Double Generative Adversarial Networks for Conditional Independence Testing

Chengchun Shi C.SHI7@LSE.AC.UK
Tianlin Xu T.XU12@LSE.AC.UK
Wicher Bergsma W.P.BERGSMA@LSE.AC.UK
Department of Statistics, London School of Economics and Political Science

Lexin Li LEXINLI@BERKELEY.EDU
Department of Biostatistics and Epidemiology, University of California at Berkeley

Editor:

Abstract

In this article, we study the problem of high-dimensional conditional independence testing, a key building block in statistics and machine learning. We propose an inferential procedure based on double generative adversarial networks (GANs). Specifically, we first introduce a double GANs framework to learn two generators of the conditional distributions. We then integrate the two generators to construct a test statistic, which takes the form of the maximum of generalized covariance measures of multiple transformation functions. We also employ data-splitting and cross-fitting to minimize the conditions on the generators to achieve the desired asymptotic properties, and employ multiplier bootstrap to obtain the corresponding p-value. We show that the constructed test statistic is doubly robust, and the resulting test both controls type-I error and has the power approaching one asymptotically. Also notably, we establish those theoretical guarantees under much weaker and practically more feasible conditions compared to the existing tests, and our proposal gives a concrete example of how to utilize some state-of-the-art deep learning tools, such as GANs, to help address a classical but challenging statistical problem. We demonstrate the efficacy of our test through both simulations and an application to an anti-cancer drug dataset.

Keywords: Conditional independence; Double-robustness; Generalized covariance measure; Generative adversarial networks; Multiplier bootstrap.

1. Introduction

Conditional independence (CI) is a fundamental concept in statistics and machine learning. Testing conditional independence is a key building block and plays a central role in a large variety of statistical learning problems, for instance, causal inference (Pearl, 2009), graphical models (Koller and Friedman, 2009), dimension reduction (Li, 2018), among many others. It is frequently used in a wide range of scientific and business applications, and we demonstrate its application with a cancer genetics example later.

In this article, we aim at testing whether two random variables X and Y are conditionally independent given a set of confounding variables Z. That is, we test the hypotheses:

$$\mathcal{H}_0 : X \perp\!\!\!\perp Y \mid Z \qquad \text{versus} \qquad \mathcal{H}_1 : X \not\perp\!\!\!\perp Y \mid Z, \tag{1}$$

given the observed data of n i.i.d. copies $\{(X_i, Y_i, Z_i)\}_{1\leq i\leq n}$ of (X, Y, Z). For our problem, X, Y and Z can all be multivariate. However, the main challenge arises when the confounding set of variables Z is multivariate and high-dimensional. As such, we primarily focus on the scenario where X and Y are univariate, and Z is multivariate and its dimension can potentially diverge to infinity. Meanwhile, our proposed method can be readily extended to the scenario of multivariate X and Y as well. Another challenge is the limited sample size compared to the dimensionality of Z. As a result, many existing tests may become ineffective, suffering from either an inflated type-I error, or not having enough power to detect the alternatives. See Section 2 for a detailed literature review.

To deal with those challenges, we propose a testing procedure based on double generative adversarial networks (GANs, Goodfellow et al., 2014) for the CI testing problem in (1). GANs have recently stood out as a powerful approach for learning and generating random samples from a complex, high-dimensional data distribution. They have been successfully applied in numerous applications, ranging from image processing and computer vision, to sequential data modeling such as natural language, music, speech, and to medical fields such as DNA design and drug discovery; see Gui et al. (2020) for a review of the GANs applications. Moreover, there have recently emerged works studying the consistency and rate of convergence of the GANs estimators; see, e.g., Liang (2018); Chen et al. (2020).

Our proposal involves two key components: a double GANs framework to learn two generators that approximate the conditional distribution of X given Z, and Y given Z, respectively, and a test statistic that is taken as the maximum of generalized covariance measures of multiple transformation functions of X and Y. We first show that our test statistic is doubly-robust, which offers an additional layer of protection against potential misspecification of the conditional distributions; see Theorems 2 and 3. We then show that the resulting test achieves a valid control of the type-I error asymptotically, and more importantly, under the set of conditions that are much weaker and practically more feasible compare to the existing tests; see Theorem 4. Besides, we prove that the power of our test approaches one asymptotically; see Theorem 5, and we demonstrate through simulations that it is more powerful than numerous competing tests empirically. In addition, we employ data splitting and cross-fitting that allow us to derive the asymptotic properties under minimal conditions on the generators, and employ multiplier bootstrap to obtain the corresponding p-value of the test. Our contributions are multi-fold. We develop a useful testing procedure for a fundamentally important statistical inference problem. We establish the statistical guarantees under much weaker conditions. We also give an example of how to utilize some state-of-the-art deep learning tools, such as GANs, to address a classical but challenging statistical problem.

The rest of the article is organized as follows. Section 2 reviews some key existing CI testing methods. Section 3 develops the double GANs-based testing procedure. Section 4 derives the theoretical properties. Section 5 presents the simulations and a cancer genetics data example. Section 6 concludes the paper. The Appendix collects all technical proofs.

2. Literature review on conditional independence testing

There has been a large and growing literature on conditional independence testing; see Li and Fan (2019) for a review. Broadly speaking, the existing tests can be cast into four main

categories, the metric-based tests (e.g., Su and White, 2007, 2014; Wang et al., 2015; Pan et al., 2017; Wang et al., 2018), the conditional randomization-based tests (e.g., Candes et al., 2018; Bellot and van der Schaar, 2019), the kernel-based tests (e.g., Fukumizu et al., 2008; Zhang et al., 2011), and the regression-based tests (e.g., Hoyer et al., 2009; Shah and Peters, 2018). There are also some other types of tests (e.g., Bergsma, 2004; Berrett et al., 2019, to name a few).

The metric-based tests typically employ some kernel smoothers to estimate the conditional characteristic function or the distribution function of Y given X and Z. Kernel smoothers, however, are known to suffer from the curse of dimensionality, and as such, these tests are usually not suitable when the dimension of Z is high. The conditional randomization-based tests require the knowledge of the conditional distribution of $X|Z$ (Candes et al., 2018). If unknown, the type-I error rates of these tests rely critically on the quality of the approximation of this conditional distribution. Kernel-based tests are built upon the notion of maximum mean discrepancy (MMD, Gretton et al., 2012), and could have inflated type-I errors. Regression-based tests have valid type-I error control, but may suffer from inadequate power.

Next, we discuss in detail the conditional randomization-based tests, in particular, the work of Bellot and van der Schaar (2019), the regression-based tests, and the MMD-based tests, as our proposal is related to and built on those methods. For each family of tests, we first lay out the main ideas, then discuss their potential limitations.

2.1 Conditional randomization-based tests

The family of conditional randomization-based tests is built upon the following basis. If the conditional distribution $P_{X|Z}$ of X given Z is known, then one can independently draw $X_i^{(1)} \sim P_{X|Z=Z_i}$, for $i = 1, \ldots, n$, where the superscript denotes the first round of draws. Besides, these samples are independent of the observed samples X_i's and Y_i's. Write $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$, $\boldsymbol{X}^{(1)} = (X_1^{(1)}, \ldots, X_n^{(1)})^\top$, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, and $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^\top$. Hereinafter we use boldface letters to denote data matrices that consist of n samples. Since the joint distributions of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ and $(\boldsymbol{X}^{(1)}, \boldsymbol{Y}, \boldsymbol{Z})$ are the same under $\mathcal{H}_0$, any large difference between the two distributions can be interpreted as evidence against $\mathcal{H}_0$. Therefore, one can repeat the sample drawing process M times, i.e., $X_i^{(m)} \sim P_{X|Z=Z_i}$, $i = 1, \ldots, n$, $m = 1, \ldots, M$. Write $\boldsymbol{X}^{(m)} = (X_1^{(m)}, \ldots, X_n^{(m)})^\top$. Then, for a given test statistic $\rho = \rho(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, the associated p-value is

$$p = \frac{1}{M}\left[\sum_{m=1}^{M} \mathbb{I}\left\{\rho(\boldsymbol{X}^{(m)}, \boldsymbol{Y}, \boldsymbol{Z}) \geq \rho(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})\right\}\right],$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Since the triplets $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}), (\boldsymbol{X}^{(1)}, \boldsymbol{Y}, \boldsymbol{Z}), \ldots,$ $(\boldsymbol{X}^{(M)}, \boldsymbol{Y}, \boldsymbol{Z})$ are exchangeable under $\mathcal{H}_0$, the above p-value is valid, in the sense that it equals the significance level under the null, i.e.,

$$\Pr\left(p \leq \alpha | \mathcal{H}_0\right) = \alpha, \quad \text{for any } \ 0 < \alpha < 1.$$

In practice, however, $P_{X|Z}$ is rarely known. Bellot and van der Schaar (2019) proposed to approximate $P_{X|Z}$ using GANs. Specifically, they learned a generator $\mathbb{G}_X(\cdot, \cdot)$ from the

observed data, then took Z_i along with an independent noise variable as the input to obtain a sample $\widetilde{X}_i^{(m)}$, which minimizes the divergence between the distributions of (X_i, Z_i) and $(\widetilde{X}_i^{(m)}, Z_i)$. They computed the p-value by replacing $\boldsymbol{X}^{(m)}$ with $\widetilde{\boldsymbol{X}}^{(m)} = (\widetilde{X}_1^{(m)}, \ldots, \widetilde{X}_n^{(m)})^\top$. They called this test the *generative conditional independence test* (GCIT). By Theorem 1 of Bellot and van der Schaar (2019), the excess type-I error of this test is upper bounded as,

$$\begin{aligned} \Pr(p \leq \alpha | \mathcal{H}_0) - \alpha &\leq \mathrm{E}\left\{ d_{\mathrm{TV}}\left(\widetilde{P}_{\boldsymbol{X}|\boldsymbol{Z}}, P_{\boldsymbol{X}|\boldsymbol{Z}} \right) \right\} \\ &= \mathrm{E}\left\{ \sup_A \left| \Pr(\boldsymbol{X} \in A | \boldsymbol{Z}) - \Pr(\widetilde{\boldsymbol{X}}^{(m)} \in A | \boldsymbol{Z}) \right| \right\} \equiv D, \end{aligned} \tag{2}$$

where d_{TV} is the total variation norm between two probability distributions P and Q such that $d_{\mathrm{TV}}(P, Q) = \sup_A |P(A) - Q(A)|$, the supremum is taken over all measurable sets A, and the expectations in (2) are taken with respect to $\boldsymbol{Z}$.

By definition, the error term D in (2) measures the quality of the conditional distribution approximation. Bellot and van der Schaar (2019) argued that this error term is negligible due to the capacity of deep neural networks in terms of estimating the conditional distribution. To the contrary, we find this approximation error is usually *not* negligible, and consequently, it may inflate the type-I error and invalidate the test. We consider a simple example to further elaborate this.

Example 1 *Suppose X is one-dimensional, and follows a simple linear regression model, $X = Z^\top \beta_0 + \varepsilon$, where the error ε is independent of Z, and $\varepsilon \sim N(0, \sigma_0^2)$ for some $\sigma_0^2 > 0$.*

Suppose we know a priori that the linear regression model holds. We thus estimate β_0 by ordinary least squares, and denote the resulting estimator by $\widehat{\beta}$. For simplicity, suppose σ_0^2 is known too. For this simple example, we have the following result regarding the approximation error D.

Proposition 1 *Suppose the linear regression model holds, the dimension of Z is much smaller than the sample size n, and the derived distribution $\widetilde{P}_{\boldsymbol{X}|\boldsymbol{Z}}$ is Normal$(\boldsymbol{Z}\widehat{\beta}, \sigma_0^2 I_n)$, where I_n is the $n \times n$ identity matrix. Then D does not decay to zero.*

To facilitate the understanding of the convergence behavior of D, we sketch a few lines of the proof of Proposition 1. The complete proof is given in the Appendix. Let $\widetilde{P}_{X|Z=Z_i}$ denote the conditional distribution of $\widetilde{X}_i^{(m)}$ given Z_i, which is Normal$(Z_i^\top \widehat{\beta}, \sigma_0^2)$ in this example. If $D = o(1)$, then,

$$\widetilde{D} \equiv n^{1/2} \sqrt{\mathrm{E}\left\{ d_{\mathrm{TV}}^2\left(\widetilde{P}_{X|Z=Z_i}, P_{X|Z=Z_i} \right) \right\}} = o(1). \tag{3}$$

In other words, in order to control the type-I error, GCIT requires the total variation distance measure in (3) to converge at a faster rate than $n^{-1/2}$. However, this rate cannot be achieved in general. In our Example 1, we have $\widetilde{D} \geq c$ for some constant $c > 0$. Consequently, D in (2) is not $o(1)$. Proposition 1 shows that, even if we know a priori that the linear model holds, D does not decay to zero as n tends to infinity. In practice, we do not have such prior model information. Then it would be even more difficult to estimate the conditional distribution $P_{X|Z}$. Therefore, using GANs to approximate $P_{X|Z}$ does not guarantee a negligible approximation error.

2.2 Regression-based tests

The family of regression-based tests is built upon the generalized covariance measure,

$$\mathrm{GCM}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left\{X_i - \widehat{\mathrm{E}}(X_i|Z_i)\right\}\left\{Y_i - \widehat{\mathrm{E}}(Y_i|Z_i)\right\},$$

where $\widehat{\mathrm{E}}(X|Z)$ and $\widehat{\mathrm{E}}(Y|Z)$ are the estimated condition means $\mathrm{E}(X|Z)$ and $\mathrm{E}(Y|Z)$, respectively, obtained by some supervised learner. When the prediction errors of $\widehat{\mathrm{E}}(X|Z)$ and $\widehat{\mathrm{E}}(Y|Z)$ satisfy certain convergence rates, Shah and Peters (2018) proved that GCM is asymptotically normal under $\mathcal{H}_0$, in which the asymptotic mean is zero, and the standard deviation can be consistently estimated by some standard error estimator, denoted by $\widehat{s}(\mathrm{GCM})$. Therefore, at level α, we reject $\mathcal{H}_0$, if $|\mathrm{GCM}|/\widehat{s}(\mathrm{GCM})$ exceeds the upper $\alpha/2$th quantile of a standard normal distribution.

Such a test can control the type-I error. Nevertheless, it may not have sufficient power to detect $\mathcal{H}_1$. Consider the asymptotic mean of GCM, which is $\mathrm{GCM}^*(X,Y) = \mathrm{E}\{X - \mathrm{E}(X|Z)\}\{Y - \mathrm{E}(Y|Z)\}$. The regression-based tests require $|\mathrm{GCM}^*|$ to be nonzero under $\mathcal{H}_1$ to have power. However, it may be difficult to satisfy such a requirement. We again consider a simple example.

Example 2 *Suppose X^*, Y and Z are independent random variables. Besides, X^* has mean zero, and $X = X^* g(Y)$ for some function g.*

For this example, we have $\mathrm{E}(X|Z) = \mathrm{E}(X)$, since both X^* and Y are independent of Z, and so is X. Besides, $\mathrm{E}(X) = \mathrm{E}(X^*)\mathrm{E}\{g(Y)\} = 0$, since X^* is independent of Y and $\mathrm{E}(X^*) = 0$. Thus $\mathrm{GCM}^*(X,Y) = \mathrm{E}\{X - \mathrm{E}(X)\}\{Y - \mathrm{E}(Y|Z)\} = 0$ for any function g. On the other hand, X and Y are conditionally dependent given Z, as long as g is not a constant function. Therefore, for this example, the regression-based tests would fail to discriminate between $\mathcal{H}_0$ and $\mathcal{H}_1$.

2.3 MMD-based tests

The family of MMD-based tests involves the maximum mean discrepancy as a measure of independence. For any two probability measures P, Q and a function space $\mathbb{F}$, define

$$\mathrm{MMD}(P,Q|\mathbb{F}) = \sup_{f\in\mathbb{F}}\left\{\mathrm{E}f(W_1) - \mathrm{E}f(W_2)\right\}, \quad \text{where } W_1 \sim P,\ W_2 \sim Q.$$

Let $\mathbb{H}_1$, $\mathbb{H}_2$ denote some function spaces of X and Y. Define

$$\phi_{XY} = \mathrm{MMD}(P_{XY}, Q_{XY} \mid \mathbb{H}_1 \otimes \mathbb{H}_2),$$

where $\otimes$ is the tensor product, P_{XY} is the joint distribution of (X,Y) whose definition does not rely on Z, and Q_{XY} is the conditionally independent distribution with the same X and Y margins as P_{XY}. Let X' and Y' be independent copies of X and Y, such that they are conditionally independent given Z. Then Q_{XY} corresponds to the joint distribution of (X',Y'). Note that, to generate (X',Y'), we need to first sample Z according to P_Z, then generate X' and Y' that follow $P_{X|Z}$ and $P_{Y|Z}$, respectively. As

such, Q_{XY} depends on Z, and ϕ_{XY} depends on Z through Q_{XY}. Furthermore, since $\mathrm{E}\{h_1(X')h_2(Y')\} = \mathrm{E}[\mathrm{E}\{h_1(X')|Z\}\mathrm{E}\{h_2(Y')|Z\}]$, we have,

$$\begin{aligned}
\phi_{XY} &= \sup_{h_1\in\mathbb{H}_1, h_2\in\mathbb{H}_2} \big[\mathrm{E}\{h_1(X)h_2(Y)\} - \mathrm{E}\{h_1(X')h_2(Y')\}\big] \\
&= \sup_{h_1\in\mathbb{H}_1, h_2\in\mathbb{H}_2} \Big(\mathrm{E}\{h_1(X)h_2(Y)\} - \mathrm{E}[\mathrm{E}\{h_1(X)|Z\}\mathrm{E}\{h_2(Y)|Z\}]\Big) \\
&= \sup_{h_1\in\mathbb{H}_1, h_2\in\mathbb{H}_2} \Big(\mathrm{E}\{h_1(X)h_2(Y)\} - \mathrm{E}[h_1(X)\mathrm{E}\{h_2(Y)|Z\}] - \mathrm{E}[\{h_1(X)|Z\}h_2(Y)] \\
&\qquad\qquad + \mathrm{E}[\mathrm{E}\{h_1(X)|Z\}\mathrm{E}\{h_2(Y)|Z\}]\Big) \\
&= \sup_{h_1\in\mathbb{H}_1, h_2\in\mathbb{H}_2} \mathrm{E}\Big[h_1(X) - E\{h_1(X)|Z\}\Big]\Big[h_2(Y) - \mathrm{E}\{h_2(Y)|Z\}\Big].
\end{aligned}$$

As such, ϕ_{XY} measures the average conditional association between X and Y given Z. Under $\mathcal{H}_0$, it equals zero, and hence an estimator of this measure can be used as a test statistic for $\mathcal{H}_0$. Moreover, if $\mathbb{H}_1$ and $\mathbb{H}_2$ are reproducing kernel Hilbert spaces (RKHSs), then ϕ_{XY} has a closed form expression in terms of the reproducing kernels of the RKHS (Doran et al., 2014; Gretton et al., 2012), which makes the tests based on an estimator of ϕ_{XY} easier to evaluate.

A notable example of this family is the kernel MMD-based test (KCIT) of Zhang et al. (2011). We next further discuss this test. To control the type-I error asymptotically, KCIT requires the dimension d_Z of Z to be fixed (Zhang et al., 2011, Proposition 5), since it uses the continuous mapping theorem to derive the limiting distribution of its test statistic. However, the continuous mapping theorem may not hold when d_Z diverges with n. In addition, KCIT requires the ℓ_1 distance between the covariance operator and its empirical estimator to decay to zero. It remains unknown whether such an assertion holds as d_Z diverges. By contrast, the test we develop allows d_Z to diverge while maintaining the asymptotic control of the type-I error. This implies that our test is expected to have a better size control than KCIT when d_Z is large. We later further verify this through numerical simulations. Moreover, the maximization of KCIT is done over unit balls in an RKHS, while our proposed test can deal with much more general function spaces such as those generated by neural networks. Consequently, the power of our test can be tailored to more general alternatives than KCIT. For instance, it is known that deep neural networks learn certain non-smooth functions at a faster rate than kernel methods (Imaizumi and Fukumizu, 2019). This implies that our test is expected to have a better power than KCIT under certain types of alternatives.

3. A new double GANs-based testing procedure

We propose a double GANs-based testing procedure for the conditional independence testing problem (1). Conceptually, our test integrates three families of tests that are based on conditional randomization, regression, and MMD. Meanwhile, our new test overcomes the limitations of the existing ones. Unlike the GCIT of Bellot and van der Schaar (2019) that only learned the conditional distribution of X given Z, we learn two generators $\mathbb{G}_X$ and $\mathbb{G}_Y$ to approximate the conditional distributions of both X given Z and Y given Z. We then integrate the two generators in an appropriate way to construct a doubly-robust

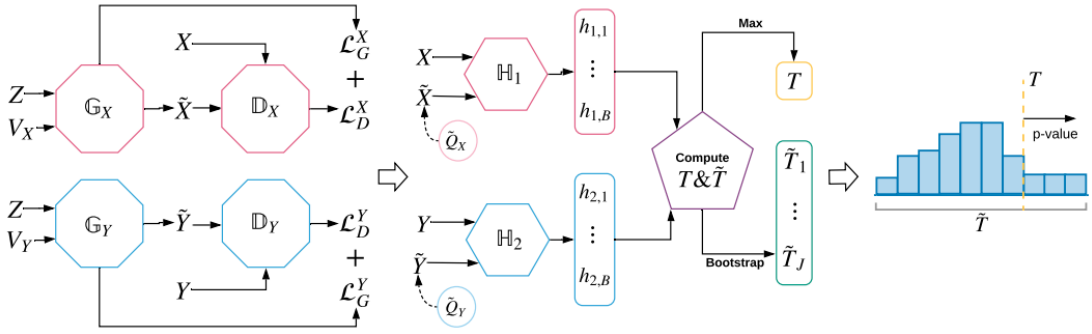


Figure 1: Illustration of the conditional independence test with double GANs.

test statistic. To ensure the theoretical properties of this test, we only require the root mean squared total variation norm to converge at a rate of $n^{-\kappa}$ for some $\kappa > 1/4$. Such a requirement is much weaker and practically more feasible than the condition in (3).

Moreover, to improve the power of the test, we consider a collection of the generalized covariance measures, $\{\mathrm{GCM}(h_1(X), h_2(Y)) : h_1, h_2\}$, for multiple combinations of transformation functions $h_1(X)$ and $h_2(Y)$. We then take the maximum of all these GCMs as our test statistic. This essentially yields a type of maximum mean discrepancy measure ϕ_{XY}. To see why this statistic can enhance the power, we quickly revisit Example 2. When g is not a constant function, there exists some nonlinear function h_1 such that $h_1^*(Y) = \mathrm{E}\{h_1(X)|Y\}$ is not a constant function of Y. Set $h_2 = h_1^*$. We then have $\mathrm{GCM}^* = \mathrm{E}[h_1\{X^* g(Y)\}\{Y - \mathrm{E}(Y)\}] = \mathrm{Var}\{h_1^*(Y)\} > 0$, which enables us to discriminate $\mathcal{H}_1$ from $\mathcal{H}_0$.

We note that the maximum of GCMs yields MMD. Instead of using kernels, we have chosen GANs, because they have been shown to give good approximations of complex distributions (Imaizumi and Fukumizu, 2019). This allows the transformation functions h_1 and h_2 to be arbitrary function spaces. We set these function spaces to the class of neural networks in our implementation. In contrast, kernel based measures such as KCIT are limited to vector spaces of functions, which can be problematic for a high-dimensional conditioning variable (Doran et al., 2014).

We also remark that, even though our proposal is built upon the existing CI tests, our test is far from a simple extension. The major challenge lies in how to properly utilize the GAN estimators for the purpose of high-dimensional conditional independence testing. Despite the fact that GANs are capable of approximating complex high-dimensional probability distributions, the GAN estimators have non-negligible bias that decays slower than the parametric root-n rate. Naively plugging the GAN estimators in the test statistic can invalidate the subsequent inference.

We give a graphical overview of our proposed testing procedure in Figure 1. We first employ double GANs to compute the test statistic that is the maximum of the GCMs over multiple transform functions. We then employ multiplier bootstrap to compute the corresponding p-value. We next detail the main components of our testing procedure.

3.1 Test statistic

We begin with two function spaces, $\mathbb{H}_1 = \left\{h_{1,\theta_1} : \theta_1 \in \mathbb{R}^{d_1}\right\}$ and $\mathbb{H}_2 = \left\{h_{2,\theta_2} : \theta_2 \in \mathbb{R}^{d_2}\right\}$, indexed by some parameters θ_1 and θ_2, respectively. In our implementation, we set $\mathbb{H}_1$ and $\mathbb{H}_2$ to the classes of neural networks with a single-hidden layer, finitely many hidden nodes, and the sigmoid activation function. However, a broad range of other function spaces may be considered, as appropriate for the application at hand. We next randomly generate B functions, $h_{1,1}, \ldots, h_{1,B} \in \mathbb{H}_1$, $h_{2,1}, \ldots, h_{2,B} \in \mathbb{H}_2$, where we independently generate i.i.d. multivariate normal variables $\theta_{1,1}, \ldots, \theta_{1,B} \sim N(0, 2I_{d_1}/d_1)$, and $\theta_{2,1}, \ldots, \theta_{2,B} \sim N(0, 2I_{d_2}/d_2)$. We then set $h_{1,b} = h_{1,\theta_{1,b}}$, and $h_{2,b} = h_{2,\theta_{2,b}}$, $b \in [B] = \{1, \ldots, B\}$. Consider the following maximum-type test statistic,

$$T = \max_{b_1,b_2 \in [B]} \widehat{\sigma}^{-1}_{b_1,b_2} \left| \frac{1}{n} \sum_{i=1}^{n} \left[h_{1,b_1}(X_i) - \widehat{\mathrm{E}}\{h_{1,b_1}(X_i)|Z_i\}\right] \left[h_{2,b_2}(Y_i) - \widehat{\mathrm{E}}\{h_{2,b_2}(Y_i)|Z_i\}\right] \right|,$$

where $\widehat{\sigma}^2_{b_1,b_2}$ is the sampling variance estimator,

$$\begin{aligned}\widehat{\sigma}^2_{b_1,b_2} = \frac{1}{n-1} \sum_{i=1}^{n} \Bigg(& \left[h_{1,b_1}(X_i) - \widehat{\mathrm{E}}\{h_{1,b_1}(X_i)|Z_i\}\right] \left[h_{2,b_2}(Y_i) - \widehat{\mathrm{E}}\{h_{2,b_2}(Y_i)|Z_i\}\right] \\ & - \frac{1}{n} \sum_{i=1}^{n} \left[h_{1,b_1}(X_i) - \widehat{\mathrm{E}}\{h_{1,b_1}(X_i)|Z_i\}\right] \left[h_{2,b_2}(Y_i) - \widehat{\mathrm{E}}\{h_{2,b_2}(Y_i)|Z_i\}\right] \Bigg)^2.\end{aligned}$$

To compute T, we need to estimate the conditional means, $\mathrm{E}\{h_{1,b_1}(X)|Z\}$ and $\mathrm{E}\{h_{2,b_2}(Y)|Z\}$, which can be done by applying some supervised learning methods. However, this needs to be performed for all $b_1, b_2 \in [B]$. In theory, B should diverge to infinity to guarantee the power property of the test. As such, this approach is computationally very expensive. Instead, we propose to implement this step based on the generators $\mathbb{G}_X$ and $\mathbb{G}_Y$ estimated using GANs, which is much more efficient computationally.

Specifically, we first randomly generate i.i.d. samples $\{v^{(m)}_{i,X}\}_{m=1}^M$, $\{v^{(m)}_{i,Y}\}_{m=1}^M$ from multivariate normal distribution, for $i = 1, \ldots, n$. We then feed Z_i and $v^{(m)}_{i,X}$ into GANs to obtain the pseudo samples $\widetilde{X}^{(m)}_i = \mathbb{G}_X(Z_i, v^{(m)}_{i,X})$, and feed Z_i and $v^{(m)}_{i,Y}$ to obtain $\widetilde{Y}^{(m)}_i = \mathbb{G}_Y(Z_i, v^{(m)}_{i,Y})$, for $i = 1, \ldots, n, m = 1, \ldots, M$. These pseudo samples approximate the conditional distribution of X_i and Y_i given Z_i, respectively. We then compute

$$\widehat{\mathrm{E}}\{h_{1,b_1}(\widetilde{X}_i)|Z_i\} = \frac{1}{M} \sum_{m=1}^{M} h_{1,b_1}(\widetilde{X}^{(m)}_i), \quad \widehat{\mathrm{E}}\{h_{2,b_2}(Y_i)|Z_i\} = \frac{1}{M} \sum_{m=1}^{M} h_{2,b_2}(\widetilde{Y}^{(m)}_i),$$

for $b_1, b_2 = 1, \ldots, B$. Plugging the estimated means into T produces the sample test statistic,

$$\widehat{T} = \max_{b_1,b_2} \left| n^{-1/2} \sum_{i=1}^{n} \psi_{b_1,b_2,i} \right|, \quad \text{where} \tag{4}$$

$$\psi_{b_1,b_2,i} = \widehat{\sigma}^{-1}_{b_1,b_2} \left\{ h_{1,b_1}(X_i) - \frac{1}{M} \sum_{m=1}^{M} h_{1,b_1}\left(\widetilde{X}^{(m)}_i\right) \right\} \left\{ h_{2,b_2}(Y_i) - \frac{1}{M} \sum_{m=1}^{M} h_{2,b_2}\left(\widetilde{Y}^{(m)}_i\right) \right\}.$$

Algorithm 1 Algorithm for computing the test statistic.

Input: The number of transformation functions B, the number of pseudo samples M, and the number of data splits L.

Step 1: Divide $\{1, \ldots, n\}$ into L folds $\mathcal{I}^{(1)}, \ldots, \mathcal{I}^{(L)}$. Denote $\mathcal{I}^{(-\ell)} = \{1, \ldots, n\} \backslash \mathcal{I}^{(\ell)}$.

Step 2: For $\ell = 1, \ldots, L$, train two generators $\mathbb{G}_X^{(\ell)}$ and $\mathbb{G}_Y^{(\ell)}$ based on $\{(X_i, Z_i)\}_{i \in \mathcal{I}^{(-\ell)}}$ and $\{(Y_i, Z_i)\}_{i \in \mathcal{I}^{(-\ell)}}$, to approximate the conditional distributions of $X|Z$ and $Y|Z$.

Step 3: For $\ell = 1, \ldots, L$ and $i \in \mathcal{I}_\ell$, generate i.i.d. random noises $\left\{v_{i,X}^{(m)}\right\}_{m=1}^{M}$, $\left\{v_{i,Y}^{(m)}\right\}_{m=1}^{M}$. Set $\widetilde{X}_i^{(m)} = \mathbb{G}_X^{(\ell)}\left(Z_i, v_{i,X}^{(m)}\right)$, and $\widetilde{Y}_i^{(m)} = \mathbb{G}_Y^{(\ell)}\left(Z_i, v_{i,Y}^{(m)}\right)$, $m = 1, \ldots, M$.

Step 4: Randomly generate $h_{1,1}, \ldots, h_{1,B} \in \mathbb{H}_1$ and $h_{2,1}, \ldots, h_{2,B} \in \mathbb{H}_2$.

Step 5: Compute the test statistic $\widehat{T}$.

To help reduce the type-I error, we further employ a data splitting and cross-fitting strategy, which has been commonly used in statistical inferences in recent years (Romano and DiCiccio, 2019). That is, we use different subsets of data samples to learn GANs and to construct the test statistic. We begin by dividing the data into L folds of equal size. We use $\mathcal{I}^{(\ell)}$ to denote the set of indices of subsamples in the ℓth fold, and $\mathcal{I}^{(-\ell)}$ its complement. We next learn two generators $\mathbb{G}_X^{(\ell)}$ and $\mathbb{G}_Y^{(\ell)}$, based on $\{(X_i, Z_i)\}_{i \in \mathcal{I}^{(-\ell)}}$ and $\{(Y_i, Z_i)\}_{i \in \mathcal{I}^{(-\ell)}}$, to approximate the conditional distributions of $X|Z$ and $Y|Z$, for $\ell = 1, \cdots, L$. Finally, for each ℓ and $i \in \mathcal{I}^{(\ell)}$, we generate the pseudo samples $\widetilde{X}_i^{(m)}$ and $\widetilde{Y}_i^{(m)}$ using $\mathbb{G}_X^{(\ell)}$ and $\mathbb{G}_Y^{(\ell)}$, and construct $\widehat{T}$ as in (4). In this way, $\widetilde{X}_i^{(m)}$ and $\widetilde{Y}_i^{(m)}$ are conditionally independent of the observations in $\mathcal{I}^{(\ell)}$ given Z_i. Such a cross-fitting strategy allows us to derive the asymptotic properties of the test under minimal conditions on the generators.

We summarize our procedure of computing the test statistic in Algorithm 1.

3.2 Approximation of conditional distribution via GANs

There are numerous GANs methods available for learning high-dimensional distributions. We adopt the proposal of Genevay et al. (2017) to learn the conditional distributions $P_{X|Z}$ and $P_{Y|Z}$ in our setting thanks to its competitive performance. Recall that $\widetilde{P}_{X|Z}$ is the distribution of pseudo outcome generated by the generator $\mathbb{G}_X$ given Z. We consider estimating $P_{X|Z}$ by optimizing

$$\min_{\mathbb{G}_X} \max_{c} \widetilde{\mathcal{D}}_{c,\epsilon}(P_{X|Z}, \widetilde{P}_{X|Z}).$$

Here $\widetilde{\mathcal{D}}_{c,\epsilon}$ denotes the Sinkhorn loss function between two probability measures with respect to some cost function c and some regularization parameter $\epsilon > 0$,

$$\begin{aligned} \widetilde{\mathcal{D}}_{c,\epsilon}(\mu, \nu) &= 2\mathcal{D}_{c,\epsilon}(\mu, \nu) - \mathcal{D}_{c,\epsilon}(\mu, \mu) - \mathcal{D}_{c,\epsilon}(\nu, \nu), \\ \mathcal{D}_{c,\epsilon}(\mu, \nu) &= \inf_{\pi \in \Pi(\mu,\nu)} \int_{x,y} \left\{c(x, y) - \epsilon H(\pi | \mu \otimes \nu)\right\} \pi(dx, dy), \end{aligned}$$

where $\Pi(\mu, \nu)$ is a set containing all probability measures π whose marginal distributions correspond to μ and ν, H is the Kullback-Leibler divergence, and $\mu \otimes \nu$ is the product

measure of μ and ν. When $\epsilon = 0$, $\mathcal{D}_{c,0}(\mu, \nu)$ measures the optimal transport of μ into ν with respect to the cost function $c(\cdot, \cdot)$ (Cuturi, 2013). When $\epsilon \neq 0$, an entropic regularization is added to this optimal transport. As such, the objective function $\mathcal{D}_{c,\epsilon}$ is a regularized optimal transport metric, and the regularization is to facilitate the computation, so that $\mathcal{D}_{c,\epsilon}$ can be efficiently evaluated.

Intuitively, the closer the two probability measures, the smaller the Sinkhorn loss. As such, maximizing the loss with respect to the cost function learns a discriminator that can better discriminate the samples generated between $P_{X|Z}$ and $\widetilde{P}_{X|Z}$. On the other hand, minimizing the maximum cost with respect to the generator $\mathbb{G}_X$ makes it closer to the true distribution $P_{X|Z}$. This yields the minimax formulation $\min_{\mathbb{G}_X} \max_c \widetilde{\mathcal{D}}_{c,\epsilon}(P_{X|Z}, \widetilde{P}_{X|Z})$ that we target. In practice, we approximate the cost and the generator based on neural networks. Integrations in the objective function $\widetilde{\mathcal{D}}_{c,\epsilon}(P_{X|Z}, \widetilde{P}_{X|Z})$ are approximated by sample averages. The conditional distribution of $P_{Y|Z}$ is estimated similarly.

3.3 Bootstrap for the p-value

Next, we propose a multiplier bootstrap method to approximate the distribution of $\widehat{T}$ under $\mathcal{H}_0$ and compute the corresponding p-value. Let $\psi_{b_1,b_2} = n^{-1} \sum_{i=1}^n \psi_{b_1,b_2,i}$. The key observation is that $\{\psi_{b_1,b_2}\}_{b_1,b_2=1}^B$ are asymptotically multivariate normal with zero mean under $\mathcal{H}_0$; see the proof of Theorem 4 for details. Consequently, $\widehat{T} = \max_{b_1,b_2} |n^{-1/2} \sum_{i=1}^n \psi_{b_1,b_2,i}|$ is to converge to a maximum of normal variables in absolute values.

To approximate this limiting distribution, we first estimate the covariance matrix of a B^2-dimensional vector formed by $\{n^{-1/2}\psi_{b_1,b_2}\}_{b_1,b_2=1}^B$ using the sample covariance matrix $\widehat{\Sigma}$, whose $\{b_1 + B(b_2 - 1), b_3 + B(b_4 - 1)\}$th entry is given by

$$\frac{1}{n} \sum_{i=1}^n (\psi_{b_1,b_2,i} - \psi_{b_1,b_2})(\psi_{b_3,b_4,i} - \psi_{b_3,b_4}), \quad b_1, b_2, b_3, b_4 = 1, \ldots, B.$$

We then generate i.i.d. random vectors with the covariance matrix equal to $\widehat{\Sigma}$. This can be achieved by generating i.i.d. standard normal variables $\{W_{i,j}\}_{i,j}$ for $1 \leq i \cdots \leq n$ and $j = 1, \cdots, J$, then compute B^2-dimensional normal random vectors $\boldsymbol{W}_j$ whose $\{b_1 + B(b_2 - 1)\}$th entry is given by $n^{-1/2} \sum_{i=1}^n (\psi_{b_1,b_2,i} - \psi_{b_1,b_2}) W_{i,j}$ for $j = 1, \cdots, J$. We next compute $\widetilde{T}_j = \|\boldsymbol{W}_j\|_\infty$, for $j = 1, \ldots, J$, where $\|\cdot\|_\infty$ is the maximum element of a vector in absolute value, and J is the number of bootstrap samples. Finally, we use these maximum absolute values to approximate the distribution of $\widehat{T}$ under the null hypothesis. This yields the p-value, $p = J^{-1} \sum_{j=1}^J \mathbb{I}(\widehat{T} \geq \widetilde{T}_j)$. We summarize this bootstrap procedure in Algorithm 2.

Algorithm 2 Algorithm for computing the p-value.

Input: The number of bootstrap samples J, and $\{\psi_{b_1,b_2,i}\}_{b_1,b_2=1,i=1}^{B,n}$.
Step 1: Generate i.i.d. standard normal variables $W_{i,j}$ for $i = 1, \cdots, n$, $j = 1, \ldots, J$.
Step 2: Compute B^2-dimensional normal random vectors $\boldsymbol{W}_j$ whose $\{b_1 + B(b_2 - 1)\}$th entry is given by $n^{-1/2} \sum_{i=1}^n (\psi_{b_1,b_2,i} - \psi_{b_1,b_2}) W_{i,j}$ and set $\widetilde{T}_j = \|\boldsymbol{W}_j\|_\infty$ for $j = 1, \cdots, J$.
Step 3: Compute the p-value, $p = J^{-1} \sum_{j=1}^J \mathbb{I}(\widehat{T} \geq \widetilde{T}_j)$.

4. Asymptotic theory

To derive the theoretical properties of the test statistic $\widehat{T}$, we first introduce the concept of the "oracle" test statistic T^*. If $P_{X|Z}$ and $P_{Y|Z}$ were known a priori, then one can draw $\{X_i^{(m)}\}_m$ and $\{Y_i^{(m)}\}_m$ from $P_{X|Z=Z_i}$ and $P_{Y|Z=Z_i}$ directly, and can compute the test statistic by replacing $\{\widetilde{X}_i^{(m)}\}_m$ and $\{\widetilde{Y}_i^{(m)}\}_m$ with $\{X_i^{(m)}\}_m$ and $\{Y_i^{(m)}\}_m$. We call the resulting T^* an "oracle" test statistic. We next establish the double-robustness property of $\widehat{T}$, which helps explain why our test can relax the requirement in (3). Roughly speaking, the double-robustness means that $\widehat{T}$ is asymptotically equivalent to T^* when either the conditional distribution of $X|Z$, or that of $Y|Z$, is well approximated by GANs. It guarantees that $\widehat{T}$ converges to T^* at a faster rate than the estimated conditional distribution. In contrast, the convergence rate of the GCIT test statistic is the same as the rate of the estimated conditional distribution. For this reason, our procedure only requires a weaker condition.

Theorem 2 (Double-robustness) *Suppose M is proportional to n, and $B = O(n^c)$ for some constant $c > 0$. Suppose $\min_{h_1\in\mathbb{H}_1, h_2\in\mathbb{H}_2} Var[\{h_1(X) - E\{h_1(X)|Z\}\}\{h_2(Y) - E\{h_2(Y)|Z\}\}] \geq c^*$ for some constant $c^* > 0$. Then, $\widehat{T} - T^* = o_p(1)$, when*

$$E\left[d_{TV}^2\left\{\widetilde{Q}_X^{(\ell)}(\cdot|Z), Q_X(\cdot|Z)\right\}\right] = o(\log^{-1} n), \text{ or } E\left[d_{TV}^2\left\{\widetilde{Q}_Y^{(\ell)}(\cdot|Z), Q_Y(\cdot|Z)\right\}\right] = o(\log^{-1} n).$$

We note that the conditions on M and B are mild, as these are user-specified parameters. As we have mentioned, when both total variation distances converge to zero, the test statistic T converges at a faster rate than those total variation distances. Therefore, we can greatly relax the condition in (3), and replace it with,

$$\left[\mathrm{E}\left\{d_{\mathrm{TV}}^2\left(\widetilde{P}_{X|Z}^{(\ell)}, P_{X|Z}\right)\right\}\right]^{1/2} = O(n^{-\kappa_x}), \text{ and } \left[\mathrm{E}\left\{d_{\mathrm{TV}}^2\left(\widetilde{P}_{Y|Z}^{(\ell)}, P_{Y|Z}\right)\right\}\right]^{1/2} = O(n^{-\kappa_y}), \quad (5)$$

for some constants $0 < \kappa_x, \kappa_y < 1/2$ and any $\ell \in [L]$, where $\widetilde{P}_{X|Z}^{(\ell)}$ and $\widetilde{P}_{Y|Z}^{(\ell)}$ denote the conditional distributions approximated via GANs trained on the ℓ-th subset of data samples. The next theorem summarizes this discussion.

Theorem 3 *Suppose the conditions in Theorem 2. Furthermore, suppose* (5) *holds. Then, $\widehat{T} - T^* = O_p\left(n^{-(\kappa_x+\kappa_y)}\log n\right)$.*

Since $\kappa_x, \kappa_y > 0$, the convergence rate of $(\widehat{T} - T^*)$ is faster than that in (5). To ensure $\sqrt{n}(T - T^*) = o_p(1)$, it suffices to require $\kappa_x + \kappa_y > 1/2$. In contrast to (3), this rate is achievable. We consider two examples in Berrett et al. (2019) to illustrate this, while the condition holds in a much wider range of settings.

Example 3 (Parametric setting) *Suppose the parametric forms of Q_X and Q_Y are correctly specified. Then under certain regularity conditions, the requirement $\kappa_x + \kappa_y > 1/2$ holds if $k_x = O(n^{t_x})$ and $k_y = O(n^{t_y})$ for some $t_x + t_y < 1/2$, where k_x and k_y are the dimensions of the parameters defining the parametric models for Q_X and Q_y, respectively.*

Example 4 (Nonparametric setting with binary data) *Suppose X, Y are binary variables. Then the requirement $\kappa_x + \kappa_y > 1/2$ holds if the mean squared prediction errors of the nonparametric estimators of the conditional means of X and Y given Z are $O(n^{-t_x})$ and $O(n^{-t_y})$ for some t_x, t_y, such that $t_x + t_y > 1/2$.*

We briefly remark that, there is no explicit specification on d_Z in the statement of Theorem 3. It is implicitly imposed due to the requirement that $\kappa_x + \kappa_y > 1/2$, and d_Z is allowed to diverge with the sample size. In addition, the condition $\kappa_x + \kappa_y > 1/2$ can be further relaxed to $\kappa_1, \kappa_2 > 0$ using the theory of higher order influence functions (Robins et al., 2008, 2017; Mukherjee et al., 2017). However, the resulting estimators would be considerably much more complicated, and thus we do not pursue those estimators.

Next, we show that our proposed test can control the type-I error asymptotically.

Theorem 4 *Suppose the conditions in Theorem 2 hold. Suppose (5) holds for some κ_x, κ_y such that $\kappa_x + \kappa_y > 1/2$. Then, the p-value from Algorithm 2 satisfies that $Pr(p \leq \alpha | \mathcal{H}_0) = \alpha + o(1)$.*

Next, to derive the asymptotic power of the test, we introduce the pair of hypotheses based on the notion of weak conditional independence (Daudin, 1980),

$$\begin{aligned} \mathcal{H}_0^* &: \mathrm{E}[\mathrm{cov}\{f(X), g(Y)|Z\}] = 0, \quad \text{for any } f \in L_X^2, g \in L_Y^2 \quad \text{versus} \\ \mathcal{H}_1^* &: \mathrm{E}[\mathrm{cov}\{f(X), g(Y)|Z\}] \neq 0, \quad \text{for some } f \in L_X^2, g \in L_Y^2, \end{aligned}$$

where L_X^2 and L_Y^2 denote the class of all squared integrable functions of X and Y, respectively. We note that conditional independence implies weak conditional independence, i.e., $\mathcal{H}_0$ implies $\mathcal{H}_0^*$, and $\mathcal{H}_1^*$ implies $\mathcal{H}_1$. We consider an example to further elaborate on the difference between weak CI and CI.

Example 5 *Let X, Y, Z be binary random variables with the distribution functions,*

$$\begin{pmatrix} Pr(X=0,Y=0|Z=0) & Pr(X=0,Y=1|Z=0) \\ Pr(X=1,Y=0|Z=0) & Pr(X=1,Y=1|Z=0) \end{pmatrix} = \begin{pmatrix} 1/6 & 1/3 \\ 1/3 & 1/6 \end{pmatrix},$$
$$\begin{pmatrix} Pr(X=0,Y=0|Z=1) & Pr(X=0,Y=1|Z=1) \\ Pr(X=1,Y=0|Z=1) & Pr(X=1,Y=1|Z=1) \end{pmatrix} = \begin{pmatrix} 1/3 & 1/6 \\ 1/6 & 1/3 \end{pmatrix},$$

and Z takes the value $\{0,1\}$ with equal probability. We can show that, for any $x, y \in \{0,1\}$,

$$\begin{aligned} E\{Pr(X=x|Z)Pr(Y=y|Z)\} &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}, \\ Pr(X=x,Y=y) &= \frac{1}{2}\Big\{Pr(X=x,Y=y|Z=0) + Pr(X=x,Y=y|Z=1)\Big\} \\ &= \frac{1}{2} \times \left(\frac{1}{6} + \frac{1}{3}\right) = \frac{1}{4}. \end{aligned}$$

By definition, this implies that X and Y are weakly conditionally independent given Z, since

$$\begin{aligned} E[cov\{f(X), g(Y)|Z\}] = \sum_{x,y} f(x)g(y)\Big\{ & Pr(X=x,Y=y) \\ & - E\big\{Pr(X=x|Z)Pr(Y=y|Z)\big\}\Big\} = 0. \end{aligned}$$

However, $Pr(X = 0, Y = 0|Z = 0) \neq Pr(X = 0|Z = 0)Pr(Y = 0|Z = 0)$, since the former equals $1/6$, and the latter equals $1/4$. As such, X and Y are not conditionally independent given Z.

The next theorem shows that our proposed test is consistent against the alternatives in $\mathcal{H}_1^*$, but not against all alternatives in $\mathcal{H}_1$.

Theorem 5 *Suppose the conditions in Theorem 4 hold, $B = c_0 n^c$ for some $c_0, c > 0$, and X, Y are bounded random variables. Then the p-value from Algorithm 2 satisfies that $Pr(p \leq \alpha|\mathcal{H}_1^*) \to 1$, as $n \to \infty$.*

Finally, we remark that our test is constructed based on ϕ_{XY}. Meanwhile, we may consider another test based on $\phi_{XYZ} = \text{MMD}(P_{XYZ}, Q_{XYZ}|\mathbb{H}_1 \otimes \mathbb{H}_2 \otimes \mathbb{H}_3)$, where P_{XYZ} is the joint distribution of (X, Y, Z), $Q_{XYZ} = P_{X|Z}P_{Y|Z}P_Z$, and $\mathbb{H}_3$ is a neural network class of functions of Z. This type of test is consistent against all alternatives in $\mathcal{H}_1$. However, in our numerical experiments, we find it less powerful compared to our test. This agrees with the observation by Li and Fan (2019) in that, even though the tests based on weak CI cannot fully characterize CI, they potentially enjoy an improved power.

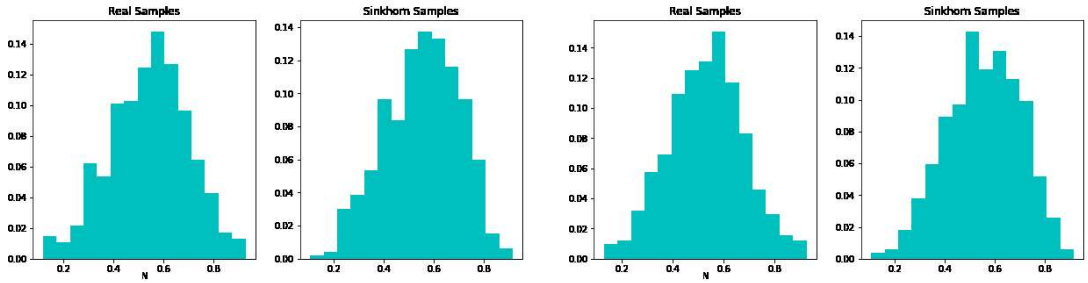
5. Numerical studies

We begin with a discussion of some implementation details. We then carry out simulations to study the empirical size and power of the proposed test, and compare with several alternative methods. We further illustrate with an application to a cancer genetics example.

5.1 Implementation details

For the number of functions B in Algorithm 2, it represents a trade-off. By Theorem 5, B should be as large as possible to guarantee a good power. In practice, the computation complexity increases as B increases. Our numerical studies suggest that the value of B between 30 and 50 achieves a good balance between the power and the computational cost, and we fix $B = 30$. For the number of pseudo samples M, and the number of sample splittings L, we find the results are not overly sensitive to their choices, and thus we fix $M = 100$ and $L = 3$. Besides, we set the number of bootstrap samples $J = 1000$.

For the GANs, we use a single-hidden layer neural network to approximate both the discriminator and the generator. The number of nodes in the hidden layer is set at 128. The dimension of the input noise $v_{i,X}^{(m)}$ and $v_{i,Y}^{(m)}$ is set at 10. These tuning parameters are chosen following the common practice in the GANs literature, and also by investigating the goodness-of-fit of the resulting generator, which can be done by comparing the conditional histogram of the generated samples to that of the true samples. In our experiments, we find such an approach yields GANs with satisfactory performances. More specifically, let d_Z denote the dimension of Z, and $\widehat{\mu}_Z$ the sample average $n^{-1}\sum_i Z_i$. Let $\widetilde{Y}_i = G_Y(Z_i, v_{i,Y})$ denote a simulated sample to approximate the distribution of $Y|Z = Z_i$ obtained by the generator G_Y. When G_Y is accurate, we expect the conditional distribution of $\widetilde{Y}_i$ and Y_i given Z_i are similar. As such, for any d_Z-dimensional vector a, the histograms $\{\widetilde{Y}_i : a^\top(\widetilde{Z}_i - \widehat{\mu}_Z) > 0\}$ and $\{Y_i : a^\top(Z_i - \widehat{\mu}_Z) > 0\}$ should be similar. We sample i.i.d. vectors



(a) One random value of a_g (b) Another random value of a_g

Figure 2: Conditional histograms. GANs are trained using data generated from the simulation study in Section 5.2.

$\{a_g\}_g$ from Normal$(0, I_{d_Z})$. For each g, we plot the histogram $\{Y_i : a_g^\top (Z_i - \widehat{\mu}_Z) > 0\}$ and $\{\widetilde{Y}_i^{(m)} : a_g^\top (Z_i - \widehat{\mu}_Z) > 0\}$. See Figures 2 (a) and (b) for the conditional histograms with two choices of a_g. It is seen that the GANs fit the conditional density reasonably well. The fitted conditional distribution for $P_{X|Z}$ can be checked in a similar fashion.

5.2 Simulations

We generate the data following the post nonlinear noise model similarly as in Zhang et al. (2011); Doran et al. (2014); Bellot and van der Schaar (2019), i.e.,

$$X = \sin(a_f^\top Z + \varepsilon_f), \quad \text{and} \quad Y = \cos(a_g^\top Z + bX + \varepsilon_g).$$

The entries of a_f, a_g are randomly and uniformly sampled from $[0, 1]$, then normalized to the unit ℓ_1 norm. The noise variables $\varepsilon_f, \varepsilon_g$ are independently sampled from a normal distribution with mean zero and variance 0.25. In this model, the parameter b determines the degree of conditional dependence. When $b = 0$, $\mathcal{H}_0$ holds, and otherwise $\mathcal{H}_1$ holds. The sample size is set at $n = 1000$.

We call our test DGCIT, short for double GANs-based conditional independence test. We compare it with the GCIT test of Bellot and van der Schaar (2019), the regression-based test (RCIT) of Shah and Peters (2018), the kernel MMD-based test (KCIT) of Zhang et al. (2011), and the classifier CI test (CCIT) of Sen et al. (2017).

We first study the empirical size when $b = 0$. We vary the dimension of Z as $d_Z = 50, 100, 150, 200, 250$, and consider two generation distributions. We first generate Z from a standard normal distribution, then from a Laplace distribution. We set the significance level at $\alpha = 0.05$ and 0.1. Figure 3 reports the empirical size of the tests aggregated over 500 data replications. We make the following observations. First, the type-I error rates of our test and RCIT are close to or below the nominal level in nearly all cases. Second, KCIT fails in that its type-I error is considerably larger than the nominal level in all cases. We suspect it is due to the high-dimensional setting where $d_Z \geq 50$. We have experimented with $d_Z = 5$, and found that KCIT is able to control the type-I error in that case. This is consistent with Proposition 5 of Zhang et al. (2011), which suggests that KCIT should work in a low-dimensional setting. Third, GCIT and CCIT both have inflated type-I errors in some cases. Take GCIT as an example. When Z is normal, $d_Z = 250$ and $\alpha = 0.1$, its

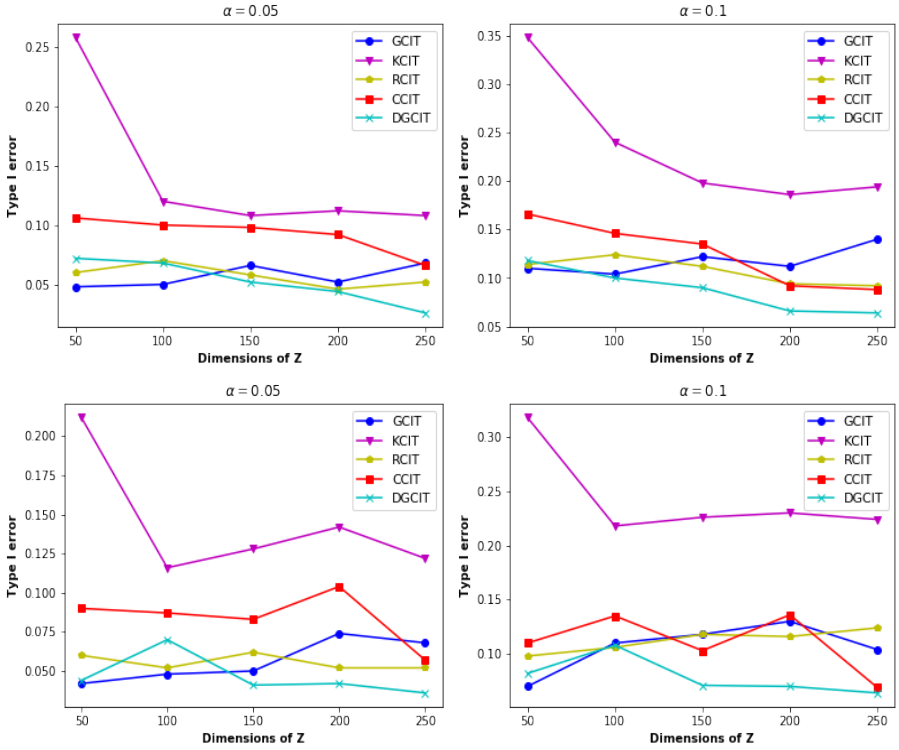


Figure 3: The empirical type-I error rate of various tests under $\mathcal{H}_0$. Left panels: $\alpha = 0.05$, right panels: $\alpha = 0.1$. Top panels: Z is normal, bottom panels: Z is Laplacian.

empirical size is close to 0.15. This is consistent with our discussion in Section 2.1, since GCIT requires a rather strong condition to control the type-I error.

We then study the empirical power when $b > 0$. We generate Z from a standard normal distribution, with $d_Z = 100, 200$, and vary the value of $b = 0.3, 0.45, 0.6, 0.75, 0.9$ that controls the magnitude of the alternative. Figure 4 reports the empirical power of the tests over 500 data replications. We observe that our test is the most powerful, and the empirical power approaches 1 as b increases to 0.9, demonstrating the consistency of the test. Meanwhile, both GCIT and RCIT have no power in all cases. We do not report the power of KCIT, because as we have shown earlier, it cannot control the size, and thus its empirical power is not meaningful.

Finally, we discuss the computation time. All experiments were run on a 16 N1 CPUs Google Cloud Computing platform. The wall clock time for running the entire GCIT test for one data replication was about 2.5 minutes. In contrast, the running time for CCIT was about 2 minutes, for KCIT about 30 seconds, and for GCIT and RCIT about 20 seconds.

5.3 Anti-cancer drug data example

We illustrate our proposed test with an anti-cancer drug dataset from the Cancer Cell Line Encyclopedia (Barretina et al., 2012). We concentrate on a subset, the CCLE data, that measures the treatment response of drug PLX4720. It is well known that the patient's

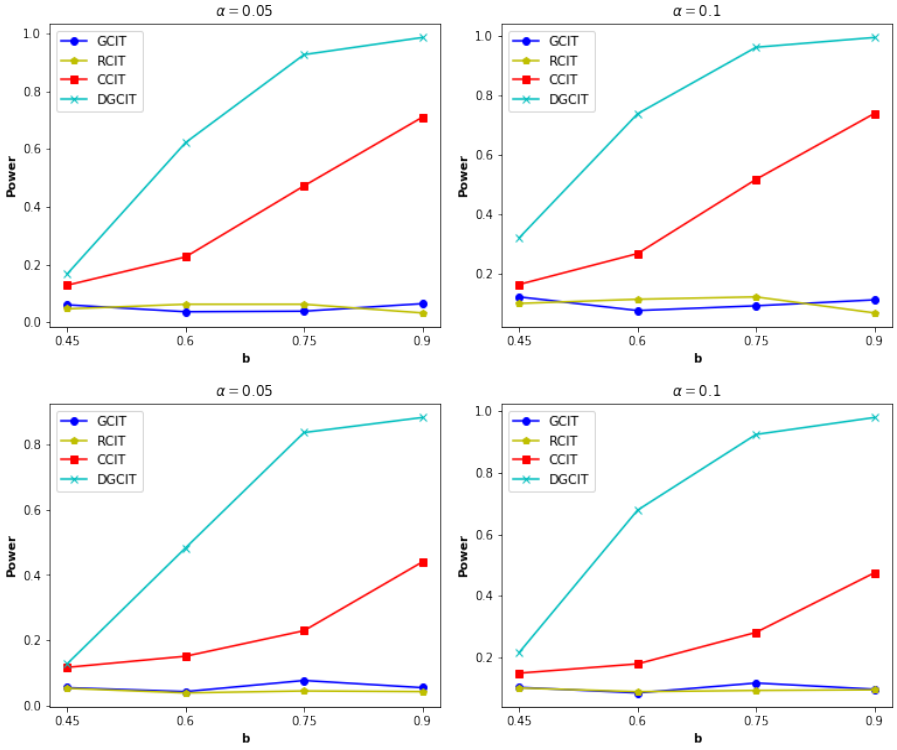


Figure 4: The empirical power of various tests under $\mathcal{H}_1$. Left panels: $\alpha = 0.05$, right panels: $\alpha = 0.1$. Top panels: $d_Z = 100$, bottom panels: $d_Z = 200$.

cancer treatment response to drug can be strongly influenced by alterations in the genome (Garnett et al., 2012). This data measures 1638 genetic mutations of $n = 472$ cell lines, and the goal of our analysis is to determine which genetic mutation is significantly correlated with the drug response after conditioning on all other mutations. The same data was also analyzed in Tansey et al. (2018) and Bellot and van der Schaar (2019). We adopt the same screening procedure as theirs to screen out irrelevant mutations, which leaves a total of 466 potential mutations for our conditional independence testing.

The ground truth is unknown for this data. Instead, we compare with the variable importance measures obtained from fitting an elastic net (EN) model and a random forest (RF) model as reported in Barretina et al. (2012). In addition, we compare with the GCIT test of Bellot and van der Schaar (2019). Table 1 reports the corresponding variable

Table 1: The variable importance measures of the elastic net and random forest models, versus the p-values of the GCIT and DGCIT tests for the anti-cancer drug example.

	BRAF.V600E	BRAF.MC	HIP1	FTL3	CDC42BPA	THBS3	DNMT1	PRKD1	PIP5K1A	MAP3K5
EN	1	3	4	5	7	8	9	10	19	78
RF	1	2	3	14	8	34	28	18	7	9
GCIT	<0.001	<0.001	0.008	0.521	0.050	0.013	0.020	0.002	0.001	<0.001
DGCIT	0	0	0	0	0	0	0	0	0	0.794

importance measures and the p-values, for 10 mutations that were also reported by Bellot and van der Schaar (2019). We see that, the p-values of the tests generally agree well with the variable important measures from the EN and RF models. Meanwhile, the two conditional independence tests agree relatively well, except for two genetic mutations, MAP3K5 and FTL3. GCIT concluded that MAP3K5 is significant ($p < 0.001$) but FTL3 is not ($p = 0.521$), whereas our test leads to the opposite conclusion that MAP3K5 is insignificant ($p = 0.794$) but FTL3 is ($p = 0$). Besides, both EN and RF place FTL3 as an important mutation. We then compare our findings with the cancer drug response literature. Actually, MAP3K5 has not been previously reported in the literature as being directly linked to the PLX4720 drug response. Meanwhile, there is strong evidence showing the connections of the FLT3 mutation with cancer response (Tsai et al., 2008; Larrosa-Garcia and Baer, 2017). Combining the existing literature with our theoretical and synthetic results, we have more confidence about the findings of our proposed test.

6. Discussion

In this article, we have developed a new inferential procedure for high-dimensional conditional independence testing, where the dimension of the conditional variables can diverge with the sample size. Our proposal utilizes a set of state-of-the-art deep learning tools to help address a classical statistics and machine learning problem. It integrates GANs, neural networks, cross-fitting and multiplier bootstrap. It achieves the asymptotic guarantees under much weaker conditions, and enjoys better empirical performances, when compared to the existing tests. As a tradeoff, our test is computationally more complicated. Nevertheless, the wall clock time for running the entire test for one data replication is in the order of a few minutes and is deemed reasonable. Finally, the computer code is publicly available on the GitHub repository: `https://github.com/tianlinxu312/dgcit`.

Appendix A. Proofs

We provide the proofs of Proposition 1, Theorems 3, 4, and 5. We omit the proof of Theorem 2, since it is similar to that of Theorem 3. We note that Theorems 2-5 are established under our choice of the function classes $\mathbb{H}_1$ and $\mathbb{H}_2$, which are set to the classes of neural networks with a single-hidden layer, finitely many hidden nodes, and the sigmoid activation function, as used in our implementation. Meanwhile, our results can be extended to more general choices of the function classes.

A.1 Proof of Proposition 1

Note that the total variation distance is bounded by 1. Suppose $\mathrm{E} d_{\mathrm{TV}}(\widetilde{P}_{\boldsymbol{X}|\boldsymbol{Z}}, P_{\boldsymbol{X}|\boldsymbol{Z}}) = o(1)$. Then we have $d_{\mathrm{TV}}(\widetilde{P}_{\boldsymbol{X}|\boldsymbol{Z}}, P_{\boldsymbol{X}|\boldsymbol{Z}}) = o_p(1)$. By the dominated convergence theorem, we have $\mathrm{E} d^2_{\mathrm{TV}}(\widetilde{P}_{\boldsymbol{X}|\boldsymbol{Z}}, P_{\boldsymbol{X}|\boldsymbol{Z}}) = o(1)$.

By Theorem 1.2 of Devroye et al. (2018), we have $d_{\mathrm{TV}}(\widetilde{P}_{\boldsymbol{X}|\boldsymbol{Z}}, P_{\boldsymbol{X}|\boldsymbol{Z}})$ is proportional to

$$\min\left[1, \sigma_0^{-1}\sqrt{\sum_{i=1}^{n}\left\{Z_i^\top(\widehat{\beta} - \beta_0)\right\}^2}\right].$$

It follows that

$$\frac{1}{\sigma_0}\mathrm{E}\sum_{i=1}^{n}\{Z_i^\top(\widehat{\beta}-\beta_0)\}^2 = o(1).$$

Applying Theorem 1.2 of Devroye et al. (2018) again, we obtain that $d_{\mathrm{TV}}(\widetilde{P}_{X|Z=Z_i}, P_{X|Z=Z_i})$ is proportional to

$$\min\left\{1, \sigma_0^{-1}|Z_i^\top(\widehat{\beta}-\beta_0)|\right\}.$$

Therefore, we obtain that,

$$\sum_{i=1}^{n}\mathrm{E}d_{\mathrm{TV}}^2\left(\widetilde{P}_{X|Z=Z_i}, P_{X|Z=Z_i}\right) = o(1).$$

Since the data is exchangeable, we have that,

$$\mathrm{E}d_{\mathrm{TV}}^2\left(\widetilde{P}_{X|Z=Z_i}, P_{X|Z=Z_i}\right) = o(n^{-1}). \tag{6}$$

This shows that when RHS of (2), i.e., $\mathrm{E}\{d_{\mathrm{TV}}(\widetilde{P}_{\boldsymbol{X}|\boldsymbol{Z}}, P_{\boldsymbol{X}|\boldsymbol{Z}})\}$ is $o(1)$, (6) holds.

Next, we show (6) is violated in the linear regression example. By the data exchangeability, it suffices to show $\sum_{i=1}^{n}\mathrm{E}d_{\mathrm{TV}}^2\{\widetilde{P}_{X|Z=Z_i}, P_{X|Z=Z_i}\}$ is not $o(1)$. With some calculations, we obtain that,

$$\begin{aligned}
&\sum_{i=1}^{n}\mathrm{E}\min\left\{1, \sigma_0^{-2}|Z_i^\top(\widehat{\beta}-\beta_0)|^2\right\} \\
=& \sum_{i=1}^{n}\mathrm{E}\sigma_0^{-2}|Z_i^\top(\widehat{\beta}-\beta_0)|^2\mathbb{I}\left\{\sigma_0^{-2}|Z_i^\top(\widehat{\beta}-\beta_0)|^2 \le 1\right\} + \sum_{i=1}^{n}\mathrm{E}\mathbb{I}\left\{\sigma_0^{-2}|Z_i^\top(\widehat{\beta}-\beta_0)|^2 > 1\right\} \\
=& \sum_{i=1}^{n}\mathrm{E}\sigma_0^{-2}|Z_i^\top(\widehat{\beta}-\beta_0)|^2 - \sum_{i=1}^{n}\mathrm{E}\left\{\sigma_0^{-2}|Z_i^\top(\widehat{\beta}-\beta_0)|^2 - 1\right\}\mathbb{I}\left\{\sigma_0^{-2}|Z_i^\top(\widehat{\beta}-\beta_0)|^2 > 1\right\}.
\end{aligned} \tag{7}$$

By the definition of $\widehat{\beta}$, we have

$$\sum_{i=1}^{n}\mathrm{E}\sigma_0^{-2}|Z_i^\top(\widehat{\beta}-\beta_0)|^2 = \frac{1}{\sigma_0^2}\mathrm{E}(\widehat{\beta}-\beta)^\top\boldsymbol{Z}^\top\boldsymbol{Z}(\widehat{\beta}-\beta) = \frac{1}{\sigma_0^2}\mathrm{E}\boldsymbol{\varepsilon}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^\top$ consist of i.i.d. copies of ε defined in Example 1. It follows that,

$$\begin{aligned}
\sum_{i=1}^{n}\mathrm{E}\sigma_0^{-2}|Z_i^\top(\widehat{\beta}-\beta_0)|^2 = \frac{1}{\sigma_0^2}\mathrm{E}\boldsymbol{\varepsilon}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^\top\boldsymbol{Z}^\top\boldsymbol{\varepsilon} &= \frac{1}{\sigma_0^2}\mathrm{trace}\left\{\mathrm{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\right\} \\
&= \mathrm{trace}\left\{\mathrm{E}\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\right\} = d_Z,
\end{aligned} \tag{8}$$

where d_Z is the dimension of Z.

Next, we show that,

$$\sum_{i=1}^{n} \mathrm{E}\sigma_0^{-2}|Z_i^\top(\widehat{\beta}-\beta_0)|^2\mathbb{I}\left\{\sigma_0^{-2}|Z_i^\top(\widehat{\beta}-\beta_0)|^2\geq 1\right\} = o(1), \tag{9}$$

or equivalently,

$$\mathrm{E}n\sigma_0^{-2}|Z_i^\top(\widehat{\beta}-\beta_0)|^2\mathbb{I}\left\{\sigma_0^{-2}|Z_i^\top(\widehat{\beta}-\beta_0)|^2\geq 1\right\} = o(1).$$

We have already shown that $\mathrm{E}n\sigma_0^{-2}|Z_i^\top(\widehat{\beta}-\beta_0)|^2 = d_Z$. By the dominated convergence theorem, it suffices to show that,

$$n\sigma_0^{-2}|Z_i^\top(\widehat{\beta}-\beta_0)|^2\mathbb{I}\left\{\sigma_0^{-2}|Z_i^\top(\widehat{\beta}-\beta_0)|^2\geq 1\right\} = o_p(1).$$

By definition, it in turn suffices to show that,

$$\Pr\left\{\sigma_0^{-2}|Z_i^\top(\widehat{\beta}-\beta_0)|^2\geq 1\right\}\to 0.$$

This holds by Markov's inequality, as

$$\mathrm{E}\sigma_0^{-2}|Z_i^\top(\widehat{\beta}-\beta_0)|^2 = \frac{d_Z}{n}\to 0.$$

Combining (9) together with (7) and (8) yields that,

$$\sum_{i=1}^{n}\mathrm{E}\min\left\{1,\sigma_0^{-2}|Z_i^\top(\widehat{\beta}-\beta_0)|^2\right\}\geq d_Z - o(1)\geq 1-o(1),$$

and hence $\sum_{i=1}^{n}\mathrm{E}d_{\mathrm{TV}}^2\{\widetilde{P}_{X|Z=Z_i}, Q_X^{(n)}(\cdot|Z_i)\}\geq 1-o(1)$.

This completes the proof of Proposition 1. □

A.2 Proof of Theorem 3

We begin by providing an upper bound for the function classes $\mathbb{H}_1$ and $\mathbb{H}_2$. Recall that both $\mathbb{H}_1$ and $\mathbb{H}_2$ are classes of neural networks with a single-hidden layer, finitely many hidden nodes, and the sigmoid activation function. Because of that, each function $h_{1,\theta_1}\in\mathbb{H}_1$ and $h_{2,\theta_2}\in\mathbb{H}_2$ can be represented as

$$h_{1,\theta_1}(x) = \sum_{j=1}^{M}\theta_{1,j}^{(1)}\mathrm{sigmoid}(x^\top\theta_{1,j}^{(2)}),\quad h_{2,\theta_2}(x) = \sum_{j=1}^{M}\theta_{2,j}^{(1)}\mathrm{sigmoid}(y^\top\theta_{2,j}^{(2)}),$$

where θ_1 and θ_2 correspond to the sets of parameters $\left\{(\theta_{1,j}^{(1)},\theta_{1,j}^{(2)}) : 1\leq j\leq M\right\}$ and $\left\{(\theta_{2,j}^{(1)},\theta_{2,j}^{(2)}) : 1\leq j\leq M\right\}$, respectively, and M is a finite integer. Note that the sigmoid function is bounded. As such, the functions h_{1,θ_1} and h_{2,θ_2} are uniformly bounded by

$\sum_{j=1}^{M} |\theta_{1,j}^{(1)}|$ and $\sum_{j=1}^{M} |\theta_{2,j}^{(2)}|$, respectively. Since we sample B many functions $\{h_{1,\theta_b}\}_{b=1}^{B}$ and $\{h_{2,\theta_b}\}_{b=1}^{B}$, these functions are uniformly bounded by

$$M \max_{b,j} \left(|\theta_{b,j}^{(1)}| + |\theta_{b,j}^{(2)}| \right).$$

Since these parameters θ_1, θ_2 are sampled from standard normal distributions, and that

$$\Pr(W > t) = \frac{1}{\sqrt{2\pi}} \int_t^\infty \exp\left(-\frac{w^2}{2}\right) dw \leq \frac{1}{\sqrt{2\pi}} \int_t^\infty w \exp\left(-\frac{w^2}{2}\right) dw = \frac{\exp(-t^2/2)}{\sqrt{2\pi}},$$

for any $t \geq 1$, we can show that $\max_{b,j} \left(|\theta_{b,j}^{(1)}| + |\theta_{b,j}^{(2)}| \right)$ is upper bounded by $\sqrt{\log B}$, with probability approaching one. Note that B grows polynomially with respect to the sample size n. Therefore, we have that the functions in $\mathbb{H}_1$ and $\mathbb{H}_2$ are upper bounded by $\log n$ in absolute values.

Define a test statistic

$$T^{**} = \max_{b_1,b_2} \widehat{\sigma}_{b_1,b_2}^{-1} \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ h_{1,b_1}(X_i) - \frac{1}{M} \sum_{m=1}^{M} h_{1,b_1}(X_i^{(m)}) \right\} \left\{ h_{2,b_2}(Y_i) - \frac{1}{M} \sum_{m=1}^{M} h_{2,b_2}(Y_i^{(m)}) \right\} \right|,$$

where the $\widehat{\sigma}_{b_1,b_2}$ is constructed based on $\{\widetilde{X}_i^{(m)}\}_m$ and $\{\widetilde{Y}_i^{(m)}\}_m$, instead of $\{X_i^{(m)}\}_m$ and $\{Y_i^{(m)}\}_m$. It suffices to show that $|\widehat{T} - T^{**}| = O_p(n^{-2\kappa} \log n)$, and $|T^* - T^{**}| = O_p(n^{-2\kappa} \log n)$.

Step 1. We first consider the difference $|\widehat{T} - T^{**}|$. For any sequences $\{a_n\}_n$, $\{b_n\}_n$, we have that,

$$|\max_n |a_n| - \max_n |b_n|| \leq \max_n |a_n - b_n|. \tag{10}$$

Consequently, we have $|\widehat{T} - T^{**}| \leq I_1 + I_2 + I_3$, where

$$\begin{aligned}
I_1 &= \max_{b_1,b_2} \widehat{\sigma}_{b_1,b_2}^{-1} \left| \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{M} \sum_{m=1}^{M} \left\{ h_{1,b_1}(X_i^{(m)}) - h_{1,b_1}(\widetilde{X}_i^{(m)}) \right\} \right] \left\{ h_{2,b_2}(Y_i) - \frac{1}{M} \sum_{m=1}^{M} h_{2,b_2}(Y_i^{(m)}) \right\} \right|, \\
I_2 &= \max_{b_1,b_2} \widehat{\sigma}_{b_1,b_2}^{-1} \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ h_{1,b_1}(X_i) - \frac{1}{M} \sum_{m=1}^{M} h_{1,b_1}(X_i^{(m)}) \right\} \left[\frac{1}{M} \sum_{m=1}^{M} \left\{ h_{2,b_2}(Y_i^{(m)}) - h_{2,b_2}(\widetilde{Y}_i^{(m)}) \right\} \right] \right|, \\
I_3 &= \max_{b_1,b_2} \widehat{\sigma}_{b_1,b_2}^{-1} \left| \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{M} \sum_{m=1}^{M} \left\{ h_{1,b_1}(X_i^{(m)}) - h_{1,b_1}(\widetilde{X}_i^{(m)}) \right\} \right] \left[\frac{1}{M} \sum_{m=1}^{M} \left\{ h_{2,b_2}(Y_i^{(m)}) - h_{2,b_2}(\widetilde{Y}_i^{(m)}) \right\} \right] \right|.
\end{aligned}$$

If $\min \widehat{\sigma}_{b_1,b_2} \geq c_0$ for some constant $c_0 > 0$, then it suffices to show that $I_j^* = O_p(n^{-(\kappa_x+\kappa_y)} \log n)$, for $j = 1, 2, 3$, where

$$\begin{aligned}
I_1^* &= \max_{b_1,b_2} \left| \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{M} \sum_{m=1}^{M} \left\{ h_{1,b_1}(X_i^{(m)}) - h_{1,b_1}(\widetilde{X}_i^{(m)}) \right\} \right] \left\{ h_{2,b_2}(Y_i) - \frac{1}{M} \sum_{m=1}^{M} h_{2,b_2}(Y_i^{(m)}) \right\} \right|, \\
I_2^* &= \max_{b_1,b_2} \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ h_{1,b_1}(X_i) - \frac{1}{M} \sum_{m=1}^{M} h_{1,b_1}(X_i^{(m)}) \right\} \left[\frac{1}{M} \sum_{m=1}^{M} \left\{ h_{2,b_2}(Y_i^{(m)}) - h_{2,b_2}(\widetilde{Y}_i^{(m)}) \right\} \right] \right|, \\
I_3^* &= \max_{b_1,b_2} \left| \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{M} \sum_{m=1}^{M} \left\{ h_{1,b_1}(X_i^{(m)}) - h_{1,b_1}(\widetilde{X}_i^{(m)}) \right\} \right] \left[\frac{1}{M} \sum_{m=1}^{M} \left\{ h_{2,b_2}(Y_i^{(m)}) - h_{2,b_2}(\widetilde{Y}_i^{(m)}) \right\} \right] \right|.
\end{aligned}$$

The number of folds L is finite, as such, it suffices to show that $I_j^{(\ell)} = O_p(n^{-(\kappa_x+\kappa_y)}\log n)$, for $j = 1, 2, 3$ and $\ell = 1, \ldots, L$, where

$$I_1^{(\ell)} = \max_{b_1,b_2}\left|\frac{1}{n}\sum_{i\in\mathcal{I}^{(\ell)}}\left[\frac{1}{M}\sum_{m=1}^{M}\left\{h_{1,b_1}(X_i^{(m)}) - h_{1,b_1}(\widetilde{X}_i^{(m)})\right\}\right]\left\{h_{2,b_2}(Y_i) - \frac{1}{M}\sum_{m=1}^{M}h_{2,b_2}(Y_i^{(m)})\right\}\right|,$$

$$I_2^{(\ell)} = \max_{b_1,b_2}\left|\frac{1}{n}\sum_{i\in\mathcal{I}^{(\ell)}}\left\{h_{1,b_1}(X_i) - \frac{1}{M}\sum_{m=1}^{M}h_{1,b_1}(X_i^{(m)})\right\}\left[\frac{1}{M}\sum_{m=1}^{M}\left\{h_{2,b_2}(Y_i^{(m)}) - h_{2,b_2}(\widetilde{Y}_i^{(m)})\right\}\right]\right|,$$

$$I_3^{(\ell)} = \max_{b_1,b_2}\left|\frac{1}{n}\sum_{i\in\mathcal{I}^{(\ell)}}\left[\frac{1}{M}\sum_{m=1}^{M}\left\{h_{1,b_1}(X_i^{(m)}) - h_{1,b_1}(\widetilde{X}_i^{(m)})\right\}\right]\left[\frac{1}{M}\sum_{m=1}^{M}\left\{h_{2,b_2}(Y_i^{(m)}) - h_{2,b_2}(\widetilde{Y}_i^{(m)})\right\}\right]\right|.$$

We divide the rest of the proof into four sub-steps. We first show that $I_j^{(\ell)} = O_p(n^{-(\kappa_x+\kappa_y)}\log n)$, for $j = 1, 2, 3$. Finally, we show $\Pr(\min\widehat{\sigma}_{b_1,b_2} \geq c_0) \to 1$ for some constant $c_0 > 0$.

Step 1.1. Recall we have shown that the functions in $\mathbb{H}_1$ and $\mathbb{H}_2$ are bounded by $\log n$ in absolute values at the beginning of the proof of Theorem 3. By Bernstein's inequality, we have that,

$$\Pr\left[\left|\sum_{m=1}^{M}h_{1,b}(X_i^{(m)}) - M\mathrm{E}\{h_{1,b}(X_i)|Z_i\}\right| \geq t\right] \leq 2\exp\left\{-\frac{t^2}{2(M\log n + t\sqrt{\log n}/3)}\right\},$$

for any b and i. Set $t = \sqrt{3(c+2)M}\log n$, where the constant c is as defined in the statement of Theorem 2. For a sufficiently large n, we have $t\sqrt{\log n}/3 \leq M\log n/2$. It follows that

$$\Pr\left[\left|\sum_{m=1}^{M}h_{1,b}(X_i^{(m)}) - M\mathrm{E}\{h_{1,b}(X_i)|Z_i\}\right| \geq \sqrt{3(c+2)M}\log n\right] \leq \frac{2}{n^{c+2}}.$$

By Bonferroni's inequality, we obtain that,

$$\begin{aligned}&\Pr\left[\max_{b\in\{1,\cdots,B\}}\max_{i\in\{1,\cdots,n\}}\left|\sum_{m=1}^{M}h_{1,b}(X_i^{(m)}) - M\mathrm{E}\{h_{1,b}(X_i)|Z_i\}\right| \geq \sqrt{3(c+2)M}\log n\right]\\ \leq\ & Bn\max_{b\in\{1,\cdots,B\}}\max_{i\in\{1,\cdots,n\}}\Pr\left[\left|\sum_{m=1}^{M}h_{1,b}(X_i^{(m)}) - M\mathrm{E}\{h_{1,b}(X_i)|Z_i\}\right| \geq \sqrt{3(c+2)M}\log n\right] \leq \frac{2Bn}{n^{c+2}}.\end{aligned}$$

Under the condition $B = O(n^c)$, we obtain with probability $1 - O(n^{-1})$ that,

$$\max_{b\in\{1,\cdots,B\}}\max_{i\in\{1,\cdots,n\}}\left|\sum_{m=1}^{M}h_{1,b}(X_i^{(m)}) - M\mathrm{E}\{h_{1,b}(X_i)|Z_i\}\right| \leq O(1)n^{-1/2}\log n, \tag{11}$$

as M is proportional to n, and $O(1)$ denotes some positive constant.

Similarly, we can show that,

$$\max_{b\in\{1,\cdots,B\}}\max_{i\in\mathcal{I}^{(\ell)}}\left|\sum_{m=1}^{M}h_{1,b}(\widetilde{X}_i^{(m)}) - M\int_x h_{1,b}(x)\widetilde{P}^{(\ell)}_{X|Z=Z_i}(dx)\right| \leq O(1)\sqrt{n}\log n,$$

with probability $1 - O(n^{-1})$. Combining this with (11), we obtain with probability $1 - O(n^{-1})$ that,

$$\max_{\substack{b\in\{1,\ldots,B\}\\ i\in\mathcal{I}^{(\ell)}}} \left| \sum_{m=1}^{M} \left\{ h_{1,b}(X_i^{(m)}) - h_{1,b}(\widetilde{X}_i^{(m)}) \right\} \right. \\ \left. -M \int_x h_{1,b}(x) \left\{ P_{X|Z=Z_i}(dx) - \widetilde{P}^{(\ell)}_{X|Z=Z_i}(dx) \right\} \right| \le O(1)\sqrt{n}\log n. \tag{12}$$

Conditional on Z_i, the expectation of $h_{2,b_2}(Y_i) - M^{-1}\sum_{m=1}^{M} h_{2,b_2}(Y_i^{(m)})$ equals zero. Under the null hypothesis, the expectation of $M^{-1}\sum_{m=1}^{M}\{h_{1,b_1}(X_i^{(m)}) - h_{1,b_1}(\widetilde{X}_i^{(m)})\}\{h_{2,b_2}(Y_i) - M^{-1}\sum_{m=1}^{M} h_{2,b_2}(Y_i^{(m)})\}$ equals zero as well. Applying Bernstein's inequality again, we can show with probability tending to 1 that,

$$I_1^{(\ell)} \le O(1)\left(\sigma n^{-1/2}\log^{3/2} n + n^{-1}\log^2 n\right), \tag{13}$$

where

$$\sigma^2 = \max_{b_1,b_2} \mathrm{E}\left| \frac{1}{M}\sum_{m=1}^{M} \left\{ h_{1,b_1}(X_i^{(m)}) - h_{1,b_1}(\widetilde{X}_i^{(m)}) \right\} \left\{ h_{2,b_2}(Y_i) - \frac{1}{M}\sum_{m=1}^{M} h_{2,b_2}(Y_i^{(m)}) \right\} \right|^2 \\ \le \max_{b_1} \mathrm{E}\left| \frac{1}{M}\sum_{m=1}^{M} \left\{ h_{1,b_1}(X_i^{(m)}) - h_{1,b_1}(\widetilde{X}_i^{(m)}) \right\} \right|^2 \log n.$$

Let $\mathcal{A}$ denote the event in (12). The last term on the second line can be bounded from above by

$$\max_{b_1,i} \mathrm{E}\left| \frac{1}{M}\sum_{m=1}^{M} \{ h_{1,b_1}(X_i^{(m)}) - h_{1,b_1}(\widetilde{X}_i^{(m)}) \} \right|^2 \mathbb{I}(\mathcal{A})\log n \tag{14}$$

$$+ \quad \max_{b_1,i} \mathrm{E}\left| \frac{1}{M}\sum_{m=1}^{M} \{ h_{1,b_1}(X_i^{(m)}) - h_{1,b_1}(\widetilde{X}_i^{(m)}) \} \right|^2 \mathbb{I}(\mathcal{A}^c)\log n. \tag{15}$$

Since M is proportional to n, by (10), (14) is upper bounded by

$$O(1)\left[n^{-1}\log^2 n + \max_{\substack{b\in\{1,\cdots,B\}\\ i\in\mathcal{I}^{(\ell)}}} \mathrm{E}\left| \int_x h_{1,b}(x)\left\{ \widetilde{P}^{(\ell)}_{X|Z=Z_i}(dx) - P_{X|Z=Z_i}(dx) \right\} \right|^2 \right] \log n.$$

By the boundedness of the function class $\mathbb{H}_1$, it can be further bounded from above by

$$O(1)\left\{ n^{-1}\log^3 n + \mathrm{E}d^2_{\mathrm{TV}}(\widetilde{P}^{(\ell)}_{X|Z}, P_{X|Z})\log^2 n \right\}. \tag{16}$$

The above quantity is of order $O(n^{-2\kappa_x}\log^2 n)$. Consequently, (14) is of the order $O(n^{-2\kappa_x}\log^2 n)$.

Note that the event $\mathcal{A}$ occurs with probability at least $1 - O(n^{-1})$. By the boundedness of the function class $\mathbb{H}_1$, (15) is of the order $O(n^{-1}\log^2 n)$.

Therefore, σ^2 is of the order $O(n^{-2\kappa_x}\log^2 n)$. This implies that $\mathcal{I}_1^{(\ell)}$ can be bounded from above by $O(n^{-1/2-\kappa_x}\log^{5/2} n)$, which in turn yields that $\mathcal{I}_1^{(\ell)} = O_p(n^{-\kappa_x-\kappa_y}\log n)$, since $\kappa_x, \kappa_y < 1/2$.

Step 1.2. This step can be proven in a similar way as Step 1.1, and is omitted.

Step 1.3. Under H_0, the expectation of

$$\frac{1}{|\mathcal{I}^{(\ell)}|}\sum_{i\in\mathcal{I}^{(\ell)}}\left[\frac{1}{M}\sum_{m=1}^{M}\left\{h_{1,b_1}(X_i^{(m)}) - h_{1,b_1}(\widetilde{X}_i^{(m)})\right\}\right]\left[\frac{1}{M}\sum_{m=1}^{M}\left\{h_{2,b_2}(Y_i^{(m)}) - h_{2,b_2}(\widetilde{Y}_i^{(m)})\right\}\right]$$

equals

$$\mathrm{E}\int_x h_{1,b_1}(x)\left\{\widetilde{P}^{(\ell)}_{X|Z}(dx) - P_{X|Z}(dx)\right\}\int_y h_{2,b_2}(y)\left\{\widetilde{P}^{(\ell)}_{Y|Z}(dy) - P_{Y|Z}(dy)\right\}.$$

Similar to (16), its absolute value can be upper bounded by

$$\mathrm{E}d_{\mathrm{TV}}\left\{\widetilde{P}^{(\ell)}_{X|Z=Z_i}, P_{X|Z}\right\} d_{\mathrm{TV}}\left\{\widetilde{P}^{(\ell)}_{Y|Z=Z_i}, P_{Y|Z}\right\}\log n.$$

Following Cauchy-Schwarz inequality, we have that,

$$\begin{aligned}&\mathrm{E}d_{\mathrm{TV}}\left\{\widetilde{P}^{(\ell)}_{X|Z=Z_i}, P_{X|Z}\right\} d_{\mathrm{TV}}\left\{\widetilde{P}^{(\ell)}_{Y|Z=Z_i}, P_{Y|Z}\right\}\\ \leq\ &\sqrt{\mathrm{E}d^2_{\mathrm{TV}}\left\{\widetilde{P}^{(\ell)}_{X|Z=Z_i}, P_{X|Z}\right\}\mathrm{E}d^2_{\mathrm{TV}}\left\{\widetilde{P}^{(\ell)}_{Y|Z=Z_i}, P_{Y|Z}\right\}} = O(n^{-(\kappa_x+\kappa_y)}).\end{aligned}$$

This yields that,

$$\max_{b_1,b_2}\left|\mathrm{E}\int_x h_{1,b_1}(x)\left\{\widetilde{P}^{(\ell)}_{X|Z}(dx) - P_{X|Z}(dx)\right\}\int_y h_{2,b_2}(y)\left\{\widetilde{P}^{(\ell)}_{Y|Z}(dy) - P_{Y|Z}(dy)\right\}\right| = O(n^{-(\kappa_x+\kappa_y)}\log n).$$

Following similar arguments as in Step 1.1, we obtain that,

$$\begin{aligned}I_3^{(\ell)} - \max_{b_1,b_2}\left|\mathrm{E}\int_x h_{1,b_1}(x)\left\{\widetilde{P}^{(\ell)}_{X|Z}(dx) - P_{X|Z}(dx)\right\}\int_y h_{2,b_2}(y)\left\{\widetilde{P}^{(\ell)}_{Y|Z}(dy) - P_{Y|Z}(dy)\right\}\right|&\\ = O_p(n^{-(\kappa_x+\kappa_y)}\log n).&\end{aligned}$$

Therefore, we obtain that $I_3^{(\ell)} = O_p(n^{-(\kappa_x+\kappa_y)}\log n)$.

Step 1.4. Recall that $\widehat{\sigma}^2_{b_1,b_2}$ is defined by

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\left[h_{1,b_1}(X_i) - \widehat{\mathrm{E}}\{h_{1,b_1}(X_i)|Z_i\}\right]\left[h_{2,b_2}(Y_i) - \widehat{\mathrm{E}}\{h_{2,b_2}(Y_i)|Z_i\}\right] - \mathrm{GCM}\{h_{1,b_1}(X), h_{2,b_2}(Y)\}\right)^2.$$

With some calculations, it is equal to

$$\begin{aligned}\frac{1}{n-1}\sum_{i=1}^{n}\left[h_{1,b_1}(X_i) - \widehat{\mathrm{E}}\{h_{1,b_1}(X_i)|Z_i\}\right]^2\left[h_{2,b_2}(Y_i) - \widehat{\mathrm{E}}\{h_{2,b_2}(Y_i)|Z_i\}\right]^2&\\ -\frac{n}{n-1}\mathrm{GCM}^2\{h_{1,b_1}(X), h_{2,b_2}(Y)\},&\end{aligned} \tag{17}$$

where the estimated conditional expectation $\widehat{\mathbb{E}}$ is computed using GANs.

Consider the second term $\mathrm{GCM}\{h_{1,b_1}(X), h_{2,b_2}(Y)\}$ in (17). Following similar arguments as in Steps 1.1 and 1.3, we have that,

$$\max_{b_1,b_2} \left|\mathrm{GCM}\{h_{1,b_1}(X), h_{2,b_2}(Y)\} - \mathrm{GCM}'\{h_{1,b_1}(X), h_{2,b_2}(Y)\}\right| = O_p(n^{-(\kappa_x+\kappa_y)} \log n),$$

where $\mathrm{GCM}'\{h_{1,b_1}(X), h_{2,b_2}(Y)\}$ equals

$$\frac{1}{n}\sum_{i=1}^{n}\left\{h_{1,b_1}(X_i) - \frac{1}{M}\sum_{m=1}^{M} h_{1,b_1}(X_i^{(m)})\right\}\left\{h_{2,b_2}(Y_i) - \frac{1}{M}\sum_{m=1}^{M} h_{2,b_2}(Y_i^{(m)})\right\}.$$

Similar to (12), we can show that,

$$\max_{b_1,b_2} \left|\mathrm{GCM}'\{h_{1,b_1}(X), h_{2,b_2}(Y)\} - \mathrm{GCM}^*\{h_{1,b_1}(X), h_{2,b_2}(Y)\}\right| = O_p\left(n^{-1/2}\sqrt{\log n}\right).$$

Consequently, we have that,

$$\max_{b_1,b_2} |\mathrm{GCM}\{h_{1,b_1}(X), h_{2,b_2}(Y)\} - \mathrm{GCM}^*\{h_{1,b_1}(X), h_{2,b_2}(Y)\}| = O_p\left(n^{-1/2}\sqrt{\log n}\right).$$

Since the function classes $\mathbb{H}_1$ and $\mathbb{H}_2$ are bounded, both GCM and GCM^* are bounded by $\log n$ in absolute values. Consequently,

$$\max_{b_1,b_2} \left|\mathrm{GCM}^2\{h_{1,b_1}(X), h_{2,b_2}(Y)\} - \mathrm{GCM}^{*2}\{h_{1,b_1}(X), h_{2,b_2}(Y)\}\right| = O_p\big(n^{-1/2}\log^{3/2} n\big). \quad (18)$$

Next, consider the first term in (17). Note that it can be represented by

$$\frac{n}{n-1}\frac{1}{L}\sum_{\ell=1}^{L}\left(\frac{1}{|\mathcal{I}^\ell|}\sum_{i\in\mathcal{I}^{(\ell)}}\left[h_{1,b_1}(X_i) - \widehat{\mathrm{E}}\{h_{1,b_1}(X_i)|Z_i\}\right]^2\left[h_{2,b_2}(Y_i) - \widehat{\mathrm{E}}\{h_{2,b_2}(Y_i)|Z_i\}\right]^2\right).$$

Similar to (12), we can show that,

$$\max_{b_1,b_2}\left|\frac{1}{|\mathcal{I}^\ell|}\sum_{i\in\mathcal{I}^{(\ell)}}\left[h_{1,b_1}(X_i) - \widehat{\mathrm{E}}\{h_{1,b_1}(X_i)|Z_i\}\right]^2\left[h_{2,b_2}(Y_i) - \widehat{\mathrm{E}}\{h_{2,b_2}(Y_i)|Z_i\}\right]^2\right.$$
$$\left. - \mathrm{E}\left[h_{1,b_1}(X_1) - \widehat{\mathrm{E}}\{h_{1,b_1}(X_1)|Z_1\}\right]^2\left[h_{2,b_2}(Y_1) - \widehat{\mathrm{E}}\{h_{2,b_2}(Y_1)|Z_1\}\right]^2\right| = O_p(n^{-1/2}\log^{3/2} n).$$

Following similar arguments as in Steps 1.1 and 1.3, we can show that,

$$\max_{b_1,b_2}\left|\mathrm{E}\left[h_{1,b_1}(X_1) - \widehat{\mathrm{E}}\{h_{1,b_1}(X_1)|Z_1\}\right]^2\left[h_{2,b_2}(Y_1) - \widehat{\mathrm{E}}\{h_{2,b_2}(Y_1)|Z_1\}\right]^2\right.$$
$$\left. - \mathrm{E}\left[h_{1,b_1}(X_1) - \mathrm{E}\{h_{1,b_1}(X_1)|Z_1\}\right]^2\left[h_{2,b_2}(Y_1) - \mathrm{E}\{h_{2,b_2}(Y_1)|Z_1\}\right]^2\right| = O_p(n^{-\bar{c}}),$$

for some constant $0 < \bar{c} < 1/2$. It follows that,

$$\max_{b_1,b_2} \Bigg| \frac{1}{|\mathcal{I}^{\ell}|} \sum_{i \in \mathcal{I}^{(\ell)}} \Big[h_{1,b_1}(X_i) - \widehat{\mathrm{E}}\{h_{1,b_1}(X_i)|Z_i\} \Big]^2 \Big[h_{2,b_2}(Y_i) - \widehat{\mathrm{E}}\{h_{2,b_2}(Y_i)|Z_i\} \Big]^2 \\ - \mathrm{E}\left[h_{1,b_1}(X) - \mathrm{E}\{h_{1,b_1}(X)|Z\} \right]^2 \left[h_{2,b_2}(Y) - \mathrm{E}\{h_{2,b_2}(Y)|Z\} \right]^2 \Bigg| = O_p(n^{-\bar{c}}),$$

and henceforth,

$$\max_{b_1,b_2} \Bigg| \frac{1}{n} \sum_{i=1}^{n} \Big[h_{1,b_1}(X_i) - \widehat{\mathrm{E}}\{h_{1,b_1}(X_i)|Z_i\} \Big]^2 \Big[h_{2,b_2}(Y_i) - \widehat{\mathrm{E}}\{h_{2,b_2}(Y_i)|Z_i\} \Big]^2 \\ - \mathrm{E}\left[h_{1,b_1}(X) - \mathrm{E}\{h_{1,b_1}(X)|Z\} \right]^2 \left[h_{2,b_2}(Y) - \mathrm{E}\{h_{2,b_2}(Y)|Z\} \right]^2 \Bigg| = O_p(n^{-\bar{c}}).$$

Combining this together with (18) yields that,

$$\max_{b_1,b_2} \left| \widehat{\sigma}^2_{b_1,b_2} - \frac{n}{n-1} \mathrm{Var}\Big(\left[h_{1,b_1}(X) - \mathrm{E}\{h_{1,b_1}(X)|Z\} \right] \left[h_{2,b_2}(Y) - \mathrm{E}\{h_{2,b_2}(Y)|Z\} \right] \Big) \right| = O_p(n^{-\bar{c}}).$$

Then, we have that,

$$\min_{b_1,b_2} \mathrm{Var}\Big(\left[h_{1,b_1}(X) - \mathrm{E}\{h_{1,b_1}(X)|Z\} \right] \left[h_{2,b_2}(Y) - \mathrm{E}\{h_{2,b_2}(Y)|Z\} \right] \Big) \geq c^*,$$

for some constant $c^* > 0$. Therefore, we have that

$$\min_{b_1,b_2} \widehat{\sigma}^2_{b_1,b_2} \geq 2^{-1} c^*,$$

with probability tending to 1.

Step 2. We next consider the difference $|T^* - T^{**}|$, and show that it is of the order $O_p(n^{-2\kappa} \log n)$. Denote by $\widehat{\sigma}^{*2}_{b_1,b_2}$ the variance estimator with $\{\widetilde{X}_i^{(m)}\}_m$ and $\{\widetilde{Y}_i^{(m)}\}_m$ replaced by $\{X_i^{(m)}\}_m$ and $\{Y_i^{(m)}\}_m$. Using (10), the difference between T^* and T^{**} is upper bounded by

$$\max_{b_1,b_2} |\widehat{\sigma}^{-1}_{b_1,b_2} - \widehat{\sigma}^{*-1}_{b_1,b_2}| \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ h_{1,b_1}(X_i) - \frac{1}{M} \sum_{m=1}^{M} h_{1,b_1}(X_i^{(m)}) \right\} \left\{ h_{2,b_2}(Y_i) - \frac{1}{M} \sum_{m=1}^{M} h_{2,b_2}(Y_i^{(m)}) \right\} \right|.$$

Under H_0, similar to (12), we can show that,

$$\max_{b_1,b_2} \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ h_{1,b_1}(X_i) - \frac{1}{M} \sum_{m=1}^{M} h_{1,b_1}(X_i^{(m)}) \right\} \left\{ h_{2,b_2}(Y_i) - \frac{1}{M} \sum_{m=1}^{M} h_{2,b_2}(Y_i^{(m)}) \right\} \right| \\ = O_p(n^{-1/2} \log^{3/2} n).$$

To show $|T^* - T^{**}| = O_p(n^{-2\kappa} \log n)$, it suffices to show that $\max_{b_1,b_2} |\widehat{\sigma}^{-1}_{b_1,b_2} - \widehat{\sigma}^{*-1}_{b_1,b_2}| = O_p(n^{-\bar{c}})$ for some constant $\bar{c} > 0$. Since both $\widehat{\sigma}^{-1}_{b_1,b_2}$ and $\widehat{\sigma}_{b_1,b_2}$ are bounded away from zero, it suffices to show that $\max_{b_1,b_2} |\widehat{\sigma}^2_{b_1,b_2} - \widehat{\sigma}^{*2}_{b_1,b_2}| = O_p(n^{-\bar{c}})$.

Following similar arguments as in Steps 1.1 and 1.3, we can show that,

$$\max_{b_1,b_2}\left|\widehat{\sigma}^2_{b_1,b_2}-\frac{n}{n-1}\mathrm{Var}\Big(\left[h_{1,b_1}(X)-\mathrm{E}\{h_{1,b_1}(X)|Z\}\right]\left[h_{2,b_2}(Y)-\mathrm{E}\{h_{2,b_2}(Y)|Z\}\right]\Big)\right|=O_p(n^{-\bar{c}}),$$

$$\max_{b_1,b_2}\left|\widehat{\sigma}^{*2}_{b_1,b_2}-\frac{n}{n-1}\mathrm{Var}\Big(\left[h_{1,b_1}(X)-\mathrm{E}\{h_{1,b_1}(X)|Z\}\right]\left[h_{2,b_2}(Y)-\mathrm{E}\{h_{2,b_2}(Y)|Z\}\right]\Big)\right|=O_p(n^{-\bar{c}}).$$

This completes the proof of Theorem 3. □

A.3 Proof of Theorem 4

In the proof of Theorem 3, we have already shown that $\widehat{T}-T^*=O_p(n^{-(\kappa_x+\kappa_y)}\log n)$. Following similar arguments as in Step 1.4, we can show that $T^*-T^{***}=O_p(n^{-(\kappa_x+\kappa_y)}\log n)$, where

$$T^{***}=\max_{b_1,b_2}\sigma^{-1}_{b_1,b_2}|\left|n^{-1}\sum_{i=1}^n\left\{h_{1,b_1}(X_i)-\frac{1}{M}\sum_{m=1}^M h_{1,b_1}(X_i^{(m)})\right\}\left\{h_{2,b_2}(Y_i)-\frac{1}{M}\sum_{m=1}^M h_{2,b_2}(Y_i^{(m)})\right\}\right|,$$

where

$$\sigma^2_{b_1,b_2}=\frac{n}{n-1}\mathrm{Var}\Big(\left[h_{1,b_1}(X)-\mathrm{E}\{h_{1,b_1}(X)|Z\}\right]\left[h_{2,b_2}(Y)-\mathrm{E}\{h_{2,b_2}(Y)|Z\}\right]\Big).$$

By (11), following similar arguments as in the proof regarding the term I_1 in Theorem 3, we can show that $T^{***}-T^{****}=O_p(n^{-(\kappa_x+\kappa_y)}\log n)$, where

$$T^{****}=\max_{b_1,b_2}\sigma^{-1}_{b_1,b_2}|\left|n^{-1}\sum_{i=1}^n\left[h_{1,b_1}(X_i)-\mathrm{E}\{h_{1,b_1}(X_i)|Z_i\}\right]\left\{h_{2,b_2}(Y_i)-\frac{1}{M}\sum_{m=1}^M h_{2,b_2}(Y_i^{(m)})\right\}\right|.$$

Similarly, we can show that $T^{****}-T_0=O_p(n^{-(\kappa_x+\kappa_y)}\log n)$, where

$$T_0=\max_{b_1,b_2}\sigma^{-1}_{b_1,b_2}|\left|n^{-1}\sum_{i=1}^n\left[h_{1,b_1}(X_i)-\mathrm{E}\{h_{1,b_1}(X_i)|Z_i\}\right]\left[h_{2,b_2}(Y_i)-\mathrm{E}\{h_{2,b_2}(Y_i)|Z_i\}\right]\right|.$$

Therefore, we have shown that $\widehat{T}-T_0=O_p(n^{-(\kappa_x+\kappa_y)}\log n)$. Since $\kappa_x+\kappa_y>1/2$, we have that,

$$\sqrt{n}(\widehat{T}-T_0)=o_p(\log^{-1/2}n). \tag{19}$$

Define a $B^2\times B^2$ matrix Σ_0 whose $\{b_1+B(b_2-1),b_3+B(b_4-1)\}$th entry is given by

$$\mathrm{cov}\Big(\sigma^{-1}_{b_1,b_2}\left[h_{1,b_1}(X_i)-\mathrm{E}\{h_{1,b_1}(X_i)|Z_i\}\right]\left[h_{2,b_2}(Y_i)-\mathrm{E}\{h_{2,b_2}(Y_i)|Z_i\}\right],$$
$$\sigma^{-1}_{b_3,b_4}\left[h_{1,b_3}(X_i)-\mathrm{E}\{h_{1,b_3}(X_i)|Z_i\}\right]\left[h_{2,b_4}(Y_i)-\mathrm{E}\{h_{2,b_4}(Y_i)|Z_i\}\right]\Big).$$

In the following, we show that,

$$\sup_t\left|\Pr\left(\sqrt{n}\widehat{T}_0\le t|\mathcal{H}_0\right)-\Pr\left(\|N(0,\Sigma_0)\|_\infty\le t\right)\right|=o(1). \tag{20}$$

When B is finite, this is implied by the classical weak convergence results. When B diverges with n, we require $B = O(n^c)$ for some constant $c > 0$. By the definition of σ_{b_1,b_2}, the variance of

$$\sigma_{b_1,b_2}^{-1} \left[h_{1,b_1}(X_i) - \mathrm{E}\{h_{1,b_1}(X_i)|Z_i\}\right] \left[h_{2,b_2}(Y_i) - \mathrm{E}\{h_{2,b_2}(Y_i)|Z_i\}\right]$$

is bounded from above by $(n-1)/n$. Moreover, combining the boundedness of the function spaces $\mathbb{H}_1$ and $\mathbb{H}_2$ together with the definition of σ_{b_1,b_2} yields that,

$$\left\{\sigma_{b_1,b_2}^{-1} \left[h_{1,b_1}(X_i) - \mathrm{E}\{h_{1,b_1}(X_i)|Z_i\}\right] \left[h_{2,b_2}(Y_i) - \mathrm{E}\{h_{2,b_2}(Y_i)|Z_i\}\right] : b_1, b_2 \in \{1, \cdots, B\}\right\}$$

are uniformly bounded from infinity by $O(\log n)$, with probability tending to 1. We can show that (20) holds. This implies that,

$$\sigma_{b_1,b_2}^{-1} n^{-1/2} \sum_{i=1}^{n} \left[h_{1,b_1}(X_i) - \mathrm{E}\{h_{1,b_1}(X_i)|Z_i\}\right] \left[h_{2,b_2}(Y_i) - \mathrm{E}\{h_{2,b_2}(Y_i)|Z_i\}\right]$$

is asymptotically normal with zero mean.

Combining (20) together with (19) yields that,

$$\begin{aligned} \Pr\left(\sqrt{n}\widehat{T} \leq t|\mathcal{H}_0\right) &\geq \Pr\left(\|N(0,\Sigma_0)\|_\infty \leq t - \epsilon_0 \log^{-1/2} n\right) - o(1), \\ \Pr\left(\sqrt{n}\widehat{T} \leq t|\mathcal{H}_0\right) &\leq \Pr\left(\|N(0,\Sigma_0)\|_\infty \leq t + \epsilon_0 \log^{-1/2} n\right) + o(1), \end{aligned} \tag{21}$$

for any sufficiently small $\epsilon_0 > 0$, where the little-o terms are uniform in t.

Following similar arguments as in Step 1.4 and Step 2 of the proof of Theorem 3, we can show that $\|\widehat{\Sigma} - \Sigma_0\|_{\infty,\infty} = O_p(n^{-\bar{c}})$ for some constant $\bar{c} > 0$. Following similar arguments for (21), we have that,

$$\begin{aligned} \Pr\left(\sqrt{n}\widehat{T} \leq t|\mathcal{H}_0\right) &\geq \Pr\left(\|N(0,\widehat{\Sigma})\|_\infty \leq t - 2\epsilon_0 \log^{-1/2} n|\widehat{\Sigma}\right) - o(1), \\ \Pr\left(\sqrt{n}\widehat{T} \leq t|\mathcal{H}_0\right) &\leq \Pr\left(\|N(0,\widehat{\Sigma})\|_\infty \leq t + 2\epsilon_0 \log^{-1/2} n|\widehat{\Sigma}\right) + o(1), \end{aligned}$$

for any sufficiently small $\epsilon_0 > 0$. Since the little-o terms are uniform in $t \in \mathbb{R}$, we obtain that,

$$\begin{aligned} &\sup_t |\Pr(\sqrt{n}\widehat{T} \leq t|\mathcal{H}_0) - \Pr(\|N(0,\widehat{\Sigma})\|_\infty \leq t|\widehat{\Sigma})| \leq o(1) \\ + \quad &\sup_t |\Pr(\|N(0,\widehat{\Sigma})\|_\infty \leq t + 2\epsilon \log^{-1/2} n|\widehat{\Sigma}) - \Pr(\|N(0,\widehat{\Sigma})\|_\infty \leq t - 2\epsilon_0 \log^{-1/2} n|\widehat{\Sigma})|. \end{aligned}$$

By Theorem 1 of Chernozhukov et al. (2017), the term on the second line can be bounded by $O(1)\epsilon_0 \log^{1/2} B \log^{-1/2} n$, where $O(1)$ denotes some positive constant. Since $B = O(n^c)$, $\log^{1/2} B \log^{-1/2} n = O(1)$. As ϵ_0 grows to zero, this term becomes negligible. Consequently, we obtain that,

$$\sup_t \left|\Pr\left(\sqrt{n}\widehat{T} \leq t|\mathcal{H}_0\right) - \Pr\left(\|N(0,\widehat{\Sigma})\|_\infty \leq t|\widehat{\Sigma}\right)\right| \leq o(1).$$

As such, the distribution of our test statistic can be well-approximated by that of the bootstrap samples. This completes the proof of Theorem 4. □

A.4 Proof of Theorem 5

We break the proof into two steps. In Step 1, we show that, under $\mathcal{H}_1^*$, there exist two neural networks functions $f(X) \in \mathbb{H}_1$ and $g(Y) \in \mathbb{H}_2$, such that

$$I(f,g) = \mathrm{E}[f(X) - \mathrm{E}\{f(X)|Z\}][g(Y) - \mathrm{E}\{g(Y)|Z\}] \neq 0,$$

In Step 2, we prove the power of our test approaches one, as the sample size diverges to infinity.

Step 1. We first observe that the measure $I(f,g) = \mathrm{E}[f(X) - \mathrm{E}\{f(X)|Z\}][g(Y) - \mathrm{E}\{g(Y)|Z\}]$ is continuous in f and g. That is, for any $f_1, f_2 \in L_X^2$ and $g_1, g_2 \in L_Y^2$, the difference $I(f_1,g_1) - I(f_2,g_2)$ decays to zero as both $\mathrm{E}|f_1(X) - f_2(X)|^2$ and $\mathrm{E}|g_1(X) - g_2(X)|^2$ decay to zero.

Under $\mathcal{H}_1^*$, there exist functions $f^* \in L_X^2$ and $g^* \in L_Y^2$, such that $I(f^*,g^*) \neq 0$. Without loss of generality, assume f^* and g^* are bounded. Otherwise, we can find sequences of bounded functions $\{f_n^*\}_n$ and $\{g_n^*\}_n$ that converge to f^* and g^* under L_2-norm, respectively. As a result, we would have $I(f_n^*, g_n^*) \neq 0$ for some n.

By Lusin's theorem, we can find a sequence of bounded and continuous functions $\{f_n^{**}\}_n$, such that $\lim_n \Pr(f_n^{**}(X) \neq f^*(X)) = 0$. By dominated convergence theorem, it follows that f_n^{**} converges to f^* under L_2-norm. Similarly, we can find a sequence of continuous functions $\{g_n^{**}\}_n$, such that g_n^{**} converges to g^* under L_2-norm. This together with the fact that $I(f,g)$ is continuous in (f,g) implies that there exist some continuous functions f^{**} and g^{**}, such that $I(f^{**}, g^{**}) \neq 0$.

A key observation here is that, the class of neural networks have universal approximation property. Since the support of X and Y are bounded, it follows from Theorem 1 of Cybenko (1989) that the class of single-layered neural networks with sigmoid activation function is dense in the class of bounded, continuous functions with a compact support. As such, we can find some neural network functions f^{***} and g^{***} such that $I(f^{***}, g^{***}) \neq 0$. We then argue that there must exist $f \in \mathbb{H}_1$ and $g \in \mathbb{H}_2$, such that $I(f,g) = 0$. Otherwise, f^{***} and g^{***} can be represented as linear combinations of neural network functions in $\mathbb{H}_1$, $\mathbb{H}_2$ with finitely many number of parameters, and we would have $I(f^{***}, g^{***}) = 0$ as a result. This completes Step 1.

Step 2. We first show that $I(h_{1,\theta_1}, h_{2,\theta_2})$ is a Lipschitz continuous function of (θ_1, θ_2). Note that $h_{1,\theta_1}(X)$ and $h_{2,\theta_2}(Y)$ are Lipschitz continuous functions of θ_1 and θ_2, respectively. For any $\theta_{1,1}, \theta_{1,2} \in \mathbb{R}^{d_1}$, $\theta_{2,1}, \theta_{2,2} \in \mathbb{R}^{d_2}$, we have that,

$$\begin{aligned}
&|I(h_{1,\theta_1}, h_{2,\theta_2}) - I(h_{1,\theta_1}, h_{2,\theta_2})| \\
\leq\ &|\mathrm{E}[h_{1,1}(X) - \mathrm{E}\{h_{1,1}(X)|Z\} - h_{1,2}(X) + \mathrm{E}\{h_{2,1}(X)|Z\}][h_{2,1}(Y) - \mathrm{E}\{h_{2,1}(Y)|Z\}]| \quad (22)\\
&+ |\mathrm{E}[h_{1,2}(X) - \mathrm{E}\{h_{1,2}(X)|Z\}][h_{2,1}(Y) - \mathrm{E}\{h_{2,1}(Y)|Z\} - h_{2,2}(Y) + \mathrm{E}\{h_{2,2}(Y)|Z\}]|\,. \quad (23)
\end{aligned}$$

Since the class of functions in $\mathbb{H}_2$ are upper bounded by $O(\sqrt{\log n})$ with probability tending to 1, the right-hand-side of (22) is bounded from above by

$$O(1)\mathrm{E}\,|h_{1,1}(X) - \mathrm{E}\{h_{1,1}(X)|Z\} - h_{1,2}(X) + \mathrm{E}\{h_{2,1}(X)|Z\}|\,\sqrt{\log n},$$

with probability tending to 1. By Jensen's inequality, the above quantity can be further bounded from above by

$$O(1)\mathrm{E}\,|h_{1,1}(X) - h_{1,2}(X)|\,2\sqrt{\log n} \leq K\|\theta_{1,1} - \theta_{1,2}\|_2\sqrt{\log n},$$

for some constant $K > 0$. Following similar arguments, we can show that the right-hand-side of (23) is bounded from above by $K\|\theta_{2,1} - \theta_{2,2}\|_2\sqrt{\log n}$, for any $\theta_{2,1}$ and $\theta_{2,2}$, with probability tending to 1. To summarize, conditional on the event that $\mathcal{H}_1$ and $\mathcal{H}_2$ are bounded function classes, we have shown that

$$|I(h_{1,\theta_1}, h_{2,\theta_2}) - I(h_{1,\theta_1}, h_{2,\theta_2})| \leq K\left(\|\theta_{1,1} - \theta_{1,2}\|_2 + \|\theta_{2,1} - \theta_{2,2}\|_2\right)\sqrt{\log n}.$$

Consequently, for any sufficiently small $\epsilon > 0$, there exists a neighborhood $\mathcal{N} = \{(\theta_1, \theta_2) : \|\theta_j - \theta_j^*\|_2 \leq \delta \log^{-1/2} n\}$ for some constant $\delta > 0$ around (θ_1^*, θ_2^*), such that $I(h_{1,\theta_1}, h_{2,\theta_2}) \geq \epsilon$ for any (θ_1, θ_2) that belongs to this neighborhood.

Since $(\theta_{1,b}, \theta_{2,b})$ are generated from the multivariate normal distribution, and the dimensions d_1 and d_2 are finite, the probability that $(\theta_{1,b}, \theta_{2,b})$ belongs to this neighborhood is strictly greater than $O(\log^{-c_1} n)$ for some constant $c_1 > 0$. Since $B = c_0 n^c$, the probability that at least one pair of parameters $(\theta_{1,b_1}, \theta_{2,b_2})$ belongs to this neighborhood approaches one. Consequently, we have that,

$$\max_{b_1,b_2} \mathrm{GCM}^*\left\{h_{1,b_1}(X), h_{2,b_2}(Y)\right\} \geq \epsilon,$$

with probability tending to 1.

Following similar arguments as in the proof of Theorems 3 and 4, we can show that $|T - \max_{b_1,b_2} \mathrm{GCM}^*\{h_{1,b_1}(X), h_{2,b_2}(Y)\}| = o_p(1)$, and $\widetilde{T}_j = o_p(1)$. Consequently, both probabilities $\Pr(T < \epsilon/2)$ and $\Pr(\widetilde{T}_j \geq \epsilon/2)$ converge to zero. Therefore, the probability that the p-value is greater than α is bounded by the probability that $\Pr(T < \epsilon/2)$, which converges to zero. This completes the proof of Theorem 5.

References

Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012.

Alexis Bellot and Mihaela van der Schaar. Conditional independence testing using generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 2199–2208, 2019.

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

Thomas B Berrett, Yi Wang, Rina Foygel Barber, and Richard J Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, accepted, 2019.

Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:model-xknockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

Minshuo Chen, Wenjing Liao, Hongyuan Zha, and Tuo Zhao. Statistical guarantees of generative adversarial networks for distribution estimation. *arXiv preprint arXiv:2002.03938*, 2020.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Detailed proof of nazarov's inequality. *arXiv preprint arXiv:1711.10696*, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

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

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians. *arXiv preprint arXiv:1810.08693*, 2018.

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

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.

Mathew Garnett, Elena Edelman, Sonja Gill, Chris Greenman, Anahita Dastur, King Lau, Patricia Greninger, Richard Thompson, Xi Luo, Jorge Soares, Qingsong Liu, Francesco Iorio, Didier Surdez, Li Chen, Randy Milano, Graham Bignell, Ah Tam, Helen Davies, Jesse Stevenson, and Cyril Benes. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483:570–5, 03 2012. doi: 10.1038/nature11005.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. *arXiv preprint arXiv:1706.00292*, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *arXiv preprint arXiv:2001.06937*, 2020.

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.

Masaaki Imaizumi and Kenji Fukumizu. Deep neural networks learn non-smooth functions effectively. In *The 22nd international conference on artificial intelligence and statistics*, pages 869–878. PMLR, 2019.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Adaptive computation and machine learning. MIT Press, 2009.

Maria Larrosa-Garcia and Maria R. Baer. Flt3 inhibitors in acute myeloid leukemia: Current status and future directions. *Molecular Cancer Therapeutics*, 16(6):991–1001, 2017.

Bing Li. *Sufficient Dimension Reduction: Methods and Applications with R*. CRC Press, 2018.

Chun Li and Xiaodan Fan. On nonparametric conditional independence tests for continuous variables. *Wiley Interdisciplinary Reviews: Computational Statistics*, page e1489, 2019.

Tengyuan Liang. On how well generative adversarial networks learn densities: Nonparametric and parametric results. *arXiv preprint arXiv:1811.03179*, 2018.

Rajarshi Mukherjee, Whitney K Newey, and James M Robins. Semiparametric efficient empirical higher order influence function estimators. *arXiv preprint arXiv:1705.07577*, 2017.

Wenliang Pan, Xueqin Wang, Canhong Wen, Martin Styner, and Hongtu Zhu. Conditional local distance correlation for manifold-valued data. In *International Conference on Information Processing in Medical Imaging*, pages 41–52. Springer, 2017.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press, 2nd Edition, 2009.

James Robins, Lingling Li, Eric Tchetgen, Aad van der Vaart, et al. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics, 2008.

James M Robins, Lingling Li, Rajarshi Mukherjee, Eric Tchetgen Tchetgen, Aad van der Vaart, et al. Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics*, 45(5):1951–1987, 2017.

J Romano and Cyrus DiCiccio. Multiple data splitting for testing. Technical report, Stanford University, 2019.

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.

Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *arXiv preprint arXiv:1804.07203*, 2018.

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

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

Wesley Tansey, Victor Veitch, Haoran Zhang, Raul Rabadan, and David M Blei. The hold-out randomization test: Principled and easy black box feature selection. *arXiv preprint arXiv:1811.00645*, 2018.

James Tsai, John T. Lee, Weiru Wang, et al. Discovery of a selective inhibitor of oncogenic b-raf kinase with potent antimelanoma activity. *Proceedings of the National Academy of Sciences*, 105(8):3041–3046, 2008. doi: 10.1073/pnas.0711741105.

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

Xueqin Wang, Wenliang Pan, Wenhao Hu, Yuan Tian, and Heping Zhang. Conditional distance correlation. *Journal of the American Statistical Association*, 110(512):1726–1734, 2015.

K Zhang, J Peters, D Janzing, and B Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 804–813. AUAI Press, 2011.