - k)}(d'y - j_y + 1) - h^{(k)}(d'x' - i_x + 1)h^{(\lfloor s \rfloor - k)}(d'y' - j_y + 1) \right| \right.$$

$$\left. + \left| h^{(k)}(d'x - i_{x'} + 1)h^{(\lfloor s \rfloor - k)}(d'y - j_{y'} + 1) - h^{(k)}(d'x' - i_{x'} + 1)h^{(\lfloor s \rfloor - k)}(d'y' - j_{y'} + 1) \right| \right].$$

Next we will handle the first expression in the bracket above:

$$\leq |h^{(k)}(d'x - i_x + 1) - h^{(k)}(d'x' - i_x + 1)||h^{(\lfloor s \rfloor - k)}(d'y - j_y + 1)|$$

$$+ |h^{(k)}(d'x' - i_x + 1)||h^{(\lfloor s \rfloor - k)}(d'y - j_y + 1) - h^{(\lfloor s \rfloor - k)}(d'y' - j_y + 1)|$$

$$\leq d' \|h^{(k+1)}\|_\infty |x - x'| \|h^{(\lfloor s \rfloor - k)}\|_\infty \wedge (2\|h^{(k)}\|_\infty \|h^{(\lfloor s \rfloor - k)}\|_\infty)$$

$$+ d' \|h^{(\lfloor s \rfloor - k) + 1}\|_\infty |y - y'| \|h^k\|_\infty \wedge (2\|h^{(\lfloor s \rfloor - k)}\|_\infty \|h^k\|_\infty)$$

$$\leq C(1 \wedge d' \sqrt{(x - x')^2 + (y - y')^2})$$

$$\leq C(d' \sqrt{(x - x')^2 + (y - y')^2})^{s - \lfloor s \rfloor},$$

where in the last inequality we used that $(1 \wedge u)^a \leq u^a$ for $u > 0$ and $0 \leq a \leq 1$. We can handle the second expression in the bracket above in a similar way.

In addition it is clear that any lower order $k \leq \lfloor s \rfloor$ partial derivatives with respect to x and y of $q(x,y|z)$ are bounded by $\sqrt{d}\rho^3 \|h\|_\infty \sqrt{d'}^2 (d')^k C$ for some constant C which will depend on the function h.

It therefore suffices that $\sqrt{d}\rho^3 (\sqrt{d'})^2 (d')^s$ to be smaller than a constant and we will have both conditions satisfied. Now we write down the likelihood ratio between the null and the alternative mixing over all choices of Rademacher vector and matrix ν, Δ:

$$W = \mathbb{E}_{\nu,\Delta} \prod_{i=1}^{n} (1 + \gamma_\Delta(X_i, Y_i)\eta_\nu(Z_i)).$$

The second moment of W is

$$\mathbb{E}W^2 = \mathbb{E}_{\nu,\nu',\Delta,\Delta'} \prod_{i=1}^{n} \mathbb{E}_0 (1 + \gamma_\Delta(X_i, Y_i)\eta_\nu(Z_i))(1 + \gamma_{\delta'}(X_i, Y_i)\eta_{\nu'}(Z_i))$$

$$= \mathbb{E}_{\nu,\nu',\Delta,\Delta'} \prod_{i=1}^{n} (1 + \mathbb{E}_0 \gamma_\Delta(X_i, Y_i)\eta_\nu(Z_i)\gamma_{\delta'}(X_i, Y_i)\eta_{\nu'}(Z_i)),$$

where the above follows from the fact that $\mathbb{E}_0 \eta_\nu(Z_i) = 0$ (and that X_i and Y_i are independent of Z_i under the null hypothesis). Continuing the identities yields

$$\mathbb{E}W^2 = \mathbb{E}_{\nu,\nu',\Delta,\Delta'} \prod_{i=1}^{n} (1 + \mathbb{E}_0 \gamma_\Delta(X_i, Y_i)\gamma_{\delta'}(X_i, Y_i)\mathbb{E}_0 \eta_\nu(Z_i)\eta_{\nu'}(Z_i))$$

$$= \mathbb{E}_{\nu,\nu',\Delta,\Delta'} \prod_{i=1}^{n} (1 + \rho^6 \langle \Delta, \Delta' \rangle \langle \nu, \nu' \rangle),$$

where $\langle \Delta, \Delta' \rangle = \operatorname{Tr}(\Delta^\top \Delta')$. From here the proof can continue as in Theorem 4.1. The final expression that needs to be smaller than a constant is $(n\rho^6)^2 d(d')^2$. Set $d \asymp n^{2s/(5s+2)}$, $d' = d^{1/s}$, $\rho^3 \asymp d^{-(3/2+1/s)}$. This results in a rate $\asymp 1/d \asymp n^{-2s/(5s+2)}$. $\qquad \square$

C   Proofs from Section 5

This section contains the proofs of Sections 5.1, 5.2 and 5.3.

C.1   Proofs from Section 5.1

Proof of Lemma 5.1. According to Section 12 of [42] the variance of the U-statistic (5.3) equals to

$$O\left(\frac{1}{\sigma}\right) \operatorname{Cov}(h_{ijkl}, h_{ij'k'l'}) + O\left(\frac{1}{\sigma^2}\right) \operatorname{Cov}(h_{ijkl}, h_{ijk'l'})$$

$$+ O\left(\frac{1}{\sigma^3}\right) \operatorname{Cov}(h_{ijkl}, h_{ijkl'}) + O\left(\frac{1}{\sigma^4}\right) \operatorname{Var}(h_{ijkl}),$$

where $i, j, k, l, j', k', l' \in [\sigma]$ are distinct indices of observations (it is ok if the sample size σ is smaller than 7, then the first terms simply do not contribute). We will now argue that

$$\operatorname{Cov}(h_{ijkl}, h_{ij'k'l'}) \le C\mathbb{E}[U(\mathcal{D})] \max(\|p_{X',Y'}\|_2, \|p_{X'}p_{Y'}\|_2),$$

and that all other covariances are bounded by $C \max(\|p_{X',Y'}\|_2^2, \|p_{X'}p_{Y'}\|_2^2)$ which will complete the proof. In order to bound the first term from above it suffices to control the following expression

$$\mathbb{E} \sum_{x,y} \phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy) \sum_{x',y'} \phi_{\pi'_1\pi'_2}(x'y')\phi_{\pi'_3\pi'_4}(x'y')$$

$$- \mathbb{E} \sum_{x,y} \phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy) \mathbb{E} \sum_{x',y'} \phi_{\pi'_1\pi'_2}(x'y')\phi_{\pi'_3\pi'_4}(x'y'),$$

where π is a permutation of i, j, k, l and π' is a permutation of i, j', k', l'. We can rewrite the above as

$$\sum_{x,y,x',y'} \left\{ \mathbb{E}\phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy)\phi_{\pi_1'\pi_2'}(x'y')\phi_{\pi_3'\pi_4'}(x'y') - (\mathbb{E}\phi_{\pi_1\pi_2}(xy))^2(\mathbb{E}\phi_{\pi_1'\pi_2'}(x'y'))^2 \right\} \qquad \text{(C.1)}$$

where we used that by independence $\mathbb{E}\phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy) = \mathbb{E}\phi_{\pi_1\pi_2}(xy)\mathbb{E}\phi_{\pi_3\pi_4}(xy) = (\mathbb{E}\phi_{\pi_1\pi_2}(xy))^2$ and similarly for π'. Without loss of generality suppose that i is some of π_1, π_2 and π_1', π_2'. Going back to equation (C.1) we have

$$\sum_{x,y,x',y'} \left\{ \mathbb{E}\phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy)\phi_{\pi_1'\pi_2'}(x'y')\phi_{\pi_3'\pi_4'}(x'y') - (\mathbb{E}\phi_{\pi_1\pi_2}(xy))^2(\mathbb{E}\phi_{\pi_1'\pi_2'}(x'y'))^2 \right\}$$

$$\leq \sum_{x,y,x',y'} \left\{ \mathbb{E}\phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy)\phi_{\pi_1'\pi_2'}(x'y')\phi_{\pi_3'\pi_4'}(x'y') \right\}$$

$$= \sum_{x,y,x',y'} \left\{ \mathbb{E}(\phi_{\pi_1\pi_2}(xy)\phi_{\pi_1'\pi_2'}(x'y'))\mathbb{E}\phi_{\pi_3\pi_4}(xy)\mathbb{E}\phi_{\pi_3'\pi_4'}(x'y') \right\}$$

$$\leq \sqrt{\sum_{x,y,x',y'} \{\mathbb{E}(\phi_{\pi_1\pi_2}(xy)\phi_{\pi_1'\pi_2'}(x'y'))\}^2} \sqrt{\sum_{x,y,x',y'} \{\mathbb{E}\phi_{\pi_3\pi_4}(xy)\}^2\{\mathbb{E}\phi_{\pi_3'\pi_4'}(x'y')\}^2}$$

$$= \sqrt{\sum_{x,y,x',y'} \{\mathbb{E}(\phi_{\pi_1\pi_2}(xy)\phi_{\pi_1'\pi_2'}(x'y'))\}^2 \mathbb{E}[U(\mathcal{D})]}$$

where the next to last step follows from Cauchy-Schwarz. To this end we formalize the following lemma.

Lemma C.1. *For two random variables $A, B \in \{\pm 1, 0\}$ we have*

$$(\mathbb{E}AB)^2 \leq \mathbb{E}[|B|\,||A| = 1](\mathbb{E}|A|)^2.$$

Now we apply Lemma C.1 to the first term on the RHS with $A = \phi_{\pi_1\pi_2}(xy)$ and $B = \phi_{\pi_1'\pi_2'}(x'y')$ noting that ϕ can only take values $\{\pm 1, 0\}$. We obtain

$$\sum_{x,y,x',y'} \{\mathbb{E}(\phi_{\pi_1\pi_2}(xy)\phi_{\pi_1'\pi_2'}(x'y'))\}^2 \leq \sum_{x,y,x',y'} \mathbb{E}[|\phi_{\pi_1'\pi_2'}(x'y')|\,||\phi_{\pi_1\pi_2}(xy)| = 1](\mathbb{E}|\phi_{\pi_1\pi_2}(xy)|)^2$$

Note that

$$|\phi_{\pi_1\pi_2}(xy)| \leq \mathbb{1}(X_{\pi_1}' = x, Y_{\pi_1}' = y) + \mathbb{1}(Y_{\pi_2}' = y)\mathbb{1}(X_{\pi_1}' = x),$$

and a similar inequality holds for $|\phi_{\pi_1'\pi_2'}(x'y')|$. Thus

$$\sum_{x',y'} |\phi_{\pi_1'\pi_2'}(x'y')| \leq \sum_{x',y'} \mathbb{1}(X_{\pi_1'}' = x', Y_{\pi_1'}' = y') + \mathbb{1}(Y_{\pi_2'}' = y')\mathbb{1}(X_{\pi_1'}' = x') \leq 2.$$

Hence

$$\sum_{x,y} \mathbb{E}[\sum_{x',y'} |\phi_{\pi_1'\pi_2'}(x'y')|\,||\phi_{\pi_1\pi_2}(xy)| = 1](\mathbb{E}|\phi_{\pi_1\pi_2}(xy)|)^2 \leq 2 \sum_{x,y} (\mathbb{E}|\phi_{\pi_1\pi_2}(xy)|)^2$$

$$\leq 2 \sum_{x,y} (p_{X',Y'}(x,y) + p_{X'}(x)p_{Y'}(y))^2 \leq 4(\|p_{X',Y'}\|_2^2 + \|p_{X'}p_{Y'}\|_2^2),$$

which is what we wanted to show. Now it remains to prove Lemma C.1.

Proof of Lemma C.1. We have

$$(\mathbb{E}AB)^2 \leq (\mathbb{E}|AB|)^2 = (\mathbb{E}[|B|||A| = 1]\mathbb{E}|A|)^2 \leq \mathbb{E}[|B|||A| = 1](\mathbb{E}|A|)^2,$$

where in the last step we used that $\mathbb{E}[|B|||A| = 1] \leq 1$. $\qquad\square$

Now we will show how to bound any higher order terms: $\text{Cov}(h_{ijkl}, h_{ijk'l'})$ where $i, j, k, l, k', l' \in [\sigma]$ (and it's possible for k', l' to be equal to k or l). To bound these terms, following the same strategy as before, it suffices to control the quantity

$$\sum_{x,y,x',y'} \left\{ \mathbb{E}\phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy)\phi_{\pi'_1\pi'_2}(x'y')\phi_{\pi'_3\pi'_4}(x'y') \right\}$$

We will now use the fact that for random variables $A, B \in \{\pm 1, 0\}$

$$\mathbb{E}AB \leq \mathbb{E}|AB| = \mathbb{E}[|B|||A| = 1]\mathbb{E}|A|,$$

where $A = \phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy)$ and $B = \phi_{\pi'_1\pi'_2}(x'y')\phi_{\pi'_3\pi'_4}(x'y')$. We have

$$\sum_{x,y,x',y'} \left\{ \mathbb{E}\phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy)\phi_{\pi'_1\pi'_2}(x'y')\phi_{\pi'_3\pi'_4}(x'y') \right\}$$

$$\leq \sum_{x,y,x',y'} \mathbb{E}[|\phi_{\pi'_1\pi'_2}(x'y')\phi_{\pi'_3\pi'_4}(x'y')|||\phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy)| = 1]\mathbb{E}|\phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy)|$$

We now use that as we saw before

$$|\phi_{\pi_1\pi_2}(xy)| \leq \mathbb{1}(X'_{\pi_1} = x, Y'_{\pi_1} = y) + \mathbb{1}(Y'_{\pi_2} = y)\mathbb{1}(X'_{\pi_1} = x),$$

and analogously for the others. Furthermore $|\phi_{\pi_1\pi_2}(xy)| \leq 1$. We have

$$\sum_{x',y'} \mathbb{E}[|\phi_{\pi'_1\pi'_2}(x'y')\phi_{\pi'_3\pi'_4}(x'y')|||\phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy)| = 1] \leq \sum_{x',y'} \mathbb{E}[|\phi_{\pi'_1\pi'_2}(x'y')||\phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy)| = 1]$$

$$\leq 2.$$

Next by independence,

$$\sum_{x,y} \mathbb{E}|\phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy)| = \sum_{x,y} \mathbb{E}|\phi_{\pi_1\pi_2}(xy)|\mathbb{E}|\phi_{\pi_3\pi_4}(xy)|$$

$$\leq \sum_{x,y} (p_{X',Y'}(x,y) + p_{X'}(x)p_{Y'}(y))^2$$

$$\leq \sum_{x,y} 2(p_{X',Y'}^2(x,y) + p_{X'}^2(x)p_{Y'}^2(y)) = 2(\|p_{X',Y'}\|_2^2 + \|p_{X'}p_{Y'}\|_2^2).$$

This completes the proof. $\qquad\square$

Below we will prove the following version of Theorem 5.2

45

Theorem C.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 $\varepsilon \geq cn^{-2/5}$, for a sufficiently large constant c (depending on ζ, L, ℓ_1, ℓ_2). Then we have that*

$$\sup_{p \in \mathcal{P}'_{0,[0,1],\mathrm{TV}^2}(L) \cup \mathcal{P}'_{0,[0,1],\mathrm{TV}}(L) \cup \mathcal{P}'_{0,[0,1],\chi^2}(L)} \sum_{k=0}^{\infty} \mathbb{P}(N = k) \mathbb{E}_p[\psi_\tau(\mathcal{D}_k)] \leq \frac{1}{10},$$

$$\sup_{p \in \{p \in \mathcal{Q}'_{0,[0,1],\mathrm{TV}}(L) : \inf_{q \in \mathcal{P}'_{0,[0,1]}} \|p-q\|_1 \geq \varepsilon\}} \sum_{k=0}^{\infty} \mathbb{P}(N = k) \mathbb{E}_p[1 - \psi_\tau(\mathcal{D}_k)] \leq \frac{1}{10}.$$

Proof of Theorem 5.2. We now derive Theorem 5.2 from Theorem C.2. Note that the test is equivalent to $\mathbb{1}(T \geq \tau)\mathbb{1}(N \leq n)$. Hence under the null hypothesis we have

$$\mathbb{E}\mathbb{1}(T \geq \tau)\mathbb{1}(N \leq n) \leq \mathbb{E}\mathbb{1}(T \geq \tau) \leq \frac{1}{10},$$

where the above expectation is with respect to the randomness of the samples, as well as the randomness of N. On the other hand under the alternative we have

$$\mathbb{E}(1 - \mathbb{1}(T \geq \tau)\mathbb{1}(N \leq n)) \leq \mathbb{E}(1 - \mathbb{1}(T \geq \tau)) + \mathbb{E}\mathbb{1}(N \geq n) \leq \frac{1}{10} + \exp(-n/8),$$

where we used the bound provided in the following note [12]. □

Proof of Theorem C.2. To prove this theorem we will assume that $N \sim Poi(n)$ instead of $Poi(n/2)$ for convenience. Since this changes only constant factors, we can do this WLOG. We define

$$q_{xy}(m) = \frac{\int_{C_m} p_{X,Y|Z}(x,y|z) dP_Z(z)}{\mathbb{P}(Z \in C_m)} = \int_{C_m} p_{X,Y|Z}(x,y|z) d\widetilde{P}_Z(z), \tag{C.2}$$

where $p_{X,Y|Z}(x,y|z)$ is the conditional distribution of $X,Y|Z = z$, and $P_Z(z)$ is the distribution of Z which is absolutely continuous with respect to the Lebesgue measure, and

$$d\widetilde{P}_Z(z) = \frac{dP_Z(z)}{\mathbb{P}(Z \in C_m)},$$

is the conditional distribution of $Z|Z \in C_m$. Further define

$$q_{x\cdot}(m) = \sum_{y \in [\ell_2]} q_{xy}(m) = \int_{C_m} p_{X|Z}(x|z) d\widetilde{P}(z), \quad q_{\cdot y}(m) = \sum_{x \in [\ell_1]} q_{xy}(m) = \int_{C_m} p_{Y|Z}(y|z) d\widetilde{P}(z), \tag{C.3}$$

Analysis of the Expectation of T. Conditioning on $\sigma = (\sigma_m)_{m \in [d]}$, and using the fact that $\mathbb{E}[U_m|\sigma_m] = \sum_{x,y}(q_{xy}(m) - q_{x\cdot}(m)q_{\cdot y}(m))^2$ is independent of σ_m, we have

$$\mathbb{E}[T] = \mathbb{E}[\mathbb{E}[T|\sigma]] = \sum_{m \in [d]} \mathbb{E}[U_m|\sigma_m]\mathbb{E}[\sigma_m \mathbb{1}(\sigma_m \geq 4)]$$

Let $p_m = \mathbb{P}(Z \in C_m)$. Since $\sigma_m \sim \text{Poi}(np_m)$, Lemma 3.1. of [13] shows that

$$\mathbb{E}[\sigma_m \mathbb{1}(\sigma_m \geq 4)] \geq \gamma \min(np_m, (np_m)^4),$$

where $\gamma = 1 - \frac{5}{2e}$. Observe that even under the null $\mathbb{E}[U_m|\sigma_m] = \sum_{x,y}(q_{xy}(m) - q_{x\cdot}(m)q_{\cdot y}(m))^2 \neq 0$ in general. We will now prove that under the null hypothesis we have:

$$\mathbb{E}[U_m|\sigma_m] \leq \frac{L^2}{d^2}. \tag{C.4}$$

Lemma C.3. *Since the distribution of (X, Y, Z) belongs to the class $\mathcal{P}'_{0,[0,1],\text{TV}^2}$ of Definition 2.1 we have* (C.4).

Proof. We have

$$\sum_{x,y}(q_{xy}(m) - q_{x\cdot}(m)q_{\cdot y}(m))^2 \leq \left(\sum_{x,y} |q_{xy}(m) - q_{x\cdot}(m)q_{\cdot y}(m)| \right)^2$$

Furthermore, the following chain of identities holds

$$\sum_{x,y} |q_{xy}(m) - q_{x\cdot}(m)q_{\cdot y}(m)|$$

$$= \sum_{x,y} \left| \int (p_{X|Z}(x|z) - \int p_{X|Z}(x|z)d\widetilde{P}_Z(z))(p_{Y|Z}(y|z) - \int p_{Y|Z}(y|z)d\widetilde{P}_Z(z))d\widetilde{P}_Z(z) \right|$$

$$\leq \int \sum_x \left| p_{X|Z}(x|z) - \int p_{X|Z}(x|z)d\widetilde{P}_Z(z) \right| \sum_y \left| p_{Y|Z}(y|z) - \int p_{Y|Z}(y|z)d\widetilde{P}_Z(z) \right| d\widetilde{P}_Z(z)$$

$$\overset{(i)}{\leq} \int \int \sum_x |p_{X|Z}(x|z) - p_{X|Z}(x|z')|d\widetilde{P}_Z(z') \int \sum_y |p_{Y|Z}(y|z) - p_{Y|Z}(y|z')|d\widetilde{P}_Z(z')d\widetilde{P}_Z(z)$$

$$= \int \int \|p_{X|Z=z} - p_{X|Z=z'}\|_1 d\widetilde{P}_Z(z') \int \|p_{Y|Z=z} - p_{Y|Z=z'}\|_1 d\widetilde{P}_Z(z')d\widetilde{P}_Z(z)$$

$$\overset{(ii)}{\leq} \int \int \sqrt{L|z-z'|}d\widetilde{P}_Z(z') \int \sqrt{L|z-z'|}d\widetilde{P}_Z(z')d\widetilde{P}_Z(z)$$

$$\leq \frac{L}{d},$$

where (i) follows by Jensen's inequality, and (ii) follows by the fact that $p \in \mathcal{P}'_{0,[0,1],\text{TV}^2}$. This completes the proof. $\qquad\qquad\square$

Hence a bound on $\mathbb{E}[T]$ is

$$\mathbb{E}[T] \leq \frac{nL^2}{d^2}, \tag{C.5}$$

since $\sum_{m \in [d]} \mathbb{E}[\sigma_m \mathbb{1}(\sigma_m \geq 4)] \leq \sum_{m \in [d]} \mathbb{E}[\sigma_m] = \sum_{m \in [d]} np_m = n$. Next we need to lower bound the expectation under the alternative. To this end consider the following

Lemma C.4. *We have that*

$$\sum_{m\in[d]}\sqrt{\mathbb{E}[U_m|\sigma_m]}p_m = \sum_{m\in[d]}\sqrt{\sum_{x,y}(q_{xy}(m)-q_{x\cdot}(m)q_{\cdot y}(m))^2}p_m$$

$$\geq \frac{\mathbb{E}_Z\|p_{X,Y|Z}-p_{X|Z}p_{Y|Z}\|_1 - 3\frac{L}{d}}{\sqrt{\ell_1\ell_2}} =: \frac{\eta}{\sqrt{\ell_1\ell_2}},$$

Proof of Lemma C.4. First we will show that the function $z \mapsto \|p_{X,Y|Z=z} - p_{X|Z=z}p_{Y|Z=z}\|_1$ is continuous. Take two values $z, z' \in C_m$ and observe that

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

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

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

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

$$\leq 3L|z - z'|,$$

where we first used the triangle inequality, next the fact that $\|\cdot\|_1$ is sub-additive on product distributions and finally we used our assumption on the distribution $p_{X,Y|Z}$ and noted that by the triangle inequality

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

Since C_m is compact it follows that the function $z \mapsto \|p_{X,Y|Z=z} - p_{X|Z=z}p_{Y|Z=z}\|_1$ achieves its maximum. Suppose that

$$z_m^* \in \underset{z\in C_m}{\operatorname{argmax}} \|p_{X,Y|Z=z} - p_{X|Z=z}p_{Y|Z=z}\|_1.$$

By Cauchy-Schwarz we have

$$\sqrt{\sum_{x,y}(q_{xy}(m) - q_{x\cdot}(m)q_{\cdot y}(m))^2} \geq \frac{\sum_{x,y}|q_{xy}(m) - q_{x\cdot}(m)q_{\cdot y}(m)|}{\sqrt{\ell_1\ell_2}}.$$

Now we apply the triangle inequality to obtain

$$\sum_{x,y}|q_{xy}(m) - q_{x\cdot}(m)q_{\cdot y}(m)| \geq \|p_{x,y|z=z_m^*} - p_{X|Z=z_m^*}p_{Y|Z=z_m^*}\|_1 - \sum_{x,y}|q_{xy}(m) - p_{X,Y|Z}(x,y|z_m^*)|$$

$$- \sum_{x,y}|q_{x\cdot}(m)(q_{\cdot y}(m) - p_{Y|Z}(y|z_m^*))| - \sum_{x,y}|(p_{Y|Z}(y|z_m^*)(q_{x\cdot}(m) - p_{X|Z}(x|z_m^*))|.$$

For the first term we use

$$\sum_{x,y}|q_{xy}(m) - p_{X,Y|Z}(x,y|z_m^*)| = \sum_{x,y}\left|\int_{C_m} p_{X,Y|Z}(x,y|z) - p_{X,Y|Z}(x,y|z_m^*)d\widetilde{P}(z)\right|$$

$$\leq \int_{C_m}\sum_{x,y}|p_{X,Y|Z}(x,y|z) - p_{X,Y|Z}(x,y|z_m^*)|d\widetilde{P}(z)$$

$$= \int_{C_m}\|p_{X,Y|Z=z} - p_{X,Y|Z=z_m^*}\|_1 d\widetilde{P}(z)$$

$$\leq \int_{C_m} L|z - z_m^*|d\widetilde{P}(z) \leq L\operatorname{diam}(C_m) = \frac{L}{d},$$

48

For the second term we have

$$\sum_{x,y} |q_{x\cdot}(m)(q_{\cdot y}(m) - p_{Y|Z}(y|z_m^*))| = \sum_{x,y} \left| \int_{C_m} q_{x\cdot}(m)(p_{Y|Z}(y|z) - p_{Y|Z}(y|z_m^*))d\widetilde{P}(z) \right|$$

$$\leq \int_{C_m} \sum_{x,y} |q_{x\cdot}(m)(p_{Y|Z}(y|z) - p_{Y|Z}(y|z_m^*))|d\widetilde{P}(z)$$

$$= \int_{C_m} \|p_{Y|Z=z} - p_{Y|Z=z_m^*}\|_1 d\widetilde{P}(z)$$

$$\leq L \operatorname{diam}(C_m) = \frac{L}{d}.$$

The last term is similar to the previous term so we conclude that

$$\sqrt{\sum_{x,y}(q_{xy}(m) - q_{x\cdot}(m)q_{\cdot y}(m))^2} \geq \frac{\|p_{x,y|z=z_m^*} - p_{X|Z=z_m^*}p_{Y|Z=z_m^*}\|_1 - 3\frac{L}{d}}{\sqrt{\ell_1 \ell_2}}$$

Summing up over m and noting that by the definition of z_m^* we obtain

$$\sum_{m \in [d]} \|p_{XY|Z=z_m^*} - p_{X|Z=z_m^*}p_{Y|Z=z_m^*}\|_1 p_m \geq \mathbb{E}_Z \|p_{X,Y|Z} - p_{X|Z}p_{Y|Z}\|_1 \geq \inf_{q \in \mathcal{P}_{0,[0,1]}} \|p_{X,Y,Z} - q\|_1$$

where we used the fact that the distribution $p_{X|Z}p_{Y|Z}p_Z$ is a conditionally independent distribution.

$\square$

Next we have

$$\sum_{m \in [d]} \mathbb{E}[U_m|\sigma_m]\mathbb{E}[\sigma_m \mathbb{1}(\sigma_m \geq 4)] \geq \gamma \sum_{m:(np_m)>1} \mathbb{E}[U_m|\sigma_m]np_m + \gamma \sum_{m:(np_m)\leq 1} \mathbb{E}[U_m|\sigma_m](np_m)^4.$$

We now consider two cases:

i. In the first case we assume

$$\sum_{m:np_m>1} \sqrt{\mathbb{E}[U_m|\sigma_m]}np_m \geq \frac{n\eta}{2\sqrt{\ell_1 \ell_2}}.$$

Then by Cauchy-Schwarz we have

$$\sum_{m:(np_m)>1} \mathbb{E}[U_m|\sigma_m]np_m \geq \frac{(\sum_{m:np_m>1} \sqrt{\mathbb{E}[U_m|\sigma_m]}np_m)^2}{\sum_{m:(np_m)>1} np_m} \geq \frac{n\eta^2}{4\ell_1 \ell_2}.$$

ii. In the second case we suppose:

$$\sum_{m:np_m\leq 1} \sqrt{\mathbb{E}[U_m|\sigma_m]}np_m \geq \frac{n\eta}{2\sqrt{\ell_1 \ell_2}}.$$

49

By Jensen's inequality we have

$$\sum_{m:np_m\leq 1} \frac{\mathbb{E}[U_m|\sigma_m]^{1/3}}{\sum_{m:np_m\leq 1}\mathbb{E}[U_m|\sigma_m]^{1/3}}\mathbb{E}[U_m|\sigma_m]^{2/3}(np_m)^4$$

$$\geq \left(\sum_{m:np_m\leq 1}\frac{\mathbb{E}[U_m|\sigma_m]^{1/3}}{\sum_{m:np_m\leq 1}\mathbb{E}[U_m|\sigma_m]^{1/3}}\mathbb{E}[U_m|\sigma_m]^{1/6}np_m\right)^4,$$

which is equivalent to

$$\left(\sum_{m:np_m\leq 1}\mathbb{E}[U_m|\sigma_m]^{1/3}\right)^3\sum_{m:np_m\leq 1}\mathbb{E}[U_m|\sigma_m](np_m)^4 \geq \left(\sum_{m:np_m\leq 1}\sqrt{\mathbb{E}[U_m|\sigma_m]}np_m\right)^4 \geq \frac{(n\eta)^4}{16\ell_1^2\ell_2^2},$$

Since $\mathbb{E}[U_m|\sigma_m]\leq\left(\sum_{x,y}|q_{xy}(m)-q_{x\cdot}(m)q_{\cdot y}(m)|\right)^2\leq 4$ we have

$$\sum_{m:np_m\leq 1}\mathbb{E}[U_m|\sigma_m](np_m)^4 \geq \frac{(n\eta)^4}{64d^3\ell_1^2\ell_2^2}.$$

We will now select a threshold at the level of $\zeta\sqrt{d}$, and will give conditions on the minimum critical radius for each of the cases. We will use $\gtrsim$ in the sense bigger up to an absolute constant. We will assume that $\varepsilon-3\frac{L}{d}\geq\varepsilon/2$ so that $\eta\geq\varepsilon/2$.

- In the first case we obtain the following bound

$$\frac{n\eta^2}{4\ell_1\ell_2}\gtrsim\zeta\sqrt{d}$$

  which is ensured when $\varepsilon\gtrsim\sqrt{\frac{\zeta\sqrt{d}\ell_1\ell_2}{n}}\vee\frac{1}{d}$

- In the second case we have

$$\frac{(n\eta)^4}{64d^3\ell_1^2\ell_2^2}\gtrsim\zeta\sqrt{d},$$

  which happens when $\varepsilon\gtrsim\frac{\zeta^{1/4}d^{7/8}\sqrt{\ell_1\ell_2}}{n}\vee\frac{1}{d}$.

It is simple to check that when $d\asymp n^{2/5}$ the bigger of the two rates is $\varepsilon\gtrsim n^{-2/5}$.

Analysis of the Variance of T. The rule of total variance ensures that

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

where $\sigma=(\sigma_m)_{m\in[d]}$. Put for brevity $T_m=\mathbb{1}(\sigma_m\geq 4)U_m\sigma_m$ so that $T=\sum_{m\in[d]}T_m$. We first handle the first term. We have

$$\mathrm{Var}[T|\sigma]=\sum_{m,k\in[d]}\mathrm{Cov}(T_m,T_k|\sigma_m,\sigma_k)=\sum_{m\in[d]}\mathrm{Var}(T_m|\sigma_m),$$

50

where we used that T_m and T_k are independent given σ_m, σ_k. Using Lemma 5.1 and the fact that $\sum_{x,y} q_{x,y}^2(m) \leq 1$ and $\sum_{x,y} q_{x\cdot}^2(m)q_{\cdot y}^2(m) \leq 1$, we have

$$\sum_{m\in[d]} \mathrm{Var}(T_m|\sigma_m) \leq \sum_{m\in[d]} \sigma_m^2 \mathbb{1}(\sigma_m \geq 4)C\left(\frac{\mathbb{E}[U_m|\sigma_m]}{\sigma_m} + \frac{1}{\sigma_m^2}\right)$$

$$= \sum_{m\in[d]} C(\mathbb{E}[\mathbb{1}(\sigma_m \geq 4)U_m\sigma_m|\sigma_m] + \mathbb{1}(\sigma_m \geq 4)).$$

Taking expectation of the expression above we end up with

$$\mathbb{E}[\mathrm{Var}[T|\sigma]] \leq C\left(\mathbb{E}[T] + \mathbb{E}\sum_{m\in[d]} \mathbb{1}(\sigma_m \geq 4)\right) \leq C(\mathbb{E}[T] + d).$$

For the second term we have

$$\mathbb{E}[T|\sigma] = \sum_{m\in[d]} \sigma_m \mathbb{1}(\sigma_m \geq 4)\mathbb{E}[U_m|\sigma_m] = \sum_{m\in[d]} \sigma_m \mathbb{1}(\sigma_m \geq 4)\sum_{x,y}(q_{xy}(m) - q_{x\cdot}(m)q_{\cdot y}(m))^2$$

Since the σ_m are independent we have

$$\mathrm{Var}[\mathbb{E}[T|\sigma]] = \sum_{m\in[d]} \mathrm{Var}[\sigma_m \mathbb{1}(\sigma_m \geq 4)]\left(\sum_{x,y}(q_{xy}(m) - q_{x\cdot}(m)q_{\cdot y}(m))^2\right)^2$$

By Claim 2.1. of [13] we have that $\mathrm{Var}[\sigma_m \mathbb{1}(\sigma_m \geq 4)] \leq C'\mathbb{E}[\sigma_m \mathbb{1}(\sigma_m \geq 4)]$, and $\sum_{x,y}(q_{xy}(m) - q_{x\cdot}(m)q_{\cdot y}(m))^2 \leq (\sum_{x,y}|q_{xy}(m) - q_{x\cdot}(m)q_{\cdot y}(m)|)^2 \leq 4$ thus

$$\mathrm{Var}[\mathbb{E}[T|\sigma]] \leq 4C'\sum_{m\in[d]} \mathbb{E}[\sigma_m \mathbb{1}(\sigma_m \geq 4)]\sum_{x,y}(q_{xy}(m) - q_{x\cdot}(m)q_{\cdot y}(m))^2 = 4C'\mathbb{E}[T].$$

Hence we conclude $\mathrm{Var}\, T \leq C(\mathbb{E}[T] + d)$.

Putting Things Together. Recall that the threshold is set as $\tau = \zeta n^{1/5}$. First we handle the null hypothesis. By Chebyshev's inequality we have

$$\mathbb{P}(|T - \mathbb{E}T| \geq \tau) \leq \frac{\mathrm{Var}\, T}{\tau^2} = \frac{C(\mathbb{E}[T] + d)}{\tau^2} \leq \frac{C(C'n^{1/5} + n^{2/5})}{\zeta n^{2/5}} \leq \frac{1}{10},$$

when ζ is large enough, where we used the bound (C.5). In this scenario we have that $T \leq \tau + \mathbb{E}T \leq 2\tau$ for large enough ζ. Under the alternative

$$\mathbb{P}(|T - \mathbb{E}T| \geq \mathbb{E}T/2) \leq \frac{4\,\mathrm{Var}\, T}{(\mathbb{E}T)^2} \leq 4C\left(\frac{d}{(\mathbb{E}T)^2} + \frac{1}{\mathbb{E}T}\right) \leq \frac{1}{10},$$

since $\mathbb{E}T \gtrsim \zeta\sqrt{d} \gtrsim \zeta n^{1/5}$. $\qquad\square$

C.2 Proofs of Section 5.2

Proof of Lemma 5.4. We will show that the variance $\mathrm{Var}[U_W(\mathcal{D})|\mathcal{D}_{X'}, \mathcal{D}_{Y'}]$ is bounded as

$$C\left(\frac{\|p_{X',Y',A} - p_{X',Y',A}^{\Pi}\|_2^2 \max(\|p_{X',Y',A}\|_2, \|p_{X',Y',A}^{\Pi}\|_2)}{\sigma} + \frac{\max(\|p_{X',Y',A}\|_2^2, \|p_{X',Y',A}^{\Pi}\|_2^2)}{\sigma^2}\right),$$

We will now complete the proof assuming this is correct. We use the triangle inequality to obtain

$$\|p_{X',Y',A}\|_2^2 \le (\|p_{X',Y',A}^{\Pi} - p_{X',Y',A}\|_2 + \|p_{X',Y',A}^{\Pi}\|_2)^2 \le 2(\|p_{X',Y',A}^{\Pi} - p_{X',Y',A}\|_2^2 + \|p_{X',Y',A}^{\Pi}\|_2^2).$$

This gives the following bound on the variance $\mathrm{Var}[U_W(\mathcal{D})|\mathcal{D}_{X'}, \mathcal{D}_{Y'}]$

$$C\left(\frac{\|p_{X',Y',A} - p_{X',Y',A}^{\Pi}\|_2^2 \|p_{X',Y',A}^{\Pi}\|_2}{\sigma} + \frac{\|p_{X',Y',A} - p_{X',Y',A}^{\Pi}\|_2^3}{\sigma}\right.$$
$$\left. + \frac{\|p_{X',Y',A}^{\Pi}\|_2^2}{\sigma^2} + \frac{\|p_{X',Y',A} - p_{X',Y',A}^{\Pi}\|_2^2}{\sigma^2}\right), \tag{C.6}$$

which is what we wanted to show. Now it remains to show the first bound. The calculation is almost identical to the one of Lemma 5.1 (but we will repeat it for the sake of completeness). Note that the sample size $2t + 4 \ge \sigma/2$ so by adjusting the constant we can get a bound with σ in place of $2t + 4$. Going back to equation (C.1)

$$\sum_{x,y,x',y'} \frac{\left\{\mathbb{E}\phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy)\phi_{\pi_1'\pi_2'}(x'y')\phi_{\pi_3'\pi_4'}(x'y') - (\mathbb{E}\phi_{\pi_1\pi_2}(xy))^2(\mathbb{E}\phi_{\pi_1'\pi_2'}(x'y'))^2\right\}}{(1 + a_{xy})(1 + a_{x'y'})}$$

$$\le \sum_{x,y,x',y'} \frac{\left\{\mathbb{E}\phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy)\phi_{\pi_1'\pi_2'}(x'y')\phi_{\pi_3'\pi_4'}(x'y')\right\}}{(1 + a_{xy})(1 + a_{x'y'})}$$

$$= \sum_{x,y,x',y'} \frac{\left\{\mathbb{E}(\phi_{\pi_1\pi_2}(xy)\phi_{\pi_1'\pi_2'}(x'y'))\mathbb{E}\phi_{\pi_3\pi_4}(xy)\mathbb{E}\phi_{\pi_3'\pi_4'}(x'y')\right\}}{(1 + a_{xy})(1 + a_{x'y'})}$$

$$\le \sqrt{\sum_{x,y,x',y'} \frac{\{\mathbb{E}(\phi_{\pi_1\pi_2}(xy)\phi_{\pi_1'\pi_2'}(x'y'))\}^2}{(1 + a_{xy})(1 + a_{x'y'})}} \sqrt{\sum_{x,y,x',y'} \frac{\{\mathbb{E}\phi_{\pi_3\pi_4}(xy)\}^2}{(1 + a_{xy})} \frac{\{\mathbb{E}\phi_{\pi_3'\pi_4'}(x'y')\}^2}{(1 + a_{x'y'})}}$$

$$= \sqrt{\sum_{x,y,x',y'} \frac{\{\mathbb{E}(\phi_{\pi_1\pi_2}(xy)\phi_{\pi_1'\pi_2'}(x'y'))\}^2}{(1 + a_{xy})(1 + a_{x'y'})}} \|p_{X',Y',A} - p_{X',Y',A}^{\Pi}\|_2^2$$

where, as before we supposed that i is some of π_1, π_2 and π_1', π_2', and the next to last step follows from Cauchy-Schwarz. Now we apply Lemma C.1 to the first term on the RHS with $A = \phi_{\pi_1\pi_2}(xy)$ and $B = \phi_{\pi_1'\pi_2'}(x'y')$. We obtain

$$\sum_{x,y,x',y'} \frac{\{\mathbb{E}(\phi_{\pi_1\pi_2}(xy)\phi_{\pi_1'\pi_2'}(x'y'))\}^2}{(1 + a_{xy})(1 + a_{x'y'})} \le \sum_{x,y,x',y'} \frac{\mathbb{E}[|\phi_{\pi_1'\pi_2'}(x'y')|\,||\phi_{\pi_1\pi_2}(xy)| = 1]}{(1 + a_{x'y'})} \frac{(\mathbb{E}|\phi_{\pi_1\pi_2}(xy)|)^2}{(1 + a_{xy})}$$

Note that as before

$$|\phi_{\pi_1\pi_2}(xy)| \leq \mathbb{1}(X_{\pi_1} = x, Y_{\pi_1} = y) + \mathbb{1}(Y_{\pi_2} = y)\mathbb{1}(X_{\pi_1} = x)$$

and a similar inequality holds for $|\phi_{\pi'_1\pi'_2}(x'y')|$. Thus

$$\sum_{x',y'} \frac{|\phi_{\pi'_1\pi'_2}(x'y')|}{1 + a_{x'y'}} \leq \sum_{x',y'} \frac{\mathbb{1}(X_{\pi_1} = x', Y_{\pi_1} = y') + \mathbb{1}(Y_{\pi_2} = y')\mathbb{1}(X_{\pi_1} = x')}{1 + a_{x'y'}} \leq 2,$$

since $a_{x'y'} \geq 0$. Thus

$$\sum_{x,y} \mathbb{E}[\sum_{x',y'} \frac{|\phi_{\pi'_1\pi'_2}(x'y'))|}{1 + a_{x'y'}} | |\phi_{\pi_1\pi_2}(xy)| = 1] \frac{(\mathbb{E}|\phi_{\pi_1\pi_2}(xy)|)^2}{1 + a_{xy}} \leq 2 \sum_{x,y} \frac{(\mathbb{E}|\phi_{\pi_1\pi_2}(xy)|)^2}{1 + a_{xy}}$$

$$\leq 2 \sum_{x,y} \frac{(p_{X',Y'}(x,y) + p_{X'}(x)p_{Y'}(y))^2}{1 + a_{xy}} \leq 4(\|p_{X',Y',A}\|_2^2 + \|p_{X',Y',A}^{\Pi}\|_2^2),$$

which is what we wanted to show.

Now we will show how to bound any higher order terms (i.e., according to [42] the variance of the next term is governed by $\mathrm{Cov}(h^a_{ijkl}, h^a_{ijk'l'})$ where it's possible for k', l' to be equal to k or l). To bound this we directly go back to the inequality

$$\leq \sum_{x,y,x',y'} \frac{\left\{ \mathbb{E}(\phi_{\pi_1\pi_2}(xy)\phi_{\pi'_1\pi'_2}(x'y'))\mathbb{E}\phi_{\pi_3\pi_4}(xy)\mathbb{E}\phi_{\pi'_3\pi'_4}(x'y') \right\}}{(1 + a_{xy})(1 + a_{x'y'})}$$

We will now use the fact that for $A, B \in \{0, \pm 1\}$

$$\mathbb{E}AB \leq \mathbb{E}|AB| = \mathbb{E}[|B| \mid |A| = 1]\mathbb{E}|A|,$$

where $A = \phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy)$ and $B = \phi_{\pi'_1\pi'_2}(x'y')\phi_{\pi'_3\pi'_4}(x'y')$. We have

$$\sum_{x,y,x',y'} \frac{\left\{ \mathbb{E}(\phi_{\pi_1\pi_2}(xy)\phi_{\pi'_1\pi'_2}(x'y'))\mathbb{E}\phi_{\pi_3\pi_4}(xy)\mathbb{E}\phi_{\pi'_3\pi'_4}(x'y') \right\}}{(1 + a_{xy})(1 + a_{x'y'})}$$

$$\leq \sum_{x,y,x',y'} \mathbb{E}\left[\frac{|\phi_{\pi'_1\pi'_2}(x'y')\phi_{\pi'_3\pi'_4}(x'y')|}{1 + a_{x'y'}} | |\phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy)| = 1 \right] \frac{\mathbb{E}|\phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy)|}{1 + a_{xy}}$$

We now use that as we saw before

$$|\phi_{\pi_1\pi_2}(xy)| \leq \mathbb{1}(X_{\pi_1} = x, Y_{\pi_1} = y) + \mathbb{1}(Y_{\pi_1} = y)\mathbb{1}(X_{\pi_2} = x),$$

and analogously for the others. Furthermore $|\phi_{\pi_1\pi_2}(xy)| \leq 1$.

$$\sum_{x',y'} \mathbb{E}\left[\frac{|\phi_{\pi'_1\pi'_2}(x'y')\phi_{\pi'_3\pi'_4}(x'y')|}{1 + a_{x'y'}} | |\phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy)| = 1 \right] \leq \sum_{x',y'} \mathbb{E}\left[\frac{\phi_{\pi'_1\pi'_2}(x'y')}{1 + a_{x'y'}} | |\phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy)| = 1 \right]$$

$$\leq 2.$$

Next by independence,

$$
\sum_{x,y} \frac{\mathbb{E}|\phi_{\pi_1\pi_2}(xy)\phi_{\pi_3\pi_4}(xy)|}{1+a_{xy}} = \sum_{x,y} \frac{\mathbb{E}|\phi_{\pi_1\pi_2}(xy)|\mathbb{E}|\phi_{\pi_3\pi_4}(xy)|}{1+a_{xy}}
$$
$$
\leq \sum_{x,y} \frac{(p_{X',Y'}(x,y)+p_{X'}(x)p_{Y'}(y))^2}{1+a_{xy}} \leq 2\sum_{x,y} \frac{p_{X',Y'}^2(x,y)+p_{X'}^2(x)p_{Y'}^2(y)}{1+a_{xy}}
$$
$$
= 2(\|p_{X',Y',A}\|_2^2 + \|p_{X',Y',A}^\Pi\|_2^2).
$$

This completes the proof. $\qquad\square$

We now state the following Poissonized version of Theorem 5.5

Theorem C.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*

$$
\sup_{p\in\mathcal{P}'_{0,[0,1],\chi^2}(L)} \sum_{k=0}^\infty \mathbb{P}(N=k)\mathbb{E}_p[\psi_\tau(\mathcal{D}_k)] \leq \frac{1}{10},
$$
$$
\sup_{p\in\{p\in\mathcal{Q}'_{0,[0,1],\mathrm{TV}}(L):\inf_{q\in\mathcal{P}'_{0,[0,1]}}\|p-q\|_1\geq\varepsilon\}} \sum_{k=0}^\infty \mathbb{P}(N=k)\mathbb{E}_p[1-\psi_\tau(\mathcal{D}_k)] \leq \frac{1}{10}.
$$

Proof of Theorem 5.5. Using Theorem C.5 the proof is the same as that of Theorem 5.2 $\qquad\square$

Proof of Theorem C.5. To prove this theorem we will assume that $N \sim Poi(n)$ instead of $Poi(n/2)$ for convenience. Since this changes only constant factors, we can do this WLOG. As discussed in the introduction of Section 5.2, we remind the reader that we focus throughout this proof, without loss of generality, on the setting where $\sqrt{\ell_1\ell_2}/n \lesssim 1$, noting that when this condition is not satisfied there is a trivial test which is minimax optimal.

For each dataset $\mathcal{D}_m$ we will index with m all the quantities defined in the main text. For example $t_{1,m}, t_{2,m}$ will refer to the sample sizes of $\mathcal{D}_{m,X}$ and $\mathcal{D}_{m,Y}$, while t_m will be such that $\sigma_m = 4 + 4t_m$. In addition a_{xy}^m, a_x^m and $a_y'^m$ will denote the weighting amounts. Furthermore, A_m will denote the multi-set of samples where (x,y) appears a_{xy}^m times. For brevity we will refer to the weighting randomness as R_m, i.e., $R_m = \{\mathcal{D}_{m,X}, \mathcal{D}_{m,Y}\}$. Furthermore we will denote $\sigma = \{\sigma_m\}_{m\in[d]}$ and $R = \{R_m\}_{m\in[d]}$.

Let us also define

$$
T_m = \mathbb{1}(\sigma_m \geq 4)\omega_m\sigma_m U_m. \tag{C.7}
$$

Recall the definitions of $q_{xy}(m)$, $q_{x\cdot}(m)$ and $q_{\cdot y}(m)$ (C.2) and (C.3). These distributions "play the role" of $p_{X',Y'}$ and $p_{X'}$ an $p_{Y'}$ from the main text. For brevity, denote the distributions with density $q_{xy}(m)$ and $q_{x\cdot}(m)q_{y\cdot}(m)$ with $q(m)$ and $q_\Pi(m)$. Denote the A_m-split distribution $q(m)$ by $q_{A_m}(m)$ and the A_m-split distribution $q_\Pi(m)$ by $q_{\Pi,A_m}(m)$.

Before we delve into the proof we give several useful definitions and results about the weighting which we take from [13].

Definition C.6 (1-Dimensional Split Distribution). *Given a discrete distribution p over $[d]$ and a multi-set S of elements of $[d]$ define the distribution p_S over $[d + |S|]$. Let $a_i = \sum_{j \in S} \mathbb{1}(j = i)$. Thus $\sum_{i \in [d]} 1 + a_i = d + |S|$. We can therefore associate elements of the set $[d + |S|]$ with elements in the set $B_S = \{(i, j) | i \in [d], 1 \le j \le 1 + a_i\}$. The split distribution p_S is supported on B_S and is obtained by sampling i from p and j uniformly from the set $[1 + a_i]$.*

Lemma C.7 (Fact 2.2 [13]). *Let p, q are distributions over $[d]$, and S is a given multi-set of $[d]$. Then we can simulate a sample from p_S or q_S by taking a single sample from p or q. It also holds that $d_{\mathrm{TV}}(p_S, q_S) = d_{\mathrm{TV}}(p, q)$.*

Lemma C.8 (Lemma 2.3 [13]). *Let p be a discrete distribution over $[d]$. Then, for any multi-sets $S \subseteq S'$ of $[d]$, $\|p_S\|_2 \le \|p_{S'}\|_2$, and if S is obtained by m independent samples from p, then $\mathbb{E}[\|p_S\|_2^2] \le \frac{1}{m+1}$.*

An important implication of the proof of this lemma is that if a_i denotes the number of samples in S which equal to i, when S is drawn as m independent samples from p:

$$\mathbb{E}\frac{1}{1 + a_i} \le \frac{1}{(m+1)p_i}. \tag{C.8}$$

Using the independence of a_x^m and a_y^m we can therefore conclude that by (C.8)

$$\mathbb{E}\left[\frac{1}{1 + a_{xy}^m}\right] = \mathbb{E}\left[\frac{1}{1 + a_x^m}\right]\mathbb{E}\left[\frac{1}{1 + a_y'^m}\right] \le \frac{1}{(1 + t_{1,m})(1 + t_{2,m})q_{x\cdot}(m)q_{\cdot y}(m)}, \tag{C.9}$$

where the expectation above is with respect to the randomness in R_m. Hence

$$\mathbb{E}\|q_{\Pi,A_m}(m)\|_2^2 = \sum_{x,y} \mathbb{E}\frac{(q_{x\cdot}(m)q_{\cdot y}(m))^2}{1 + a_{xy}^m} \le \sum_{x,y} \frac{q_{x\cdot}(m)q_{\cdot y}(m)}{(1 + t_{1,m})(1 + t_{2,m})} = \frac{1}{(1 + t_{1,m})(1 + t_{2,m})}. \tag{C.10}$$

Analysis of the Expectation. Recall that the test statistic is

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

Recall that we denote the randomness of the flattening with R_m for bin m, the sample size within each bin as σ_m, and let us denote the randomness associated with the estimator U_m with K_m. We have

$$\mathbb{E}[T_m | \sigma_m, R_m] = \sigma_m \omega_m \|q_{A_m}(m) - q_{\Pi,A_m}(m)\|_2^2 \mathbb{1}(\sigma_m \ge 4).$$

Here we use the following bound

$$\|q_{A_m}(m) - q_{\Pi,A_m}(m)\|_2 \ge \frac{\|q_{A_m}(m) - q_{\Pi,A_m}(m)\|_1}{\sqrt{(\ell_1 + t_{1,m})(\ell_2 + t_{2,m})}} = \frac{\|q(m) - q_\Pi(m)\|_1}{\sqrt{(\ell_1 + t_{1,m})(\ell_2 + t_{2,m})}}$$

$$\ge \frac{\|q(m) - q_\Pi(m)\|_1}{2\sqrt{\ell_1 \ell_2}},$$

where the first inequality follows from Cauchy-Schwarz (or simply by the T2 Lemma) and the fact that $\sum_{x,y} 1 + a_{xy}^m = \sum_{x,y}(1 + a_x^m)(1 + a_y'^m) = (\ell_1 + t_{1,m})(\ell_2 + t_{2,m})$, and the second identity follows

55

from Lemma C.7. Denote by $\varepsilon_m = d_{\mathrm{TV}}(q(m), q_\Pi(m)) = \frac{\|q(m)-q_\Pi(m)\|_1}{2}$ for convenience. We have that

$$\mathbb{E}[T_m|\sigma_m, R_m] \geq \sigma_m \omega_m \mathbb{1}(\sigma_m \geq 4)\frac{\varepsilon_m^2}{\ell_1 \ell_2}. \tag{C.11}$$

Denote by $\alpha_m = np_m$ where $p_m = \mathbb{P}(Z \in C_m)$. We have the following lemma

Lemma C.9. *The following inequality holds*

$$\mathbb{E} \sum_{m \in [d]} \sigma_m \omega_m \mathbb{1}(\sigma_m \geq 4)\frac{\varepsilon_m^2}{\ell_1 \ell_2} \geq \gamma \sum_{m \in [d]} \frac{\varepsilon_m^2}{\ell_1 \ell_2} \min(\alpha_m \beta_m, \alpha_m^4), \tag{C.12}$$

where $\beta_m = \sqrt{\min(\alpha_m, \ell_1)\min(\alpha_m, \ell_2)}$ and γ is some absolute constant.

Proof. We have that

$$\mathbb{E} \sum_{m \in [d]} \sigma_m \omega_m \mathbb{1}(\sigma_m \geq 4)\frac{\varepsilon_m^2}{\ell_1 \ell_2} = \sum_{m \in [d]} \mathbb{E}[\sigma_m \omega_m \mathbb{1}(\sigma_m \geq 4)]\frac{\varepsilon_m^2}{\ell_1 \ell_2}$$

By Claim 2.3 of [13] we have that for $X \sim \mathrm{Poi}(\lambda)$:

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

Thus

$$\mathbb{E}[\sigma_m \omega_m \mathbb{1}(\sigma_m \geq 4)] \geq \gamma \min(\alpha_m\sqrt{\min(\alpha_m, \ell_1)\min(\alpha_m, \ell_2)}, \alpha_m^4)$$

which completes the proof. $\qquad\square$

Next suppose that $\inf_{q \in \mathcal{P}'_{0,[0,1]}} \|p_{X,Y,Z} - q\|_1 > \varepsilon$. We want to show some lower bounds on the RHS of (C.12). We start by looking into the expression

$$\sum_{m \in [d]} \frac{\varepsilon_m}{\sqrt{\ell_1 \ell_2}}\alpha_m. \tag{C.13}$$

Recall that $\varepsilon_m = d_{\mathrm{TV}}(q(m), q_\Pi(m))$. To obtain a lower bound we take

$$z_m^* \in \underset{z \in C_m}{\operatorname{argmax}} \|p_{X,Y|Z=z} - p_{X|Z=z}p_{Y|Z=z}\|_1,$$

where just as before in the proof of Lemma C.4 we can show that the map $z \mapsto \operatorname{argmax}_{z \in C_m} \|p_{X,Y|Z=z} - p_{X|Z=z}p_{Y|Z=z}\|_1$ is continuous. The rest of the proof is very similar to Lemma C.4 but we provide full details for completeness. By the triangle inequality we have

$$2\varepsilon_m \geq \|p_{X,Y|Z=z_m^*} - p_{X|Z=z_m^*}p_{Y|Z=z_m^*}\|_1 - \|q(m) - p_{X,Y|Z=z_m^*}\|_1$$
$$- \|q_{X\cdot}(m)(q_{\cdot Y}(m) - p_{Y|Z=z_m^*})\|_1 - \|p_{Y|Z=z_m^*}(q_{X\cdot}(m) - p_{X|Z=z_m^*})\|_1,$$

where we denoted the distributions with density $q_{x.}(m)$: $q_{X.}(m)$ and similarly for $q_{.Y}(m)$. Now we will bound the three terms. We start with

$$\|q(m) - p_{X,Y|Z=z_m^*}\|_1 = \sum_{x,y} \left| \int_{C_m} p_{X,Y|Z}(x,y|z) - p_{X,Y|Z}(x,y|z_m^*) d\widetilde{P}(z) \right|$$

$$\leq \int_{C_m} \sum_{x,y} |p_{X,Y|Z}(x,y|z) - p_{X,Y|Z}(x,y|z_m^*)| d\widetilde{P}(z)$$

$$\leq \int_{C_m} L|z - z_m^*| d\widetilde{P}(z) \leq L \operatorname{diam}(C_m) = \frac{L}{d},$$

where $d\widetilde{P}(z) = \frac{dP(z)}{\mathbb{P}(Z \in A_m)}$, is the conditional distribution of $Z \in C_m$. Similarly

$$\|q_{X.}(m)(q_{.Y}(m) - p_{Y|z_m^*})\|_1 = \sum_{x,y} \left| \int_{C_m} q_{x.}(m)(p_{Y|Z}(y|z) - p_{Y|Z}(y|z_m^*)) d\widetilde{P}(z) \right|$$

$$\leq \int_{C_m} \sum_{x,y} q_{x.}(m)|p_{Y|Z}(y|z) - p_{Y|Z}(y|z_m^*)| d\widetilde{P}(z)$$

$$\leq L \operatorname{diam}(C_m) = \frac{L}{d}.$$

The last term is similar to the previous term so we conclude that

$$\varepsilon_m \geq d_{\mathrm{TV}}(p_{X,Y|Z=z_m^*}, p_{X|Z=z_m^*} p_{Y|Z=z_m^*}) - \frac{3}{2}\frac{L}{d}.$$

Therefore

$$\sum_{m \in [d]} \frac{\varepsilon_m}{\sqrt{\ell_1 \ell_2}} \alpha_m \geq \frac{n}{\sqrt{\ell_1 \ell_2}} \left(\sum_m p_m d_{\mathrm{TV}}(p_{X,Y|Z=z_m^*}, p_{X|Z=z_m^*} p_{Y|Z=z_m^*}) - \frac{3}{2}\frac{L}{d} \right)$$

$$\geq \frac{n}{\sqrt{\ell_1 \ell_2}} \left(\frac{1}{2} \mathbb{E}_Z \|p_{XY|Z} - p_{X|Z} p_{Y|Z}\|_1 - \frac{3}{2}\frac{L}{d} \right)$$

$$=: \frac{n}{\sqrt{\ell_1 \ell_2}} \eta \ {}^{7}$$

Now, as in the proof of Theorem 5.2, the analysis is partitioned into two parts. The first part takes the set $M_H = \{m | \alpha_m^3 \geq \beta_m\}$ and $M_L = \{m | \alpha_m^3 < \beta_m\}$. By the above analysis we know that either $\sum_{m \in M_H} \frac{\varepsilon_m}{\sqrt{\ell_1 \ell_2}} \alpha_m \geq \frac{n}{2\sqrt{\ell_1 \ell_2}} \eta$ or $\sum_{m \in M_L} \frac{\varepsilon_m}{\sqrt{\ell_1 \ell_2}} \alpha_m \geq \frac{n}{2\sqrt{\ell_1 \ell_2}} \eta$. In the first case we want to lower bound

$$\sum_{m \in M_H} \frac{\varepsilon_m^2}{\ell_1 \ell_2} \alpha_m \beta_m.$$

In order to determine the value of β_m we consider three more cases. Suppose without loss of generality that $\ell_2 \leq \ell_1$. Define the three sets $M_{H,1} = \{m | \ell_2 \leq \ell_1 \leq \alpha_m\}$, $M_{H,2} = \{m | \ell_2 \leq \alpha_m \leq \ell_1\}$, $M_{H,3} = \{m | \alpha_m \leq \ell_2 \leq \ell_1\}$. We have that

$$\max_{i \in [3]} \sum_{m \in M_{H,i}} \frac{\varepsilon_m}{\sqrt{\ell_1 \ell_2}} \alpha_m \geq \frac{n}{6\sqrt{\ell_1 \ell_2}} \eta$$

We now analyze the three cases depending on where the maximum above is achieved.

---

[7]Note here that $\mathbb{E}_Z \|p_{X,Y|Z} - p_{X|Z} p_{Y|Z}\|_1 = \|p_{X,Y,Z} - p_{X|Z} p_{Y|Z} p_Z\|_1 \geq \varepsilon$

- Suppose that the maximum is achieved at $i = 1$. Thus

$$\sum_{m \in M_{H,1}} \frac{\varepsilon_m}{\sqrt{\ell_1 \ell_2}} \alpha_m \geq \frac{n}{6\sqrt{\ell_1 \ell_2}} \eta.$$

We have in this subcase that $\beta_m = \sqrt{\ell_1 \ell_2}$. Therefore

$$\sum_{m \in M_{H,1}} \frac{\varepsilon_m^2}{\ell_1 \ell_2} \alpha_m \beta_m = \sqrt{\ell_1 \ell_2} \sum_{m \in M_{H,1}} \frac{\varepsilon_m^2}{\ell_1 \ell_2} \alpha_m \geq \sqrt{\ell_1 \ell_2} \frac{\left(\sum_{m \in M_{H,1}} \frac{\varepsilon_m \alpha_m}{\sqrt{\ell_1 \ell_2}}\right)^2}{\sum_{m \in M_{H,1}} \alpha_m}$$

$$\geq \sqrt{\ell_1 \ell_2} \frac{n^2 \eta^2}{36 \ell_1 \ell_2 (\sum_{m \in M_{H,1}} \alpha_m)}.$$

Now we have that $\sum_{m \in M_{H,1}} \alpha_m = \sum_{m \in M_{H,1}} np_m \leq n \sum_{m \in [d]} p_m = n$. Thus we conclude

$$\sum_{m \in M_{H,1}} \frac{\varepsilon_m^2}{\ell_1 \ell_2} \alpha_m \beta_m \geq \frac{n \eta^2}{36\sqrt{\ell_1 \ell_2}}.$$

- In the second case we have that the maximum is achieved in $i = 2$ which means

$$\sum_{m \in M_{H,2}} \frac{\varepsilon_m}{\sqrt{\ell_1 \ell_2}} \alpha_m \geq \frac{n}{6\sqrt{\ell_1 \ell_2}} \eta.$$

In this case $\beta_m = \sqrt{\ell_2} \sqrt{\alpha_m}$, so we have

$$\sum_{m \in M_{H,2}} \frac{\varepsilon_m^2}{\ell_1 \ell_2} \alpha_m \beta_m = \sqrt{\ell_2} \sum_{m \in M_{H,2}} \frac{\varepsilon_m^2}{\ell_1 \ell_2} \alpha_m^{3/2} \geq \sqrt{\ell_2} \frac{\left(\sum_{m \in M_{H,2}} \frac{\varepsilon_m}{\sqrt{\ell_1 \ell_2}} \alpha_m\right)^2}{\sum_{m \in M_{H,2}} \sqrt{\alpha_m}}.$$

To bound $\sum_{m \in M_{H,2}} \sqrt{\alpha_m}$ we note that

$$\sum_{m \in M_{H,2}} \sqrt{\alpha_m} \leq \sum_{m \in [d]} \sqrt{n} \sqrt{p_m} \leq \sqrt{nd}.$$

On the other hand, we also have that $\ell_2 \leq \alpha_m \leq \ell_1$ so that $n \geq \sum_{m \in M_{H,2}} \alpha_m \geq \ell_2 |M_{H,2}|$, so that $|M_{H,2}| \leq \frac{n}{\ell_2}$. Therefore $\sum_{m \in M_{H,2}} \sqrt{\alpha_m} \leq |M_{H,2}| \sqrt{\ell_1} \leq \frac{n\sqrt{\ell_1}}{\ell_2}$. Thus

$$\sum_{m \in M_{H,2}} \frac{\varepsilon_m^2}{\ell_1 \ell_2} \alpha_m \beta_m \geq \sqrt{\ell_2} \frac{\left(\sum_{m \in M_{H,2}} \frac{\varepsilon_m}{\sqrt{\ell_1 \ell_2}} \alpha_m\right)^2}{\min(\sqrt{nd}, \frac{n\sqrt{\ell_1}}{\ell_2})} \geq \frac{n^{3/2} \eta^2}{36 \ell_1 \sqrt{\ell_2} \min(\sqrt{d}, \frac{\sqrt{n\ell_1}}{\ell_2})}.$$

Note that in this case we have to have

$$\frac{n\eta}{6\sqrt{\ell_1 \ell_2}} \leq \sum_{m \in M_{H,2}} \frac{\varepsilon_m}{\sqrt{\ell_1 \ell_2}} \alpha_m \leq \frac{|M_{H,2}|}{\sqrt{\ell_1 \ell_2}} \ell_1 \leq \frac{d\ell_1}{\sqrt{\ell_1 \ell_2}},$$

where we used the fact that $\varepsilon_m \leq 1$. Thus this case cannot happen when $n\eta \geq 6d\ell_1$.

58

- In the last sub-case we have that $\beta_m = \alpha_m$, and the maximum is achieved at $i = 3$ so that we have

$$\sum_{m \in M_{H,3}} \frac{\varepsilon_m}{\sqrt{\ell_1 \ell_2}} \alpha_m \geq \frac{n}{6\sqrt{\ell_1 \ell_2}} \eta.$$

By the AM-GM inequality we have that

$$\sum_{m \in M_{H,3}} \frac{\varepsilon_m^2}{\ell_1 \ell_2} \alpha_m \beta_m = \sum_{m \in M_{H,3}} \frac{\varepsilon_m^2}{\ell_1 \ell_2} \alpha_m^2 \geq \frac{\left(\sum_{m \in M_{H,3}} \frac{\varepsilon_m \alpha_m}{\sqrt{\ell_1 \ell_2}} \right)^2}{|M_{H,3}|} \geq \frac{n^2 \eta^2}{36 \ell_1 \ell_2 |M_{H,3}|}.$$

We have a simple bound on the cardinality $|M_{H,3}| \leq d$ (since it cannot be more than the total number of categories). Thus we conclude the bound

$$\sum_{m \in M_{H,3}} \frac{\varepsilon_m^2}{\ell_1 \ell_2} \alpha_m \beta_m = \sum_{m \in M_{H,3}} \frac{\varepsilon_m^2}{\ell_1 \ell_2} \alpha_m^2 \geq \frac{n^2 \eta^2}{36 \ell_1 \ell_2 d}.$$

Similarly to before in this case we must have

$$\frac{n\eta}{6\sqrt{\ell_1 \ell_2}} \leq \sum_{m \in M_{H,3}} \frac{\varepsilon_m}{\sqrt{\ell_1 \ell_2}} \alpha_m \leq \frac{|M_{H,3}|}{\sqrt{\ell_1 \ell_2}} \ell_2 \leq \frac{d\ell_2}{\sqrt{\ell_1 \ell_2}},$$

i.e., when $n\eta \geq 6d\ell_2$.

Finally we handle the last case in which we have

$$\sum_{m \in M_L} \frac{\varepsilon_m}{\sqrt{\ell_1 \ell_2}} \alpha_m \geq \frac{n}{2\sqrt{\ell_1 \ell_2}} \eta.$$

Here we have that, by Jensen's inequality:

$$\frac{1}{\ell_1 \ell_2} \sum_{m \in M_L} \varepsilon_m^2 \alpha_m^4 \geq \frac{1}{\ell_1 \ell_2} \frac{\left(\sum_{m \in M_L} \varepsilon_m \alpha_m \right)^4}{\left(\sum_{m \in M_L} \varepsilon_m^{2/3} \right)^3} \geq \frac{n^4 \eta^4}{\ell_1 \ell_2 16 \left(\sum_{m \in M_L} \varepsilon_m^{2/3} \right)^3}.$$

Now using that $\varepsilon_m \leq 1$ we have $\sum_{m \in M_L} \varepsilon_m^{2/3} \leq |M_L| \leq d$ thus we conclude that

$$\frac{1}{\ell_1 \ell_2} \sum_{m \in M_L} \varepsilon_m^2 \alpha_m^4 \geq \frac{n^4 \eta^4}{16 \ell_1 \ell_2 d^3}.$$

Finally in this case we have to have

$$\frac{n\eta}{6\sqrt{\ell_1 \ell_2}} \leq \sum_{m \in M_L} \frac{\varepsilon_m}{\sqrt{\ell_1 \ell_2}} \alpha_m \leq \frac{|M_L|}{\sqrt{\ell_1 \ell_2}} 1 \leq \frac{d}{\sqrt{\ell_1 \ell_2}},$$

or equivalently when $n\eta \geq 6d$ where we used that $\alpha_m^3 \leq \beta_m$ implies that $\alpha_m \leq 1$.

We will now select a threshold at the level of $\zeta\sqrt{d}$, and will give conditions on the minimum sample size for each of the cases. We will use $\gtrsim$ in the sense bigger up to an absolute constant. We will assume that $\frac{\varepsilon}{2} - \frac{3}{2}\frac{L}{d} \geq \frac{\varepsilon}{4}$ so that $\eta \geq \frac{\varepsilon}{4}$.

- In the first sub-case we have to satisfy $\frac{n\eta^2}{\sqrt{\ell_1\ell_2}} \gtrsim \sqrt{\zeta d}$. This is ensured when

$$\varepsilon \gtrsim \sqrt{\frac{\sqrt{\zeta d \ell_1 \ell_2}}{n}} \vee \frac{1}{d}. \tag{C.14}$$

- In the second sub-case we have $\frac{n^{3/2}\eta^2}{\ell_1\sqrt{\ell_2}\frac{\sqrt{n\ell_1}}{\ell_2}} \gtrsim \sqrt{\zeta d}$ or $\frac{n^{3/2}\eta^2}{\ell_1\sqrt{\ell_2}\sqrt{d}} \gtrsim \sqrt{\zeta d}$, This is implied when

$$\varepsilon \gtrsim \min\left(\sqrt{\frac{\sqrt{\zeta d}\ell_1^{3/2}}{\sqrt{\ell_2}n}}, \sqrt{\frac{\sqrt{\zeta d}\ell_1\sqrt{\ell_2}}{n^{3/2}}}, \frac{d\ell_1}{n}\right) \vee \frac{1}{d}. \tag{C.15}$$

- In the third sub-case case we need $\frac{n^2\eta^2}{36\ell_1\ell_2 d} \gtrsim \sqrt{\zeta d}$ which happens when

$$\varepsilon \gtrsim \min\left(\frac{\zeta^{1/4}d^{3/4}\sqrt{\ell_1\ell_2}}{n}, \frac{d\ell_2}{n}\right) \vee \frac{1}{d}, \tag{C.16}$$

  where the last condition enforces when this case is not feasible.

- Finally in the second case we need $\frac{n^4\eta^4}{16\ell_1\ell_2 d^3} \gtrsim \sqrt{\zeta d}$, which is implied when

$$\varepsilon \gtrsim \min\left(\frac{\zeta^{1/8}d^{7/8}(\ell_1\ell_2)^{1/4}}{n}, \frac{d}{n}\right) \vee \frac{1}{d} \tag{C.17}$$

Analysis of the Variance. Now we derive a bound on the variance of the statistic

$$\sum_{m\in[d]} \sigma_m \omega_m \mathbb{1}(\sigma_m \geq 4)\frac{\varepsilon_m^2}{\ell_1\ell_2}$$

Recall now that $\sigma_m \sim \text{Poi}(\alpha_m)$ are independent and therefore

$$\text{Var}\left[\sum_{m\in[d]} \sigma_m \omega_m \mathbb{1}(\sigma_m \geq 4)\frac{\varepsilon_m^2}{\ell_1\ell_2}\right] = \sum \text{Var}\left[\sigma_m \omega_m \mathbb{1}(\sigma_m \geq 4)\frac{\varepsilon_m^2}{\ell_1\ell_2}\right]$$

$$= \sum_{m\in[d]} \frac{\varepsilon_m^4}{\ell_1^2\ell_2^2} \text{Var}(\sigma_m \omega_m \mathbb{1}(\sigma_m \geq 4))$$

$$\leq \frac{C'}{\ell_1\ell_2} \sum_{m\in[d]} \frac{\varepsilon_m^2}{\ell_1\ell_2} \mathbb{E}(\sigma_m \omega_m \mathbb{1}(\sigma_m \geq 4))$$

$$\leq \frac{C'}{\ell_1\ell_2} \mathbb{E}\left[\sum_{m\in[d]} \sigma_m \omega_m \mathbb{1}(\sigma_m \geq 4)\frac{\varepsilon_m^2}{\ell_1\ell_2}\right], \tag{C.18}$$

where in the next to last inequality we used Claim 2.2. of [13], and C' is an absolute constant described in that claim.

We now bound the variance of the statistic T (recall the definition (5.6)). Since T_m (recall definition C.7) are independent given σ_m, R_m we have that

$$\text{Var}[T|\sigma, R] = \sum_{m\in[d]} \text{Var}[T_m|\sigma_m, R_m].$$

Next, by definition of T_m we have that $\mathrm{Var}[T_m|\sigma_m, R_m] = \sigma_m^2 \omega_m^2 \mathbb{1}(\sigma_m \geq 4)\,\mathrm{Var}[U_m|R_m]$. Using the bound on the variance $\mathrm{Var}[U_m|R_m]$ (C.6), we have to control four terms. We do so below. Denote

$$E := \sum_{m \in [d]} \omega_m^2 \|q_{\Pi, A_m}(m)\|_2^2 \mathbb{1}(\sigma_m \geq 4)$$

The first term we need to control is

$$\sum_{m \in [d]} \sigma_m^2 \omega_m^2 \mathbb{1}(\sigma_m \geq 4) \frac{\|q_{A_m}(m) - q_{\Pi, A_m}(m)\|_2^2 \|q_{\Pi, A_m}(m)\|_2}{\sigma_m}$$

$$\leq \sqrt{\left(\sum_{m \in [d]} \omega_m^2 \|q_{\Pi, A_m}(m)\|_2^2 \mathbb{1}(\sigma_m \geq 4) \right)} \sqrt{\sum_{m \in [d]} (\sigma_m \omega_m \|q_{A_m}(m) - q_{\Pi, A_m}(m)\|_2^2)^2 \mathbb{1}(\sigma_m \geq 4)}$$

$$\leq E^{1/2} \sum_{m \in [d]} \sigma_m \omega_m \|q_{A_m}(m) - q_{\Pi, A_m}(m)\|_2^2 \mathbb{1}(\sigma_m \geq 4)$$

$$= E^{1/2} \mathbb{E}[T|\sigma, R],$$

where we used Cauchy-Schwarz and the monotonicity of L_p norms. The second term is

$$\sum_{m \in [d]} \sigma_m^2 \omega_m^2 \mathbb{1}(\sigma_m \geq 4) \frac{\|q_{A_m}(m) - q_{\Pi, A_m}(m)\|_2^3}{\sigma_m} = \sum_{m \in [d]} \sqrt{\frac{\omega_m}{\sigma_m}} \sigma_m^{3/2} \omega_m^{3/2} \mathbb{1}(\sigma_m \geq 4) \|q_{A_m}(m) - q_{\Pi, A_m}(m)\|_2^3$$

$$\leq \sum_{m \in [d]} (\sigma_m \omega_m \mathbb{1}(\sigma_m \geq 4) \|q_{A_m}(m) - q_{\Pi, A_m}(m)\|_2^2)^{3/2}$$

$$\leq \left(\sum_{m \in [d]} \sigma_m \omega_m \mathbb{1}(\sigma_m \geq 4) \|q_{A_m}(m) - q_{\Pi, A_m}(m)\|_2^2 \right)^{3/2}$$

$$= \mathbb{E}[T|\sigma, R]^{3/2},$$

where we used that $\omega_m \leq \sigma_m$ by definition and the monotonicity of the L_p norms. The third term is

$$\sum_{m \in [d]} \sigma_m^2 \omega_m^2 \mathbb{1}(\sigma_m \geq 4) \frac{\|q_{\Pi, A_m}(m)\|_2^2}{\sigma_m^2} = E.$$

Finally the fourth term is

$$\sum_{m \in [d]} \sigma_m^2 \omega_m^2 \mathbb{1}(\sigma_m \geq 4) \frac{\|q_{A_m}(m) - q_{\Pi, A_m}(m)\|_2^2}{\sigma_m^2} \leq \sum_{m \in [d]} \omega_m \sigma_m \mathbb{1}(\sigma_m \geq 4) \|q_{A_m}(m) - q_{\Pi, A_m}(m)\|_2^2$$

$$= \mathbb{E}[T|\sigma, R].$$

We conclude that

$$\mathrm{Var}[T|\sigma, R] \leq C(E + (E^{1/2} + 1)\mathbb{E}[T|\sigma, R] + \mathbb{E}[T|\sigma, R]^{3/2}). \tag{C.19}$$

Now we will show that $\mathbb{E}[E|\sigma] = O(\min(d, N))$. We start by analyzing the expectation of one term from E below.

$$\mathbb{E}[\omega_m^2 \|q_{\Pi, A_m}(m)\|_2^2 \mathbb{1}(\sigma_m \geq 4)|\sigma_m] = \omega_m^2 \mathbb{1}(\sigma_m \geq 4)\mathbb{E}[\|q_{\Pi, A_m}(m)\|_2^2 |\sigma_m] \leq \frac{\omega_m^2 \mathbb{1}(\sigma_m \geq 4)}{(1 + t_{1,m})(1 + t_{2,m})},$$

where we applied (C.10). Recall that $t_{i,m} = \min((\sigma_m - 4)/4, \ell_i)$ and $\omega_m^2 = \min(\sigma_m, \ell_1)\min(\sigma_m, \ell_2)$. Thus

$$\frac{\omega_m^2}{(1 + t_{1,m})(1 + t_{2,m})} \leq O(1).$$

We conclude that

$$\mathbb{E}[E|\sigma] = \mathbb{E}[\sum_{m \in [d]} \omega_m^2 \|q_{\Pi, A_m}(m)\|_2^2 \mathbb{1}(\sigma_m \geq 4)|\sigma_m] \leq O(1) \sum_{m \in [d]} \mathbb{1}(\sigma_m \geq 4) \leq O(1)\min(d, N).$$

$$(\text{C.20})$$

We have the following result

Lemma C.10. *Suppose* $\inf_{q \in \mathcal{P}'_{0,[0,1]}} \|p_{X,Y,Z} - q\|_1 \geq \varepsilon$, *where* $\varepsilon \geq 3\frac{L}{d}$ *and it satisfies conditions* (C.14), (C.15), (C.16) *and* (C.17). *Then with probability at least* $19/20$ *over* σ, R *we have* $\mathbb{E}[T|\sigma, R] = \Omega(\sqrt{\zeta d})$ *and*

$$\mathrm{Var}[T|\sigma, R] \leq O(d + (\sqrt{d} + 1)\mathbb{E}[T|\sigma, R] + \mathbb{E}[T|\sigma, R]^{3/2}). \tag{C.21}$$

Proof. Let

$$D = \sum_{m \in [d]} \sigma_m \omega_m \mathbb{1}(\sigma_m \geq 4)\frac{\varepsilon_m^2}{\ell_1 \ell_2}.$$

We first showed that $\mathbb{E}[T|\sigma, R] \geq D$ for all σ, R (C.11). We also derived that $\mathrm{Var}[D] \leq O(\mathbb{E}[D]/(\ell_1 \ell_2))$ (C.18), and that for the selected regimes of sample size $\mathbb{E}[D] \gtrsim \sqrt{\zeta d}$. Therefore we have

$$\mathbb{P}_{\sigma, R}\left(\mathbb{E}[T|\sigma, R] \leq \kappa\sqrt{\zeta d}\right) \leq \mathbb{P}_{\sigma, R}(D \leq O(\mathbb{E}[D])) \leq O\left(\frac{\mathrm{Var}[D]}{(\mathbb{E}[D])^2}\right) = O(1/(\sqrt{\zeta d}\ell_1 \ell_2)) \leq 1/40,$$

for some small enough absolute constant κ. For the second statement we will use bound (C.19). By (C.20) we have

$$\mathbb{E}[\mathbb{E}[E|\sigma]] \leq O(1)\mathbb{E}\min(d, N) \leq O(1)\min(d, n).$$

Thus by Markov's inequality $E \leq 200\mathbb{E}[E] = O(1)d$ with probability at least $39/40$. Therefore

$$\mathbb{P}_{\sigma, R}(\mathrm{Var}[T|\sigma, R] \geq \kappa'(d + (\sqrt{d} + 1)\mathbb{E}[T|\sigma, R] + \mathbb{E}[T|\sigma, R]^{3/2})) \leq 1/40.$$

A union bound over the two events completes the proof. $\square$

We now turn to bound the expectation and variance under the null hypothesis.

Lemma C.11. *Suppose* $p_{X,Y,Z} \in \mathcal{P}'_{0,[0,1],\chi^2}(L)$. *Let further* $\ell_1 \geq \ell_2$ *be such that* $\ell_1 d \lesssim n$ (*here* $\lesssim$ *means smaller up to an absolute constant*). *Then with probability at least* $19/20$ *we have* $\mathbb{E}[T|\sigma, R] \leq Cn\frac{L^2}{\sqrt{\ell_1 \ell_2}d^2}$ *and the variance* $\mathrm{Var}[T|\sigma, R]$ *satisfies* (C.21).

Proof. Let us start with bounding $\mathbb{E}[T|\sigma, R]$ from above. Recall that

$$\mathbb{E}[T|\sigma, R] = \sum_{m \in [d]} \sigma_m \omega_m \|q_{A_m}(m) - q_{\Pi, A_m}(m)\|_2^2 \mathbb{1}(\sigma_m \geq 4)$$

We will now control $\mathbb{E}[\|q_{A_m}(m) - q_{\Pi, A_m}(m)\|_2^2 | \sigma_m]$. Recall that

$$\|q_{A_m}(m) - q_{\Pi, A_m}(m)\|_2^2 = \sum_{x,y} \frac{(q_{xy}(m) - q_{x\cdot}(m)q_{\cdot y}(m))^2}{1 + a_{xy}^m}. \tag{C.22}$$

Since by (C.9) we have

$$\mathbb{E}_{A_m} \frac{1}{1 + a_{xy}^m} \leq \frac{1}{(1 + t_{1,m})(1 + t_{2,m})q_{x\cdot}(m)q_{\cdot y}(m)} = \frac{O(1)}{\omega_m^2 q_{x\cdot}(m)q_{\cdot y}(m)},$$

we have that

$$\mathbb{E}_{A_m} \|q_{A_m}(m) - q_{\Pi, A_m}(m)\|_2^2 \leq \frac{O(1)}{\omega_m^2} \sum_{x,y} \frac{(q_{xy}(m) - q_{x\cdot}(m)q_{\cdot y}(m))^2}{q_{x\cdot}(m)q_{\cdot y}(m)}. \tag{C.23}$$

We will now focus on controlling the RHS. Using the fact that under the null hypothesis $p_{x,y|z=z} = p_{X|Z=z}p_{Y|Z=z}$ we have

$$\sum_{x,y} \frac{(q_{xy}(m) - q_{x\cdot}(m)q_{\cdot y}(m))^2}{q_{x\cdot}(m)q_{\cdot y}(m)}$$

$$= \sum_{x,y} \frac{(\int (p_{X|Z}(x|z) - \int p_{X|Z}(x|z)d\widetilde{P}(z))(p_{Y|Z}(y|z) - \int p_{Y|Z}(y|z)d\widetilde{P}(z))d\widetilde{P}(z))^2}{\int p_{X|Z}(x|z)d\widetilde{P}(z) \int p_{Y|Z}(y|z)d\widetilde{P}(z)}$$

$$\leq \sum_x \frac{\int p_{X|Z}^2(x|z)d\widetilde{P}(z) - (\int p_{X|Z}(x|z)d\widetilde{P}(z))^2}{\int p_{X|Z}(x|z)d\widetilde{P}(z)} \sum_y \frac{\int p_{Y|Z}^2(y|z)d\widetilde{P}(z) - (\int p_{Y|Z}(y|z)d\widetilde{P}(z))^2}{\int p_{Y|Z}(y|z)d\widetilde{P}(z)}.$$

We now handle the first term on the RHS, the second one being analogous.

$$\sum_x \frac{\int p_{X|Z}^2(x|z)d\widetilde{P}(z) - (\int p_{X|Z}(x|z)d\widetilde{P}(z))^2}{\int p_{X|Z}(x|z)d\widetilde{P}(z)} = \sum_x \frac{\int p_{X|Z}^2(x|z)d\widetilde{P}(z)}{\int p_{X|Z}(x|z)d\widetilde{P}(z)} - 1$$

$$= \int \sum_x \frac{p_{X|Z}^2(x|z)}{\int p_{X|Z}(x|z)d\widetilde{P}(z)} d\widetilde{P}(z) - 1.$$

We now show that the function $z \mapsto \sum_x \frac{p_{X|Z}^2(x|z)}{\int p_{X|Z}(x|z)d\widetilde{P}(z)}$ is continuous. This follows from the fact that each of $z \mapsto p_{X|Z}(x|z)$ is continuous. To see why $p_{X|Z}(x|z)$ is continuous first recall that $\|p_{X|Z=z} - p_{X|Z=z'}\|_1^2 \leq d_{\chi^2}(p_{X|z}, p_{X|z'}) \leq L|z - z'|$, which shows the continuity of each of $p_{X|Z}(x|z)$ for all x. Hence by the mean value theorem we have that

$$\sum_x \frac{\int p_{X|Z}^2(x|z)d\widetilde{P}(z)}{\int p_{X|Z}(x|z)d\widetilde{P}(z)} - 1 = \int \sum_x \frac{p_{X|Z}^2(x|z)}{\int p_{X|Z}(x|z)d\widetilde{P}(z)}d\widetilde{P}(z) - 1 = \sum_x \frac{p_{X|Z}^2(x|\widetilde{z})}{\int p_{X|Z}(x|z)d\widetilde{P}(z)} - 1,$$

63

for some $\widetilde{z} \in C_m$. Next since $x \mapsto \frac{1}{x}$ is convex on the positive reals, by Jensen's inequality we have

$$\sum_x \frac{p_{X|Z}^2(x|\widetilde{z})}{\int p_{X|Z}(x|z)d\widetilde{P}(z)} - 1 \leq \int \sum_x \frac{p_{X|Z}^2(x|\widetilde{z})}{p_{X|Z}(x|z)} - 1 d\widetilde{P}(z) = \int d_{\chi^2}(p_{x|\widetilde{z}}, p_{x|z})d\widetilde{P}(z)$$

$$\leq L \operatorname{diam}(C_m) = \frac{L}{d}.$$

We get a similar bound on the second term which implies that

$$\sum_{x,y} \frac{(q_{xy}(m) - q_{x\cdot}(m)q_{\cdot y}(m))^2}{q_{x\cdot}(m)q_{\cdot y}(m)} \leq \frac{L^2}{d^2}.$$

Hence, combining the last observation with (C.23) we have

$$\mathbb{E}[T|\sigma] = \sum_{m \in [d]} \sigma_m \omega_m \mathbb{1}(\sigma_m \geq 4)\mathbb{E}_{A_m}\|q_{A_m}(m) - q_{\Pi, A_m}(m)\|_2^2$$

$$\leq C\frac{L^2}{d^2} \sum_{m \in [d]} \frac{\sigma_m}{\omega_m}\mathbb{1}(\sigma_m \geq 4).$$

Recall that $\omega_m = \sqrt{\min(\sigma_m, \ell_1)\min(\sigma_m, \ell_2)}$. Recall that we are supposing (without loss of generality) $\ell_1 \geq \ell_2$. We therefore have

$$\sum_{m \in [d]} \frac{\sigma_m}{\omega_m}\mathbb{1}(\sigma_m \geq 4) \leq \sum_{m:\sigma_m \leq \ell_2} \mathbb{1}(\sigma_m \geq 4) + \sum_{m:\ell_2 < \sigma_m \leq \ell_1} \frac{\sqrt{\sigma_m}}{\sqrt{\ell_2}}\mathbb{1}(\sigma_m \geq 4) + \sum_{m:\sigma_m > \ell_1} \frac{\sigma_m}{\sqrt{\ell_1\ell_2}}\mathbb{1}(\sigma_m \geq 4)$$

$$\leq d + d\sqrt{\frac{\ell_1}{\ell_2}} + \frac{N}{\sqrt{\ell_1\ell_2}}$$

Taking expectation it follows that

$$\mathbb{E}[T] \leq C\frac{L^2}{d^2}\left(d + d\sqrt{\frac{\ell_1}{\ell_2}} + \frac{n}{\sqrt{\ell_1\ell_2}}\right) \leq C\frac{L^2}{d^2}\left(2d\sqrt{\frac{\ell_1}{\ell_2}} + \frac{n}{\sqrt{\ell_1\ell_2}}\right) \leq \frac{C'n}{d^2\sqrt{\ell_1\ell_2}},$$

for some constant C' which depends on L, and we used the fact that $d\ell_1 \lesssim n$. It follows from Markov's inequality that

$$\mathbb{P}_{\sigma,R}\left(\mathbb{E}[T|\sigma, R] \geq 40\frac{C'n}{d^2\sqrt{\ell_1\ell_2}}\right) \leq \frac{1}{40}.$$

The second part follows directly by Lemma C.10. $\qquad \square$

Putting Things Together. For what follows suppose that d is selected so that

$$\frac{n}{d^2\sqrt{\ell_1\ell_2}} \asymp \sqrt{d}. \tag{C.24}$$

Lemma C.12. *If $p_{X,Y,Z} \in \mathcal{P}'_{0,[0,1],\chi^2}(L)$ and that (C.24) holds. Then for a sufficiently large absolute constant α we have*

$$\mathbb{P}\left(T \geq (\alpha + 1)\frac{C'n}{d^2\sqrt{\ell_1\ell_2}}\right) \leq \frac{1}{10}.$$

64

Proof. Let $T' = (T|\sigma, R)$. Denote the event from Lemma C.11 with $\mathcal{E}$. Then we have

$$\mathbb{P}\left(T \geq (\alpha+1)\frac{C'n}{d^2\sqrt{\ell_1\ell_2}}\right) = \mathbb{P}\left(T' \geq (\alpha+1)\frac{C'n}{\sqrt{\ell_1\ell_2}d^2}\right)$$

$$\leq \mathbb{P}\left(T' \geq (\alpha+1)\frac{C'n}{\sqrt{\ell_1\ell_2}d^2}\bigg|\mathcal{E}\right) + \mathbb{P}(\mathcal{E}^c).$$

Now we have

$$\mathbb{P}\left(T' \geq (\alpha+1)\frac{C'n}{\sqrt{\ell_1\ell_2}d^2}\bigg|\mathcal{E}\right) \leq \mathbb{P}\left(T' - \mathbb{E}[T|\sigma, R] \geq \alpha\frac{C'n}{\sqrt{\ell_1\ell_2}d^2}\bigg|\mathcal{E}\right) \leq \frac{\mathrm{Var}[T'|\mathcal{E}]}{(\alpha\frac{C'n}{\sqrt{\ell_1\ell_2}d^2})^2}$$

$$\leq \frac{O(d + (\sqrt{d}+1)\mathbb{E}[T|\sigma, R] + \mathbb{E}[T|\sigma, R]^{3/2})}{(\alpha\frac{C'n}{\sqrt{\ell_1\ell_2}d^2})^2}$$

$$\leq \frac{O(d + (\sqrt{d}+1)(\frac{C'n}{\sqrt{\ell_1\ell_2}d^2}) + (\frac{C'n}{\sqrt{\ell_1\ell_2}d^2})^{3/2})}{(\alpha\frac{C'n}{\sqrt{\ell_1\ell_2}d^2})^2}$$

$$\leq \frac{1}{20}$$

where the above holds when $\frac{n}{\sqrt{\ell_1\ell_2}d^2} \asymp \sqrt{d}$ for a large enough α. $\qquad\square$

Lemma C.13. *If $p_{X,Y,Z}$ is such that $\inf_{q\in\mathcal{P}'_{0,[0,1]}} \|p_{X,Y,Z} - q\|_1 \geq \varepsilon$, and the conditions of Lemma C.10 hold. Then for a small enough absolute constant κ we have that*

$$\mathbb{P}(T \leq \kappa\sqrt{\zeta d}) \leq \frac{1}{10}.$$

Proof. We apply Chebyshev's inequality to $T' = (T|\sigma, R)$. Let $\mathcal{E}$ be the event of Lemma C.10. Set $\tau = \kappa\sqrt{\zeta d}$ for some small enough absolute constant κ.

$$\mathbb{P}(T \leq \tau) = \mathbb{P}(T' \leq \tau) \leq \mathbb{P}(T' \leq \tau|\mathcal{E}) + \mathbb{P}(\mathcal{E}^c).$$

$$\leq \mathbb{P}\left(|T' - \mathbb{E}[T|\sigma, R]| \geq \frac{1}{2}\mathbb{E}[T|\sigma, R]|\mathcal{E}\right) + \frac{1}{20}.$$

Next

$$\mathbb{P}\left(|T' - \mathbb{E}[T|\sigma, R]| \geq \frac{1}{2}\mathbb{E}[T|\sigma, R]|\mathcal{E}\right) \leq O\left(\frac{d + (\sqrt{d}+1)\mathbb{E}[T|\sigma, R] + \mathbb{E}[T|\sigma, R]^{3/2}}{\mathbb{E}[T|\sigma, R]^2}\right)$$

$$= O(1/\zeta^{1/2}) \leq \frac{1}{20},$$

for a large enough value of ζ. $\qquad\square$

Combining Lemmas C.12 and G.12 we have that if $\kappa\sqrt{\zeta d} \geq (\alpha+1)nC'/(\sqrt{\ell_1\ell_2}d^2) \asymp (\alpha+1)C'\sqrt{d}$, there will be a gap between the values under the null and the alternative hypothesis. This happens when ζ is large enough. Notice that $\frac{nC'}{\sqrt{\ell_1\ell_2}d^2} \asymp \sqrt{d}$ is equivalent to $d \asymp \frac{C'^{2/5}n^{2/5}}{(\ell_1\ell_2)^{1/5}}$. Plugging this in

all the inequalities (C.14), (C.15), (C.16) and (C.17), results in new inequalities that need to hold. We list those below, and in addition we recall that $d\ell_1 \lesssim n$. Condition (C.14) is equivalent to

$$\varepsilon \gtrsim \frac{(\ell_1\ell_2)^{1/5}}{n^{2/5}}.$$

Taking the second term of (C.15), and using the assumption that $d\ell_1 \lesssim n$ we have

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

since $\sqrt{\frac{d\ell_1\sqrt{\ell_2}}{n^{3/2}}} = \sqrt[4]{\frac{d\ell_1}{n}}\sqrt{\frac{\sqrt{d\ell_1\ell_2}}{n}}$. Taking (C.16), we have

$$\varepsilon \gtrsim \left(\frac{\sqrt{\ell_1\ell_2}}{n}\right)^{7/10},$$

which is of smaller order than $\frac{(\ell_1\ell_2)^{1/5}}{n^{2/5}}$ whenever $\frac{\sqrt{\ell_1\ell_2}}{n} \lesssim 1$. Finally for (C.17) we have

$$\varepsilon \gtrsim \left(\frac{\sqrt{\ell_1\ell_2}}{n}\right)^{13/20}\frac{1}{\sqrt[4]{\ell_1\ell_2}},$$

which is also of smaller order than $\frac{(\ell_1\ell_2)^{1/5}}{n^{2/5}}$ whenever $\frac{\sqrt{\ell_1\ell_2}}{n} \lesssim 1$. This completes the proof. $\qquad\square$

C.3 Proofs from Section 5.3

We start this section with showing the following result:

Lemma C.14. *There exists a constant C depending only on (s, L) such that if $p_{X,Y|Z}(x, y|z) \in \mathcal{H}^{2,s}(L)$ it follows that $p_{X|Z}(x|z)p_{Y|Z}(y|z) \in \mathcal{H}^{2,s}(C)$.*

Proof. We will start by showing that $p_{X|Z}(x|z)$ is Hölder smooth in the x dimension. We have, by the Leibniz rule and Jensen's inequality

$$\left|\frac{\partial^{\lfloor s\rfloor}}{\partial x^{\lfloor s\rfloor}}p_{X|Z}(x|z) - \frac{\partial^{\lfloor s\rfloor}}{\partial x^{\lfloor s\rfloor}}p_{X|Z}(x'|z)\right| \leq \int_0^1 \left|\frac{\partial^{\lfloor s\rfloor}}{\partial x^{\lfloor s\rfloor}}p_{XY|Z}(x, y|z) - \frac{\partial^{\lfloor s\rfloor}}{\partial x^{\lfloor s\rfloor}}p_{XY|Z}(x', y|z)\right|dy \leq L|x - x'|^{s-\lfloor s\rfloor}.$$

In addition any derivative of a lower or equal to $0 \leq k \leq \lfloor s\rfloor$ order is bounded since (by the Leibniz rule and Jensen's inequality)

$$\left|\frac{\partial^k}{\partial x^k}p_{X|Z}(x|z)\right| \leq \int_0^1 \left|\frac{\partial^k}{\partial x^k}p_{XY|Z}(x, y|z)\right|dy \leq L.$$

By symmetry the same statement holds for $p_{Y|Z}(y|z)$. Next we show that the product $p_{X|Z}(x|z)p_{Y|Z}(y|z) \in \mathcal{H}^{2,s}(C)$ for a sufficiently large C. First we argue that the derivatives are bounded. We have

$$\left|\frac{\partial^k}{\partial x^k}\frac{\partial^{\lfloor s\rfloor-k}}{\partial y^{\lfloor s\rfloor-k}}p_{X|Z}(x|z)p_{Y|Z}(y|z)\right| = \left|\frac{\partial^k}{\partial x^k}p_{X|Z}(x|z)\right|\left|\frac{\partial^{\lfloor s\rfloor-k}}{\partial y^{\lfloor s\rfloor-k}}p_{Y|Z}(y|z)\right| \leq L^2.$$

66

Next, take any $0 < k < \lfloor s \rfloor$. We have

$$\left| \frac{\partial^k}{\partial x^k} \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} p_{X|Z}(x|z) p_{Y|Z}(y|z) - \frac{\partial^k}{\partial x^k} \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} p_{X|Z}(x'|z) p_{Y|Z}(y'|z) \right|$$

$$\leq \left| \frac{\partial^k}{\partial x^k} p_{X|Z}(x|z) - \frac{\partial^k}{\partial x^k} p_{X|Z}(x'|z) \right| \left| \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} p_{Y|Z}(y|z) \right|$$

$$+ \left| \frac{\partial^k}{\partial x^k} p_{X|Z}(x'|z) \right| \left| \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} p_{Y|Z}(y|z) - \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} p_{Y|Z}(y'|z) \right|$$

$$\leq 2L^2(1 \wedge |x - x'|) + 2L^2(1 \wedge |y - y'|)$$

$$\leq 2L^2(|x - x'|^{s - \lfloor s \rfloor} + |y - y'|^{s - \lfloor s \rfloor})$$

$$\leq L^2 4 \left(\frac{|x - x'|^2 + |y - y'|^2}{2} \right)^{\frac{s - \lfloor s \rfloor}{2}},$$

where the last inequality follows by the generalized means inequality and the third inequality follows because we can bound

$$\left| \frac{\partial^k}{\partial x^k} p_{X|Z}(x|z) - \frac{\partial^k}{\partial x^k} p_{X|Z}(x'|z) \right| \leq 2L \wedge L|x - x'| \leq 2L(1 \wedge |x - x'|).$$

Similarly when $k = \lfloor s \rfloor$ we have

$$\left| \frac{\partial^{\lfloor s \rfloor}}{\partial x^{\lfloor s \rfloor}} p_{X|Z}(x|z) p_{Y|Z}(y|z) - \frac{\partial^{\lfloor s \rfloor}}{\partial x^{\lfloor s \rfloor}} p_{X|Z}(x'|z) p_{Y|Z}(y'|z) \right|$$

$$\leq \left| \frac{\partial^{\lfloor s \rfloor}}{\partial x^{\lfloor s \rfloor}} p_{X|Z}(x|z) - \frac{\partial^{\lfloor s \rfloor}}{\partial x^{\lfloor s \rfloor}} p_{X|Z}(x'|z) \right| |p_{Y|Z}(y|z)| + |p_{Y|Z}(y|z) - p_{Y|Z}(y'|z)| \left| \frac{\partial^{\lfloor s \rfloor}}{\partial x^{\lfloor s \rfloor}} p_{X|Z}(x'|z) \right|$$

$$\leq L^2 |x - x'|^{s - \lfloor s \rfloor} + 2L^2(1 \wedge |y - y'|)$$

$$\leq 2L^2(|x - x'|^{s - \lfloor s \rfloor} + |y - y'|^{s - \lfloor s \rfloor})$$

$$\leq L^2 4 \left(\frac{|x - x'|^2 + |y - y'|^2}{2} \right)^{\frac{s - \lfloor s \rfloor}{2}}.$$

Similar logic shows that the same holds when $k = 0$. This completes the proof. $\square$

We now formulate a Poissonized version of Theorem 5.6.

Theorem C.15 (Continuous X, Y, Z Upper Bound). *Set $d = \lceil n^{2s/(5s+2)} \rceil$, $d' = \lceil d^{1/s} \rceil$ and set the threshold $\tau = \sqrt{\zeta d}$ for a sufficiently large ζ (depending on L). Define $\mathcal{H}_0(s) = \mathcal{P}_{0,[0,1]^3,\mathrm{TV}}(L) \cup \mathcal{P}_{0,[0,1]^3,\chi^2}(L)$ 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*

$$\sup_{p \in \mathcal{H}_0(s)} \sum_{k=0}^{\infty} \mathbb{P}(N = k) \mathbb{E}_p[\psi_\tau(\mathcal{D}'_k)] \leq \frac{1}{10},$$

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

67

Proof of Theorem 5.6. The proof is identical to that of Theorem 5.2. $\qquad\square$

Proof of Theorem C.15. To prove this theorem we will assume that $N \sim Poi(n)$ instead of $Poi(n/2)$ for convenience. Since this changes only constant factors, we can do this WLOG.

Denote by

$$q_{ij}(m) := \mathbb{P}(X \in C_i', Y \in C_j' | Z \in C_m) = \int_{C_m} \int_{C_i' \times C_j'} p_{X,Y|Z}(x,y|z) dx dy d\widetilde{P}(z),$$

where $d\widetilde{P}(z) = dP(z)/\mathbb{P}(Z \in C_m)$. Denote by

$$q_{i\cdot}(m) := \sum_{j \in [d]} q_{ij}(m) = \int_{C_m} \sum_{j \in [d]} \int_{C_i' \times C_j'} p_{X,Y|Z}(x,y|z) dx dy d\widetilde{P}(z) = \int_{C_m} \int_{C_i' \times [0,1]} p_{X,Y|Z}(x,y|z) dx dy d\widetilde{P}(z).$$

Similarly define $q_{\cdot j}(m)$. Inspection of the proof of Theorem 5.5, reveals that we need to re-prove several facts and the proof will hold.

First we need to upper bound (C.22) or (C.23), where we now index by i, j instead of x, y for convenience. We have the following two results which control this expression for each of the two null hypothesis respectively.

Lemma C.16. *Suppose that* $p_{X,Y,Z} \in \mathcal{P}_{0,[0,1]^3,\mathrm{TV}}(L)$. *Then*

$$\|q_{A_m} - q_{\Pi,A_m}\|_2^2 \le \left(\frac{L}{d}\right)^4.$$

Proof. Since $a_{ij}^m \ge 0$ we have that $\|q_{A_m} - q_{\Pi,A_m}\|_2^2 \le \|q - q_\Pi\|_2^2$. Next by the monotonicity of the L_p norms we have

$$\|q - q_\Pi\|_2^2 = \sum_{i,j} (q_{ij}(m) - q_{i\cdot}(m) q_{\cdot j}(m))^2 \le \left(\sum_{i,j} |q_{ij}(m) - q_{i\cdot}(m) q_{\cdot j}(m)|\right)^2.$$

Recall that $p \in \mathcal{P}_{0,[0,1]^3,\mathrm{TV}}(L)$ implies that $\|p_{X|Z=z} - p_{X|Z=z'}\|_1 \le L|z-z'|$ and $\|p_{Y|Z=z} - p_{Y|Z=z'}\|_1 \le L|z-z'|$. We have

$$\sum_{i,j} \left| \int_{C_m} \int_{C_i' \times C_j'} p_{X|Z}(x|z) p_{Y|Z}(y|z) dx dy d\widetilde{P}(z) - \int_{C_m} \int_{C_i'} p_{X|Z}(x|z) dx d\widetilde{P}(z) \int_{C_m} \int_{C_j'} p_{Y|Z}(y|z) dy d\widetilde{P}(z) \right|$$

$$\le \int_{C_m} \sum_i \left| \int_{C_i'} p_{X|Z}(x|z) dx - \int_{C_m} \int_{C_i'} p_{X|Z}(x|z) dx d\widetilde{P}(z) \right| \times$$

$$\sum_j \left| \int_{C_j'} p_{Y|Z}(y|z) dy - \int_{C_m} \int_{C_j'} p_{Y|Z}(y|z) dy d\widetilde{P}(z) \right| d\widetilde{P}(z)$$

Take the first summation, and apply Jensen's inequality to conclude

$$
\sum_i \left| \int_{C_i'} p_{X|Z}(x|z)dx - \int_{C_m} \int_{C_i'} p_{X|Z}(x|z)dx d\widetilde{P}(z) \right|
$$

$$
\leq \int_{C_m} \sum_i \left| \int_{C_i'} p_{X|Z}(x|z)dx - \int_{C_i'} p_{X|Z}(x|z')dx \right| d\widetilde{P}(z')
$$

$$
\leq \int_{C_m} \sum_i \int_{C_i'} |p_{X|Z}(x|z) - p_{X|Z}(x|z')| dx d\widetilde{P}(z')
$$

$$
\leq \int L|z - z'| d\widetilde{P}(z') \leq L \operatorname{diam}(C_m) = \frac{L}{d}.
$$

Using the same strategy for the second summation completes the proof, i.e. we establish

$$
\sum_{i,j} (q_{ij}(m) - q_{i\cdot}(m)q_{\cdot j}(m))^2 \leq \left(\frac{L}{d}\right)^4.
$$

$\square$

Lemma C.17. *Suppose that $p_{X,Y,Z} \in \mathcal{P}_{0,[0,1]^3,\chi^2}(L)$. Then*

$$
\mathbb{E}_{A_m} \|q_{A_m} - q_{\Pi,A_m}\|_2^2 \leq \frac{O(1)}{\omega_m^2} \left(\frac{L}{d}\right)^2
$$

Proof. Using inequality (C.23) it suffices to directly control the quantity:

$$
\sum_{i,j} \frac{(q_{ij}(m) - q_{i\cdot}(m)q_{\cdot j}(m))^2}{q_{i\cdot}(m)q_{\cdot j}(m)}.
$$

By definition we have

$$
\sum_{i,j} \frac{(q_{ij}(m) - q_{i\cdot}(m)q_{\cdot j}(m))^2}{q_{i\cdot}(m)q_{\cdot j}(m)}
$$

$$
= \sum_{i,j} \frac{\left(\int_{C_m} \int_{C_i' \times C_j'} p_{X,Y|Z}(x,y|z)dxdyd\widetilde{P}(z) - \int_{C_m} \int_{C_i'} p_{X|Z}(x|z)dxd\widetilde{P}(z) \int_{C_m} \int_{C_j'} p_{Y|Z}(y|z)dyd\widetilde{P}(z) \right)^2}{\int_{C_m} \int_{C_i'} p_{X|Z}(x|z)dxd\widetilde{P}(z) \int_{C_m} \int_{C_j'} p_{Y|Z}(y|z)dyd\widetilde{P}(z)}
$$

Using the fact that $p_{X,Y|Z}(x,y|z) = p_{X,Z}(x|z)p_{Y|Z}(y|z)$ and Cauchy-Schwarz we obtain

$$
\sum_{i,j} \frac{(q_{ij}(m) - q_{i\cdot}(m)q_{\cdot j}(m))^2}{q_{i\cdot}(m)q_{\cdot j}(m)}
$$

$$
\leq \left(\sum_i \frac{\int_{C_m} \left(\int_{C_i'} p_{X|Z}(x|z)dx \right)^2 d\widetilde{P}(z)}{\int_{C_m} \int_{C_i'} p_{X|Z}(x|z)dxd\widetilde{P}(z)} - 1 \right) \left(\sum_j \frac{\int_{C_m} \left(\int_{C_j'} p_{Y|Z}(y|z)dy \right)^2 d\widetilde{P}(z)}{\int_{C_m} \int_{C_j'} p_{Y|Z}(y|z)dyd\widetilde{P}(z)} - 1 \right).
$$

69

We will handle each of these terms individually. We first note that the function $\int_{C_i} p(x|z)dx$ is continuous in z. To see this recall that (by Cauchy-Schwarz)

$$\left(\sum_i \int_{C_i'} |p_{X|Z}(x|z) - p_{X|Z}(x|z')|dx \right)^2 \leq \int_{[0,1]} \frac{p_{X|Z}^2(x|z')}{p_{X|Z}(x|z)} dx - 1 \leq L|z - z'|.$$

It follows that $\left| \int_{C_i'} p_{X|Z}(x|z)dx - \int_{C_i'} p_{X|Z}(x|z')dx \right| \leq \sqrt{L|z - z'|}$. It therefore follows by the mean value theorem (as in the discrete case) that for some $z' \in C_m$:

$$\sum_i \frac{\int_{C_m} \left(\int_{C_i'} p_{X|Z}(x|z)dx \right)^2 d\widetilde{P}(z)}{\int_{C_m} \int_{C_i'} p_{X|Z}(x|z)dxd\widetilde{P}(z)} - 1 = \sum_i \frac{\left(\int_{C_i'} p_{X|Z}(x|z')dx \right)^2}{\int_{C_m} \int_{C_i'} p_{X|Z}(x|z)dxd\widetilde{P}(z)} - 1$$

$$\leq \int_{C_m} \left(\sum_i \frac{\left(\int_{C_i'} p_{X|Z}(x|z')dx \right)^2}{\int_{C_i'} p_{X|Z}(x|z)dx} - 1 \right) d\widetilde{P}(z),$$

where we used Jensen's inequality and the fact that $x \mapsto 1/x$ is convex in the last inequality. Now by Cauchy-Schwarz we have

$$\frac{\left(\int_{C_i'} p_{X|Z}(x|z')dx \right)^2}{\int_{C_i'} p_{X|Z}(x|z)dx} \leq \int_{C_i'} \frac{p_{X|Z}^2(x|z')}{p_{X|Z}(x|z)} dx.$$

Hence we conclude

$$\sum_i \frac{\int_{C_m} \left(\int_{C_i'} p_{X|Z}(x|z)dx \right)^2 d\widetilde{P}(z)}{\int_{C_m} \int_{C_i'} p_{X|Z}(x|z)dxd\widetilde{P}(z)} - 1 \leq \int_{C_m} \int_{[0,1]} \frac{p_{X|Z}^2(x|z')}{p_{X|Z}(x|z)} dx - 1 d\widetilde{P}(z) \leq L\operatorname{diam}(C_m) = \frac{L}{d},$$

by assumption. Handling the second term in the same way warrants the desired conclusion. $\square$

Next we need to lower bound (C.13). For this it suffices to lower bound the distance $d_{\mathrm{TV}}(q(m), q_\Pi(m))$, where $q(m)$ is the distribution with "density" $q_{ij}(m)$ while $q_\Pi(m)$ is the distribution with density $q_{i\cdot}(m)q_{\cdot j}(m)$. We have

Lemma C.18. *There exist constants $b_1, b_2 > 0$ depending only on (s, L), such that the following bound holds*

$$\|q(m) - q_\Pi(m)\|_1 \geq b_1 \sup_{z \in C_m} \int_{[0,1]^2} |p_{X,Y|Z}(x, y|z) - p_{X|Z}(x|z)p_{Y|Z}(y|z)|dxdy - \frac{b_2}{d}.$$

Proof. Take a point $z^* \in C_m$ such that it maximizes the function $z \mapsto \int_{[0,1]^2} |p_{X,Y|Z}(x, y|z) - p_{X|Z}(x|z)p_{Y|Z}(y|z)|dxdy$. If such a point does not exist, take a sequence of points that converge to

the supremum. By the triangle inequality we have

$$\sum_{i,j} |q_{ij}(m) - q_{i\cdot}(m)q_{\cdot j}(m)|$$

$$\geq \sum_{i,j} \left| \int_{C_i' \times C_j'} p_{X,Y|Z}(x,y|z^*)dxdy - \int_{C_i' \times [0,1]} p_{X,Y|Z}(x,y|z^*)dxdy \int_{[0,1] \times C_j'} p_{X,Y|Z}(x,y|z^*)dxdy \right|$$

$$- \sum_{i,j} \left| \int_{C_m} \int_{C_i' \times C_j'} p_{X,Y|Z}(x,y|z) - p_{X,Y|Z}(x,y|z^*)dxdyd\widetilde{P}(z) \right|$$

$$- \sum_{i,j} \left| q_{i\cdot}(m) \left(\int_{C_m} \int_{[0,1] \times C_j'} p_{X,Y|Z}(x,y|z) - p_{X,Y|Z}(x,y|z^*)dxdyd\widetilde{P}(z) \right) \right|$$

$$- \sum_{i,j} \left| \int_{[0,1] \times C_j'} p_{X,Y|Z}(x,y|z^*)dxdy \left(\int_{C_m} \int_{C_i' \times [0,1]} p_{X,Y|Z}(x,y|z) - p_{X,Y|Z}(x,y|z^*)dxdyd\widetilde{P}(z) \right) \right|$$

We will first handle the last three terms, starting with the first one. Next we obtain

$$\sum_{i,j} \left| \int_{C_m} \int_{C_i' \times C_j'} p_{X,Y|Z}(x,y|z) - p_{X,Y|Z}(x,y|z^*)dxdyd\widetilde{P}(z) \right|$$

$$\leq \int_{C_m} \sum_{i,j} \int_{C_i' \times C_j'} |p_{X,Y|Z}(x,y|z) - p_{X,Y|Z}(x,y|z^*)|dxdyd\widetilde{P}(z)$$

$$\leq \int_{C_m} L|z - z^*|d\widetilde{P}(z) \leq \frac{L}{d}.$$

For the second term we have

$$\sum_{i,j} \left| q_{i\cdot}(m) \left(\int_{C_m} \int_{[0,1] \times C_j'} p_{X,Y|Z}(x,y|z) - p_{X,Y|Z}(x,y|z^*)dxdyd\widetilde{P}(z) \right) \right|$$

$$\leq \int_{C_m} \sum_i q_{i\cdot}(m) \sum_j \int_{[0,1] \times C_j'} |p_{X,Y|Z}(x,y|z) - p_{X,Y|Z}(x,y|z^*)|dxdyd\widetilde{P}(z)$$

$$\leq \int_{C_m} L|z - z^*|d\widetilde{P}(z) \leq \frac{L}{d}.$$

The analysis of the third term is the same. We will finally need a lower bound on

$$\sum_{i,j} \left| \int_{C_i' \times C_j'} p_{X,Y|Z}(x,y|z^*)dxdy - \int_{C_i' \times [0,1]} p_{X,Y|Z}(x,y|z^*)dxdy \int_{[0,1] \times C_j'} p_{X,Y|Z}(x,y|z^*)dxdy \right|$$

(C.25)

We will compare this term to the following

$$\int_{[0,1]^2} |p_{X,Y|Z}(x,y|z^*) - p_{X|Z}(x|z^*)p_{Y|Z}(y|z^*)|dxdy,$$

(C.26)

where $p_{X|Z}(x|z^*) = \int_{[0,1]} p_{X,Y|Z}(x,y|z^*)dy$ and similarly for $p_{Y|Z}(y|z^*)$. Note that in terms of this notation we may rewrite the term in question as

$$\sum_{i,j} \left| \int_{C_i' \times C_j'} p_{X,Y|Z}(x,y|z^*)dxdy - \int_{C_i'} p_{X|Z}(x|z^*)dx \int_{C_j'} p_{Y|Z}(y|z^*)dy \right| =$$

$$\sum_{i,j} \left| \int_{C_i' \times C_j'} [p_{X,Y|Z}(x,y|z^*) - p_{X|Z}(x|z^*)p_{Y|Z}(y|z^*)]dxdy \right|.$$

For what follows let μ denote the Lebesgue measure on $[0,1]$. We now need the following Lemma

Lemma C.19. *Suppose that $h \in \mathcal{H}^{2,s}(L)$. For an integer $\kappa \in \mathbb{N}$, take a decomposition of $[0,1]^2 = \prod_{i,j \in [\kappa]} B_i \times B_j$ where the length of each $\mu(B_i) = \kappa^{-1}$ and B_i form a decomposition of $[0,1]$. Let $W_\kappa h = \sum_{i,j \in [\kappa]} \kappa^2 (\int_{B_i \times B_j} h) \mathbb{1}_{B_i \times B_j}$. There exist constants $b_1, b_2 > 0$ depending only on (s, L) such that*

$$\|W_\kappa(h)\|_1 \geq b_1 \|h\|_1 - b_2 \kappa^{-s}, \qquad \forall h \in \mathcal{H}^{2,s}(L)$$

Proof. The proof of this Lemma is extremely similar to the proof of Lemma 3 of [2] modulo changes from L_2 distance to L_1 distance which do not modify the arguments. Hence we omit the details. We also remark that the proof of this Lemma in the one-dimensional case can be found in Proposition 2.16 of [24]. $\qquad \square$

In order to use Lemma C.19, we note that the difference $p_{X,Y|Z}(x,y|z^*) - p_{X|Z}(x|z^*)p_{Y|Z}(y|z^*)$ belongs to the Hölder class $\mathcal{H}^{2,s}(2L)$ by definition.Using Lemma C.19, we can write the following bound

$$\sum_{i,j} \left| \int_{C_i' \times C_j'} [p_{X,Y|Z}(x,y|z^*) - p_{X|Z}(x|z^*)p_{Y|Z}(y|z^*)]dxdy \right| \geq$$

$$b_1 \int_{[0,1]^2} |p_{X,Y|Z}(x,y|z^*) - p_{X|Z}(x|z^*)p_{Y|Z}(y|z^*)|dxdy - b_2 \frac{1}{(d^{1/s})^s}$$

Putting everything together we conclude that

$$\sum_{i,j} |q_{ij}(m) - q_{i\cdot}(m)q_{\cdot j}(m))| \geq b_1 \sup_{z \in C_m} \int_{[0,1]^2} |p_{X,Y|Z}(x,y|z) - p_{X|Z}(x|z)p_{Y|Z}(y|z)|dxdy - \frac{b_2}{d},$$

completing the proof. $\qquad \square$

After having these three results, the remaining details of the proof follow closely that of Theorem 5.5 so we omit the details. $\qquad \square$

D   Unbounded Z, Discrete X, Y Non-Scaling

Define the set of distributions $\mathcal{E}'$ 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 potentially unbounded support 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 by $\mathcal{P}'_0 \subset \mathcal{E}'$ the set of null distributions (i.e. distributions such that $X \perp\!\!\!\perp Y|Z$) and let $\mathcal{Q}'_0 = \mathcal{E}' \setminus \mathcal{P}'_0$.

Definition D.1 (Null Lipschitzness)**.**

1. *Null TV Lipschitzness:* Let $\mathcal{P}'_{0,\mathrm{TV}}(L) \subset \mathcal{P}'_0$ be the collection of distributions $p_{X,Y,Z}$ such that for all z, z' we have:

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

   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 $\mathcal{P}'_{0,\chi^2}(L) \subset \mathcal{P}'_0$ be the collection of distributions $p_{X,Y,Z}$ such that for all z, z' we have:

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

   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 $\mathcal{P}'_{0,\mathrm{TV}^2}(L) \subset \mathcal{P}'_0$ be the collection of distributions $p_{X,Y,Z}$ such that for all z, z' we have:

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

   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.

Definition D.2 (Alternative TV Lipschitzness)**.** *Let $\mathcal{Q}'_{0,\mathrm{TV}}(L) \subset \mathcal{Q}'_0$ be the collection of distributions $p_{X,Y,Z}$ such that for all z, z' 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 $p_{X,Y,Z}$.

First it is straightforward to generalize the results of Section 5, to the case where Z is supported on an interval $[A, B]$ with $A < B$. The proof needs to be modified to respect the fact that the bins have a size of $\frac{B-A}{d}$ instead of size d. The final rate that one obtains is $\frac{(B-A)^{1/5}}{n^{2/5}} \vee \frac{(B-A)^{7/15}}{n^{8/15}}$.

Lemma D.3. *Suppose that $\inf_{p \in \mathcal{P}'_0} \|q_{X,Y,Z} - p\|_1 = \varepsilon$. Define the distribution $\widetilde{q}_{X,Y,Z} = q_{X,Y|Z}\widetilde{q}_Z$ where $\widetilde{q}_Z$ is the restriction of q_Z on an interval S_η with the property that $\mathbb{P}(Z \in S_\eta) \geq 1 - \eta$. Then $\inf_{p \in \mathcal{P}'_0} \|\widetilde{q}_{X,Y,Z} - p\|_1 \geq \frac{1}{6}\frac{\varepsilon - 2\eta}{1 - \eta}$.*

Proof of Lemma D.3. By Lemma B.4 (whose proof does not depend on the bounded Z support) we know that

$$\inf_{p \in \mathcal{P}'_0} \|q_{X,Y,Z} - p\|_1 \geq \|\widetilde{q}_{X,Y,Z} - \widetilde{q}_{X|Z}\widetilde{q}_{Y|Z}\widetilde{q}_Z\|_1/6$$

Next

$$\|\widetilde{q}_{X,Y,Z} - \widetilde{q}_{X|Z}\widetilde{q}_{Y|Z}\widetilde{q}_Z\|_1 = \int_{S_\eta} \sum_{x,y} |q_{X,Y|Z}(x,y|z) - q_{X|Z}(x|z)q_{Y|Z}(y|z)| \frac{q_Z(z)}{\mathbb{P}(Z \in S_\eta)} dz$$

On the other hand

$$
\begin{aligned}
\varepsilon \leq \|q_{X,Y,Z} - q_{X|Z}q_{Y|Z}q_Z\|_1 &= \int_{S_\eta} \sum_{x,y} |q_{X,Y|Z}(x,y|z) - q_{X|Z}(x|z)q_{Y|Z}(y|z)| q_Z(z) dz \\
&+ \int_{S_\eta^c} \sum_{x,y} |q_{X,Y|Z}(x,y|z) - q_{X|Z}(x|z)q_{Y|Z}(y|z)| q_Z(z) dz \\
&\leq \int_{S_\eta} \sum_{x,y} |q_{X,Y|Z}(x,y|z) - q_{X|Z}(x|z)q_{Y|Z}(y|z)| q_Z(z) dz \\
&+ 2\mathbb{P}(Z \in S_\eta^c)
\end{aligned}
$$

We conclude that

$$\inf_{p \in \mathcal{P}_0'} \|q_{X,Y,Z} - p\|_1 \geq \frac{1}{6} \frac{\varepsilon - 2\eta}{1 - \eta},$$

as claimed.

$\square$

Next suppose we split the sample into two samples of equal size. Assume for simplicity that initially we had $2n$ samples. Take the second data, and construct the shortest interval $\widehat{S}_\eta$ containing $n(1 - \eta) + C\sqrt{n \log n}$ of the samples $\{Z_1, \ldots, Z_n\}$, where C is an absolute constant. We have the following result.

Lemma D.4. *For a sufficiently large C we have*

$$\mathbb{P}(Z \in \widehat{S}_\eta) \geq 1 - \eta,$$

with probability at least $1 - \frac{1}{n}$. In addition $\mu(\widehat{S}_\eta) \leq \mu(S_{\eta/2})$ with probability at least $1 - \frac{1}{n}$, where μ is the Lebesgue measure, $\eta \geq 4C\sqrt{\frac{\log n}{n}}$, and $S_{\eta/2}$ is the shortest interval such that $\mathbb{P}(Z \in S_{\eta/2}) \geq 1 - \eta/2$.

Proof of Lemma D.4. By the VC inequality [15, cf. Theorem 12.5], and the fact that closed intervals on the real line have VC dimension equal to 2, we have that

$$\mathbb{P}\left(\left|\mathbb{P}(Z \in \widehat{S}_\eta) - \left((1 - \eta) + C\sqrt{\frac{\log n}{n}}\right)\right| \geq t\right) \leq 8(\frac{ne}{2})^2 \exp(-32nt^2)$$

Now set $t = C\sqrt{\frac{\log n}{n}}$. For large enough C, the above probability is bounded by $\frac{1}{n}$ which is what we claimed.

By the VC inequality

$$\mathbb{P}\left(\left|\mathbb{P}(Z \in S_{\eta/2}) - n^{-1}\sum_{n}\mathbb{1}(Z_i \in S_{\eta/2})\right| \geq t\right) \leq 8\left(\frac{ne}{2}\right)^2 \exp(-32nt^2)$$

Thus with a choice of $t = C\sqrt{\frac{\log n}{n}}$ we obtain

$$n^{-1}\sum_{n}\mathbb{1}(Z_i \in S_{\eta/2}) \geq 1 - \eta/2 - C\sqrt{\frac{\log n}{n}} > 1 - \eta + C\sqrt{\frac{\log n}{n}},$$

with probability $\frac{1}{n}$, under our assumption on η. Thus if $\mu(S_{\eta/2}) \leq \mu(\widehat{S}_\eta)$ that would be a contradiction.

$\square$

Next we describe the testing procedure. Suppose we have $2n$ observations. Split the data equally, and estimate $\widehat{S}_\eta$, where $1 \geq \eta \geq C\sqrt{\frac{\log n}{n}}$ on the second half as discussed above.

Draw $N \sim Poi(\frac{n}{2})$. If $N > n$ accept the null hypothesis. If $N \leq n$ take arbitrary N out of the n samples from the first half of the data and work with them. The next step is to discretize the variable Z into d bins of equal size. Denote those bins with $\{C_1, \ldots, C_d\}$, so that $\cup_{i \in [d]} C_i = \widehat{S}_\eta$, and each C_i is an interval of length $\frac{\mu(\widehat{S}_\eta)}{d}$. 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 $\sigma_m \geq 4$ 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. [13] and consider the following test statistic

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

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

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

where $\mathcal{D}_N = \{(X_1, Y_1, Z_1), \ldots, (X_N, Y_N, Z_N)\}$. Recall the definitions of the null Lipschitzness classes defined above $\mathcal{P}'_{0,\mathrm{TV}}(L), \mathcal{P}'_{0,\chi^2}(L), \mathcal{P}'_{0,\mathrm{TV}^2}(L)$ and the alternative Lipschitzness classes $\mathcal{Q}'_{0,\mathrm{TV}}(L)$. We have

Theorem D.5 (Finite Discrete X, Y Upper Bound)**.** *Set $d = \lceil \mu(\widehat{S}_\eta)^{4/5} n^{2/5}\rceil \wedge \lceil \mu(\widehat{S}_\eta)^{8/15} n^{8/15}\rceil$ and let $\tau = \zeta\sqrt{d}$ for a sufficiently large absolute constant ζ (depending on L). Finally, suppose that $\varepsilon \geq c \max\left(\frac{\mu(S_{\eta/2})^{1/5}}{n^{2/5}}, \frac{\mu(S_{\eta/2})^{7/15}}{n^{8/15}}, \eta\right)$, for a sufficiently large constant c (depending on ζ, L, ℓ_1, ℓ_2). Then we have that*

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

$$\sup_{p \in \{p \in \mathcal{Q}'_{0,\mathrm{TV}}(L) : \inf_{q \in \mathcal{P}'_0}\|p-q\|_1 \geq \varepsilon\}} \mathbb{E}_p[1 - \psi_\tau(\mathcal{D}_N)] \leq \frac{1}{10} + \exp(-n/8).$$

Proof. The proof of this result follows directly from the previous discussion and the proof of Theorem 5.2. We omit the details. □

Remark D.6. *In addition we point out that for sub-Gaussian distributions, the above strategy can lead to rates which coincide with the rates described in the bounded support case up to logarithmic factors. To see this it suffices to select $S_\eta = [Z_{(1)}, Z_{[n]}]$ where $Z_{(i)}$ denote the order statistics on a second dataset; by sub-Gaussianity one has with high probability that $|Z_{(n)}|, |Z_{(1)}| \lesssim \sqrt{\log n}$. Furthermore $\mathbb{E}\mathbb{P}(Z \geq Z_{(n)}) = \mathbb{E}F_Z(Z)^n = \int_0^1 u^n du = \frac{1}{n+1}$. Thus by Markov's inequality $\mathbb{P}(\mathbb{P}(Z \geq Z_{(n)}) \geq a_n) \leq \frac{1}{(n+1)a_n}$. Thus if $a_n n \to \infty$ with high probability $\mathbb{P}(Z \geq Z_{(n)}) \leq a_n$. By symmetry the same argument is valid for $\mathbb{P}(Z \leq Z_{(1)}) \leq a_n$. Thus it suffices that the critical radius scales as*

$$\varepsilon \gtrsim \frac{(\sqrt{\log n})^{1/5}}{n^{2/5}}.$$

E   Proofs from Section 6

Proof of Lemma 6.8. Note that condition (6.3) ensures the second part of the definition of the set $\mathcal{Q}_{0,[0,1]^3,\text{TV}}$ hence we only need to prove that $\|p_{X,Y|Z=z} - p_{X,Y|Z=z'}\|_1$ is bounded by $(e^L - 1)|z - z'|$.

$$\int\int |p_{X,Y|Z}(x,y|z) - p_{X,Y|Z}(x,y|z')|dxdy$$

$$= \int\int \left(\frac{\max(p_{X,Y|Z}(x,y|z), p_{X,Y|Z}(x,y|z'))}{\min(p_{X,Y|Z}(x,y|z), p_{X,Y|Z}(x,y|z'))} - 1\right) \min(p_{X,Y|Z}(x,y|z), p_{X,Y|Z}(x,y|z'))dxdy$$

$$\leq \int\int \left(\frac{\max(p_{X,Y|Z}(x,y|z), p_{X,Y|Z}(x,y|z'))}{\min(p_{X,Y|Z}(x,y|z), p_{X,Y|Z}(x,y|z'))} - 1\right) p_{X,Y|Z}(x,y|z)dxdy.$$

By the Lipschitz property of $\log p_{X,Y|Z}(x,y|z)$ we have

$$\frac{\max(p_{X,Y|Z}(x,y|z), p_{X,Y|Z}(x,y|z'))}{\min(p_{X,Y|Z}(x,y|z), p_{X,Y|Z}(x,y|z'))} - 1 \leq \exp(L|z - z'|) - 1.$$

From here the proof can continue as in the proof of Lemma 6.5.

The fact that $p_{X,Y|Z}(x,y|z)$ is Hölder in x,y for every z with $s = 1$ is clear since $C|x - x'| + |y - y'| \leq \sqrt{2}C((x - x')^2 + (y - y')^2)^{1/2}$.

This completes the proof. □

F   Proofs from Section 7

In this section we prove that the the generation mechanism of Section 7 is indeed "TV smooth". We first start by the example $X = \frac{U_1+Z}{2}$ and $Y = \frac{U_2+Z}{2}$. We want to show that $\|p_{X|Z=z} - p_{X|Z=z'}\|_1 \leq L|z - z'|$ for an appropriate L (and similarly for Y). In this example it is simple to verify that

$$\|p_{X|Z=z} - p_{X|Z=z'}\|_1 = \int_0^1 2|\mathbb{1}_{[\frac{z}{2}, \frac{z+1}{2}]}(x) - \mathbb{1}_{[\frac{z'}{2}, \frac{z'+1}{2}]}(x)|dx = 2|z - z'|.$$

This implies that the so generated data belongs to the set $\mathcal{P}_{0,[0,1]^3,\text{TV}}(2)$.

Next we consider $X = \frac{U_1 + U + Z}{3}$ and $Y = \frac{U_2 + U + Z}{3}$. We will argue that $\|p_{X,Y|Z=z} - p_{X,Y|Z=z'}\|_1 \leq L|z - z'|$ for an appropriate L. Let Σ be the Borel σ-field on $[0,1]^2$. We note the following bounds

$$
d_{\mathrm{TV}}(p_{X,Y|Z=z}, p_{X,Y|Z=z'})
$$
$$
= \sup_{A \in \Sigma} \left| \mathbb{P}\left(\left(\frac{U_1 + U + z}{3}, \frac{U_2 + U + z}{3} \right) \in A \right) - \mathbb{P}\left(\left(\frac{U_1 + U + z'}{3}, \frac{U_2 + U + z'}{3} \right) \in A \right) \right|
$$
$$
= \sup_{A \in \Sigma} \left| \int_0^1 \mathbb{P}\left(\left(\frac{U_1 + U + z}{3}, \frac{U_2 + U + z}{3} \right) \in A \middle| U = u \right) du \right.
$$
$$
\left. - \int_0^1 \mathbb{P}\left(\left(\frac{U_1 + U + z'}{3}, \frac{U_2 + U + z'}{3} \right) \in A \middle| U = u \right) du \right|
$$
$$
\leq \int_0^1 \sup_{A \in \Sigma} \left| \mathbb{P}\left(\left(\frac{U_1 + u + z}{3}, \frac{U_2 + u + z}{3} \right) \in A \right) - \mathbb{P}\left(\left(\frac{U_1 + u + z'}{3}, \frac{U_2 + u + z'}{3} \right) \in A \right) \right| du
$$

The expression in the above integral is nothing but the total variation between the law of $(\frac{U_1 + u + z}{3}, \frac{U_2 + u + z}{3})$ and $(\frac{U_1 + u + z'}{3}, \frac{U_2 + u + z'}{3})$ which is $\frac{1}{2}$ of the L_1 distance. Since U_1 and U_2 are independent it is simple to see that for a fixed u the above is equivalent to

$$
\sup_{A \in \Sigma} \left| \mathbb{P}\left(\left(\frac{U_1 + u + z}{3}, \frac{U_2 + u + z}{3} \right) \in A \right) - \mathbb{P}\left(\left(\frac{U_1 + u + z'}{3}, \frac{U_2 + u + z'}{3} \right) \in A \right) \right|
$$
$$
= \frac{9}{2} \int_0^1 \int_0^1 \left| \mathbb{1}_{[\frac{u+z}{3}, \frac{1+u+z}{3}]}(u_1) \mathbb{1}_{[\frac{u+z}{3}, \frac{1+u+z}{3}]}(u_2) - \mathbb{1}_{[\frac{u+z'}{3}, \frac{1+u+z'}{3}]}(u_1) \mathbb{1}_{[\frac{u+z'}{3}, \frac{1+u+z'}{3}]}(u_2) \right| du_1 du_2.
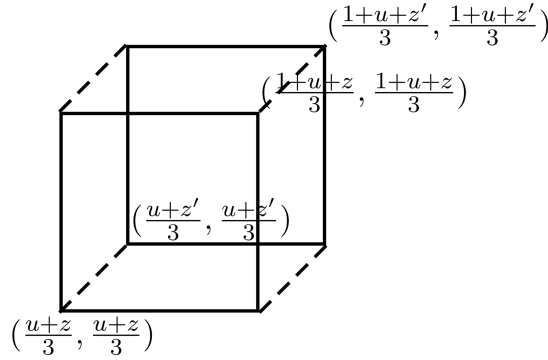$$
$$
\tag{F.1}
$$

Figure 5: The two rectangles given in equation (F.1) are shown above. The area of the symmetric difference is bounded by 4 times the area of a parallelogram.

Using Figure 5 we can bound the integral in (F.1) by 4 times the area of a parallelogram, which is further bounded by $\sqrt{2}/9|z - z'|$. We conclude that

$$
\frac{1}{2}\|p_{X,Y|Z=z} - p_{X,Y|Z=z'}\|_1 = d_{\mathrm{TV}}(p_{X,Y|Z=z}, p_{X,Y|Z=z'}) \leq 2\sqrt{2}|z - z'|,
$$

which is what we wanted to show.

The fact that the density $p_{X,Y|Z}(x,y|z)$ is Lipschitz in x, y follows by a straightforward but tedious calculation which is omitted. The idea is to calculate the cdf by conditioning on U and then take derivatives with respect to x and y.

G Extension to a Multivariate Z

G.1 Discrete X, Y Case with Fixed Many Categories

Let $Z \in [0,1]^{d_Z}$. Define $\mathcal{P}'_{0,[0,1]^{d_Z}}$ and $\mathcal{Q}'_{0,[0,1]^{d_Z}}$ analogously to the sets $\mathcal{P}'_{0,[0,1]}$ and $\mathcal{Q}'_{0,[0,1]}$. Analogously to the Null and Alternative TV Lipschitzness define the following classes.

Definition G.1 (Null Lipschitzness)**.**

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

$$\|p_{X|Z=z} - p_{X|Z=z'}\|_1 \le L\|z - z'\|_2 \text{ and } \|p_{Y|Z=z} - p_{Y|Z=z'}\|_1 \le L\|z - z'\|_2,$$

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

$$d_{\chi^2}(p_{X|Z=z}, p_{X|Z=z'}) \le L\|z - z'\|_2 \text{ and } d_{\chi^2}(p_{Y|Z=z}, p_{Y|Z=z'}) \le L\|z - z'\|_2,$$

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

$$\|p_{X|Z=z} - p_{X|Z=z'}\|_1 \le \sqrt{L\|z - z'\|_2} \text{ and } \|p_{Y|Z=z} - p_{Y|Z=z'}\|_1 \le \sqrt{L\|z - z'\|_2},$$

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

$$\|p_{X,Y|Z=z} - p_{X,Y|Z=z'}\|_1 \le L\|z - z'\|_2,$$

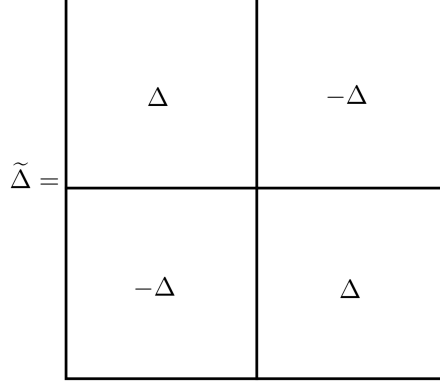*where $p_{X,Y|Z=z}$ denotes the conditional distribution of $X, Y|Z = z$ under $p_{X,Y,Z}$.*

Figure 6: The construction of the matrix $\widetilde{\Delta}$ from Δ.

Theorem G.3 (Critical Radius Lower Bound)**.** *Let* $\overline{\mathcal{H}}_0 = \mathcal{P}'_{0,[0,1]^{d_Z}}$. *Suppose that* $\mathcal{H}_0$ *is either of* $\mathcal{P}'_{0,[0,1]^{d_Z},\mathrm{TV}}(L)$, $\mathcal{P}'_{0,[0,1]^{d_Z},\mathrm{TV}^2}(L)$ *or* $\mathcal{P}'_{0,[0,1]^{d_Z},\chi^2}(L)$, *while* $\mathcal{H}_1 = \mathcal{Q}'_{0,[0,1]^{d_Z},\mathrm{TV}}(L)$ *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*

$$\varepsilon_n(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1) \geq c_0 \left(\frac{(\ell_1 \ell_2)^{1/(d_Z+4)}}{n^{2/(d_Z+4)}} \wedge 1 \right).$$

Proof. The proof is identical with the one of Theorem 4.1. For completeness we provide full details below. To derive a lower bound we will first show how to obtain multiple distributions which are far from independent by perturbing the uniform discrete distribution. Suppose for simplicity that $\ell_1 = 2\ell'_1$ and $\ell_2 = 2\ell'_2$ for some integers ℓ'_1 and ℓ'_2 (although this simplifies our calculation we will remark how to fix the calculation for the odd case as well).

We first construct a single null hypothesis distribution. Suppose that $Z \sim U[0,1]$. Let the basic null distribution be given by the density $p_{X,Y|Z}(x,y|z) = \frac{1}{\ell_1 \ell_2}$ for each $x, y \in [\ell_1] \times [\ell_2]$ and $z \in [0,1]^{d_Z}$. Clearly, this distribution belongs to all three sets of null distributions $\mathcal{P}'_{0,[0,1]^{d_Z},\mathrm{TV}}(L)$, $\mathcal{P}'_{0,[0,1]^{d_Z},\mathrm{TV}^2}(L)$ and $\mathcal{P}'_{0,[0,1]^{d_Z},\chi^2}(L)$.

Next we will perturb this null distribution p to obtain alternative distributions. Let $\Delta = (\delta_{xy})_{x \in [\ell'_1], y \in [\ell'_2]}$ be a matrix of ± 1 numbers δ_{xy}. We create the $\ell_1 \times \ell_2$ matrix $\widetilde{\Delta}$ so that

$$\widetilde{\delta}_{xy} = \delta_{xy} \text{ for } x, y \in [\ell'_1] \times [\ell'_2],$$
$$\widetilde{\delta}_{xy} = -\delta_{(x-\ell'_1)y} \text{ for } x > \ell'_1, y \in [\ell'_2],$$
$$\widetilde{\delta}_{xy} = -\delta_{x(y-\ell'_2)} \text{ for } x \in [\ell'_1], y > \ell'_2,$$
$$\widetilde{\delta}_{xy} = \delta_{(x-\ell'_1)(y-\ell'_2)} \text{ for } x > \ell'_1, y > \ell'_2.$$

It is simple to check that all row sums and column sums of the matrix $\widetilde{\Delta}$ are 0 (see also Figure 6). We perturb the null distribution p using the following procedure: for $x, y \in [\ell_1] \times [\ell_2]$ we take $q_{X,Y|Z}(x,y|z) = \frac{1}{\ell_1 \ell_2} + \widetilde{\delta}_{xy} \eta_\nu(z)$, where

$$\eta_\nu(z) = \rho \sum_{j \in [d]} \nu_{j_1, \dots j_{d_Z}} \prod_{k=1}^{d_Z} h_{j_k, d}(z_k),$$

79

where $\rho > 0$ is a constant, $d \in \mathbb{N}$, $\nu_{j_1,\dots j_{d_Z}} \in \{-1, +1\}$, and $h_{j,d}(z) = \sqrt{d}h(dz - j + 1)$ for $z \in [(j-1)/d, j/d]$, and h is an infinitely differentiable function supported on $[0, 1]$ such that $\int h(z)dz = 0$ and $\int h^2(z)dz = 1$. Since the row sums and column sums of $\widetilde{\Delta}$ are 0, it is simple to verify that the marginals of the distribution remain unchanged under this perturbation, i.e. $q_{X|Z}(x|z) = \sum_y q_{X,Y|Z}(x, y|z) = p_{X|Z}(x|z) = \frac{1}{\ell_1}$ and similarly $q_{Y|Z}(y|z) = \sum_x q_{X,Y|Z}(x, y|z) = p_{Y|Z}(y|z) = \frac{1}{\ell_2}$. We note that in the case that one of ℓ_1 or ℓ_2 or both is odd, the fix is to add one row and/or column to the matrix Δ to be fixed, and reason as in the even case.

When perturbing, in order to ensure that we create valid probability distributions, we need to satisfy the conditions that

$$\frac{1}{\ell_1\ell_2} - \rho(\sqrt{d})^{d_Z}\|h\|_\infty^{d_Z} \geq 0,$$

and

$$\frac{1}{\ell_1\ell_2} + \rho(\sqrt{d})^{d_Z}\|h\|_\infty^{d_Z} \leq 1.$$

We will ensure this by our choices of ρ and d. Next, we need to verify that $q_{X,Y,Z} \in \mathcal{Q}'_{0,[0,1]^{d_Z},\text{TV}}(L)$. We start by showing that $\|q_{X,Y|Z=z} - q_{X,Y|Z=z'}\|_1 \leq L\|z - z'\|_2$. We have

$$\|q_{X,Y|Z=z} - q_{X,Y|Z=z'}\|_1 = \ell_1\ell_2|\eta_\nu(z) - \eta_\nu(z')|.$$

Using telescoping arguments we can bound

$$\|q_{X,Y|Z=z} - q_{X,Y|Z=z'}\|_1 \lesssim \ell_1\ell_2 d_Z\rho d(\sqrt{d})^{d_Z}\|h'\|_\infty\|h\|_\infty^{d_Z-1} \leq L.$$

We let $Z \sim U([0, 1])$. Next we will show that the constructed distributions $q_{X,Y,Z}$ are ε far from being independent, that is we will show that

$$\inf_{p \in \mathcal{P}'_{0,[0,1]}} \|q_{X,Y,Z} - p\|_1 \geq \varepsilon,$$

for some $\varepsilon > 0$.

Next, suppose that h satisfies $\int |h(z)|dz = c$ for some $0 < c < 1$. We then have that the L_1 distance between $q_{X,Y,Z}$ and $q_{X|Z}q_{Y|Z}q_Z$ satisfies

$$\varepsilon := \mathbb{E}_Z\ell_1\ell_2|\eta_\nu(Z)| = \ell_1\ell_2\rho \sum_{\boldsymbol{j}\in[d]^{d_Z}} \int \prod_{k=1}^{d_Z} |h_{j_k,d}(z_k)|dz_k, = \ell_1\ell_2\rho(\sqrt{d})^{d_Z}c^{d_Z},$$

where in the above we used that $h_{j,d}$ have disjoint support. By Lemma B.4 this shows that $q_{X,Y,Z}$ is at least $\varepsilon/6$ from any conditionally independent distribution in L_1 distance.

Next, we put uniform priors over ν and Δ, i.e., the random variables $(\nu_i)_{i\in[d]}$ and $(\delta_{xy})_{x,y\in[\ell_1]\times[\ell_2]}$ are taken as i.i.d. Rademachers. The likelihood ratio is

$$W = \mathbb{E}_{\nu,\Delta} \prod_{i=1}^n (1 + \widetilde{\delta}_{X_i,Y_i}\bar{\eta}_\nu(Z_i))$$

where $\bar{\eta}_\nu(z) = \ell_1\ell_2\eta_\nu(z)$ and the expectation is taken over all Rademacher sequences ν and Δ. By a standard argument [2, 5, 23] the risk of the likelihood ratio (which is the optimal test by

Neyman-Pearson's Lemma) is bounded from below by $1 - \frac{1}{2}\sqrt{\mathrm{Var}_0 W}$. Hence it suffices to study $\mathbb{E}_0(W^2) - 1$ (here $\mathbb{E}_0$ is the expectation under the null hypothesis). We have

$$
\begin{aligned}
\mathbb{E}_0 W^2 &= \mathbb{E}_{\nu,\nu',\Delta,\Delta'} \prod_{i=1}^n \mathbb{E}_0 (1 + \widetilde{\delta}_{X_i,Y_i} \bar{\eta}_\nu(Z_i))(1 + \widetilde{\delta}'_{X_i,Y_i} \bar{\eta}_{\nu'}(Z_i)) \\
&= \mathbb{E}_{\nu,\nu',\Delta,\Delta'} \prod_{i=1}^n \mathbb{E}_{Z_i} \sum_{(x,y)\in[\ell_1]\times[\ell_2]} \frac{(1 + \widetilde{\delta}_{xy}\bar{\eta}_\nu(Z_i))(1 + \widetilde{\delta}'_{xy}\bar{\eta}_{\nu'}(Z_i))}{\ell_1\ell_2} \\
&= \mathbb{E}_{\nu,\nu',\Delta,\Delta'} \prod_{i=1}^n \left(1 + \frac{\mathbb{E}_{Z_i}\bar{\eta}_\nu(Z_i)\bar{\eta}_{\nu'}(Z_i)}{\ell_1\ell_2} \sum_{(x,y)\in[\ell_1]\times[\ell_2]} \widetilde{\delta}_{xy}\widetilde{\delta}'_{xy} \right) \\
&= \mathbb{E}_{\nu,\nu',\Delta,\Delta'} \prod_{i=1}^n (1 + 4\ell_1\ell_2\rho^2 \langle \nu,\nu'\rangle \langle \Delta,\Delta'\rangle) \\
&\leq \mathbb{E}_{\nu,\nu',\Delta,\Delta'} \exp(4n\ell_1\ell_2\rho^2 \langle \nu,\nu'\rangle \langle \Delta,\Delta'\rangle)
\end{aligned}
$$

In the above $\langle \Delta,\Delta'\rangle = \mathrm{Tr}(\Delta^\top \Delta')$ is the standard matrix dot product, while $\langle \nu,\nu'\rangle$ is the standard vector dot product and $\mathbb{E}_{\nu,\nu',\Delta,\Delta'}$ is the expectation with respect to independent Rademacher draws of ν,ν',Δ,Δ'. Thus,

$$
\begin{aligned}
\mathbb{E}W^2 &\leq \mathbb{E}_{\Delta,\Delta'}\mathbb{E}_{\nu,\nu'}[\exp(4n\ell_1\ell_2\rho^2\langle\nu,\nu'\rangle\langle\Delta,\Delta'\rangle)] = \mathbb{E}_{\Delta,\Delta'}\cosh(4n\ell_1\ell_2\rho^2\langle\Delta,\Delta'\rangle)^d \\
&\leq \mathbb{E}_{\Delta,\Delta'}\exp((4n\ell_1\ell_2\rho^2\langle\Delta,\Delta'\rangle)^2 d^{d_Z}/2),
\end{aligned}
$$

where we used the inequality $\cosh(x) \leq \exp(x^2/2)$, which can be verified by a Taylor expansion. Next, since when we condition on one value of Δ' all values of $\langle \Delta,\Delta'\rangle$ happen with the same probability as if we conditioned on any other value of Δ' we have the identity

$$
\mathbb{E}_{\Delta,\Delta'}\exp((4n\ell_1\ell_2\rho^2\langle\Delta,\Delta'\rangle)^2 d^{d_Z}/2) = \mathbb{E}_\Delta \exp((4n\ell_1\ell_2\rho^2 \sum_{xy\in[\ell_1']\times[\ell_2']} \delta_{xy})^2 d^{d_Z}/2) \tag{G.1}
$$

Note that if one has i.i.d. Rademacher random variables δ_{xy} and i.i.d. standard normal variables W_{xy} for any nonnegative integers a_{xy} for $x,y \in [\ell_1'] \times [\ell_2']$ one has

$$
\mathbb{E} \prod_{xy\in[\ell_1']\times[\ell_2']} \delta_{xy}^{a_{xy}} \leq \mathbb{E} \prod_{xy\in[\ell_1']\times[\ell_2']} W_{xy}^{a_{xy}}
$$

Expanding the exponential function in (G.1) one can control all moments of δ_{xy} using the inequality above with corresponding moments of W_{xy}. Thus we conclude

$$
\mathbb{E}_\Delta \exp((4n\ell_1\ell_2\rho^2 \sum_{xy\in[\ell_1']\times[\ell_2']} \delta_{xy})^2 d^{d_Z}/2) \leq \mathbb{E}_{\boldsymbol{W}} \exp((4n\ell_1\ell_2\rho^2 \sum_{xy\in[\ell_1']\times[\ell_2']} W_{xy})^2 d^{d_Z}/2) \tag{G.2}
$$

The random variable $\sum_{xy\in[\ell_1']\times[\ell_2']} W_{xy} \sim N(0,\ell_1'\ell_2')$ and therefore $(\sum_{xy\in[\ell_1']\times[\ell_2']} W_{xy})^2/\ell_1'\ell_2' := \chi^2$ has a $\chi^2(1)$ distribution. We have

$$
\mathbb{E}_{\boldsymbol{W}} \exp((4n\ell_1\ell_2\rho^2 \sum_{xy\in[\ell_1']\times[\ell_2']} W_{xy})^2 d^{d_Z}/2) \leq \mathbb{E}_{\chi^2} \exp((4n\ell_1\ell_2\rho^2)^2 \ell_1'\ell_2'\chi^2 d^{d_Z}/2). \tag{G.3}
$$

81

Suppose now that $(4n\ell_1\ell_2\rho^2)^2 d^{d_Z}\ell_1'\ell_2' < 1$. The above is the mgf of a chi-squared random variable and hence equals to

$$\sqrt{\frac{1}{1 - (4n\ell_1\ell_2\rho^2)^2 d^{d_Z}\ell_1'\ell_2'}}.$$

This quantity can be made arbitrarily close to 1 provided that $(4n\ell_1\ell_2\rho^2)^2 d^{d_Z}\ell_1'\ell_2'$ is small. Based on this select $\frac{1}{d} \asymp \frac{(\ell_1\ell_2)^{1/(d_Z+4)}}{n^{2/(4+d_Z)}} \wedge 1$, $\rho \asymp \frac{1}{\ell_1\ell_2 d^{d_Z/2+1}}$ for some sufficiently small constants. This ensures $\frac{1}{\ell_1\ell_2} - \rho\sqrt{d}\|h\|_\infty \geq 0$ and $\ell_1\ell_2\rho d^{d_Z/2}d \leq C$. With these choices we obtain that the critical radius is bounded from below by a constant times $\frac{1}{d} \asymp \frac{(\ell_1\ell_2)^{1/(d_Z+4)}}{n^{2/(4+d_Z)}} \wedge 1$. $\qquad\square$

Next we show a matching upper bound in the case when ℓ_1 and ℓ_2 are not allowed to scale with n.

Theorem G.4 (Finite Discrete X, Y Upper Bound). *Set $d = \lceil n^{2/(4+d_Z)}\rceil$ and let $\tau = \zeta d^{d_Z/2}$ for a sufficiently large absolute constant ζ (depending on L). Finally, suppose that $\varepsilon \geq cn^{-2/(4+d_Z)}$, for a sufficiently large constant c (depending on ζ, L, ℓ_1, ℓ_2). Then we have that*

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

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

Proof. The proof follows very closely that of Theorem 5.2 so we omit most details. We only note that in the variance calculations the d has to be substituted with d^{d_Z} and the threshold needs to be substituted from $\sqrt{d}$ to $\sqrt{d^{d_Z}}$. This yields the following two rates

$$\sqrt{\frac{\sqrt{d^{d_Z}}}{n}} \vee \frac{1}{d},$$

and

$$\frac{\sqrt[4]{\sqrt{d^{d_Z}}d^3}}{n} \vee \frac{1}{d},$$

It turns out that when $d \asymp n^{2/(4+d_Z)}$ the second rate coincides with the first one, hence the proof is complete. $\qquad\square$

G.2 Continuous Case

Define $\mathcal{P}_{0,[0,1]^{2+d_Z}}$ and $\mathcal{Q}_{0,[0,1]^{2+d_Z}}$ analogously to $\mathcal{P}_{0,[0,1]^3}$ and $\mathcal{Q}_{0,[0,1]^3}$.

Definition G.5 (Null Lipschitzness). *Let $\mathcal{P}_{0,[0,1]^{2+d_Z},\mathrm{TV}}(L) \subset \mathcal{P}_{0,[0,1]^{2+d_Z}}$ be the collection of distributions $p_{X,Y,Z}$ such that for all $z, z' \in [0,1]$ we have:*

$$\|p_{X|Z=z} - p_{X|Z=z'}\|_1 \leq L\|z - z'\|_2 \ and \ \|p_{Y|Z=z} - p_{Y|Z=z'}\|_1 \leq L\|z - z'\|_2,$$

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.

Definition G.6 (Alternative Lipschitzness). *Let $\mathcal{Q}_{0,[0,1]^{2+d_Z},\mathrm{TV}}(L,s) \subset \mathcal{Q}_{0,[0,1]^{2+d_Z}}$ be the collection of distributions $p_{X,Y,Z}$ such that for all $z, z' \in [0,1]^{d_Z}$ we have*

$$\|p_{X,Y|Z=z} - p_{X,Y|Z=z'}\|_1 \le L\|z - z'\|_2,$$

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 $z, x, y \in [0,1]$: $p_{X,Y|Z}(x,y|z) \in \mathcal{H}^{2,s}(L)$.

Theorem G.7 (Critical Radius Lower Bound). *Let $\overline{\mathcal{H}}_0 = \mathcal{P}_{0,[0,1]^{2+d_Z}}$. Suppose that $\mathcal{H}_0$ is $\mathcal{P}_{0,[0,1]^{2+d_Z},\mathrm{TV}}(L)$, and $\mathcal{H}_1 = \mathcal{Q}_{0,[0,1]^{2+d_Z},\mathrm{TV}}(L,s)$ for some fixed $L, s \in \mathbb{R}^+$. Then we have that for some absolute constant $c_0 > 0$,*

$$\varepsilon_n(\mathcal{H}_0, \overline{\mathcal{H}}_0, \mathcal{H}_1) \ge \frac{c_0}{n^{2s/((4+d_Z)s+2)}}.$$

Proof of Theorem G.7. Suppose $(X, Y, Z) \in [0,1]^{2+d_Z}$ are variables with a joint density with respect to the Lebesgue measure in $[0,1]^{2+d_Z}$. Under the null hypothesis we specify the distribution as $p_{X,Y,Z}(x, y, z) = 1$ for all $(x, y, z) \in [0,1]^{2+d_Z}$, or in other words the variables have independent uniform distributions on $[0,1]$. Clearly this distribution belongs to the sets $\mathcal{P}_{0,[0,1]^{2+d_Z},\mathrm{TV}}(L)$ and $\mathcal{P}_{0,[0,1]^{2+d_Z},\chi^2}(L)$. Under the alternative hypothesis we specify the distribution as

$$q_{X,Y|Z}(x, y|z) = 1 + \gamma_\Delta(x,y)\eta_\nu(z),$$

where as in the proof of Theorem 4.1

$$\eta_\nu(z) = \rho \sum_{j \in [d]} \nu_{j_1,\dots j_{d_Z}} \prod_{k=1}^{d_Z} h_{j_k, d}(z_k),$$

where $\rho > 0$ is a constant, $d \in \mathbb{N}$, $\nu_{j_1,\dots j_{d_Z}} \in \{-1, +1\}$,and $h_{j,d}(z) = \sqrt{d}h(dz - j + 1)$ for $z \in [(j-1)/d, j/d]$, and h is an infinitely differentiable function supported on $[0,1]$ such that $\int h(z)dz = 0$ and $\int h^2(z)dz = 1$. Furthermore we take

$$\gamma_\Delta(x,y) = \rho^2 \sum_{j \in [d']} \sum_{i \in [d']} \delta_{ij} h_{i,d'}(x) h_{j,d'}(y), [8]$$

and we let the marginal distribution of Z be uniform on $[0,1]^{d_Z}$ (i.e. we let $q_Z = p_Z \equiv 1$). In order for this perturbation to be meaningful we need that $1 \ge \sqrt{(d')^2}\sqrt{d}^{d_Z}\|h\|_\infty^{2+d_Z}\rho^3$. It is simple to check that $\int_{[0,1]^2} q_{X,Y|Z}(x,y|z)dxdy = 1$. Let us now check what are the marginals of such a distribution conditioned on z. We have

$$q_{X|Z}(x|z) = \int_{[0,1]} q_{X,Y|Z}(x,y|z)dy = 1 + \eta_\nu(z)\int_{[0,1]} \gamma_\Delta(x,y)dy = 1.$$

Similarly $\int_{[0,1]} q_{X,Y|Z}(x,y|z)dx = 1$. It is therefore clear that $q_{X|Z}(x|z)q_{Y|Z}(y|z) \in \mathcal{H}^{2,s}(L)$ for any L. We now check how far away is the distribution $q_{X,Y,Z}(x,y,z) = q_{X,Y|Z}(x,y|z)q_Z(z) = $

---

[8]Here $\Delta = \{\delta_{ij}\}_{i,j \in [d]}$.

$q_{X,Y|Z}(x,y|z)p_Z(z) = q_{X,Y|Z}(x,y|z)$ with respect to $p_{X,Y,Z}(x,y,z)$ (note that $p_{X,Y,Z} = q_{X|Z}q_{Y|Z}q_Z$) in total variation. We have

$$\|q_{X,Y|Z}q_Z - p_{X,Y,Z}\|_1 = \int_{[0,1]^{2+d_Z}} |\gamma_\Delta(x,y)\eta_\nu(z)|dxdydz$$

$$= \int_{[0,1]} \rho \sum_{j\in[d']} |h_{j,d'}(x)|dx \int_{[0,1]} \rho \sum_{j\in[d']} |h_{j,d'}(y)|dy \int_{[0,1]^{d_Z}} \rho \sum_{j\in[d]} \prod_{k=1}^{d_Z} |h_{j_k,d}(z_k)| \prod_{k=1}^{d_Z} dz_k.$$

Calculating each of the above integrals and multiplying them yields

$$\|q_{X,Y|Z}p_Z - p_{X,Y,Z}\|_1 = \|q_{X,Y,Z} - q_{X|Z}q_{Y|Z}q_Z\|_1 = \sqrt{d}^{d_Z}\sqrt{(d')^2}\rho^3 c^{2+d_Z},$$

where $c = \int_{[0,1]} |h(x)|dx$. Using Lemma B.4 (here we use this lemma with a slight abuse of notation since the lemma is only valid for discrete X, Y and continuous Z, but the same proof extends to the continuous case) we have that

$$\inf_{p\in\mathcal{P}_{0,[0,1]^{2+d_Z}}} \|q - p\|_1 \geq \frac{\sqrt{d}^{d_Z}\sqrt{(d')^2}\rho^3 c^{2+d_Z}}{6}.$$

Next we check that the TV between the distributions $(X,Y|Z=z)$ and $(X,Y|Z=z')$ is Lipschitz.

$$\int_{[0,1]^2} |\gamma_\Delta(x,y)||\eta_\nu(z) - \eta_\nu(z')|dxdy = \sqrt{(d')^2}\rho^2 c^2|\eta_\nu(z) - \eta_\nu(z')|.$$

It is simple to see that by telescoping and triangle inequality we have

$$\left| \rho \sum_{j\in[d]} \nu_{j_1,\dots j_{d_Z}} \prod_{k=1}^{d_Z} h_{j_k,d}(z_k) - \rho \sum_{j\in[d]} \nu_{j_1,\dots j_{d_Z}} \prod_{k=1}^{d_Z} h_{j_k,d}(z_k') \right| \leq \rho\sqrt{d}^{d_Z}\|h\|_\infty^{d_Z-1}d\|h'\|_\infty \sqrt{d_Z}\|\mathbf{z} - \mathbf{z}'\|_2,$$

and therefore we need $\sqrt{d}^{d_Z}d\sqrt{(d')^2}\rho^3$ to be smaller than some constant. Next we check that $q(x,y|z)$ belongs to the Hölder class in x and y. We have that

$$\left| \frac{\partial^k}{\partial x^k}\frac{\partial^{\lfloor s\rfloor-k}}{\partial y^{\lfloor s\rfloor-k}}\gamma_\Delta(x,y)\eta_\nu(z) - \frac{\partial^k}{\partial x^k}\frac{\partial^{\lfloor s\rfloor-k}}{\partial y^{\lfloor s\rfloor-k}}\gamma_\Delta(x',y')\eta_\nu(z) \right|$$

$$\leq \sqrt{d}\rho\|h\|_\infty \left| \frac{\partial^k}{\partial x^k}\frac{\partial^{\lfloor s\rfloor-k}}{\partial y^{\lfloor s\rfloor-k}}\gamma_\Delta(x,y) - \frac{\partial^k}{\partial x^k}\frac{\partial^{\lfloor s\rfloor-k}}{\partial y^{\lfloor s\rfloor-k}}\gamma_\Delta(x',y') \right|$$

$$= \sqrt{d}\rho\|h\|_\infty\rho^2(\sqrt{d'})^2(d')^{\lfloor s\rfloor} \left| \sum_{i,j\in[d']} \delta_{ij}\left[h^{(k)}(d'x - i + 1)h^{(\lfloor s\rfloor-k)}(d'y - j + 1) \right. \right.$$

$$\left. \left. - h^{(k)}(d'x' - i + 1)h^{(\lfloor s\rfloor-k)}(d'y' - j + 1) \right] \right|.$$

Suppose now that $x \in \left[\frac{i_x-1}{d'}, \frac{i_x}{d'}\right], y \in \left[\frac{j_y-1}{d'}, \frac{j_y}{d'}\right]$, and $x' \in \left[\frac{i_{x'}-1}{d'}, \frac{i_{x'}}{d'}\right], y' \in \left[\frac{j_{y'}-1}{d'}, \frac{j_{y'}}{d'}\right]$. Therefore the above summation can be bounded as

$$\sqrt{d}\rho\|h\|_\infty\rho^2(\sqrt{d'})^2(d')^{\lfloor s\rfloor}\left[\left| h^{(k)}(d'x - i_x + 1)h^{(\lfloor s\rfloor-k)}(d'y - j_y + 1) - h^{(k)}(d'x' - i_x + 1)h^{(\lfloor s\rfloor-k)}(d'y' - j_y + 1) \right| \right.$$

$$\left. + \left| h^{(k)}(d'x - i_{x'} + 1)h^{(\lfloor s\rfloor-k)}(d'y - j_{y'} + 1) - h^{(k)}(d'x' - i_{x'} + 1)h^{(\lfloor s\rfloor-k)}(d'y' - j_{y'} + 1) \right| \right].$$

Next we will handle the first expression in the bracket above:

$$
\begin{aligned}
&\leq |h^{(k)}(d'x - i_x + 1) - h^{(k)}(d'x' - i_x + 1)||h^{(\lfloor s\rfloor - k)}(d'y - j_y + 1)| \\
&+ |h^{(k)}(d'x' - i_x + 1)||h^{(\lfloor s\rfloor - k)}(d'y - j_y + 1) - h^{(\lfloor s\rfloor - k)}(d'y' - j_y + 1)| \\
&\leq d'\|h^{(k+1)}\|_\infty |x - x'|\|h^{(\lfloor s\rfloor - k)}\|_\infty \wedge (2\|h^{(k)}\|_\infty \|h^{(\lfloor s\rfloor - k)}\|_\infty) \\
&+ d'\|h^{(\lfloor s\rfloor - k)+1}\|_\infty |y - y'|\|h^k\|_\infty \wedge (2\|h^{(\lfloor s\rfloor - k)}\|_\infty \|h^k\|_\infty) \\
&\leq C(1 \wedge d'\sqrt{(x - x')^2 + (y - y')^2}) \\
&\leq C(d'\sqrt{(x - x')^2 + (y - y')^2})^{s - \lfloor s\rfloor},
\end{aligned}
$$

where in the last inequality we used that $(1 \wedge u)^a \leq u^a$ for $u > 0$ and $0 \leq a \leq 1$. We can handle the second expression in the bracket above in a similar way.

In addition it is clear that any lower order $k \leq \lfloor s\rfloor$ partial derivatives with respect to x and y of $q(x, y|z)$ are bounded by $(\sqrt{d})^{d_Z}\rho^3\|h\|_\infty^{d_Z}\sqrt{d'}^2(d')^k C$ for some constant C which will depend on the function h.

It therefore suffices that $\sqrt{d}^{d_Z}\rho^3(\sqrt{d'})^2(d')^s$ to be smaller than a constant and we will have both conditions satisfied. Now we write down the likelihood ratio between the null and the alternative mixing over all choices of Rademacher vector and matrix ν, Δ:

$$
W = \mathbb{E}_{\nu,\Delta}\prod_{i=1}^n (1 + \gamma_\Delta(X_i, Y_i)\eta_\nu(Z_i)).
$$

The second moment of W is

$$
\begin{aligned}
\mathbb{E}W^2 &= \mathbb{E}_{\nu,\nu',\Delta,\Delta'}\prod_{i=1}^n \mathbb{E}_0(1 + \gamma_\Delta(X_i, Y_i)\eta_\nu(Z_i))(1 + \gamma_{\delta'}(X_i, Y_i)\eta_{\nu'}(Z_i)) \\
&= \mathbb{E}_{\nu,\nu',\Delta,\Delta'}\prod_{i=1}^n (1 + \mathbb{E}_0\gamma_\Delta(X_i, Y_i)\eta_\nu(Z_i)\gamma_{\delta'}(X_i, Y_i)\eta_{\nu'}(Z_i)),
\end{aligned}
$$

where the above follows from the fact that $\mathbb{E}_0\eta_\nu(Z_i) = 0$ (and that X_i and Y_i are independent of Z_i under the null hypothesis). Continuing the identities yields

$$
\begin{aligned}
\mathbb{E}W^2 &= \mathbb{E}_{\nu,\nu',\Delta,\Delta'}\prod_{i=1}^n (1 + \mathbb{E}_0\gamma_\Delta(X_i, Y_i)\gamma_{\delta'}(X_i, Y_i)\mathbb{E}_0\eta_\nu(Z_i)\eta_{\nu'}(Z_i)) \\
&= \mathbb{E}_{\nu,\nu',\Delta,\Delta'}\prod_{i=1}^n (1 + \rho^6\langle\Delta, \Delta'\rangle\langle\nu, \nu'\rangle),
\end{aligned}
$$

where $\langle\Delta, \Delta'\rangle = \mathrm{Tr}(\Delta^\top\Delta')$. From here the proof can continue as in Theorem 4.1. The final expression that needs to be smaller than a constant is $(n\rho^6)^2 d^{d_Z}(d')^2$. Set $d \asymp n^{2s/((4+d_Z)s+2)}$, $d' = d^{1/s}$, $\rho^3 \asymp d^{-(1+d_Z/2+1/s)}$. This results in a rate $\asymp 1/d \asymp n^{-2s/((4+d_Z)s+2)}$. $\qquad\square$

We will now sketch the details of the following theorem:

Theorem G.8 (Continuous X, Y, Z Upper Bound). *Suppose $s \geq 1$. Set $d = \lceil n^{2s/((4+d_Z)s+2)} \rceil$, $d' = d^{1/s}$ and set the threshold $\tau = \zeta(\sqrt{d})^{d_Z}$ for a sufficiently large ζ (depending on L). Then, for a sufficiently large absolute constant c (depending on ζ, L), when $\varepsilon \geq cn^{-2s/((4+d_Z)s+2)}$, we have that*

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

$$\sup_{p \in \{p \in \mathcal{Q}_{0,[0,1]^{2+d_Z},\mathrm{TV}}(L,s):\inf_{q \in \mathcal{P}_{0,[0,1]^{2+d_Z}}} \|p-q\|_1 \geq \varepsilon\}} \mathbb{E}_p[1 - \psi_\tau(\mathcal{D}'_k)] \leq \frac{1}{10} + \exp(-n/8).$$

Proof of Theorem G.8. For what follows set for brevity $\ell_1 = \ell_2 = d^{1/s}$. We only need minor modifications in the variance calculation to track the dimension of Z.

We will now select a threshold at the level of $\zeta(\sqrt{d})^{d_Z}$, and will give conditions on the minimum sample size for each of the cases. We will use $\gtrsim$ in the sense bigger up to an absolute constant. We will assume that $\frac{\varepsilon}{2} - \frac{C}{d} \geq \frac{\varepsilon}{4}$ so that $\eta \geq \frac{\varepsilon}{4}$.

- In the first sub-case we have to satisfy $\frac{n\eta^2}{\sqrt{\ell_1\ell_2}} \gtrsim \zeta(\sqrt{d})^{d_Z}$. This is ensured when

$$\varepsilon \gtrsim \sqrt{\frac{(\sqrt{d})^{d_Z}\sqrt{\ell_1\ell_2}}{n}} \vee \frac{1}{d}. \tag{G.4}$$

- In the second sub-case we have $\frac{n^{3/2}\eta^2}{\ell_1\sqrt{\ell_2}\sqrt{\frac{n\ell_1}{\ell_2}}} \gtrsim \zeta(\sqrt{d})^{d_Z}$ or $\frac{n^{3/2}\eta^2}{\ell_1\sqrt{\ell_2}\sqrt{d}} \gtrsim \zeta(\sqrt{d})^{d_Z}$, This is implied when

$$\varepsilon \gtrsim \min\left(\sqrt{\frac{(\sqrt{d})^{d_Z}\ell_1^{3/2}}{\sqrt{\ell_2}n}}, \sqrt{\frac{\sqrt{d}^{d_Z}\sqrt{d}\ell_1\sqrt{\ell_2}}{n^{3/2}}}, \frac{d^{d_Z}\ell_1}{n}\right) \vee \frac{1}{d}. \tag{G.5}$$

- In the third sub-case case we need $\frac{n^2\eta^2}{36\ell_1\ell_2 d} \gtrsim (\sqrt{d})^{d_Z}$ which happens when

$$\varepsilon \gtrsim \min\left(\frac{\zeta^{1/4}\sqrt{d\sqrt{d}^{d_Z}\ell_1\ell_2}}{n}, \frac{d^{d_Z}\ell_2}{n}\right) \vee \frac{1}{d}, \tag{G.6}$$

where the last condition enforces when this case is not feasible.

- Finally in the second case we need $\frac{n^4\eta^4}{16\ell_1\ell_2 d^3} \gtrsim \sqrt{\zeta d}$, which is implied when

$$\varepsilon \gtrsim \min\left(\frac{\zeta^{1/8}(d^3\sqrt{d}^{d_Z}\ell_1\ell_2)^{1/4}}{n}, \frac{d^{d_Z}}{n}\right) \vee \frac{1}{d} \tag{G.7}$$

Analysis of the Variance.

Now we derive a bound on the variance of the statistic

$$\sum_{m \in [d^{d_Z}]} \sigma_m \omega_m \mathbb{1}(\sigma_m \geq 4)\frac{\varepsilon_m^2}{\ell_1\ell_2}$$

Recall now that $\sigma_m \sim \mathrm{Poi}(\alpha_m)$ are independent and therefore

$$\mathrm{Var}\left[\sum_{m\in[d^d Z]}\sigma_m\omega_m\mathbb{1}(\sigma_m\geq 4)\frac{\varepsilon_m^2}{\ell_1\ell_2}\right] = \sum\mathrm{Var}\left[\sigma_m\omega_m\mathbb{1}(\sigma_m\geq 4)\frac{\varepsilon_m^2}{\ell_1\ell_2}\right]$$

$$= \sum_{m\in[d^d Z]}\frac{\varepsilon_m^4}{\ell_1^2\ell_2^2}\,\mathrm{Var}(\sigma_m\omega_m\mathbb{1}(\sigma_m\geq 4))$$

$$\leq \frac{C'}{\ell_1\ell_2}\sum_{m\in[d^d Z]}\frac{\varepsilon_m^2}{\ell_1\ell_2}\mathbb{E}(\sigma_m\omega_m\mathbb{1}(\sigma_m\geq 4))$$

$$\leq \frac{C'}{\ell_1\ell_2}\mathbb{E}\left[\sum_{m\in[d^d Z]}\sigma_m\omega_m\mathbb{1}(\sigma_m\geq 4)\frac{\varepsilon_m^2}{\ell_1\ell_2}\right], \qquad (G.8)$$

where in the next to last inequality we used Claim 2.2. of [13], and C' is an absolute constant described in that claim.

We now bound the variance of the statistic T (recall the definition (5.6)). Since T_m (recall definition C.7) are independent given σ_m, R_m we have that

$$\mathrm{Var}[T|\sigma, R] = \sum_{m\in[d^d Z]}\mathrm{Var}[T_m|\sigma_m, R_m].$$

Next, by definition of T_m we have that $\mathrm{Var}[T_m|\sigma_m, R_m] = \sigma_m^2\omega_m^2\mathbb{1}(\sigma_m\geq 4)\,\mathrm{Var}[U_m|R_m]$. Using the bound on the variance $\mathrm{Var}[U_m|R_m]$ (C.6), we have to control four terms. We do so below. Denote

$$E := \sum_{m\in[d^d Z]}\omega_m^2\|q_{\Pi, A_m}(m)\|_2^2\mathbb{1}(\sigma_m\geq 4)$$

The first term we need to control is

$$\sum_{m\in[d]}\sigma_m^2\omega_m^2\mathbb{1}(\sigma_m\geq 4)\frac{\|q_{A_m}(m)-q_{\Pi, A_m}(m)\|_2^2\|q_{\Pi, A_m}(m)\|_2}{\sigma_m}$$

$$\leq \sqrt{\left(\sum_{m\in[d]}\omega_m^2\|q_{\Pi, A_m}(m)\|_2^2\mathbb{1}(\sigma_m\geq 4)\right)}\sqrt{\sum_{m\in[d]}(\sigma_m\omega_m\|q_{A_m}(m)-q_{\Pi, A_m}(m)\|_2^2)^2\mathbb{1}(\sigma_m\geq 4)}$$

$$\leq E^{1/2}\sum_{m\in[d^d Z]}\sigma_m\omega_m\|q_{A_m}(m)-q_{\Pi, A_m}(m)\|_2^2\mathbb{1}(\sigma_m\geq 4)$$

$$= E^{1/2}\mathbb{E}[T|\sigma, R],$$

where we used Cauchy-Schwarz and the monotonicity of L_p norms. The second term is

$$\sum_{m\in[d^dZ]} \sigma_m^2\omega_m^2 \mathbb{1}(\sigma_m \geq 4)\frac{\|q_{A_m}(m) - q_{\Pi,A_m}(m)\|_2^3}{\sigma_m} = \sum_{m\in[d^dZ]} \sqrt{\frac{\omega_m}{\sigma_m}}\sigma_m^{3/2}\omega_m^{3/2}\mathbb{1}(\sigma_m \geq 4)\|q_{A_m}(m) - q_{\Pi,A_m}(m)\|_2^3$$

$$\leq \sum_{m\in[d^dZ]} (\sigma_m\omega_m\mathbb{1}(\sigma_m \geq 4)\|q_{A_m}(m) - q_{\Pi,A_m}(m)\|_2^2)^{3/2}$$

$$\leq \left(\sum_{m\in[d^dZ]} \sigma_m\omega_m\mathbb{1}(\sigma_m \geq 4)\|q_{A_m}(m) - q_{\Pi,A_m}(m)\|_2^2\right)^{3/2}$$

$$= \mathbb{E}[T|\sigma, R]^{3/2},$$

where we used that $\omega_m \leq \sigma_m$ by definition and the monotonicity of the L_p norms. The third term is

$$\sum_{m\in[d^dZ]} \sigma_m^2\omega_m^2 \mathbb{1}(\sigma_m \geq 4)\frac{\|q_{\Pi,A_m}(m)\|_2^2}{\sigma_m^2} = E.$$

Finally the fourth term is

$$\sum_{m\in[d^dZ]} \sigma_m^2\omega_m^2 \mathbb{1}(\sigma_m \geq 4)\frac{\|q_{A_m}(m) - q_{\Pi,A_m}(m)\|_2^2}{\sigma_m^2} \leq \sum_{m\in[d^dZ]} \omega_m\sigma_m\mathbb{1}(\sigma_m \geq 4)\|q_{A_m}(m) - q_{\Pi,A_m}(m)\|_2^2$$

$$= \mathbb{E}[T|\sigma, R].$$

We conclude that

$$\mathrm{Var}[T|\sigma, R] \leq C(E + (E^{1/2} + 1)\mathbb{E}[T|\sigma, R] + \mathbb{E}[T|\sigma, R]^{3/2}). \tag{G.9}$$

Now we will show that $\mathbb{E}[E|\sigma] = O(\min(d^{dZ}, N))$. We start by analyzing the expectation of one term from E below.

$$\mathbb{E}[\omega_m^2\|q_{\Pi,A_m}(m)\|_2^2\mathbb{1}(\sigma_m \geq 4)|\sigma_m] = \omega_m^2\mathbb{1}(\sigma_m \geq 4)\mathbb{E}[\|q_{\Pi,A_m}(m)\|_2^2|\sigma_m] \leq \frac{\omega_m^2\mathbb{1}(\sigma_m \geq 4)}{(1 + t_{1,m})(1 + t_{2,m})},$$

where we applied (C.10). Recall that $t_{i,m} = \min((\sigma_m - 4)/4, \ell_i)$ and $\omega_m^2 = \min(\sigma_m, \ell_1)\min(\sigma_m, \ell_2)$. Thus

$$\frac{\omega_m^2}{(1 + t_{1,m})(1 + t_{2,m})} \leq O(1).$$

We conclude that

$$\mathbb{E}[E|\sigma] = \mathbb{E}[\sum_{m\in[d^dZ]} \omega_m^2\|q_{\Pi,A_m}(m)\|_2^2\mathbb{1}(\sigma_m \geq 4)|\sigma_m] \leq O(1)\sum_{m\in[d^dZ]} \mathbb{1}(\sigma_m \geq 4) \leq O(1)\min(d^{dZ}, N).$$

$$\tag{G.10}$$

We have the following result

Lemma G.9. *Suppose* $\inf_{q \in \mathcal{P}'_{0,[0,1]}} \|p_{X,Y,Z} - q\|_1 \geq \varepsilon$, *where* $\varepsilon \geq C\frac{L}{d}$ *and it satisfies conditions* (G.4), (G.5), (G.6) *and* (G.7). *Then with probability at least* $19/20$ *over* σ, R *we have* $\mathbb{E}[T|\sigma, R] = \Omega(\sqrt{\zeta d^{dz}})$ *and*

$$\mathrm{Var}[T|\sigma, R] \leq O(d^{dz} + (\sqrt{d^{dz}} + 1)\mathbb{E}[T|\sigma, R] + \mathbb{E}[T|\sigma, R]^{3/2}). \tag{G.11}$$

Proof. Let

$$D = \sum_{m \in [d^{dZ}]} \sigma_m \omega_m \mathbb{1}(\sigma_m \geq 4) \frac{\varepsilon_m^2}{\ell_1 \ell_2}.$$

We first showed that $\mathbb{E}[T|\sigma, R] \geq D$ for all σ, R (C.11). We also derived that $\mathrm{Var}[D] \leq O(\mathbb{E}[D]/(\ell_1 \ell_2))$ (G.8), and that for the selected regimes of sample size $\mathbb{E}[D] \gtrsim \sqrt{\zeta d^{dz}}$. Therefore we have

$$\mathbb{P}_{\sigma,R}\left(\mathbb{E}[T|\sigma, R] \leq \kappa \sqrt{\zeta d^{dz}}\right) \leq \mathbb{P}_{\sigma,R}(D \leq O(\mathbb{E}[D])) \leq O\left(\frac{\mathrm{Var}[D]}{(\mathbb{E}[D])^2}\right) = O(1/(\sqrt{\zeta d^{dz}}\ell_1 \ell_2)) \leq 1/40,$$

for some small enough absolute constant κ. For the second statement we will use bound (G.9). By (G.10) we have

$$\mathbb{E}[\mathbb{E}[E|\sigma]] \leq O(1)\mathbb{E}\min(d^{dz}, N) \leq O(1)\min(d, n).$$

Thus by Markov's inequality $E \leq 200\mathbb{E}[E] = O(1)d^{dz}$ with probability at least $39/40$. Therefore

$$\mathbb{P}_{\sigma,R}(\mathrm{Var}[T|\sigma, R] \geq \kappa'(d^{dz} + (\sqrt{d^{dz}} + 1)\mathbb{E}[T|\sigma, R] + \mathbb{E}[T|\sigma, R]^{3/2})) \leq 1/40.$$

A union bound over the two events completes the proof. $\qquad\square$

We now turn to bound the expectation and variance under the null hypothesis.

Lemma G.10. *Suppose* $p_{X,Y,Z} \in \mathcal{P}_{0,[0,1]^3,\mathrm{TV}}(L)$. *Then with probability at least* $19/20$ *we have* $\mathbb{E}[T|\sigma, R] \leq C\frac{L^2 n}{d^2 \sqrt{\ell_1 \ell_2}}$ *and the variance* $\mathrm{Var}[T|\sigma, R]$ *satisfies* (G.11).

Proof. Let us start with bounding $\mathbb{E}[T|\sigma, R]$ from above. Recall that

$$\mathbb{E}[T|\sigma, R] = \sum_{m \in [d]} \sigma_m \omega_m \|q_{A_m}(m) - q_{\Pi, A_m}(m)\|_2^2 \mathbb{1}(\sigma_m \geq 4)$$

We will now control $\mathbb{E}[\|q_{A_m}(m) - q_{\Pi, A_m}(m)\|_2^2 | \sigma_m]$. Recall that

$$\|q_{A_m}(m) - q_{\Pi, A_m}(m)\|_2^2 = \sum_{x,y} \frac{(q_{xy}(m) - q_{x\cdot}(m)q_{\cdot y}(m))^2}{1 + a_{xy}^m} \tag{G.12}$$
$$\leq \sum_{x,y} (q_{xy}(m) - q_{x\cdot}(m)q_{\cdot y}(m))^2$$

Using the strategy of Lemma C.16 we can bound the above by C/d^4 for some constant depending on L. Therefore,

$$\mathbb{E}[T|\sigma] = \sum_{m \in [d^{dZ}]} \sigma_m \omega_m \mathbb{1}(\sigma_m \geq 4) \mathbb{E}_{A_m} \|q_{A_m}(m) - q_{\Pi, A_m}(m)\|_2^2$$
$$\leq C/d^4 \sum_{m \in [d^{dZ}]} \sigma_m \omega_m \mathbb{1}(\sigma_m \geq 4) \leq C/d^4 N \sqrt{\ell_1 \ell_2}.$$

Hence $\mathbb{E}T \leq C\frac{n\sqrt{\ell_1\ell_2}}{d^4}$. By Markov's inequality we therefore have

$$\mathbb{P}_{\sigma,R}\left(\mathbb{E}[T|\sigma,R] \geq 40C\frac{n\sqrt{\ell_1\ell_2}}{d^4}\right) \leq \frac{1}{40}.$$

Since $\ell_1 = \ell_2 = d^{1/s}$ and $s \geq 1$we have that the above is smaller than $C\frac{n\sqrt{\ell_1\ell_2}}{d^4} \leq C\frac{n}{d^2\sqrt{\ell_1\ell_2}}$, which completes the proof. $\qquad\square$

Putting Things Together. For what follows suppose that d is selected so that

$$\frac{n}{d^2\sqrt{\ell_1\ell_2}} \asymp (\sqrt{d})^{d_Z}. \tag{G.13}$$

Lemma G.11. *If $p_{X,Y,Z} \in \mathcal{P}_{0,[0,1]^3,\mathrm{TV}}(L)$ and that (G.13) holds. Then for a sufficiently large absolute constant α we have*

$$\mathbb{P}\left(T \geq (\alpha+1)\frac{C'n}{d^2\sqrt{\ell_1\ell_2}}\right) \leq \frac{1}{10}.$$

Proof. Let $T' = (T|\sigma,R)$. Denote the event from Lemma G.10 with $\mathcal{E}$. Then we have

$$\mathbb{P}\left(T \geq (\alpha+1)\frac{C'n}{d^2\sqrt{\ell_1\ell_2}}\right) = \mathbb{P}\left(T' \geq (\alpha+1)\frac{C'n}{\sqrt{\ell_1\ell_2}d^2}\right)$$
$$\leq \mathbb{P}\left(T' \geq (\alpha+1)\frac{C'n}{\sqrt{\ell_1\ell_2}d^2}\,\Big|\,\mathcal{E}\right) + \mathbb{P}(\mathcal{E}^c).$$

Now we have

$$\mathbb{P}\left(T' \geq (\alpha+1)\frac{C'n}{\sqrt{\ell_1\ell_2}d^2}\,\Big|\,\mathcal{E}\right) \leq \mathbb{P}\left(T' - \mathbb{E}[T|\sigma,R] \geq \alpha\frac{C'n}{\sqrt{\ell_1\ell_2}d^2}\,\Big|\,\mathcal{E}\right) \leq \frac{\mathrm{Var}[T'|\mathcal{E}]}{(\alpha\frac{C'n}{\sqrt{\ell_1\ell_2}d^2})^2}$$
$$\leq \frac{O(d^{d_Z} + (\sqrt{d^{d_Z}}+1)\mathbb{E}[T|\sigma,R] + \mathbb{E}[T|\sigma,R]^{3/2})}{(\alpha\frac{C'n}{\sqrt{\ell_1\ell_2}d^2})^2}$$
$$\leq \frac{O(d^{d_Z} + (\sqrt{d^{d_Z}}+1)(\frac{C'n}{\sqrt{\ell_1\ell_2}d^2}) + (\frac{C'n}{\sqrt{\ell_1\ell_2}d^2})^{3/2})}{(\alpha\frac{C'n}{\sqrt{\ell_1\ell_2}d^2})^2}$$
$$\leq \frac{1}{20}$$

where the above holds when $\frac{n}{\sqrt{\ell_1\ell_2}d^2} \asymp (\sqrt{d})^{d_Z}$ for a large enough α. $\qquad\square$

Lemma G.12. *If $p_{X,Y,Z}$ is such that $\inf_{q\in\mathcal{P}'_{0,[0,1]}} \|p_{X,Y,Z} - q\|_1 \geq \varepsilon$, and the conditions of Lemma G.9 hold. Then for a small enough absolute constant κ we have that*

$$\mathbb{P}(T \leq \kappa\zeta(\sqrt{d})^{d_Z}) \leq \frac{1}{10}.$$

Proof. We apply Chebyshev's inequality to $T' = (T|\sigma, R)$. Let $\mathcal{E}$ be the event of Lemma G.9. Set $\tau = \kappa\zeta(\sqrt{d})^{d_Z}$ for some small enough absolute constant κ.

$$\mathbb{P}(T \leq \tau) = \mathbb{P}(T' \leq \tau) \leq \mathbb{P}(T' \leq \tau | \mathcal{E}) + \mathbb{P}(\mathcal{E}^c).$$
$$\leq \mathbb{P}\left(|T' - \mathbb{E}[T|\sigma, R]| \geq \frac{1}{2}\mathbb{E}[T|\sigma, R]|\mathcal{E}\right) + \frac{1}{20}.$$

Next

$$\mathbb{P}\left(|T' - \mathbb{E}[T|\sigma, R]| \geq \frac{1}{2}\mathbb{E}[T|\sigma, R]|\mathcal{E}\right) \leq O\left(\frac{d^{d_Z} + (\sqrt{d^{d_Z}} + 1)\mathbb{E}[T|\sigma, R] + \mathbb{E}[T|\sigma, R]^{3/2}}{\mathbb{E}[T|\sigma, R]^2}\right)$$
$$= O(1/\zeta^{1/2}) \leq \frac{1}{20},$$

for a large enough value of ζ. $\qquad\square$

Combining Lemmas G.11 and G.12 we have that if $\kappa\zeta(\sqrt{d})^{d_Z} \geq (\alpha + 1)nC'/(\sqrt{\ell_1\ell_2}d^2) \asymp (\alpha + 1)C'(\sqrt{d})^{d_Z}$, there will be a gap between the values under the null and the alternative hypothesis. This happens when ζ is large enough. Notice that $\frac{nC'}{\sqrt{\ell_1\ell_2}d^2} \asymp (\sqrt{d})^{d_Z}$ is equivalent to $d \asymp n^{\frac{2s}{(4+d_Z)s+2}}$. Plugging this in all the inequalities (G.4), (G.5), (G.6) and (G.7), results in new inequalities that need to hold. Condition (G.4) is equivalent to

$$\varepsilon \gtrsim n^{-\frac{2s}{(4+d_Z)s+2}}.$$

Taking the second term of (G.5), and using the fact that $d\ell_1 \lesssim n$ we have

$$\varepsilon \gtrsim n^{-\frac{2s}{(4+d_Z)s+2}},$$

since $\sqrt{\frac{\sqrt{d^{d_Z}}\sqrt{d}\ell_1\sqrt{\ell_2}}{n^{3/2}}} = \sqrt[4]{\frac{d\ell_1}{n}}\sqrt{\frac{\sqrt{d^{d_Z}}\sqrt{\ell_1\ell_2}}{n}}$. Taking (G.6), we have

$$\varepsilon \gtrsim n^{-\frac{2s}{(4+d_Z)s+2}}$$

$\frac{\sqrt{d\sqrt{d^{d_Z}}\ell_1\ell_2}}{n} = \sqrt{d\sqrt{\ell_2\ell_2}/n}\frac{\sqrt{\sqrt{d^{d_Z}}\sqrt{\ell_1\ell_2}}}{n} \leq \frac{\sqrt{\sqrt{d^{d_Z}}\sqrt{\ell_1\ell_2}}}{n}$, when $d\ell_1 \lesssim n$ (which is the case). Finally for (G.7) we have

$$\varepsilon \gtrsim n^{-\frac{2s}{(4+d_Z)s+2}},$$

which can be checked with a direct calculation.

The rest of the proof can continue as in the proof of Theorem 5.6 and Theorem C.15. We omit the details. $\qquad\square$