Artificial Intelligence 347 (2025) 104391

Contents lists available at ScienceDirect

Artificial Intelligence

journal homepage: www.elsevier.com/locate/artint

Regression-based conditional independence test with adaptive kernels

Yixin Ren [a,1], Juncai Zhang [b,1], Yewei Xia [a], Ruxin Wang [b], Feng Xie [c], Jihong Guan [d], Hao Zhang [b,*], Shuigeng Zhou [a,*]

[a] Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, 2005 Songhu Road, Shanghai, 200438, China
[b] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, 1068 Xueyuan Avenue, Shenzhen, 518000, China
[c] Department of Applied Statistics, Beijing Technology and Business University, 33 Fucheng Road, Beijing, 102488, China
[d] Department of Computer Science and Technology, Tongji University, 4800 Cao'an Highway, Shanghai, 201804, China

ARTICLE INFO

Keywords:
Conditional independence test
Independence test
Adaptive kernel

ABSTRACT

We propose a novel framework for regression-based conditional independence (CI) test with adaptive kernels, where the task of CI test is reduced to regression and statistical independence test while proving that the test power of CI can be maximized by adaptively learning parameterized kernels of the independence test if the consistency of regression can be guaranteed. For the adaptively learning kernel of independence test, we first address the pitfall inherent in the existing signal-to-noise ratio criterion by modeling the change of the null distribution during the learning process, then design a new class of kernels that can adaptively focus on the significant dimensions of variables to judge independence, which makes the tests more flexible than using simple kernels that are adaptive only in length-scale, and especially suitable for high-dimensional complex data. Theoretically, we demonstrate the consistency of the proposed tests, and show that the non-convex objective function used for learning fits the L-smoothing condition, thus benefiting the optimization. Experimental results on both synthetic and real data show the superiority of our method. The source code and datasets are available at https://github.com/hzsiat/AdaRCIT.

1. Introduction

The conditional independence (CI) test is a fundamental problem that has wide applications throughout statistics and plays a central role in the fields of artificial intelligence, machine learning and related areas. Most of the early works of CI tests were built with the assumption of joint Gaussianity, under which CI can be reduced to zero partial correlation. Since partial correlation is easier to estimate, the Gaussian assumption gives a shortcut to CI test, but if the ground truth distribution is non-Gaussian, it can lead to misleading conclusions. For an example in the task of causal discovery, an edge might be incorrectly deleted due to a zero partial correlation between two target variables but in fact they are conditionally dependent.

* Corresponding authors.
E-mail addresses: h.zhang10@siat.ac.cn (H. Zhang), sgzhou@fudan.edu.cn (S. Zhou).
[1] Equal contribution.

https://doi.org/10.1016/j.artint.2025.104391
Received 24 December 2024; Received in revised form 1 April 2025; Accepted 23 June 2025

Available online 1 July 2025

In the last decade, a series of CI tests applicable to non-Gaussian assumption have been proposed, most of these tests are based on kernel trick or (deep) neural network. Neural network-based CI tests are usually better able to cope with more complex scenarios, but suffer from higher complexity, easy overfitting with small samples, and difficult tuning of hyperparameters. Kernel-based CI tests are more efficient and stable in low-sample cases, but it also suffers from the problem that the kernel parameters are not well optimized. Even in the simpler scenario of CI tests where the controlling set is empty, i.e. independence test, the existing kernel-based methods either presuppose the kernel function [12] (e.g. the Gaussian kernel with median bandwidth) or use a randomized feature mapping [56], thus have limited flexibility and cannot capture the differences in the distributions of complex structures.

In this work, we consider CI testing from a causal perspective. Following [8,51,52], given three sets of random variables $(X, Y, Z) \in \mathbb{R}^{d_X+d_Y+d_Z}$ (generally $d_X \geq 1, d_Y \geq 1, d_Z \geq 1$), the CI of X and Y given Z can be relaxed to a set of regressions and independence tests, i.e., $X \perp\!\!\!\perp Y | Z \Longleftrightarrow X - \mathbb{E}[X|Z] \perp\!\!\!\perp Y - \mathbb{E}[Y|Z]$ under causal faithfulness assumption, the Markov condition and the additive noise model assumption [8]. This motivates us to relax the task of adaptive kernel learning of CI tests to a much simpler form corresponding to independence tests, if we can ensure that the regression errors do not have too much effect on the adaptive kernel learning of independence tests (with some mild assumptions). Suppose that this separation can be achieved, we still have to separately face the tasks of regression and independence test. As there are many well-established strategies for optimizing regression of kernel/non-kernel, we mainly focus on the adaptive kernel learning of independence tests.

Kernel-based independence tests stand for a class of non-parametric tests that are widely used. Their criteria are mainly derived from the cross-covariance operators in the reproducing kernel Hilbert space (RKHS). The kernel canonical correlation (KCC) [2] and the constrained covariance (COCO) [13] are the pioneers. KCC uses the maximal correlation between the feature maps to measure dependency and COCO drops the normalization. As one of the most popular kernel-based dependence measures, the Hilbert-Schmidt Independence Criterion (HSIC) [12] uses the squared Hilbert-Schmidt norm to detect dependence. The HSIC-based independence tests [12,56] outperform the other kernel-based measures in performance and can handle different data scenarios through appropriate kernels.

However, these kernel-based methods either presuppose the kernel function [12] (e.g. the Gaussian kernel with median bandwidth) or use a randomized feature mapping [56], thus have limited flexibility and cannot capture the differences in the distributions of complex structures. To solve this, some works [16,24] tried to learn the kernels/features to maximize the power of hypothesis tests. [17] proposed a method to obtain features by optimizing the lower bound of testing power. Nevertheless, this method requires a large number of samples to ensure effectiveness, as well as a pre-set test location parameter, which is not easy to apply in new scenarios. Besides, [24,25] learned the kernels using a criterion called the signal-to-noise ratio [21] as the optimization objective in the two-sample test problem [13]. However, our study shows that this criterion may result in wrong solutions when learning kernels for independence tests since the change of the null distribution is ignored. In this work, we solve this problem by proposing a novel framework that models the null distribution change during the learning process. The proposed framework enables the design of flexible kernels for specific scenarios to make the tests more powerful.

Contributions. In summary, our contributions are as follows:

- We propose a novel framework for regression-based CI test (RCIT) with adaptive kernels, where the task of CI test is reduced to regression and statistical independence test while proving that the test power of CI can be maximized by adaptively learning parameterized kernels of the independence test if the consistency of regression can be guaranteed.
- We propose a novel framework for kernel-based statistical independence tests that enable adaptively learning parameterized kernels to maximize test power. Our framework overcomes the pitfall of the learning criterion in existing work by modeling the change of null distribution during the learning process.
- We further design a new class of kernels that can adaptively focus on the significant dimensions of variables for judging independence, which makes the tests more flexible than using simple kernels that are adaptive only in length-scale, and especially suitable for high-dimensional complex data.
- We theoretically demonstrate the consistency of our method and show that the non-convex objective function fits the L-smoothing condition, thus benefiting the optimization.

Outline. The rest of the paper is organized as follows: Sec. 2 reviews the related work on CI tests. Sec. 3 introduces the preliminaries. Sec. 4 first describes the pitfall of the learning criterion with existing works, and then introduces a novel framework to solve it. Furthermore, a class of importance-weighted kernels is designed. Additionally, theoretical analyses for independence tests are given. Sec. 5 shows how to extend the results of learning kernels for independence tests to CI tests, and the asymptotic validity as well as consistency of the test is demonstrated. Sec. 6 is performance evaluation. We conclude the paper in Sec. 7. All theoretical proofs are given in *Appendix*.

2. Related work

Kernel functions can represent high-order moments by evaluating the similarity of high-dimensional implicit functions and mapping variables into reproducing kernel Hilbert spaces (RKHSs), which allows us to infer properties of random variables such as independence and CI [11] from the distributions. As a result, a series of kernel-based CI tests were presented, in which RCIT is the representative CI test of high-performance under causal additive noise model assumption [8,51,52] and is thus widely used in causal discovery.

Generally, RCIT can be separated into two stages, regression and independence test. An indispensable assumption for RCIT is that the common information between X and Y caused by Z can be removed from either X or Y by regression. As we know, this assumption is not always true but RCIT works well in general continuous cases. In particular, RCIT generally perform better than other CI tests when the assumption holds.

[14] transforms the CI relationship of $X \perp\!\!\!\perp Y|Z$ into independence relationship between $X - \psi(Z)$ and (Y, Z). [52] tests $X \perp\!\!\!\perp Y|Z$ through $X - \psi(Z) \perp\!\!\!\perp (Y - \phi(Z), Z)$. In both methods, ψ (or ϕ) is obtained by regressing X (or Y) on Z, then CI test can be reduced to a set of regression and independence tests. In practice, $X - \psi(Z) \perp\!\!\!\perp Z$ is a strong condition, as $X - \mathbb{E}(X|Z) \perp\!\!\!\perp Z \Rightarrow Z \to X$ in many causal discovery scenarios where Z forms the minimal d-separator [54]. [8] showed that given structural faithfulness and Markov assumptions [29], if Z causes X or Y, then $X \perp\!\!\!\perp Y|Z \Longleftrightarrow X - \mathbb{E}(X|Z) \perp\!\!\!\perp Y - \mathbb{E}(Y|Z)$. Similarly, here a strong condition that Z causes X or Y is assumed, hence it is easy to derive the corresponding causal relations.

Moreover, faithfulness condition means that $X \perp\!\!\!\perp Y|Z \Longrightarrow X \perp\!\!\!\perp_d Y|Z$ ($\perp\!\!\!\perp_d$ denotes d-separation), and Markov condition implies that $X \perp\!\!\!\perp_d Y|Z \Longrightarrow X \perp\!\!\!\perp Y|Z$, so CI is relaxed to d-separation given the faithfulness and Markov assumptions. However, CI is neither sufficient nor necessary for d-separation. In practice, given the faithfulness assumption, $X - \mathbb{E}(X|Z) \perp\!\!\!\perp Y - \mathbb{E}(Y|Z)$ and $X \perp\!\!\!\perp Y|Z$ have significant correlations. For example, in [32], the authors suggested to use $X - \mathbb{E}(X|Z) \perp\!\!\!\perp Y - \mathbb{E}(Y|Z)$ to test $X \perp\!\!\!\perp Y|Z$ under the faithfulness assumption. In [52], the authors further conjectured that $X - \psi(Z) \perp\!\!\!\perp Y - \phi(Z)$ can lead to $X \perp\!\!\!\perp Y|Z$ under non-linear and faithfulness conditions, where ψ and ϕ are non-linear functions, X, Y and Z are generated by non-linear additive noise model. [50] showed that $X - \mathbb{E}(X|Z) \perp\!\!\!\perp Y - \mathbb{E}(Y|Z)$ is sufficient to support $X \perp\!\!\!\perp Y|Z$ if the data is generated by following the linear non-Gaussian structural equation model (SEM) under the faithfulness assumption.

Notice that in the above cases, $cov(X - \mathbb{E}(X|Z), Y - \mathbb{E}(Y|Z)) = 0$ often holds. Therefore, it is difficult to detect the common component shared by $X - \mathbb{E}(X|Z)$ and $Y - \mathbb{E}(Y|Z)$. To get best performance, most of these methods use Kernel-based independence test to achieve this goal.

Kernel-based independence test aims to compare the embedding difference of distributions between the joint distribution and the product of marginals in the RKHS. HSIC [12] is recognized as one of the most powerful test among them. In addition, some variants [56] utilize kernel approximation algorithms such as random Fourier features to further improve the efficiency of HSIC, which may lose power if the random mappings are insufficient. A closely related independence test method is based on distance covariance [41,42] which utilizes characteristic functions to measure and test dependence. In fact, distance-based methods are equivalent to the HSIC based methods with specific kernels [36]. However, these methods require predefined kernel functions or distance functions and thus lack the flexibility to handle complex situations. To solve this issue, our proposed scheme attempts to learn parameterized kernels adaptively in a data-driven way to make the test more powerful.

Learning kernels to maximize the power of the test has also been extensively studied in different applications (e.g. two-sample tests [40], independence tests [1], and goodness-of-fit tests [35]), and many methods have been proposed. Depending on the way of kernel learning, we can categorize them into two main directions. The first is to learn the parameters of the (single) kernels, which assumes a fixed form of the kernel and then optimizing the parameters. Optimizing the scale of Gaussian kernels [22] and learning deep kernels [25] are representative examples of this direction. The second is called kernel selection, which selects one or combines several from a set of predefined kernels (e.g. a set of kernels with different bandwidths). The representative methods include aggregated kernel tests [1]. Our scheme can be implemented in both directions. In this paper, we focus on the first direction (i.e. optimizing the Gaussian kernel bandwidth and learning the importance-weighted kernel).

3. Preliminaries

3.1. Conditional independence test

Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ be separable metric spaces, typically $\mathbb{R}^{d_x}, \mathbb{R}^{d_y}, \mathbb{R}^{d_z}$, $\mathbb{P}_{XYZ}$ be Borel probability measure defined on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, $\mathbb{P}_X, \mathbb{P}_Y$ and $\mathbb{P}_Z$ be the respective marginal distributions on $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$. Given n independent and identically distributed (i.i.d.) samples $\mathcal{D} = \{(x_i, y_i, z_i)\}_{i=1}^n \sim \mathbb{P}_{XYZ}$, our goal to test whether $\mathbb{P}_{XYZ}$ can be factorized into $\mathbb{P}_{X|Z}\mathbb{P}_{Y|Z}$: does $\mathbb{P}_{XYZ} = \mathbb{P}_{X|Z}\mathbb{P}_{Y|Z}$ (i.e. $X \perp\!\!\!\perp Y|Z$)? The hypothesis testing framework is as follows,

$$\mathcal{H}_0 : \mathbb{P}_{XYZ} = \mathbb{P}_{X|Z}\mathbb{P}_{Y|Z} \quad \text{versus} \quad \mathcal{H}_1 : \mathbb{P}_{XYZ} \neq \mathbb{P}_{X|Z}\mathbb{P}_{Y|Z}. \tag{1}$$

Hypothesis testing for CI is performed in the following steps. First, state the statistic T and calculate its observed value with the samples. Then, select a significance level α (typically taken as 0.05). After that, obtain the p-value, which is the probability that the sampling of T under $\mathcal{H}_0$ is at least as extreme as the observed value. Finally, the null hypothesis $\mathcal{H}_0$ is rejected if the p-value $\leq \alpha$. There are two types of errors that can occur during hypothesis testing, one is termed Type I error meaning the false rejection of $\mathcal{H}_0$, and the second is termed Type II error where $\mathcal{H}_0$ is incorrect but not rejected. A good CI test requires that Type I error rate is upper bounded by α simultaneously, while Type II error is minimized [55]. In particular, we call the CI test an independence test if the controlling set $Z = \emptyset$.

In the scenario of RCIT, the CI test can be relaxed to a simpler form under some assumptions,

$$X \perp\!\!\!\perp Y|Z \overset{assum.}{\Longleftrightarrow} X - \mathbb{E}(X|Z) \perp\!\!\!\perp Y - \mathbb{E}(Y|Z). \tag{2}$$

Generally, the required assumptions are causal faithfulness and additive noise [51]. It can be seen that CI test is separated into regression and independence test, as there are many well-established approaches for regression, such as least squares regression [46],

kernel ridge regression [8], lasso regression [44], and Gaussian process regression [8], we can therefore pay more attention to independence testing.

3.2. Hilbert-Schmidt independence criterion

In this work, we consider the independence test by the Hilbert-Schmidt Independence Criterion (HSIC).

Definition 1. [12]. Let $\mathcal{F}$ be a RKHS with kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, and let $\mathcal{G}$ be another RKHS on $\mathcal{Y}$ with kernel $l : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$. The HSIC between X and Y is denoted by

$$\mathrm{HSIC}(X,Y) := \mathbb{E}_{XX'YY'}\left[k(X,X')l(Y,Y')\right] + \mathbb{E}_{XX'}\left[k(X,X')\right]\mathbb{E}_{YY'}\left[l(Y,Y')\right] - 2\mathbb{E}_{X'Y'}\left[\mathbb{E}_X k(X,X')\mathbb{E}_Y l(Y,Y')\right], \tag{3}$$

where (X', Y') is a independent copy of (X, Y).

For characteristic kernels [10], $X \perp\!\!\!\perp Y \iff \mathrm{HSIC}(X,Y) = 0$. Given n i.i.d. samples $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n \sim \mathbb{P}_{XY}$, an observation of $\mathrm{HSIC}(X,Y)$, denoted as $\mathrm{HSIC}_b(\mathcal{S})$, is given by

$$\mathrm{HSIC}_b(\mathcal{S}) := \frac{1}{n^2}\sum_{i,j} k_{ij}l_{ij} + \frac{1}{n^4}\sum_{i,j,q,r} k_{ij}l_{qr} - 2\frac{1}{n^3}\sum_{i,j,q} k_{ij}l_{iq}, \tag{4}$$

where $k_{ij} := k(x_i, x_j)$, and $l_{ij} := l(y_i, y_j)$. This estimate can also be easily expressed by $\frac{1}{n^2}\mathrm{Tr}(\mathbf{KHLH})$, where $\mathbf{K}$ is the $n \times n$ matrix with entries k_{ij}, $\mathbf{L}$ is the $n \times n$ matrix with entries l_{ij}, $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is the centering matrix with $\mathbf{1}$ the $n \times 1$ vector of ones.

3.3. Asymptotics of HSIC

The asymptotic distribution of the statistic under the null hypothesis as well as the alternative hypothesis can be established by the following proposition [12, Theorem 1, 2].

Proposition 1. (Asymptotics of $\mathrm{HSIC}_b(\mathcal{S})$). *Let $h_{ijqr} = \frac{1}{4!}\sum_{(t,u,v,w)}^{(i,j,q,r)} k_{tu}l_{tu} + k_{tu}l_{vw} - 2k_{uv}l_{tv}$, where the sum represents all ordered quadruples (t,u,v,w) drawn without replacement from (i,j,q,r) and assume that kernels k, l are bounded. Then, under the null hypothesis $\mathcal{H}_0$, $HSIC_b(\mathcal{S})$ coverages in distribution as*

$$nHSIC_b(\mathcal{S}) \xrightarrow{d} \sum_{l=1}^{\infty} \lambda_l z_l^2, \tag{5}$$

where $z_l \sim \mathcal{N}(0,1)$ i.i.d and λ_l is the solution to the eigenvalue problem $\lambda_l\psi_l(z_l) = \int h_{ijqr}\psi_l(z_l)dF_{i,q,r}$, where the integral is over the distribution of variables z_i, z_q, z_r. And under the alternative hypothesis $\mathcal{H}_1$, $HSIC_b(\mathcal{S})$ converges in distribution to a Gaussian variable

$$n^{\frac{1}{2}}\Big(HSIC_b(\mathcal{S}) - HSIC(X,Y)\Big) \xrightarrow{d} \mathcal{N}(0, \sigma_u^2), \tag{6}$$

where the variance is given by

$$\sigma_u^2 = 16\big(\mathbb{E}_i(\mathbb{E}_{j,q,r}h_{ijqr})^2 - HSIC(X,Y)^2\big), \tag{7}$$

with the simplified notation $\mathbb{E}_{j,q,r} := \mathbb{E}_{z_j,z_q,z_r}$.

4. Learning kernels for independence test

In this section, we present a novel approach to kernel learning for independence testing. We begin by identifying the limitations of current learning criteria and propose an improved scheme to address these shortcomings. This leads to a comprehensive kernel learning framework, for which we further theoretically analyze its properties.

4.1. The pitfall of the signal-to-noise ratio criterion

The power of the test is equal to 1 − Type II error rate, which measures the efficacy of the hypothesis test.
According to Proposition 1, the power of the test with HSIC can be formulated by

$$\mathbb{P}_{\mathcal{H}_1}\left(n\mathrm{HSIC}_b(\mathcal{S}) > r\right) \to \Phi\left(\frac{n\mathrm{HSIC}(X,Y) - r}{\sqrt{n}\sigma_u}\right), \tag{8}$$

where Φ is the standard normal CDF and r is the threshold, i.e., the $(1-\alpha)$-quantile of distribution of Eq. (42) that controls Type I error rate to be $< \alpha$. To maximize the test power, the term without the threshold corresponds to $\frac{n\mathrm{HSIC}(X,Y)}{\sigma_u}$, is the popular choice [24,25] in kernel learning for the two-sample test problem [11], called signal-to-noise ratio [21]. Note that the criterion is theoretically biased

as long as r in Eq. (8) is not 0. Also, in practical applications of independence tests, learning the kernels with this criterion may lead to undesired and even catastrophic solutions. To explain this issue, we give an example as follows:

Example. We consider the case that k, l are Gaussian kernels, i.e., $k(x,x') := \exp\left(-\frac{\|x-x'\|^2}{2\omega_x^2}\right), l(y,y') := \exp\left(-\frac{\|y-y'\|^2}{2\omega_y^2}\right)$, where ω_x, ω_y are the width parameters. The estimate of signal-to-noise is taken as

$$J_{w/o} := \frac{n\mathrm{HSIC}_b(S)}{\widehat{\sigma}_u(S)}, \tag{9}$$

where $\widehat{\sigma}_u^2(S)$ is a estimator of σ_u^2 with S, given by

$$16 \cdot \left(\frac{1}{n}\sum_i \left(\frac{1}{n^3}\sum_{j,q,r} h_{ijqr}\right)^2 - \left(\mathrm{HSIC}_b(S)\right)^2\right). \tag{10}$$

For fixed samples S of sample size n and fixed width $\omega_y > 0$, we explore the behavior of the criterion $J_{w/o}$ when ω_x is close to zero. Assume that $\|x_i - x_j\|^2 \neq 0$ for all $i \neq j$, then we have the following results:

$$[\mathbf{K}]\Big|_{\omega_x=0^+} = \mathbf{I}_n,\ [n\mathrm{HSIC}_b(S)]\Big|_{\omega_x=0^+} = \frac{1}{n}\mathrm{Tr}(\mathbf{L_c}),$$
$$[\widehat{\sigma}_u^2(S)]\Big|_{\omega_x=0^+} = \frac{4}{n^2}\left[\frac{\mathrm{Tr}\left[(\mathbf{L_c})^{\cdot 2}\right]}{n} - \left(\frac{\mathrm{Tr}(\mathbf{L_c})}{n}\right)^2\right],$$

where $\mathbf{L_c} := \mathbf{HLH}$ and $()^{\cdot 2}$ is the entrywise matrix power. As a result,

$$J_{w/o}\Big|_{\omega_x=0^+} = \frac{n}{2}\cdot\frac{\mathrm{Tr}(\mathbf{L_c})/n}{\sqrt{\mathrm{Tr}\left[(\mathbf{L_c})^{\cdot 2}\right]/n - \left[\mathrm{Tr}(\mathbf{L_c})/n\right]^2}}.$$

As a comparison, we study a criterion with threshold estimation. We obtain an estimate of r by the moments of the distribution under $\mathcal{H}_0$. The moments are given as follows [12]:

Proposition 2. (Moments of Null Distribution). *Under $\mathcal{H}_0$, the estimation of mean with bias of $\mathcal{O}(n^{-1})$ to $\mathbb{E}[n\mathit{HSIC}_b(S)]$ can be given by*

$$\mathcal{E}_0 := 1 + \widehat{\|\mu_x\|^2}\widehat{\|\mu_y\|^2} - \widehat{\|\mu_y\|^2} - \widehat{\|\mu_x\|^2}, \tag{11}$$

where we assume $k_{ii} = l_{ii} = 1$ and the terms $\widehat{\|\mu_x\|^2} = \frac{1}{(n)_2}\sum_{(i,j)\in i_2^n} k_{ij}$, $\widehat{\|\mu_y\|^2} = \frac{1}{(n)_2}\sum_{(i,j)\in i_2^n} l_{ij}$. Also, the estimation of variance with bias of $\mathcal{O}(n^{-1})$ to $\mathbf{Var}[n\mathit{HSIC}_b(S)]$ can be given by

$$\mathcal{V}_0 = \frac{2(n-4)(n-5)}{(n-1)^2(n-2)(n-3)}\mathbf{1}^T\left(\mathbf{B} - \mathit{diag}(\mathbf{B})\right)\mathbf{1}, \tag{12}$$

where $\mathbf{B} = \left((\mathbf{HKH}) \odot (\mathbf{HLH})\right)^{\cdot 2} = (\mathbf{K}_c \odot \mathbf{L}_c)^{\cdot 2}$ and $\odot$ is the entrywise matrix product.

The limit of these two moments can be calculated by

$$\mathcal{E}_0\Big|_{\omega_x=0^+} = \frac{1}{n-1}\mathrm{Tr}\mathbf{L_c},$$
$$\mathcal{V}_0\Big|_{\omega_x=0^+} = \frac{2(n-4)(n-5)}{n^2(n-1)^2(n-2)(n-3)}\sum_{i\neq j}(\mathbf{L_c})_{ij}^2. \tag{13}$$

Since the variance is $\mathcal{O}(n^{-2})$ as in Eq. (13), according to Chebyshev's inequality, the distribution is concentrated around $\mathcal{E}_0$. Hence, we can use $\mathcal{E}_0$ as an estimator of r when $\omega_x = 0^+$. Let the criterion with $\mathcal{E}_0$ as

$$J_{w/,\mathcal{E}_0} := \frac{n\mathrm{HSIC}_b(S) - \mathcal{E}_0}{\widehat{\sigma}_u(S)}, \tag{14}$$

such that $J_{w/,\mathcal{E}_0}\Big|_{\omega_x=0^+} = -\frac{1}{n-1}J_{w/o}\Big|_{\omega_x=0^+}$.

In conclusion, we have shown that ignoring the threshold causes the criterion to differ by a factor of $-(n-1)$ from the true power estimate when $\omega_x = 0^+$. This will result in a very different behavior, as illustrated in Fig. 1. When ω_x is close to zero, $J_{w/o}$ takes a very large value (maximum in this case) compared to the value with the threshold. This can lead to the wrong maximum point ($\omega_x = 0$ in this case) of test power with $J_{w/o}$, resulting in a catastrophic wrong solution.

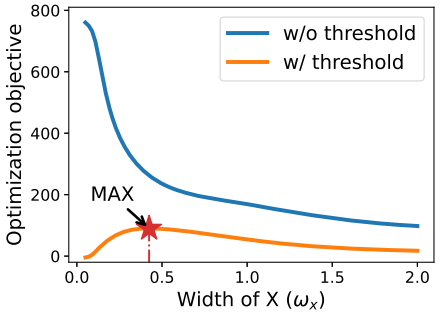

Fig. 1. The values of optimization objective for different ω_x on the ISA dataset under the setting $n = 250$, $d = 3$, $\theta = \pi/10$, $\omega_y = 1.0$. The "w/o threshold" line corresponds to Eq. (9) and the other to Eq. (17). (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Remark. Although our example is based on Gaussian kernels, the more general case also holds as long as there exists parameter k such that the kernel matrix K approaches I (such as the Laplace kernel with a width close to 0) and the fixed L is appropriate, the rest of the analysis is similar. If interested, the readers can refer to more examples as well as analysis given in *Appendix* L.1.

In the next section, we will resolve this pitfall by designing a differentiable criterion that takes into account the variation of the threshold under the null hypothesis during the optimization process.

4.2. Kernel learning framework

Using a permutation test to construct the estimator of the threshold is a possible way. That is, permuting sample Y repeatedly while that of X is kept fixed to directly simulate the null distribution. However, this process is expensive due to the significant number of permutations. Even if a parallel scheme can be adopted to improve the computational efficiency of this process, the required memory is heavily positively correlated with the number of permutations required. It is therefore not desirable in certain resource-constrained scenarios. Here, we consider a gamma approximation method [12], which instead requires only a single pass calculation. The idea is to use a two-parameter gamma distribution to approximate the infinite sum of χ^2 variables as in Eq. (42). The first two moments of Eqs. (11) and (12) are used to determine the two parameters, i.e.,

$$n\mathrm{HSIC}_b(S) \sim \frac{x^{\gamma-1}e^{-x/\beta}}{\beta^{\gamma}\Gamma(\gamma)}, \gamma = \frac{\mathcal{E}_0^2}{\mathcal{V}_0},\ \beta = \frac{\mathcal{V}_0}{\mathcal{E}_0}, \tag{15}$$

where $\Gamma(\cdot)$ is the gamma function. The estimate of the threshold, denoted as $\widehat{c_\alpha}$, can be given by the $(1-\alpha)$-quantile of this gamma distribution, i.e.,

$$\int_0^{\frac{\widehat{c_\alpha}}{\beta}} \frac{x^{\gamma-1}e^{-x}}{\Gamma(\gamma)} dx = 1-\alpha. \tag{16}$$

And the criterion can be obtained by [2]

$$J_{w/,\widehat{c_\alpha}} := \frac{n\mathrm{HSIC}_b(S) - \widehat{c_\alpha}}{\widehat{\sigma}_u(S)}. \tag{17}$$

Then, $J_{w/,\widehat{c_\alpha}}$ can be used to learn kernels (e.g. Gaussian kernels with learnable bandwidth) to maximize the testing power. We aim to optimize this objective function with any commonly used optimizer such as Adam [18]. However, the gradient of $J_{w/,\widehat{c_\alpha}}$ cannot be obtained explicitly because it is related to the parameters of the kernel through the implicit functions. Let the parameter spaces of kernels k, l be Ω_0, Ω_1, at the point $\omega_* \in \Omega_0 \times \Omega_1$, we get the gradient in two steps. First, we estimate the partial derivative $\partial_\beta \widehat{c_\alpha}$ and $\partial_\gamma \widehat{c_\alpha}$. The first term at ω_* can be directly calculated by $\partial_\beta \widehat{c_\alpha}\big|_{\omega_*} = \frac{\widehat{c_\alpha}}{\beta}\big|_{\omega_*}$ according to Eq. (16). For the second term, which cannot be easily calculated due to the presence of the Gamma function, we use the finite differences to estimate it numerically, i.e., calculating $\partial_\gamma \widehat{c_\alpha} = \lim_{\delta\to 0} \frac{\widehat{c_\alpha}(\gamma+\delta) - \widehat{c_\alpha}(\gamma)}{\delta}$. Then, we can get the gradient of

$$\frac{n\mathrm{HSIC}_b(S) - \widehat{c_\alpha}\big|_{\omega_*}}{\widehat{\sigma}_u(S)} - \frac{\left(\partial_\beta \widehat{c_\alpha}\big|_{\omega_*}\right)\cdot\beta + \left(\partial_\gamma \widehat{c_\alpha}\big|_{\omega_*}\right)\cdot\gamma}{\widehat{\sigma}_u(S)\big|_{\omega_*}} \tag{18}$$

by combining with an automatic differentiation framework such as PyTorch [28] to estimate the gradient $\partial_\omega J_{w/,\widehat{c_\alpha}}$ at the point $\omega = \omega_*$.

[2] In practice, we add a small constant to the denominator as suggested in [25].

Data split. Given samples $\mathcal{S}$, the above process allows the kernels to be learned end-to-end and then used for test. However, this would lead to uncontrollable Type I error [20]. Here we adapt the technique used in a variety of tests [17,25]: splitting the data into disjoint training and test data. The split ratio is heuristically set to 0.5 since how to set the optimal split ratio in practice remains an open problem.

Algorithm. Our algorithm is outlined in Algorithm 1. As a pre-processing step, we split the data to training data $\mathcal{S}^{tr}$ and test data $\mathcal{S}^{te}$ (Line 1). The test contains two phases: 1) We learn the kernels with Adam optimizer using full batches on $\mathcal{S}^{tr}$ (Lines 2-9). 2) With the learned kernels, we calculate the test statistic and threshold (Lines 12-13) to determine the independence (Lines 14) on $\mathcal{S}^{te}$. The overall time complexity is $\mathcal{O}\big(Tn^2(d_x + d_y)\big)$, where d_x and d_y are the dimensions of X and Y respectively.

Algorithm 1 The learning and testing framework.

Input: Samples S of X, Y, significance level α.
Output: $X \perp\!\!\!\perp Y$ or $X \not\perp\!\!\!\perp Y$.

1: Split the data as $S = S^{tr} \cup S^{te}$.
2: ◁ **Learning kernels on S^{tr}:**
3: Initialize parameters of kernels, set learning rate ϵ, and set iteration steps T.
4: **for** $t = 1, 2, ..., T$ **do**
5: $k_{\omega_0}, l_{\omega_1} \leftarrow$ Kernels with parameters ω_0, ω_1.
6: $J_{tr/.\hat{c}_\alpha} \leftarrow$ Calculate Eq. (17) with $k_{\omega_0}, l_{\omega_1}$.
7: $\nabla_{(\omega_0,\omega_1)} J_{tr/.\hat{c}_\alpha} \leftarrow$ Estimate using Eq. (18).
8: $(\omega_0, \omega_1) \leftarrow (\omega_0, \omega_1) + \epsilon \nabla_{(\omega_0,\omega_1)} J_{tr/.\hat{c}_\alpha}$.
9: **end for**
10: After training, use the learned kernels for testing.
11: ◁ **Testing on S^{te} with learned kernels:**
12: $n^{te}\mathrm{HSIC}_b(S^{te}) \leftarrow$ Estimate the statistic.
13: $\hat{c}_\alpha(S^{te}) \leftarrow$ Calculate the threshold on S^{te}.
14: **if** $\hat{c}_\alpha(S^{te}) \le n^{te}\mathrm{HSIC}_b(S^{te})$ **then**
15: **Return:** $X \not\perp\!\!\!\perp Y$.
16: **else**
17: **Return:** $X \perp\!\!\!\perp Y$.
18: **end if**

4.3. Importance-weighted kernels

The Gaussian kernel has only one parameter. It assigns equal weight to the distance measure on each dimension of the multivariate variable. Here, we consider a class of more general Gaussian kernels with the following form:

$$k(x, x') = \exp\big(-(x - x')\Sigma_x(x - x')\big), \tag{19}$$

where Σ_x is a positive definite matrix and $x \in \mathbb{R}^{d_x}$ of d_x dimensions. Since this kernel is translation invariant, i.e., $k(x, x') = k(x - t, x' - t)$ for any $t \in \mathbb{R}^{d_x}$, it can be shown to be characteristic [9,10]. This class of kernels models correlations between each two dimensions and hence is more generic. However, due to the positive definite constraints on the matrix Σ_x, it is not easy to maintain while learning the kernels. Here we consider the case that it is a diagonal positive definite matrix, i.e., assigning different positive weights to the distances on different dimensions. In this case, the kernels are referred to as the ARD kernels [47]. Here we rephrase this class of kernels as importance-weighted kernels to emphasize the role that enables higher weights on important dimensions to enhance the test power. Formally,

$$k(x, x') := \prod_{i=1}^{d_x} \exp\Big(-\frac{w_i(x_i - x_i')^2}{2\omega_x^2}\Big), w_i \in (0, 1), \tag{20}$$

where x_i is the i-th dimension of x, w_i is the importance weight of the i-th dimension, and ω_x is the overall bandwidth among all dimensions (we add it to keep the form of Gaussian kernel). In conjunction with the proposed framework, importance weights can be learned end-to-end. This is very crucial for high-dimensional complex data, as in most cases, each dimension is not equally important.

Interpretability. Larger weights indicate more important dimensions for the power of independence testing. This contributes to the interpretability of the results. An example is given in Sec. 6.1.4.

4.4. Theoretical analysis for independence test

In this section, we prove the uniform convergence bound of the kernel learning criterion, as well as its smoothness which guarantees the effectiveness of kernel learning. In addition, we give the consistency of independence test, and in the next section we further extend the consistency result to the CI test. We require the following assumptions for this part of proof:

(i) The kernels k_{ω_0} and l_{ω_1} are uniformly bounded:

$$\sup_{\omega_0\in\Omega_0}\sup_{x\in\mathcal{X}} k_{\omega_0}(x,x)\le \nu,\quad \sup_{\omega_1\in\Omega_1}\sup_{y\in\mathcal{Y}} l_{\omega_1}(y,y)\le \nu.$$

(ii) The kernel parameters ω_0,ω_1 lie in Banach spaces of dimension D_0 and D_1 respectively. Furthermore, the set of possible kernel parameters, Ω_0 and Ω_1, are separately bounded by R_{ω_0} and R_{ω_1}, respectively, i.e.,

$$\Omega_0 \subseteq \left\{\omega_0 \middle|\ \|\omega_0\|\le R_{\Omega_0}\right\}, \Omega_1 \subseteq \left\{\omega_1 \middle|\ \|\omega_1\|\le R_{\Omega_1}\right\}.$$

(iii) The kernel parameterizations are Lipschitz continuous, i.e. for all $x,x'\in\mathcal{X}$, $\omega_0,\omega_0'\in\Omega_0$,

$$\left|k_{\omega_0}(x,x')-k_{\omega_0'}(x,x')\right|\le L_k\cdot\|\omega_0-\omega_0'\|$$

and for all $y,y'\in\mathcal{Y}$, $\omega_1,\omega_1'\in\Omega_1$,

$$\left|l_{\omega_1}(y,y')-l_{\omega_1'}(y,y')\right|\le L_l\cdot\|\omega_1-\omega_1'\|$$

with the nonnegative Lipschitz constant L_k, L_l.

Remark. The assumptions (i), (ii), and (iii) do not restrict the specific form of the kernels, and the kernels used in our paper satisfy these properties.

We first give the uniform bound results for the kernel learning criterion as follows:

Theorem 1 *(Uniform Bound). Let ω_0,ω_1 parameterize uniformly bounded kernels $k_{\omega_0}, l_{\omega_1}$ in Banach spaces of dimension D_0, D_1. And $k_{\omega_0}, l_{\omega_1}$ satisfy the Lipschitz condition in ω_0,ω_1 with the Lipschitz constant L_k, L_l. Let $\overline{\Omega}_c := \overline{\Omega}_0\times\overline{\Omega}_1$ be a set of (ω_0,ω_1) for which $\sigma_u\ge c>0$ with small constant c and $\|\omega_0\|\le R_{\Omega_0}, \|\omega_1\|\le R_{\Omega_1}$. Let r denote the threshold, i.e., $(1-\alpha)$-quantile for the asymptotic distribution in Eq. (42) and $r^{(n)}$ be the threshold with kernels of size n. Under Assumptions (i) to (iii), then with probability at least $1-\delta$,*

$$\sup_{(\omega_0,\omega_1)\in\overline{\Omega}_c}\left|\frac{HSIC_b(S)-r^{(n)}/n}{\widehat{\sigma}_u(S)}-\frac{HSIC(X,Y)-r/n}{\sigma_u}\right|\sim\mathcal{O}\left(\frac{1}{c^3}\left[\sqrt{\frac{1}{n}\log\frac{1}{\delta}+(D_0+D_1)\frac{\log n}{n}}+\frac{L_k+L_l}{\sqrt{n}}\right]\right).$$

The proof procedure is given in *Appendix* D. The theorem extends the result in [25] since our criterion considers the threshold and removes the need for regular constants. This result shows that with sufficient samples, our criterion converges to the ground truth power criterion (for any kernel parameters), i.e., the error due to the estimation is reduced to 0. Thus, optimizing this criterion results in a generalizable (not just overfitting to the training set) solution. As a result, if the optimization process with our criterion is successful, we can obtain a solution that maximizes the test power. Next, we show the consistency of the tests, i.e., the test power tends to 1 as the sample size increases.

Proposition 3 *(Consistency of Independence Test). Let ω_0^*,ω_1^* be the kernel parameters after learning, S^{te} be the testing samples of size m, then the probability of Type II error*

$$\mathbb{P}_{\mathcal{H}_1}\left(mHSIC_b(S^{te})\le r^{(m)}|\omega_0^*,\omega_1^*\right)\sim\mathcal{O}(m^{-1/2}). \tag{21}$$

The result above focuses on the asymptotic behavior. The following result instead shows the property of the objective function under practical settings. Our proof focuses on the Gaussian kernel (but keep in mind that it holds for the Laplace kernel as well as the importance-weighted kernel). As a start, we need to attach some weak assumptions (which usually hold in practice, see *Appendix* E for a detailed discussion). The assumptions are

1. The domain $\mathcal{X}$ is Euclidean and bounded, $\mathcal{X}\subseteq\{x\in\mathbb{R}^d : \|x\|\le R_{\mathcal{X}}/2\}$ for constant $R_{\mathcal{X}}<\infty$.
2. The non-diagonal elements of center matrices $\mathbf{K}_c$, $\mathbf{L}_c$ are not zero, i.e. $(\mathbf{K}_c)_{ij}^2>0, (\mathbf{L}_c)_{ij}^2>0$ for all $i\ne j$ when the kernel widths $(\omega_x,\omega_y)\in[\omega_{xl},\omega_{xu}]\times[\omega_{yl},\omega_{yu}]$ with given positive constants ω_{xl}, ω_{xu}, ω_{yl}, ω_{yu}.
3. The distributions of data are continuous. Hence $\|x_i-x_j\|\ne 0, \|y_i-y_j\|\ne 0$ for all $i\ne j$.

Based on the above assumptions, we have the following theorem holds.

Theorem 2 *(Smoothness of Objective Function). Let $k_{\omega_x}, l_{\omega_y}$ be Gaussian kernels with bandwidth parameter ω_x,ω_y, for fixed samples S of size n, the objective function we used in practice*

$$J_\lambda(S) := \frac{nHSIC_b(S)-\widehat{c_\alpha}}{\sqrt{\widehat{\sigma}_u^2(S)+\lambda}},\ \lambda>0$$

satisfies the L-smoothing condition, i.e., its gradients of ω_x, ω_y are Lipschitz continuous on the compact domain $(\omega_x, \omega_y) \in [\omega_{xl}, \omega_{xu}] \times [\omega_{yl}, \omega_{yu}]$, for all positive constants ω_{xl}, ω_{xu}, ω_{yl}, ω_{yu}.

The L-smoothing condition benefits the optimization [57] in practice. The proof procedure is given in *Appendix* H.

5. Regression-based conditional independence test

In this section, we introduce our regression-based conditional independence test (RCIT) and give the theoretical guarantee of consistency. RCIT consists of two major stages: the regression stage and the residual independence testing stage. In what follows, we first present an overview of the CI testing procedure and summarize the key concepts, then we introduce the theoretical results.

5.1. Framework of RCIT

Here, we extend the results of kernel learning for independence testing to CI testing by simplifying CI test to RCIT. Two representative results corresponding to linear and the non-linear cases are given, respectively.

Theorem 3. *[52, Theorem 2] Define $m+2$ random variables X, Y and $Z = \{z_1, \cdots, z_m\}$ generated by following a l-dimensional linear structural equation model satisfying faithfulness condition, if all the external influences $s_i (i = 1, \cdots, l)$ are non-Gaussian, then $X \perp\!\!\!\perp Y | Z$ if and only if $X - \mathbb{E}(X|Z) \perp\!\!\!\perp Y - \mathbb{E}(Y|Z)$.*

Theorem 4. *[8, Theorem 4.1] Given structural assumptions of faithfulness and the Markov assumptions, and assuming that we have consistent regressors with an additive noise model, whenever Z is a cause of X or Y, it follows that $X \perp\!\!\!\perp Y | Z$ if and only if $X - \mathbb{E}(X|Z) \perp\!\!\!\perp Y - \mathbb{E}(Y|Z)$.*

Theorems 3 and 4 formulate the condition of reducing a CI test to an unconditional independence test. Throughout this work, we can test CI by first performing the regression to get two residuals and then performing the kernel learning procedure to test the independence of the two residuals. In what follows, we provide a more detailed explanation of the two-stage process in the overall CI test framework.

1. **Regression stage**: Let the dataset be $\{x_i, y_i, z_i\}_{i=1}^n \sim (X, Y, Z)$, where X, Y, Z are the corresponding random variables. The key step in this stage is to estimate the conditional expectation $\mathbb{E}(X|Z)$. A common approach is non-parametric least-squares regression, which provides a regularized estimate $\widehat{\mu}_{X|Z}(z) \approx \mathbb{E}[X|Z=z]$ using n samples. This serves as an approximation of the true conditional expectation $\mu_{X|Z}(z) = \mathbb{E}[X|Z=z]$. Under certain assumptions, we can prove that the estimation error is bounded (Theorem 5).
2. **Residual independence testing stage**: The residual independence testing stage consists of two steps: a training step and a testing step. We focus primarily on the testing step since the training step is used to learn the kernel parameters as introduced in Sec. 4. After training, the kernel parameters learned are used for the testing step. Assume there are m samples used in the testing stage: $\{x_i, y_i, z_i\}_{i=1}^m \sim (X, Y, Z)$, then we can calculate the residuals, which are the differences between the observed values and their estimated conditional means:

$$\widehat{R}_{X|Z}^{(m)} = \{\hat{r}_{x;i}\}_{i=1}^m,\ \hat{r}_{x;i} = x_i - \widehat{\mu}_{X|Z}(z_i);\quad \widehat{R}_{Y|Z}^{(m)} = \{\hat{r}_{y;i}\}_{i=1}^m,\ \hat{r}_{y;i} = y_i - \widehat{\mu}_{Y|Z}(z_i). \tag{22}$$

The corresponding random variables are defined as $\widehat{R}_{X|Z} := X - \widehat{\mu}_{X|Z}(Z)$ and $\widehat{R}_{Y|Z} := Y - \widehat{\mu}_{Y|Z}(Z)$. The test statistic follows the Hilbert-Schmidt Independence Criterion (HSIC), which measures the dependence between the residuals of X and Y given Z. Specifically, we compute the V-statistic:

$$\mathrm{HSIC}_b(\widehat{R}_{X|Z}^{(m)}, \widehat{R}_{Y|Z}^{(m)}) = \frac{1}{n^2}\sum_{i,j}^{n} \breve{k}_{ij}\breve{l}_{ij} + \frac{1}{n^4}\sum_{i,j,q,r}^{n} \breve{k}_{ij}\breve{l}_{qr} - 2\frac{1}{n^3}\sum_{i,j,q}^{n} \breve{k}_{ij}\breve{l}_{iq}, \tag{23}$$

where $\breve{k}_{ij} := k(\hat{r}_{x;i}, \hat{r}_{x;j})$ and $\breve{l}_{ij} := l(\hat{r}_{y;i}, \hat{r}_{y;j})$.

Before introducing the two-stage CI test framework, we define the final CI test statistic, denoted by $\mathrm{RCIT}_b(\mathcal{D})$, given i.i.d. samples $\mathcal{D} = \{x_i, y_i, z_i\}_{i=1}^m$ as follows:

$$\mathrm{RCIT}_b(\mathcal{D}) = \mathrm{HSIC}_b(\widehat{R}_{X|Z}^{(m)}, \widehat{R}_{Y|Z}^{(m)}). \tag{24}$$

Next, we analyze the properties of $\mathrm{RCIT}_b(\mathcal{D})$. Note that it contains two sources of bias: (1) regression error and (2) finite-sample estimation error. We analyze them separately below.

- **Effect of regression error.** First, consider the ideal case where the regression functions are perfectly estimated. The residuals are given as

$$R_{X|Z}^{(m)} = \{r_{x;i}\}_{i=1}^m,\ r_{x;i} = x_i - \mu_{X|Z}(z_i);\quad R_{Y|Z}^{(m)} = \{r_{y;i}\}_{i=1}^m,\ r_{y;i} = y_i - \mu_{Y|Z}(z_i). \tag{25}$$

In this case, the corresponding test statistic becomes

$$\mathrm{HSIC}_b(R^{(m)}_{X|Z}, R^{(m)}_{Y|Z}) = \frac{1}{n^2}\sum_{i,j}^{n} \tilde{k}_{ij}\tilde{l}_{ij} + \frac{1}{n^4}\sum_{i,j,q,r}^{n} \tilde{k}_{ij}\tilde{l}_{qr} - 2\frac{1}{n^3}\sum_{i,j,q}^{n} \tilde{k}_{ij}\tilde{l}_{iq}, \tag{26}$$

where we $\tilde{k}_{ij} := k(r_{x;i}, r_{x;j})$ and $\tilde{l}_{ij} := l(r_{y;i}, r_{y;j})$. Note that the difference between $\mathrm{HSIC}_b(\widehat{R}^{(m)}_{X|Z}, \widehat{R}^{(m)}_{Y|Z})$ and $\mathrm{HSIC}_b(R^{(m)}_{X|Z}, R^{(m)}_{Y|Z})$ is entirely due to regression estimation error. In Theorem 5, we establish a bound for this bias. Additionally, the corresponding U-statistic $\mathrm{HSIC}_u(R^{(m)}_{X|Z}, R^{(m)}_{Y|Z})$ is different from $\mathrm{HSIC}_b(R^{(m)}_{X|Z}, R^{(m)}_{Y|Z})$ by $\mathcal{O}(1/m)$, which can be shown in the same way as Lemma 2.

- **Finite-sample estimation error.** Next, we define the CI criterion as

$$\mathrm{HSIC}(R_{X|Z}, R_{Y|Z}) = \mathbb{E}\left[\mathrm{HSIC}_u(R^{(m)}_{X|Z}, R^{(m)}_{Y|Z})\right]. \tag{27}$$

This term characterizes CI according to Theorems 3 and 4:

$$\mathrm{HSIC}(R_{X|Z}, R_{Y|Z}) = 0 \Leftrightarrow X \perp\!\!\!\perp Y \,|\, Z. \tag{28}$$

Since $\mathrm{HSIC}_u(R^{(m)}_{X|Z}, R^{(m)}_{Y|Z})$ is an unbiased estimation, its value closely approximates $\mathrm{HSIC}(R_{X|Z}, R_{Y|Z})$ for sufficiently large m. Consequently, it serves as a reliable measure of CI. Furthermore, since we have analyzed the relationship between $\mathrm{HSIC}_b(\widehat{R}^{(m)}_{X|Z}, \widehat{R}^{(m)}_{Y|Z})$ and $\mathrm{HSIC}_u(R^{(m)}_{X|Z}, R^{(m)}_{Y|Z})$, the test with statistic $\mathrm{HSIC}_b(\widehat{R}^{(m)}_{X|Z}, \widehat{R}^{(m)}_{Y|Z})$ is valid and consistent if the errors from both stages are well-controlled.

In the next section, we will also derive a bound for the difference between $\mathrm{HSIC}_b(\widehat{R}^{(m)}_{X|Z}, \widehat{R}^{(m)}_{Y|Z})$ and $\mathrm{HSIC}_b(R^{(m)}_{X|Z}, R^{(m)}_{Y|Z})$. This allows us to provide a bound for the two types of error (Theorem 6) of RCIT.

5.2. Theoretical analysis for conditional independence test

In this section, we introduce the theoretical results of RCIT. To establish the theoretical properties of RCIT, we need to account for the errors introduced by both the regression stages (i.e., the estimation of the conditional mean) and the finite sample size in the testing stage.

5.2.1. Results for the regression stage

We first establish the bound for the conditional mean operator, given the following assumptions:

(iv) **Boundedness.** The residuals $R_{X|Z}$, $R_{Y|Z}$ and their estimates $\hat{R}_{X|Z}$, $\hat{R}_{Y|Z}$ are bounded:

$$\sup_{(x,z)\in(\mathcal{X},\mathcal{Z})} \left|R_{X|Z}\right| \le \mu, \quad \sup_{(y,z)\in(\mathcal{Y},\mathcal{Z})} \left|R_{Y|Z}\right| \le \mu, \quad \sup_{(x,z)\in(\mathcal{X},\mathcal{Z})} \left|\hat{R}_{X|Z}\right| \le \mu, \quad \sup_{(y,z)\in(\mathcal{Y},\mathcal{Z})} \left|\hat{R}_{Y|Z}\right| \le \mu.$$

(v) **Eigenvalue decay (EVD).** Consider that $(\lambda_i)_{i\in I}$ stands for the eigenvalues of the integral operator [7], which is a covariance-like operator derived from the kernel function used in the regression model. For some $C_\lambda > 0$ and $p \in (0,1]$ and for all $i \in I$, $\lambda_i \le C_\lambda i^{-1/p}$.

(vi) **Source condition (SRC).** There exists a smoothness parameter $1 < \ell \le 2$ such that

$$\mu_{X|Z} \in [\mathrm{HS}_{xz}]^{\ell}, \ \mu_{Y|Z} \in [\mathrm{HS}_{yz}]^{\ell}.$$

Above, $[\mathrm{HS}_{xz}]^{\ell}$ refers to the interpolation space of the original Hilbert-Schmidt space (also written as $[\mathrm{HS}]^1$). The eigenvalue decay of functions in $[\mathrm{HS}]^{\ell}$ is bounded by ℓ.

Remark. The assumption (iv) is not overly restrictive when using appropriate regression algorithms. The assumptions (v) and (vi), which are well-established (see more details in [7,23]), are typically applied to evaluating the convergence properties of regression methods. In this work, to match the required Hilbert space norm and the well-specified case, we adopt the approach of [31], where the regression is well-defined, with $\ell \in (1,2]$ and the necessary Hilbert space norm. The parameter ℓ indicates the smoothness of the regression.

Theorem 5 *([7,23]). Under the assumptions (iv) to (vi), for the specialized case with the smoothness parameter $\ell \in (1,2]$, there is a constant $K > 0$ independent of $n \ge 1$ and $\tau \ge 1$ such that for all $z \in \mathcal{Z}$,*

$$\|\widehat{\mu}_{X|Z}(z) - \mu_{X|Z}(z)\|^2 \le \tau^2 K n^{-\frac{\ell-1}{\ell+p}} \tag{29}$$

holds for sufficiently large $n \ge 1$ with $\mathbb{P}^n$-probability not less than $1 - 4e^{-\tau}$.

5.2.2. Results for residual independence testing stage

Here, we go ahead to analyze the results of the residual independence testing stage, given the following assumption:

(vii) The kernels k, l with parameter ω_1, ω_2 are assumed to be Lipschitz continuous w.r.t the residual variables. Specifically, for any (x_1, y_1, z_1), (x_2, y_2, z_2), $(x_3, y_3, z_3) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, the resulting residual variables $r_{x;i} = x_i - \mathbb{E}[X|z_i]$, $r_{y;i} = y_i - \mathbb{E}[Y|z_i]$ with $i = 1, 2, 3$, we assume that

$$\left|k_{\omega_1}(r_{x;1}, r_{x;2}) - k_{\omega_1}(r_{x;1}, r_{x;3})\right| \leq C_k \cdot \|r_{x;2} - r_{x;3}\|, \quad \left|l_{\omega_2}(r_{y;1}, r_{r;2}) - l_{\omega_2}(r_{y;1}, r_{y;3})\right| \leq C_l \cdot \|r_{y;2} - r_{y;3}\|,$$

where C_k and C_l are nonnegative Lipschitz constants.

The assumption (vii) does not restrict the specific form of the kernels, and the kernels used in our paper satisfy this condition. In the next three steps, we obtain the bound for $\mathrm{RCIT}_b(\mathcal{D}) = \mathrm{HSIC}_b(\hat{R}^{(m)}_{X|Z}, \hat{R}^{(m)}_{Y|Z})$.

1. **Bound for residuals.** According to Theorem 5 and the definitions of residuals in Eqs. (22) and (25), for all i, we have

$$\|\hat{r}_{x;i} - r_{x;i}\| \leq \tau_x K_x^{1/2} n^{-\frac{\ell-1}{2(\ell+p)}}, \tag{30}$$

with probability at least $1 - 4e^{-\tau_x}$. For simplicity, we denote $C(\delta_{x;r}) := \tau_x K_x^{1/2}$ and $\delta_{x;r} := 4e^{-\tau_x}$. Thus, we conclude that $\|\hat{r}_{x;i} - r_{x;i}\| \leq C(\delta_{x;r}) \cdot n^{-\frac{\ell-1}{2(\ell+p)}}$ with probability at least $1 - \delta_{x;r}$. And similarly for Y, we have $\|\hat{r}_{y;i} - r_{y;i}\| \leq C(\delta_{y;r}) \cdot n^{-\frac{\ell-1}{2(\ell+p)}}$ with probability at least $1 - \delta_{y;r}$ with $\delta_{y;r} := 4e^{-\tau_y}$.

2. **Bound for kernels.** Under the assumption (vii). It can be shown that

$$\begin{aligned}\left\|k(\hat{r}_{x;i}, \hat{r}_{x;j}) - k(r_{x;i}, r_{x;j})\right\| \leq& \left\|k(\hat{r}_{x;i}, \hat{r}_{x;j}) - k(\hat{r}_{x;i}, r_{x;j})\right\| + \left\|k(\hat{r}_{x;i}, r_{x;j}) - k(r_{x;i}, r_{x;j})\right\| \\ \leq& C_k \cdot \left\|\hat{r}_{x;j} - r_{x;j}\right\| + C_k \cdot \left\|\hat{r}_{x;i} - r_{x;i}\right\| \leq 2C_k C(\delta_{x;r}) \cdot n^{-\frac{\ell-1}{2(\ell+p)}},\end{aligned} \tag{31}$$

with probability at least $1 - \delta_{x;r}$. In the same way, we can show that

$$\left\|l(\hat{r}_{y;i}, \hat{r}_{y;j}) - l(r_{y;i}, r_{y;j})\right\| \leq 2C_l C(\delta_{y;r}) \cdot n^{-\frac{\ell-1}{2(\ell+p)}}, \tag{32}$$

with probability at least $1 - \delta_{y;r}$.

3. **Bound for statistic.** According to the notations in Eqs. (23) and (26), by comparing each sub term and using the bounds for kernels, the following bound can be obtained:

$$\left|\mathrm{HSIC}_b(\hat{R}^{(m)}_{X|Z}, \hat{R}^{(m)}_{Y|Z}) - \mathrm{HSIC}_b(R^{(m)}_{X|Z}, R^{(m)}_{Y|Z})\right| \leq 4\nu \cdot \left(\left\|\check{k}_{ij} - \tilde{k}_{ij}\right\| + \left\|\check{l}_{ij} - \tilde{l}_{ij}\right\|\right) \leq C(\delta) \cdot n^{-\frac{\ell-1}{2(\ell+p)}}, \tag{33}$$

where $C(\delta) = 8\nu C_k C(\delta_{x;r}) + 8\nu C_l C(\delta_{y;r})$, with probability $1 - \delta$ with $\delta := \delta_{x;r} + \delta_{y;r} - \delta_{x;r}\delta_{y;r}$.

Above, we show that the error introduced by regression function estimation is bounded by $n^{-\frac{\ell-1}{2(\ell+p)}}$. This allows us to analyze the two types of errors in RCIT, leveraging the well-characterized behavior of $\mathrm{HSIC}_b(R^{(m)}_{X|Z}, R^{(m)}_{Y|Z})$ in the independence test. Since $\mathrm{HSIC}_b(R^{(m)}_{X|Z}, R^{(m)}_{Y|Z})$ is free from regression error, its behavior closely resembles that of HSIC, with the only difference being the input transformation to residuals. We summarize the major results below. Formally, let $\tilde{r}^{(m)}$ denote the threshold that controls the Type I error at a given significance level α, ensuring that

$$\mathbb{P}_{\mathcal{H}_0}\left(m\mathrm{HSIC}_b(R^{(m)}_{X|Z}, R^{(m)}_{Y|Z}) > \tilde{r}^{(m)}\right) \leq \alpha. \tag{34}$$

For Type II error, Proposition 3 establishes that

$$\mathbb{P}_{\mathcal{H}_1}\left(m\mathrm{HSIC}_b(R^{(m)}_{X|Z}, R^{(m)}_{Y|Z}) \leq \tilde{r}^{(m)}\right) \sim \mathcal{O}(m^{-1/2}). \tag{35}$$

Above, Eqs. (34) and (35) provide the error bounds for two types of errors in CI testing using the statistic $\mathrm{HSIC}_b(R^{(m)}_{X|Z}, R^{(m)}_{Y|Z})$. Further considering the error bound in Eq. (33), we can derive the corresponding error bounds for RCIT. The final result is presented in the following theorem.

Theorem 6. (Type I and II error bounds for RCIT). *Under the assumptions (iv) to (vii), assume that n grow strictly faster than* $\omega\left(m^{\frac{2(\beta+p)}{\beta-1}}\right)$, *and let* $\check{r}^{(m)}$ *be the threshold with kernels of size m that is obtained by permutation testing or wild bootstrapping, then with high probability, the bound for Type I error*

$$\mathbb{P}(\textit{Type I error}) = \mathbb{P}_{\mathcal{H}_0}\left(mHSIC_b(\hat{R}^{(m)}_{X|Z}, \hat{R}^{(m)}_{Y|Z}) > \check{r}^{(m)}\right) \leq \alpha + o(1) \tag{36}$$

and the probability for the Type II error

$$\mathbb{P}(\textit{Type II error}) = \mathbb{P}_{\mathcal{H}_1}\left(mHSIC_b(\hat{R}^{(m)}_{X|Z}, \hat{R}^{(m)}_{Y|Z}) \leq \check{r}^{(m)}\right) \sim \mathcal{O}(m^{-1/2}) \tag{37}$$

hold for any sufficiently large $m \geq 1$.

The proof procedure is given in *Appendix* I.

Analysis. Theorem 6 establishes bounds for the two types of errors in CI test while accounting for errors in both the regression and residual independence testing stages simultaneously. The key intuition is as follows:

- Error in the regression stage generally increases the likelihood of rejecting $\mathcal{H}_0$ (CI) in the second stage, thereby boosting the test's power. Consequently, the consistency of the test and its rate of convergence remain unaffected.
- However, regression error impacts Type I error (false positive rate), meaning that we must carefully balance the sample sizes in both stages to ensure that Type I error remains controlled. Our theorem quantifies the necessary relationship between the number of samples used in the two stages by leveraging existing bounds on regression error. This allows us to prove that, under appropriate conditions, the Type I error remains asymptotically controlled, ensuring the validity of the overall test.

Remark. While our theoretical results suggest that the regression stage typically requires a larger sample size than the test stage, which models a worst-case scenario. In practice, the theoretical bounds are conservative, and empirical evaluations provide a more accurate picture of Type I error control. Indeed, our experimental results demonstrate that the test maintains a well-controlled Type I error even if we use the same sample size for both two stages. Additionally, to maximize sample efficiency, though our theoretical analysis assumes the samples in the two stages are independent to each other, we often reuse the same dataset for both the regression and testing steps — a common practice in existing CI methods to enhance power [38,55].

6. Performance evaluation

In this section, we evaluate the performance of our method, which consists of three parts. We first evaluate the performance of our method on the independence testing task, then we assess the performance on conditional independence testing, and finally we apply our method to the causal discovery task.

6.1. Independence test

Here, we evaluate the performance of our method on the independence testing task.

6.1.1. Compared methods

We compare the following tests on several datasets.

- **The randomized dependence coefficient (RDC)** [26]. A state-of-the-art method based on the canonical correlation between a finite set of random Fourier features.
- **The HSIC with random Fourier feature (FHSIC)** [56]. A variant of HSIC that uses finite-dimensional random Fourier feature mappings to approximate kernels.
- **The normalized version of the Finite Set Independence Criterion (NFSIC)** [17]. A state-of-the-art adaptive test by choosing features on a hold-out validation set to optimize a lower bound on the test power.
- **HSIC-M** [12]. The original HSIC with the kernel width being set to the Euclidean distance median of the samples.
- **HSIC-O (Ours).** HSIC with the Gaussian kernel whose bandwidth (length-scale) is optimized.
- **HSIC-W (Ours).** HSIC with importance-weighted kernels as described in Sec. 4.3.

Following are the default settings unless stated otherwise. We use Gaussian kernels for both X and Y in all methods. We set the number of random mappings in RDC and FHSIC to 10, the test location parameter J of NFHSIC to 10, which are the recommended settings in [17]. For RDC and FHSIC, we permute the samples 100 times to simulate the null distribution and compute the threshold. The thresholds for the remaining methods are obtained by asymptotic null distribution, i.e., we set the test threshold to the $(1-\alpha)$-quantile of $\chi^2(J)$ for NFSIC and obtain the test threshold of HSIC-M/O/W by gamma approximation. The significance level α is set to 0.05. In the optimization step, for stabilizing the training, in the implementation of NFSIC we determine the initial bandwidth by searching the best from 25 bandwidth combinations (including the median bandwidth combination). For HSIC-O/W, to be fair, we perform the same grid search on the benchmark datasets. In other experiments, we still use the median bandwidth as initialization for kernel width. Also, the maximum number of iterations for the optimization is set to 100 for NFSIC and HSIC-O/W. For synthetic data, we set the split ratio to 0.5 for NFSIC and HSIC-O/W, i.e., we randomly sample half of the data for training and use the remaining for independence testing, while the other methods use all data for testing. For real data, we divide a small portion of the data for training and then extract 100 random subsets of the remaining data (disjoint from the training set) for evaluation. Results for more settings (e.g. Laplace kernel setting) and more compared (not only kernel-based) methods are given in *Appendix* L.2.

6.1.2. Results on simulation

We consider the benchmarks from [17,56] and the application on independent subspace analysis from [12]. We also conduct experiments on high-dimensional data based on 3Dshapes [5]. We use the following three benchmarks:

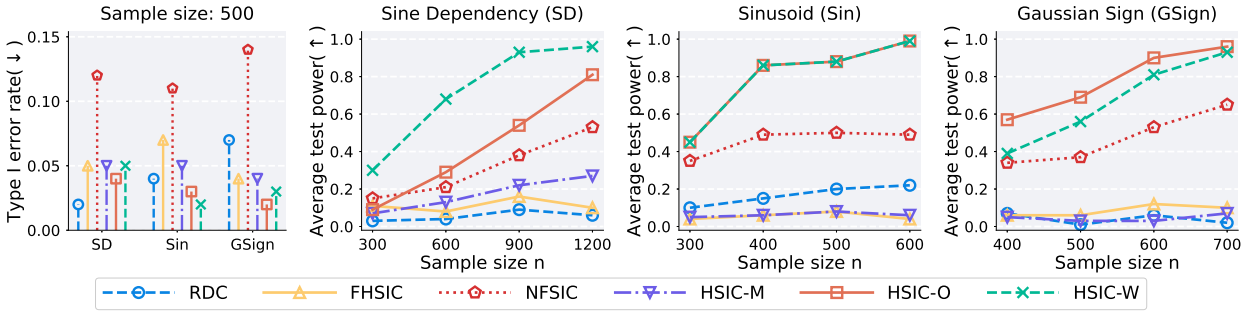

Fig. 2. Left: Results of Type I error rate on the three benchmarks with sample size $n = 500$. The other three plots: the results of average test power on the three benchmarks.

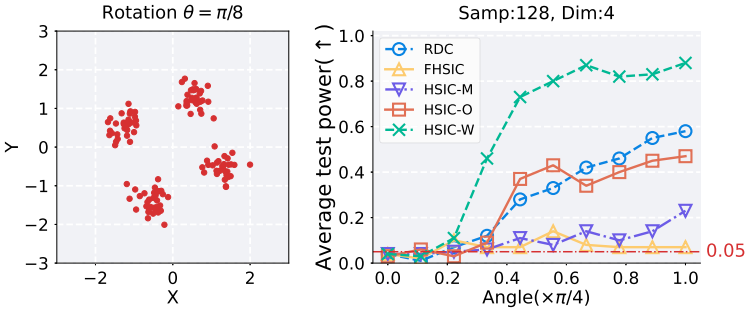

Fig. 3. Left: Example dataset for $d = 1, n = 128$ and rotation angles $\theta = \pi/8$. Right: The average test power v.s. the rotation angle of each method.

Sine Dependency (SD). We begin with a non-linear dependence model. Concretely,

$$X \sim \mathcal{N}_d(0, I_d), Y = 20 \sin\left(4\pi(X_1^2 + X_2^2)\right) + Z, \tag{38}$$

where X_i is the i-th dimension of X, d is the dimension of X, and $Z \sim \mathcal{N}(0,1)$ is independent with X. When $d \geq 2$, there is a non-linear dependence of Y on X in some local dimensions.

Sinusoid (Sin). We then consider the Sinusoid model that has local change in the probability density function. Concretely, let $\mathbb{P}_{xy}$ be the probability density function on $\mathcal{X} \times \mathcal{Y} := [-\pi, \pi]^2$, i.e.,

$$(X, Y) \sim \mathbb{P}_{xy}(x, y) \propto 1 + \sin(\omega x)\sin(\omega y), \tag{39}$$

where ω is frequency. Higher frequency makes the drawn samples more similar to those drawn from Uniform($[-\pi, \pi]^2$) [36], thus more difficult to detect dependency for small sample sizes.

Gaussian Sign (GSign). Next, we consider the Gaussian Sign model, i.e.,

$$X \sim \mathcal{N}_d(0, I_d), Y = |Z| \prod_{i=1}^{d} \operatorname{sgn}(X_i), \tag{40}$$

where $\operatorname{sgn}(\cdot)$ is the sign function, X_i is the i-th dimension of X, d is the dimensionality of X, and $Z \sim \mathcal{N}(0,1)$ is independent with X. The challenge lies in that Y is independent of any proper subset of X, but is dependent on X. Therefore, considering all dimensions of X simultaneously is crucial to independence testing.

The experimental setup is as follows: For SD, we set $d = 3$ and sample size $n \in \{300, 400, 500, 600\}$. For Sin, we set $\omega = 3$ and sample size $n \in \{300, 600, 900, 1200\}$. For GSign, we set $d = 4$ and sample size $n \in \{400, 500, 600, 700\}$. For each setup, we perform 100 repeated randomized experiments and report the average result of test power. For the evaluation of Type I error, we set the sample size to 500, permute the samples randomly to obtain new independent samples, then perform the full independence test. The results are shown in Fig. 2.

Results and analysis. HSIC-M/O/W succeeds in controlling Type I error rate below 0.05 on all three datasets, RDC and FHSIC also succeed in controlling Type I error rate around 0.05, while NFSIC has a relatively large Type I error rate (around 0.1). As for the testing power, HSIC-O/W and NFSIC perform better on the three benchmarks compared to the other methods, which confirms the need for kernel learning. The performance of HSIC-O/W is stably improved as the number of samples increases, which corroborates the consistency of our proposed test. Besides, it is worth noting that HSIC-W does not always obtain superior performance over HSIC-O. This is due to the additional risk that HSIC-W may face when having a poor estimate of the important weight w_i.

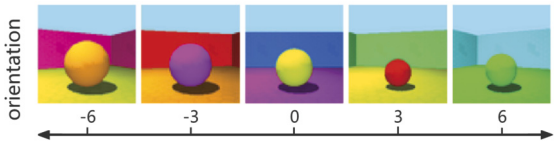

Fig. 4. Examples of images generated by varying the orientation factor while fixing the object shape.

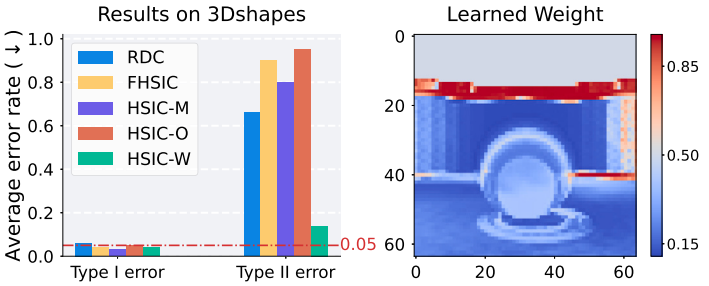

Fig. 5. Left: The average rates of the two types of error on the 3Dshapes dataset. Right: The visualization of the learned weights of HSIC-W.

On Sin, the results are the same due to the data being one-dimensional. On SD, HSIC-W gets better results since $d = 3$ and Y is only dependent on the first two dimensions of X. While on GSign, HSIC-O performs better. The reason is that each dimension of X is equally important in generating Y, making it in fact no need to learn the importance weight w_i. Imprecise estimation results of w_i due to insufficient samples cause the performance degradation of HSIC-W. As the number of samples gradually increases, more accurate estimations of importance weight narrow this gap.

6.1.3. *Independence of subspaces*

One important application of independence testing is to determine the convergence of algorithms for independent component analysis (ICA) [12], which involves separating random variables from their linear mixtures. We construct the data as follows: First, generating n i.i.d samples of two univariate random variables with the distribution $\frac{1}{2}\mathcal{N}(-1, 0.01) + \frac{1}{2}\mathcal{N}(1, 0.01)$. Second, mixing these random variables using a rotation matrix parameterized by an angle θ, varying from $[0, \pi/4]$ (a zero angle means the data are independent, while a larger angle leads to stronger dependency. See the left in Fig. 3 for an example. Third, appending noise of distribution $\mathcal{N}_{d-1}(0, I_{d-1})$ to each of the mixtures. Finally, multiplying an independent random d-dimensional orthogonal matrix, to obtain vectors dependent across all observed dimensions. The resulting random variables X and Y are dependent but uncorrelated. When d is greater than 1, the problem is associated with the independent subspace analysis (ISA) problem [43]. We set $d = 4$, sample size $n = 128$, then evaluate the average test power with $\theta \in [0, \pi/4]$. Recall that according to the default settings, we take 64/64 samples for training/testing for the methods of learning kernels. Unfortunately, NFSIC faces optimization issues in this setting and cannot successfully control Type I error, so we only present the results for the remaining five methods, as shown in the right of Fig. 3.

Results. The results obtained at $\theta = 0$ reflect Type I error rate, as the variables are independent in this case. All methods successfully control Type I error ≤ 0.05. HSIC-W stably outperforms the other methods significantly as the angle increases, while HSIC-M fails to capture the dependence with a sample size of 128, which shows the importance of kernel learning.

6.1.4. *High-dimensional data*

We consider a challenge setting on high-dimensional image data. 3DShapes [5] is a dataset of 3D scenes with additional features such as shadows and background (sky). There are 6 ground-truth independent latent factors including floor hue, wall hue, object hue, object scale, object shape and orientation, which can be controlled to generate corresponding images. We consider orientation as a dependency factor for independence testing, i.e., let X be the image, Y be the corresponding angle of orientation, and test the dependency between X and Y. To be more challenging, we fix the shape of the object to be a ball thereby (compared to a square etc.) reducing the apparent orientation feature and randomizing the other factors. Some generated examples are shown in Fig. 4, where the numbers indicate the relative orientation angles. For the experimental setup, we vectorize X to obtain a random vector with dimension $64 \times 64 \times 3 = 12,288$. The sample size is set as 64. As NFSIC cannot handle such high-dimensional input, we use the other methods for testing. Type I error rate is evaluated by the samples obtained by permutation. The results are shown in Fig. 5.

Results and analysis. All methods are successful in controlling Type I error rate under 0.05. For Type II error, the result of HSIC-W is significantly better than the other methods. HSIC-O obtains worse performance than HSIC-M due to the fact that the amount of data evaluated is half the size, thus losing some of the test power. A visualization of the weights learned by HSIC-W is shown in the right of Fig. 5, from which we can see that the channels (edges) decided by the orientation receive more attention.

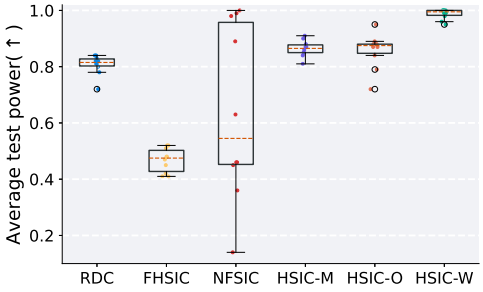

Fig. 6. The average test power of 6 methods. The dashed line in each box is the mid-point.

6.1.5. Results on real data

As for real data testing, we consider the subset of the Million Song Data[3] [4]. This dataset contains 515,345 songs with 91-dimensional features. The first dimension is the release year of each song, which we take as the variable Y. The remaining features (e.g., timbre average and timbre covariance of each song) are taken as the variable X. The goal is to detect the dependency between X and Y. For the experimental setup, we follow the recommended settings of NFSIC, for which we use permutation to ensure that Type I error is controlled. To be fair, HSIC-M/O/W are also evaluated using the permutation scheme, with the number of permutations set to 100. Note that when training, HSIC-O/W still use the gamma approximation to compute the threshold, corresponding to Algorithm 1. Recall that we randomly select a small portion ($n = 500$) of data as the training set, and use the rest for evaluation. In order to fully utilize the data, we randomly sample 500 data from the remaining data each time during the evaluation and obtain the average result of 100 times. The above training and testing processes are repeated 10 times to evaluate the robustness of the optimization scheme. Other settings are the same as before.

The final results are in Fig. 6. Compared to the other methods, HSIC-W achieves a test power close to 1 with a very small variance. RDC and HSIC-M/O achieve a test power above 0.8. As a comparison, NFSIC has a large variance. These results corroborate the robustness of the optimization approach of our method, which benefits from the design of our criterion and the theoretical guarantee of smoothness.

6.2. Conditional independence test

Here, we evaluate the performance of our method on the CI testing task.

6.2.1. Results on simulation

In this section, we evaluate the performance of RCIT based on HSIC-M/O/W in simulation and simply denote the three CI tests as RCIT-M/O/W. Specially, we conduct experiments on both linear and non-linear settings. The data generation process is described as follows.

$$x = f(Za_x) + \epsilon_x, \quad y = g(Za_y) + \epsilon_y. \tag{41}$$

In linear settings, f and g are identical transformation. In non-linear settings, f and g are randomly selected from sin, cos, tanh, exponential, or x^α, where the exponent α is randomly chosen from $\{2, 3\}$. In our setup, $Z \sim \mathcal{U}_d(0, 1)$, a_x is a $d \times d$ matrix and a_y is a $d \times 1$ vector, both with the entries being drawn from $\mathcal{U}(0.2, 1)$. The noise terms are defined as $\epsilon_x \sim \mathcal{U}_d(-0.5, 0.5)$ and $\epsilon_y = 20sin\left(4\pi(\epsilon_{x,1})^2\right) + \gamma$, where $\epsilon_{x,1}$ is the first dimension of ϵ_x, and $\gamma \sim \mathcal{U}(-0.5, 0.5)$ is an independent random variable. Notably, for the cases $d \geq 1$, non-linear dependence exists between ϵ_y and ϵ_x in certain local dimensions.

Experimental setup: to ensure fairness, HSIC-M/O/W are evaluated by using the permutation scheme, with the number of permutations 100. It is important to note that during training, HSIC-O/W still use the gamma approximation to compute the threshold, as described in Algorithm 1. Other settings remain consistent with previous experiments. We conduct 100 repeated randomized experiments and report the average results. We set the sample sizes to {50, 100, 200, 300, 500}, and for each sample size, different values for $d = \{1, 2, 3, 4, 5\}$ are used. The samples are then randomly permuted to generate new independent samples, followed by a full independence test. The results are presented in Figs. 7 and 8.

Results of linear cases. From an overall perspective, RCIT-O/W effectively controls the Type I error rate at approximately 0.05 across all samples, except in certain dimensions, demonstrating the stronger robustness in controlling Type I errors, which validates the significance of kernel learning. On the contrast, RCIT-M exhibits relatively higher Type I error rates across various sample sizes, with particularly noticeable fluctuations as the dimension increases. This indicates that RCIT-M is more prone to misjudging the CI test when dealing with high-dimensional data, leading to an increased error rate.

As for type II error rate, when the sample size is 50 and 100, RCIT-O performs worse than RCIT-M in some dimensions. This is due to the smaller sample size, which leads to suboptimal optimization. On the other hand, RCIT-W performs well even with

[3] Million Song Data subset: https://archive.ics.uci.edu/dataset/203/yearpredictionmsd.

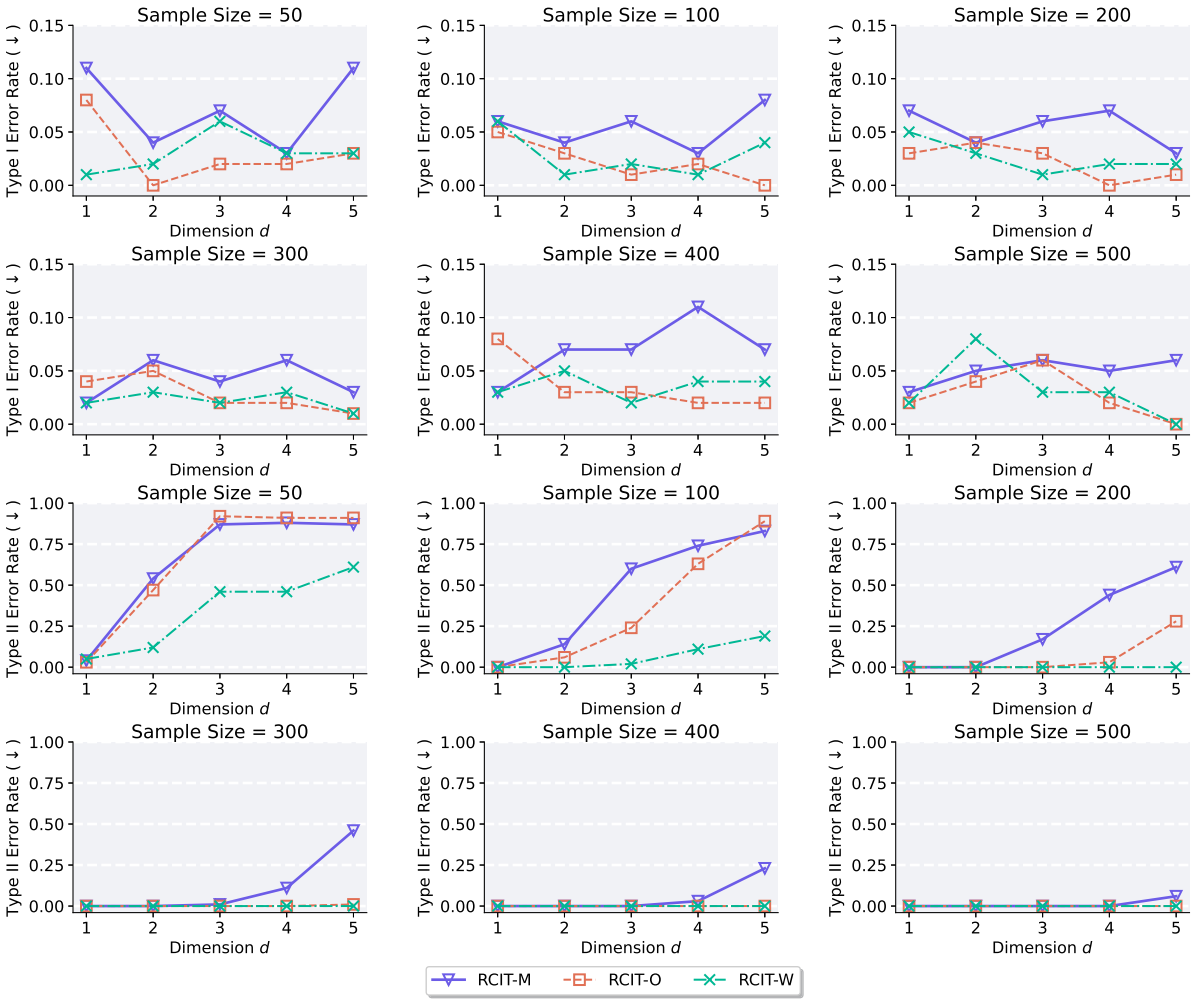

Fig. 7. Performance of RCIT-M/O/W on Type I and Type II error rates under the linear settings with the sample sizes {50, 100, 200, 300, 400, 500} and the conditioning sets of size $|Z| = \{1, 2, 3, 4, 5\}$.

low sample sizes, as the learning weights compensate for the inadequacies in bandwidth optimization caused by insufficient sample size. From a global perspective, RCIT-M shows a significant increase in the false negative rate as the dimension increases, especially with smaller sample sizes where this rate escalates sharply. This implies that RCIT-M lacks sufficient sensitivity in detecting actual dependencies, potentially leading to the omission of a substantial amount of valid information. Conversely, RCIT-O demonstrates a relatively balanced Type II error rate at medium sample sizes, although its performance deteriorates slightly in higher dimensions. RCIT-W, however, consistently achieves the lowest Type II error rates in most scenarios, particularly excelling at large sample sizes and high dimensions. This evidences RCIT-W superior capability in detecting dependencies under complex data conditions.

Results of non-linear cases. Overall, RCIT-O/W consistently maintains the Type I error rate near the nominal level of 0.05 across most sample sizes, with only minor deviations at certain dimensions. This demonstrates the robustness of RCIT-O/W in controlling Type I error and highlights the advantages of kernel learning. In contrast, RCIT-M exhibits higher variability in Type I error rate across different samples and dimensions, indicating lower stability.

Regarding the Type II error rate, when the sample size is 50 and 100, RCIT-O/W performs worse than RCIT-M in most dimensions, mainly due to the limited data available for effective kernel learning. Furthermore, it should be noted that when the dimensionality is one, the dimension of $\hat{R}^{(m)}_{X|Z}$ is equally important for $\hat{R}^{(m)}_{Y|Z}$. In such scenarios, RCIT-W tends to underperform relatively to RCIT-O/W. Nevertheless, as the dimension increases, RCIT-W achieves improved performance even at lower sample sizes, benefiting from importance-weighted kernels that effectively mitigate the drawbacks of suboptimal kernel learning caused by insufficient sample size. Globally, RCIT-M/O shows a significant increase in the false negative rate as the dimension increases, especially with smaller sample sizes where this rate escalates sharply. This implies that RCIT-M/O lacks sufficient sensitivity in detecting actual dependencies under the non-linear settings, potentially leading to the omission of a substantial amount of valid information. However, as the sample size increases, RCIT-O performs better than RCIT-M due to kernel learning. Notably, RCIT-W, consistently achieves the lowest Type II error

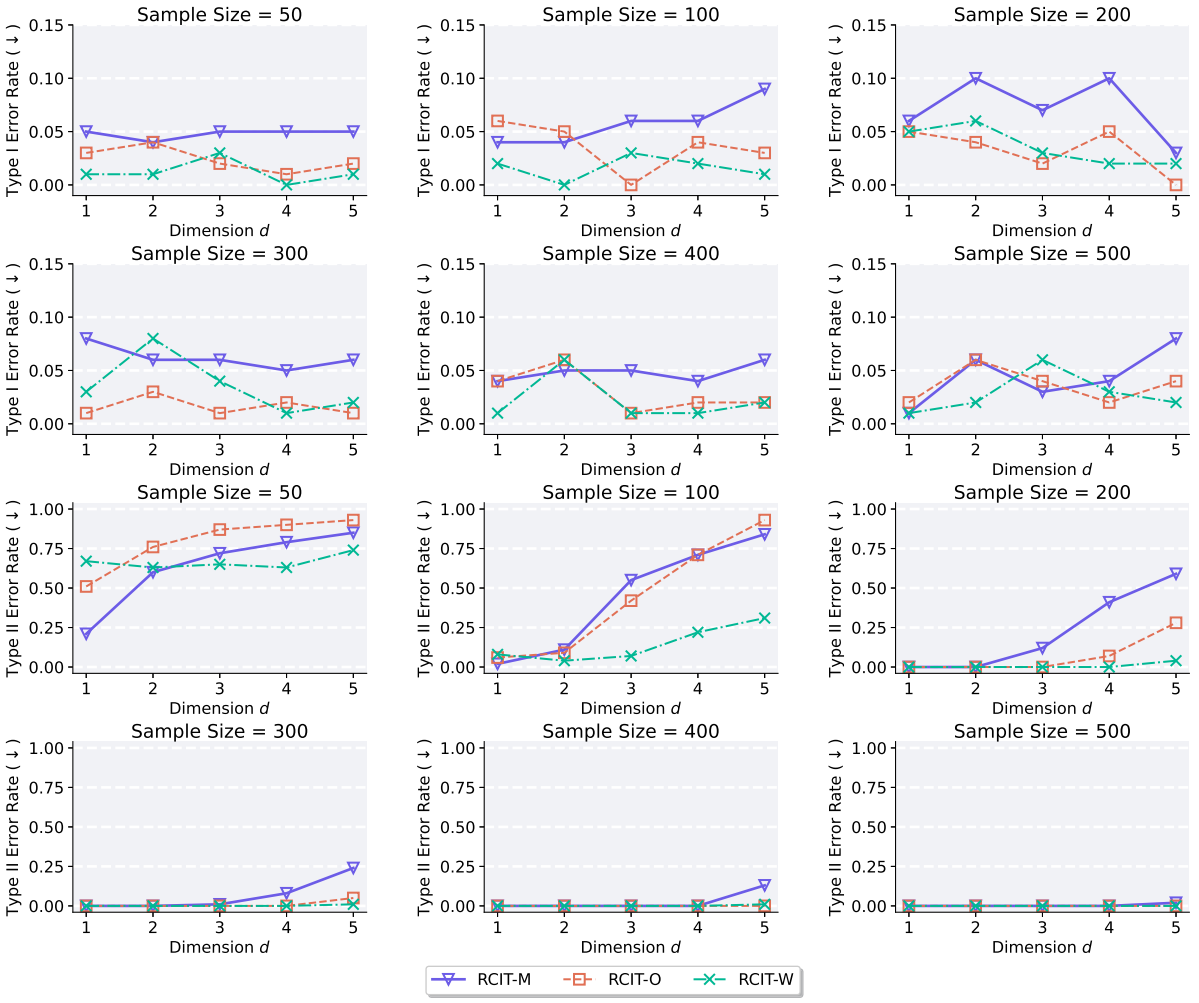

Fig. 8. Performance of RCIT-M/O/W on Type I and Type II error rates under the non-linear settings with sample sizes {50, 100, 200, 300, 400, 500} and the conditioning set of size $|Z| = \{1, 2, 3, 4, 5\}$.

rate in most scenarios, particularly excelling at large sample sizes and high dimensions. This evidences RCIT-W superior capability in detecting dependencies under complex data conditions.

Evaluation summary. With the experiments on either linear or non-linear settings, RCIT-W demonstrates strong performance across different sample sizes and dimensions, achieving low Type I and Type II error rates. This suggests that RCIT-W is well-suited for conditional independence testing in complex data settings, reinforcing the importance of kernel learning. On the other hand, RCIT-M performs less reliably in high-dimensional settings due to elevated Type I and Type II error rates, which may limit its practical applicability. While RCIT-O performs reasonably well in most cases, there is still room for improvement in handling high-dimensional data.

6.2.2. Results on real data

The above experiments demonstrate the effectiveness of our proposed method for CI test on simulations. To provide a more comprehensive evaluation, in this section, we further compare additional state-of-the-art methods on a real-world dataset, including WIT [33], DARLING [15], KCIT [55], SCIT [53], FRCIT [49], RDC [26], and PaCoT. For clarity, we refer to RDC as RDCPT and DARLING as NRIT in the subsequent discussion. We provide detailed descriptions and experimental settings for these methods in the *Appendix* K.

We employ the well-known Sachs dataset [34], commonly used in causal discovery studies [15]. This dataset comprises 853 samples and a corresponding causal graph[4] with 11 nodes and 18 arcs. These nodes form 55 possible pairs, with 18 pairs being independent and the remaining 37 not. Our objective is to evaluate the performance of various methods in testing independence

[4] Sachs causal graph: https://www.bnlearn.com/bnrepository/.

Table 1
Results of CI Test on the Sachs Dataset.

Algorithm	Type I Error	Type II Error	Average Error
WIT	0.114 ± 0.0064	0.742 ± 0.0014	0.428 ± 0.0032
NRIT	$\mathbf{0.010 \pm 0.0006}$	0.782 ± 0.0007	0.391 ± 0.0004
KCIT	0.111 ± 0.0000	0.730 ± 0.0000	0.420 ± 0.0000
SCIT	0.074 ± 0.0047	0.884 ± 0.0018	0.479 ± 0.0025
RDCPT	0.104 ± 0.0026	0.725 ± 0.0011	0.414 ± 0.0015
FRCIT	0.011 ± 0.0022	0.893 ± 0.0024	0.452 ± 0.0016
PaCoT	0.111 ± 0.0000	0.730 ± 0.0000	0.420 ± 0.0000
RCIT-M	0.111 ± 0.0000	0.730 ± 0.0000	0.420 ± 0.0000
RCIT-O (**Ours**)	0.077 ± 0.0055	0.701 ± 0.0026	0.389 ± 0.0031
RCIT-W (**Ours**)	0.073 ± 0.0053	$\mathbf{0.698 \pm 0.0026}$	$\mathbf{0.385 \pm 0.0031}$

by examining both Type I error rate, Type II error rate, and Average Error. The Average Error, is defined as the arithmetic mean of the Type I error and Type II error rates, providing a single metric that captures the overall performance of the test in balancing false positives and false negatives, computed as Average Error = (Type I Error + Type II Error)/2. For each experiment, we randomly repeat 100 times and report the average test result. All results are presented in Table 1.

Results and analysis. Overall, RCIT-O/W demonstrate superior average performance on this dataset. While their Type I error rates are not the lowest, they remain very close to the significance level, ensuring controlled false positives while maintaining a relatively low Type II error rate. Notably, RCIT-W, which optimizes both bandwidth and weighting, achieves the best balance between Type I and Type II errors, leading to improved performance on complex datasets. This highlights the substantial benefit of kernel learning. It is worth noting that Type II error rate is higher than Type I error rats. This discrepancy suggests that real-world data poses a greater challenge than simulations, underscoring the necessity of minimizing Type II error in CI testing for reliable causal inference.

6.3. Causal discovery

In this section, we evaluate the performance of our method on the causal discovery task.

6.3.1. Results on cancer graph

In the aforementioned experiments, we compared RCIT-M/O/W in terms of their performance in CI test. In this section, we evaluate these three methods by using a causal graph specifically related to cancer,[5] which involves 5 nodes and 4 arcs, using RCIT-M/O/W in conjunction with the PC algorithm [39], denoted as $\text{PC}_{\text{RCIT-M/O/W}}$, to discover the causal graph G. The graph G consists of five nodes: Pollution (v_1), Smoker (v_2), Cancer (v_3), X-ray (v_4), and Dyspnoea (v_5).

In our experiment, nodes v_1 (Pollution) and v_2 (Smoker) are sampled from a uniform distribution $\mathcal{U}(0,1)$ with dimension $d = 3$. The subsequent node v_3 (Cancer) is generated according to the additive noise model, expressed as $v_3 = v_1 a_i + v_2 a_j + \epsilon$, where $a_i, a_j \sim \mathcal{U}(0.2,1)$ are 3×3 dimensional matrices, and $\epsilon \sim \mathcal{U}(-0.2,0.2)$ with dimension $d = 3$. Node v_4 (X-ray) is defined as $v_4 = v_3 a_q + \eta$ and node v_5 (Dyspnoea) is defined as $v_5 = v_3 a_r + \gamma$, where $a_q \sim \mathcal{U}(0.2,1)$ is a 3×3 matrix, $a_r \sim \mathcal{U}(0.2,1)$ is a 3×1 dimensional matrix, η is generated by the first equation in Eq. (41) with dimension $d = 3$, and γ is generated by the second equation in Eq. (41).

For significance level 0.05 and sample sizes from {200, 500, 1000}, we evaluate the performance of the three methods that $\text{PC}_{\text{RCIT-M/O/W}}$ on discovering causal structure and skeleton. The results are shown in Fig. 9.

Results and analysis. Overall, $\text{PC}_{\text{RCIT-W}}$ consistently outperforms both $\text{PC}_{\text{RCIT-O}}$ and $\text{PC}_{\text{RCIT-M}}$ across all sample sizes and performance metrics. Although $\text{PC}_{\text{RCIT-M}}$ shows notable improvements as the sample size increases, narrowing the gap with $\text{PC}_{\text{RCIT-O}}$, it still consistently underperforms, particularly in Recall and F1-Score. These results show the superiority $\text{PC}_{\text{RCIT-W}}$, highlighting the significant advantage of kernel learning.

6.3.2. Results on real data

In our previous experiments, we demonstrate the effectiveness of our method for causal discovery on the cancer graph. Since real-world experiments inherently pose a greater challenge than simulations, to further validate our proposal, we compare our approach against additional state-of-the-art CI methods, as introduced in Sec. 6.2.2, in conjunction with the PC algorithm to demonstrate its superiority. For our experiments, we use the previously introduced Sachs dataset [34], a well-established real-world dataset for causal protein signaling networks. Similarly, we perform each experiment 100 times randomly and report the average test result. All the results are summarized in Table 2.

Results and analysis. The experimental results demonstrate that our proposed method outperforms the other methods overall in terms of average performance recall and F1-Score. For the other evaluation metric, our proposed method also shows competitive performance, further emphasizing the importance of kernel learning. In particular, $\text{PC}_{\text{RCIT-W}}$, which incorporates bandwidth and weighting learning, achieves better results, highlighting the significance of learning importance-weighted kernels.

[5] Cancer graph: https://www.bnlearn.com/bnrepository/discrete-small.html%23cancer.

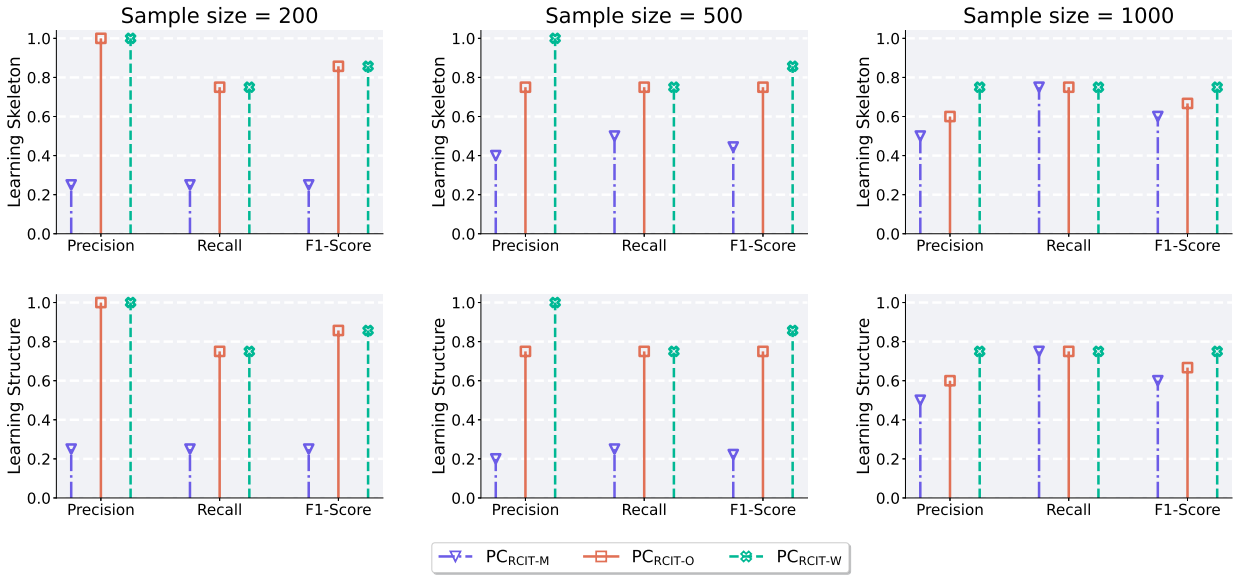

Fig. 9. Performance of $PC_{RCIT\text{-}M/O/W}$ with sample sizes $\{200, 500, 1000\}$ on causal skeleton learning and causal structure learning.

Table 2
Results of Causal Discovery on Sachs Dataset.

Algorithm	Recall	Precision	F1-Score
PC_{WIT}	0.5294	0.7857	0.6308
PC_{NRIT}	0.4706	**0.9911**	0.6380
PC_{KCIT}	0.5294	0.7500	0.6207
PC_{SCIT}	0.1394	0.4260	0.2085
PC_{RDCPT}	0.5324	0.7526	0.6232
PC_{FRCIT}	0.1853	0.7590	0.2954
PC_{PaCoT}	0.5294	0.7500	0.6207
$PC_{RCIT\text{-}M}$	0.5294	0.7500	0.6207
$PC_{RCIT\text{-}O}$ **(Ours)**	0.5518	0.7627	0.6385
$PC_{RCIT\text{-}W}$ **(Ours)**	**0.5595**	0.7645	**0.6444**

7. Conclusion

In this paper, we propose a novel framework for kernel-based CI tests that enable adaptively learning parameterized kernels to maximize the independence test power. The framework enables the design of flexible kernels, concretely, importance-weighted kernels, which can focus on the significant dimensions of variables for judging independence, thus making the tests powerful. Both theoretical analysis and experimental results show the effectiveness of our method. Future work will focus on applying our framework to more settings including multiple kernel learning.

CRediT authorship contribution statement

Yixin Ren: Writing – original draft, Methodology, Conceptualization. **Juncai Zhang:** Methodology, Data curation, Conceptualization. **Yewei Xia:** Data curation. **Ruxin Wang:** Data curation. **Feng Xie:** Methodology. **Jihong Guan:** Data curation. **Hao Zhang:** Supervision, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Shuigeng Zhou:** Supervision.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Acknowledgements

This work was supported by National Natural Science Foundation of China (NSFC) (62372116, 12471308, 62306019, 62472415).

Appendix A. List of symbols and notations

Symblos	Notations
$\mathcal{O}, o$	big O notion, small O notion
Ω, ω	big Omega notion, small Omega notion
$i.i.d.$	independent and identically distributed
$\mathbb{R}$	the set of real numbers
$\mathcal{B}(\mathbb{R})$	Borel σ-algebra on $\mathbb{R}$
$RV(s)$	random variable(s)
$\mathbb{P}_X$	marginal distribution of X
$\mathbb{P}_{XY}$	joint distribution of X, Y
$\mathbb{E}[X]$	expectation of X
$\mathrm{Var}(X)$	variance of X
$\mathrm{Cov}(X, Y)$	covariance of X, Y
$X \perp\!\!\!\perp Y$	random variables X, Y are independent
$\mathbf{i}_r^n$	the set of all r-tuples drawn without replacement from the set $\{1, ..., m\}$
$\binom{n}{k}$	number of k-combinations of n elements
$(n)_k$	number of permutations, define as $\frac{n!}{(n-k)!}$
$\mathrm{Tr}(\cdot)$	the trace of a square matrix
$\mathbf{K}, \mathbf{L}$	kernel matrix with entries k_{ij}, l_{ij}
$\mathbf{1}$	an vector of all ones
$\mathbf{H}$	centering matrix define as $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$
$\odot$	element-wise product
$()^{\cdot 2}$	element-wise power
$\mathcal{N}(\Omega, r)$	covering number with radii r for Ω
$\xrightarrow{\mathrm{d}}$	convergence in distribution.

Appendix B. Preliminaries

In this preliminary section, we give the detailed derivation of some of the formulas in the main paper. We first restate the results of asymptotic distributions as a reference, and next, give the procedure for calculating the moments of the null and alternative distributions.

B.1. Asymptotics distribution

We restate the results of asymptotic distributions here.

Proposition 1. (Asymptotics of $\mathrm{HSIC}_b(S)$). *Let $h_{ijqr} = \frac{1}{4!}\sum_{(t,u,v,w)}^{(i,j,q,r)} k_{tu}l_{tu} + k_{tu}l_{vw} - 2k_{uv}l_{tv}$, where the sum represents all ordered quadruples (t, u, v, w) drawn without replacement from (i, j, q, r) and assume that kernels k, l are bounded. Then, Under the null hypothesis $\mathcal{H}_0$, $HSIC_b(S)$ converges in distribution as*

$$nHSIC_b(S) \xrightarrow{d} \sum_{l=1}^{\infty} \lambda_l z_l^2, \tag{42}$$

where $z_l \sim \mathcal{N}(0, 1)$ $i.i.d$ and λ_l is the solution to the eigenvalue problem $\lambda_l \psi_l(z_l) = \int h_{ijqr}\psi_l(z_l) dF_{i,q,r}$, where the integral is over the distribution of variables z_i, z_q, z_r. And under the alternative $\mathcal{H}_1$, $HSIC_b(S)$ converges in distribution to a Gaussian variable

$$n^{\frac{1}{2}}\Big(HSIC_b(S) - HSIC(X, Y)\Big) \xrightarrow{d} \mathcal{N}(0, \sigma_u^2), \tag{43}$$

where the variance is given by

$$\sigma_u^2 = 16\big(\mathbb{E}_i(\mathbb{E}_{j,q,r}h_{ijqr})^2 - HSIC(X, Y)^2\big), \tag{44}$$

with the simplified notation $\mathbb{E}_{j,q,r} := \mathbb{E}_{z_j, z_q, z_r}$.

B.2. Statistic under $\mathcal{H}_0$

We give the procedure for calculating the first two moments of the null distribution.

B.2.1. Mean of $HSIC_u(S)$ under $\mathcal{H}_0$

An unbiased estimate of $\mathrm{HSIC}(X, Y)$, denoted by $\mathrm{HSIC}_u(S)$, is a sum of three U-statistics

$$\mathrm{HSIC}_u(S) := \frac{1}{(n)_2}\sum_{(i,j)\in\mathbf{i}_2^n} k_{ij}l_{ij} + \frac{1}{(n)_4}\sum_{(i,j,q,r)\in\mathbf{i}_4^n} k_{ij}l_{qr} - 2\frac{1}{(n)_3}\sum_{(i,j,q)\in\mathbf{i}_3^n} k_{ij}l_{iq}, \tag{45}$$

which has $\mathbb{E}[\mathrm{HSIC}_u(S)] = \mathbb{E}[\mathrm{HSIC}(X,Y)] = 0$ under $\mathcal{H}_0$.

B.2.2. Mean of $\mathrm{HSIC}_b(S)$ under $\mathcal{H}_0$

The complete proof is given in [12]. We show only some of the key steps here. The biased estimate of $\mathrm{HSIC}(X,Y)$, denote as $\mathrm{HSIC}_b(S)$, is a sum of three V-statistics

$$\mathrm{HSIC}_b(S) := \frac{1}{n^2}\sum_{i,j}^{n} k_{ij}l_{ij} + \frac{1}{n^4}\sum_{i,j,q,r}^{n} k_{ij}l_{qr} - 2\frac{1}{n^3}\sum_{i,j,q}^{n} k_{ij}l_{iq}. \tag{46}$$

First, we can show that the difference can be calculated by

$$\begin{aligned} n\big(\mathrm{HSIC}_b(S) - \mathrm{HSIC}_u(S)\big) = &\frac{1}{n}\sum_{i} k_{ii}l_{ii} - \frac{2}{n^2}\sum_{(i,j)\in \mathbf{i}_2^n} (k_{ii}l_{ij} + k_{ij}l_{ii}) \\ &+ \frac{1}{n^3}\sum_{(i,j,q)\in \mathbf{i}_3^n} (k_{ii}l_{jq} + k_{ij}l_{qq}) - \frac{3}{(n)_2}\sum_{(i,j)\in \mathbf{i}_2^n} k_{ij}l_{ij} \\ &+ \frac{10}{(n)_3}\sum_{(i,j,q)\in \mathbf{i}_3^n} k_{ij}l_{iq} - \frac{6}{(n)_4}\sum_{(i,j,q,r)\in \mathbf{i}_4^n} k_{ij}l_{qr} + \mathcal{O}\big(n^{-1}\big), \end{aligned} \tag{47}$$

when we assume the kernel is bounded. Secondly, we take the expectation of the last equation. To simplify, we use the notation $\mathbb{E}_{xyy'}kl = \mathbb{E}_{xyy'}k(x,x)l(y,y')$ (and so on for the rest), then

$$\begin{aligned} n\big(\mathbb{E}[\mathrm{HSIC}_b(S)] - \mathbb{E}[\mathrm{HSIC}_u(S)]\big) = &\mathbb{E}_{xy}kl - 2(\mathbb{E}_{xyy'}kl + \mathbb{E}_{xx'y}kl) \\ &+ \mathbb{E}_{xy'y''}kl + \mathbb{E}_{xx'y''}kl - 3\mathbb{E}_{xx'yy'}kl \\ &+ 10\mathbb{E}_{xx'yy''}kl - 6\mathbb{E}_{xx'}k\mathbb{E}_{yy'}l + \mathcal{O}\big(n^{-1}\big). \end{aligned}$$

Under $\mathcal{H}_0$, x is independent with y, thus we can draw the conclusions that $\mathbb{E}_{xyy'}kl = \mathbb{E}_{xy'y''}kl, \mathbb{E}_{xx'y}kl = \mathbb{E}_{xx'y''}kl$ and $\mathbb{E}_{xx'yy'}kl = \mathbb{E}_{xx'yy''}kl = \mathbb{E}_{xx'}k\mathbb{E}_{yy'}l$. Combining with $\mathbb{E}[\mathrm{HSIC}_b(S)] = 0$, we obtain that

$$\mathbb{E}[\mathrm{HSIC}_b(S)] = \frac{1}{n}\Big(\mathbb{E}_{xy}kl + \|\mu_x\|^2\|\mu_y\|^2 - \mathbb{E}_x k\|\mu_y\|^2 - \mathbb{E}_y l\|\mu_x\|^2\Big) + \mathcal{O}\big(n^{-2}\big), \tag{48}$$

where $\mu_x := \mathbb{E}_x\phi(x), \mu_x := \mathbb{E}_y\phi(y)$. And when we assume that $k_{ii} = l_{ii} = 1$, an empirical estimate can be obtained by replacing the term above with

$$\widehat{\|\mu_x\|^2} = \frac{1}{(n)_2}\sum_{(i,j)\in i_2^n} k_{ij},\ \widehat{\|\mu_y\|^2} = \frac{1}{(n)_2}\sum_{(i,j)\in \mathbf{i}_2^n} l_{ij}. \tag{49}$$

The obtained estimate

$$\mathbb{E}[n\mathrm{HSIC}_b(S)] = 1 + \widehat{\|\mu_x\|^2}\widehat{\|\mu_y\|^2} - \widehat{\|\mu_y\|^2} - \widehat{\|\mu_x\|^2} \tag{50}$$

results in a (generally negligible) bias of $\mathcal{O}(n^{-1})$ and can be calculated within the time cost $\mathcal{O}(n^2)$.

B.2.3. Variance of $\mathrm{HSIC}_u(S)$ under $\mathcal{H}_0$

The complete proof is given in [12]. We show only some of the key steps here. According to [37, Section 5.2.1], the variance of the U-statistic with the kernel can be calculated by

$$\mathbf{Var}[\mathrm{HSIC}_u(S)] = \binom{n}{4}^{-1}\sum_{c=1}^{4}\binom{4}{c}\binom{n-4}{4-c}\zeta_c = \frac{4\binom{n-4}{3}}{\binom{n}{4}}\zeta_1 + \frac{6\binom{n-4}{2}}{\binom{n}{4}}\zeta_2 + \mathcal{O}(n^{-3}), \tag{51}$$

where we only need to consider the dominant term

$$\zeta_2 = \mathbb{E}_{i,j}\Big[(\mathbb{E}_{q,r}h_{ijqr})\Big]^2 - \underbrace{[\mathbb{E}\mathrm{HSIC}_u(S)]^2}_{0 \text{ under } \mathcal{H}_0}, \tag{52}$$

using degeneracy ($\zeta_1 = 0$) under $\mathcal{H}_0$. Under $\mathcal{H}_0$, using x, y are independent, we have

$$\mathbb{E}_{q,r}h_{ijqr} = \frac{1}{6}(k_{ij} + \mathbb{E}_{xx'}k - \mathbb{E}_x k_i - \mathbb{E}_x k_j)(l_{ij} + \mathbb{E}_{yy'}l - \mathbb{E}_y l_i - \mathbb{E}_y l_j). \tag{53}$$

Combining with the results

$$\begin{aligned} \mathbb{E}_{ij}(k_{ij} + \mathbb{E}_{xx'}k - \mathbb{E}_x k_i - \mathbb{E}_x k_j)^2 &= \mathbb{E}_{ij}\langle\phi(x_i) - \mu_x, \phi(x_j) - \mu_x\rangle^2 \\ &= \mathbb{E}_{ij}\langle(\phi(x_i) - \mu_x)\otimes(\phi(x_i) - \mu_x), (\phi(x_j) - \mu_x)\otimes(\phi(x_j) - \mu_x)\rangle_{\mathrm{HS}} := \|C_{xx}\|^2, \end{aligned} \tag{54}$$

then the variance of the statistic is obtained by

$$\mathbf{Var}[\mathrm{HSIC}_u(S)] = \frac{2(n-4)(n-5)}{(n)_4}\|C_{xx}\|_{\mathrm{HS}}^2\|C_{yy}\|_{\mathrm{HS}}^2 + \mathcal{O}(n^{-3}), \tag{55}$$

where $\|\cdot\|_{\mathrm{HS}}^2$ is the Hilbert-Schmidt norm. An empirical estimate of the product of Hilbert-Schmidt norms $\|C_{xx}\|_{\mathrm{HS}}^2\|C_{yy}\|_{\mathrm{HS}}^2$ is given by

$$\frac{\mathbf{1}^T\big(\mathbf{B} - \mathrm{diag}(\mathbf{B})\big)\mathbf{1}}{n(n-1)}, \text{with } \mathbf{B} = \big((\mathbf{HKH}) \odot (\mathbf{HLH})\big)^{\cdot 2}, \tag{56}$$

where $\odot$ is the entrywise matrix product and $()^{\cdot 2}$ is the entrywise matrix power. The estimate in Eq. (55) has a bias of $\mathcal{O}(n^{-3})$ and can be calculated within time cost $\mathcal{O}(n^2)$.

B.2.4. Variance of $\mathrm{HSIC}_b(S)$ under $\mathcal{H}_0$

Since the additional terms of the bias vanish faster than Eq. (55), the result is identical to the case of unbiased.

B.3. Statistic under $\mathcal{H}_1$

We give the procedure for calculating the first two moments of the alternative distribution.

B.3.1. Mean of $\mathrm{HSIC}_u(S)$ and $\mathrm{HSIC}_b(S)$

By definition of unbiased estimator $\mathrm{HSIC}_u(S)$, we have $\mathbb{E}\mathrm{HSIC}_u(S) = \mathrm{HSIC}(X,Y)$, i.e., the mean of $\mathbb{E}\mathrm{HSIC}_u(S)$ is equal to the population mean $\mathrm{HSIC}(X,Y)$. And for the mean of $\mathrm{HSIC}_b(S)$, the result is $\mathrm{HSIC}(X,Y)+\mathcal{O}(n^{-1})$ since the difference between $\mathrm{HSIC}_u(S)$ and $\mathrm{HSIC}_b(S)$ is $\mathcal{O}(n^{-1})$ according to Eq. (47).

B.3.2. Variance of $\mathrm{HSIC}_u(S)$ and $\mathrm{HSIC}_b(S)$

Under $\mathcal{H}_1$, the term ζ_1 in Eq. (51) become positive. In this case, the variance becomes

$$\mathbf{Var}[\mathrm{HSIC}_u(S)] = \frac{16}{n}\zeta_1 + \mathcal{O}(n^{-2}) = \frac{16}{n}\Big(\mathbb{E}_i\big(\mathbb{E}_{j,q,r}h_{ijqr}\big)^2 - \mathrm{HSIC}(X,Y)^2\Big) + \mathcal{O}(n^{-2}). \tag{57}$$

In this paper, we denote

$$\sigma_u^2 := 16\Big(\mathbb{E}_i\big(\mathbb{E}_{j,q,r}h_{ijqr}\big)^2 - \mathrm{HSIC}(X,Y)^2\Big) \tag{58}$$

as the variance of $\sqrt{n}\mathrm{HSIC}_u(S)$. The variance of $\sqrt{n}\mathrm{HSIC}_b(S)$ are the same since the difference between them is given in Eq. (47) hence $\sqrt{n}(\mathrm{HSIC}_b(S) - \mathrm{HSIC}_u(S)) \sim \mathcal{O}(n^{-1/2})$. The estimator of Eq. (58) can be taken as

$$16 \cdot \Big(\frac{1}{n}\sum_i\big(\frac{1}{n^3}\sum_{j,q,r}h_{ijqr}\big)^2 - \big(\mathrm{HSIC}_b(S)\big)^2\Big). \tag{59}$$

The terms in Eq. (59) can be calculated within the time cost $\mathcal{O}(n^2)$. We mainly explain the calculation of $\sum_{j,q,r}h_{ijqr}$ here. We can express it with matrices by

$$\begin{aligned}
\sum_{j,q,r}h_{ijqr} &= \sum_{j,q,r}\frac{1}{4!}\sum_{(t,u,v,w)}^{(i,j,q,r)} k_{tu}l_{tu} + k_{tu}l_{vw} - 2k_{tu}l_{tv} \\
&= \frac{1}{4!}\sum_{j,q,r}\sum_{(u,v,w)}^{(j,q,r)}(k_{iu}l_{iu} + k_{iu}l_{vw} - 2k_{iu}l_{iv}) + \frac{1}{4!}\sum_{j,q,r}\sum_{(t,v,w)}^{(j,q,r)}(k_{ti}l_{ti} + k_{ti}l_{vw} - 2k_{ti}l_{tv}) \\
&\quad + \frac{1}{4!}\sum_{j,q,r}\sum_{(t,u,w)}^{(j,q,r)}(k_{tu}l_{tu} + k_{tu}l_{iw} - 2k_{tu}l_{ti})] + \frac{1}{4!}\sum_{j,q,r}\sum_{(t,u,v)}^{(j,q,r)}(k_{tu}l_{tu} + k_{tu}l_{vi} - 2k_{tu}l_{tv}) \\
&= \frac{1}{4!}\sum_{j,q,r}\sum_{(u,v,w)}^{(j,q,r)}(2k_{iu}l_{iu} + 2k_{iu}l_{vw} - 2k_{iu}l_{iv}) - \frac{1}{4!}\sum_{j,q,r}\sum_{(t,v,w)}^{(j,q,r)}(2k_{ti}l_{tv}) \\
&\quad + \frac{1}{4!}\sum_{j,q,r}\sum_{(t,u,w)}^{(j,q,r)}(2k_{tu}l_{tu} + 2k_{tu}l_{iw} - 2k_{tu}l_{ti}) - \frac{1}{4!}\sum_{j,q,r}\sum_{(t,u,v)}^{(j,q,r)}(2k_{tu}l_{tv}) \\
&= \frac{1}{2}\Big[n^2(\mathbf{KL})_{i,i} + (\mathbf{K1})_i(\mathbf{1}^T\mathbf{L1}) - n[(\mathbf{K1}) \odot (\mathbf{L1})]_i - n(\mathbf{KL1})_i \\
&\quad + n\mathrm{Tr}(\mathbf{KL}) + (\mathbf{L1})_i(\mathbf{1}^T\mathbf{K1}) - n(\mathbf{LK1})_i - (\mathbf{1}^T\mathbf{KL1})\Big],
\end{aligned} \tag{60}$$

where each term can be calculated within the time cost $\mathcal{O}(n^2)$.

B.4. Summary section

In the previous parts, we have given the asymptotic distribution (mainly Proposition 1) and the first two moments of the null and alternative distributions. In addition, we have explained that the first two moments of null and alternative distributions can be computed within time cost $\mathcal{O}(n^2)$. Here, we restate some of the important results for convenient reference in the following sections, as shown in the following.

Proposition 2. (Moments of Null Distribution). *Under $\mathcal{H}_0$, the estimation of mean with bias of $\mathcal{O}(n^{-1})$ to $\mathbb{E}[nHSIC_b(S)]$ can be given by*

$$\mathcal{E}_0 := 1 + \widehat{\|\mu_x\|^2}\widehat{\|\mu_y\|^2} - \widehat{\|\mu_y\|^2} - \widehat{\|\mu_x\|^2}, \tag{61}$$

where we assume $k_{ii} = l_{ii} = 1$ and the terms $\widehat{\|\mu_x\|^2} = \frac{1}{(n)_2}\sum_{(i,j)\in i_2^n} k_{ij}$, $\widehat{\|\mu_y\|^2} = \frac{1}{(n)_2}\sum_{(i,j)\in i_2^n} l_{ij}$. Also, the estimation of variance with bias of $\mathcal{O}(n^{-1})$ to $\boldsymbol{Var}[nHSIC_b(S)]$ can be given by

$$\mathcal{V}_0 = \frac{2(n-4)(n-5)}{(n-1)^2(n-2)(n-3)}\mathbf{1}^T\big(\mathbf{B} - diag(\mathbf{B})\big)\mathbf{1}, \tag{62}$$

where $\mathbf{B} = \big((\mathbf{HKH}) \odot (\mathbf{HLH})\big)^{\cdot 2} = (\mathbf{K}_c \odot \mathbf{L}_c)^{\cdot 2}$ and $\odot$ is the entrywise matrix product.

Appendix C. Assumptions

Below are some assumptions we required.

(i) The kernels k_{ω_0} and l_{ω_1} are uniformly bounded:

$$\sup_{\omega_0\in\Omega_0}\sup_{x\in\mathcal{X}} k_{\omega_0}(x,x) \le \nu, \quad \sup_{\omega_1\in\Omega_1}\sup_{y\in\mathcal{Y}} l_{\omega_1}(y,y) \le \nu.$$

(ii) The kernel parameters ω_0, ω_1 lie in Banach spaces of dimension D_0 and D_1 respectively. Furthermore, the set of possible kernel parameters, Ω_0 and Ω_1, are separately bounded by R_{ω_0} and R_{ω_1}, respectively, i.e.,

$$\Omega_0 \subseteq \left\{\omega_0 \middle|\ \|\omega_0\| \le R_{\Omega_0}\right\}, \Omega_1 \subseteq \left\{\omega_1 \middle|\ \|\omega_1\| \le R_{\Omega_1}\right\}.$$

(iii) The kernel parameterizations are Lipschitz continuous, i.e. for all $x, x' \in \mathcal{X}$, $\omega_0, \omega_0' \in \Omega_0$,

$$\left|k_{\omega_0}(x,x') - k_{\omega_0'}(x,x')\right| \le L_k \cdot \|\omega_0 - \omega_0'\|$$

and for all $y, y' \in \mathcal{Y}$, $\omega_1, \omega_1' \in \Omega_1$,

$$\left|l_{\omega_1}(y,y') - l_{\omega_1'}(y,y')\right| \le L_l \cdot \|\omega_1 - \omega_1'\|$$

with the nonnegative Lipschitz constant L_k, L_l.

(iv) **Boundedness.** The residuals $R_{X|Z}$, $R_{Y|Z}$ and their estimates $\hat{R}_{X|Z}$, $\hat{R}_{Y|Z}$ are bounded:

$$\sup_{(x,z)\in(\mathcal{X},\mathcal{Z})}\left|R_{X|Z}\right| \le \mu, \quad \sup_{(y,z)\in(\mathcal{Y},\mathcal{Z})}\left|R_{Y|Z}\right| \le \mu, \quad \sup_{(x,z)\in(\mathcal{X},\mathcal{Z})}\left|\hat{R}_{X|Z}\right| \le \mu, \quad \sup_{(y,z)\in(\mathcal{Y},\mathcal{Z})}\left|\hat{R}_{Y|Z}\right| \le \mu.$$

(v) **Eigenvalue decay (EVD).** Consider that $(\lambda_i)_{i\in I}$ is the eigenvalues of the integral operator [7] which is a covariance-like operator derived from the kernel function used in the regression model. For some $C_\lambda > 0$ and $p \in (0,1]$ and for all $i \in I$, $\lambda_i \le C_\lambda i^{-1/p}$.

(vi) **Source condition (SRC).** There exists a smoothness parameter $1 < \ell \le 2$ such that

$$\mu_{X|Z} \in [\mathrm{HS}_{xz}]^{\ell}, \ \mu_{Y|Z} \in [\mathrm{HS}_{yz}]^{\ell}.$$

Above, $[\mathrm{HS}_{xz}]^{\ell}$ refers to the interpolation space of the original HS space (also written as $[\mathrm{HS}]^1$). The eigenvalue decay of functions in $[\mathrm{HS}]^{\ell}$ is bounded by ℓ.

(vii) The kernels k, l with parameter ω_1, ω_2 are assumed to be Lipschitz continuous w.r.t the residual variables. Specifically, for any (x_1,y_1,z_1), (x_2,y_2,z_2), $(x_3,y_3,z_3) \in \mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$, the resulting residual variables $r_{x;i} = x_i - \mathbb{E}[X|z_i]$, $r_{y;i} = y_i - \mathbb{E}[Y|z_i]$ with $i = 1,2,3$, we assume that

$$\left|k_{\omega_1}(r_{x;1}, r_{x;2}) - k_{\omega_1}(r_{x;1}, r_{x;3})\right| \le C_k \cdot \|r_{x;2} - r_{x;3}\|, \quad \left|l_{\omega_2}(r_{y;1}, r_{r;2}) - l_{\omega_2}(r_{y;1}, r_{y;3})\right| \le C_l \cdot \|r_{y;2} - r_{y;3}\|,$$

where C_k and C_l are nonnegative Lipschitz constants.

The assumptions (i), (ii), (iii), and (vii) do not restrict the specific form of the kernels, and the kernels used in our paper satisfy these properties. The assumption (iv) are not overly restrictive when using appropriate regression algorithms. Assumptions (v) and (vi), which are well-established (see more details in [7,23]), are typically applied to evaluate the convergence properties of regression

methods. In this work, to match the required HS norm and the well-specified case, we adopt the approach of [31], where the regression is well-defined, with $\ell \in (1,2)$ and the necessary Hilbert space norm. The parameter ℓ indicates the smoothness of the regression.

Appendix D. Proof of Theorem 1

We restate the Theorem 1 here. The proof procedure is given in the order of convergence results 1-3.

Theorem 1 *(Uniform Bound). Let ω_0, ω_1 parameterize uniformly bounded kernels $k_{\omega_0}, l_{\omega_1}$ in Banach spaces of dimension D_0, D_1. And $k_{\omega_0}, l_{\omega_1}$ satisfy the Lipschitz condition in ω_0, ω_1 with the Lipschitz constant L_k, L_l. Let $\overline{\Omega}_c := \overline{\Omega}_0 \times \overline{\Omega}_1$ be a set of (ω_0, ω_1) for which $\sigma_u \geq c > 0$ with small constant c and $\|\omega_0\| \leq R_{\Omega_0}, \|\omega_1\| \leq R_{\Omega_1}$. Let r denote the threshold, i.e., $(1-\alpha)$-quantile for the asymptotic distribution in Eq. (42) and $r^{(n)}$ be the threshold with kernels of size n. Under Assumptions (i) to (iii), then with probability at least $1-\delta$,*

$$\sup_{(\omega_0,\omega_1)\in\overline{\Omega}_c} \left| \frac{HSIC_b(S) - r^{(n)}/n}{\widehat{\sigma}_u(S)} - \frac{HSIC(X,Y) - r/n}{\sigma_u} \right| \sim \mathcal{O}\left(\frac{1}{c^3} \left[\sqrt{\frac{1}{n}\log\frac{1}{\delta} + (D_0+D_1)\frac{\log n}{n}} + \frac{L_k + L_l}{\sqrt{n}} \right] \right).$$

D.1. Convergence results 1

Lemma 1. *Let ξ_{ω_0,ω_1} denote HSIC(X,Y) with the kernel parameters ω_0, ω_1, $\hat{\xi}^{(u)}_{\omega_0,\omega_1}$ denote the corresponding (unbiased) estimator of ξ_{ω_0,ω_1}, $\Delta^{(u)}_{\xi}(\omega_0,\omega_1) := \hat{\xi}^{(u)}_{\omega_0,\omega_1} - \xi_{\omega_0,\omega_1}$ represent random error function. $\hat{\xi}^{(b)}_{\omega_0,\omega_1}$ and $\Delta^{(b)}_{\xi}(\omega_0,\omega_1)$ are their biased counterparts. Under Assumptions (i) to (iii), then we have that with probability at least $1-\delta$,*

$$\sup_{\omega_0,\omega_1} \left| \Delta^{(b)}_{\xi}(\omega_0,\omega_1) \right| \leq 6\nu^2 \sqrt{\frac{2}{n}\log\frac{2}{\delta} + (D_0+D_1)\frac{\log n}{n}} + \frac{12\nu}{\sqrt{n}}(L_k \cdot R_{\Omega_0} + L_l \cdot R_{\Omega_1}). \tag{63}$$

Proof. We use McDiarmid's inequality to obtain the bound.

First, for fixed ω_0, ω_1, we show that $\Delta_{\xi}(\omega_0,\omega_1)$ fits the bounded differences property. Since we fix the kernel parameters in this part, for simplicity we omit the subscript ω_0, ω_1 from the statistics, e.g. shorten $\hat{\xi}^{(u)}_{\omega_0,\omega_1}$ to $\hat{\xi}$.

Then we replace (x_1, y_1) with (x_1', y_1') and keep the remaining samples the same. The newly obtained samples are named as S'. The difference between

$$\hat{\xi} := \frac{1}{(n)_2} \sum_{(i,j)\in \mathbf{i}_2^n} k_{ij} l_{ij} + \frac{1}{(n)_4} \sum_{(i,j,q,r)\in \mathbf{i}_4^n} k_{ij} l_{qr} - \frac{2}{(n)_3} \sum_{(i,j,q)\in \mathbf{i}_3^n} k_{ij} l_{iq} \tag{64}$$

and the new substitution $\hat{\xi}' := \mathrm{HSIC}_u(S')$ can be given by

$$\begin{aligned} \left| \hat{\xi} - \hat{\xi}' \right| \leq & \frac{1}{(n)_2} \sum_{(i,j)\in \mathbf{i}_2^n, ij:1\in\{i,j\}} \left| k_{ij} l_{ij} - k'_{ij} l'_{ij} \right| + \frac{1}{(n)_4} \sum_{(i,j,q,r)\in \mathbf{i}_4^n, ijqr:1\in\{i,j,q,r\}} \left| k_{ij} l_{qr} - k'_{ij} l'_{qr} \right| \\ & + \frac{2}{(n)_3} \sum_{(i,j,q)\in \mathbf{i}_3^n, ijq:1\in\{i,j,q\}} \left| k_{ij} l_{iq} - k'_{ij} l'_{iq} \right| \leq \frac{(n-1)_1}{(n)_2}\nu^2 + \frac{2(n-1)_3}{(n)_4}\nu^2 + \frac{3(n-1)_2}{(n)_3}\nu^2 = \frac{12\nu^2}{n}, \end{aligned} \tag{65}$$

since for all i, j, the term $k_{ij}, l_{ij}, k'_{ij}, l'_{ij}$ are all in the range $[0, \nu]$ by assumption (i) and notice that all the term that none of i, j, q, r are one is zero. Now using McDiarmid's inequality, we finish the first part of the proof, that is, for fixed ω_0, ω_1, with probability at least $1-\delta$,

$$\left| \Delta^{(u)}_{\xi}(\omega_0,\omega_1) \right| \leq 6\nu^2 \sqrt{\frac{2}{n}\log\frac{2}{\delta}}. \tag{66}$$

Next, we consider the case where ω_0, ω_1 changes. Take the parameter space Ω_0 as an example. Firstly since the parameter space is a compact Euclidean space, the covering number $\mathcal{N}(\Omega_0, r)$, defined as the smallest number of closed balls with centers in Ω_0 and radii r whose union covers Ω_0, is finite. According to [45, Proposition 4.2.12], by comparing the volumes, we have

$$\mathcal{N}(\Omega_0, r_0) \leq \left(\frac{2R_{\Omega_0}}{r_0} + 1 \right)^{D_0} \underbrace{\leq \left(\frac{3R_{\Omega_0}}{r_0} \right)^{D_0}}_{\text{when } r_0 \leq R_{\Omega_0}}. \tag{67}$$

As for Ω_1, we can lead a similar conclusion such that for given radii r_1, $\mathcal{N}(\Omega_1, r_1) \leq (3R_{\Omega_1}/r_1)^{D_1}$. Also, combining with the assumption (iii), we have for any two $\omega_0, \omega_0' \in \Omega_0$,

$$\begin{aligned}\left|\hat{\xi}^{(u)}_{\omega_0,\omega_1}-\hat{\xi}^{(u)}_{\omega_0',\omega_1}\right|\leq&\frac{\nu}{(n)_2}\sum_{(i,j)\in\mathbf{i}_2^n}|k_{ij}^{\omega_0}-k_{ij}^{\omega_0'}|+\frac{\nu}{(n)_4}\sum_{(i,j,q,r)\in\mathbf{i}_4^n}|k_{ij}^{\omega_0}-k_{ij}^{\omega_0'}|\\&+\frac{2\nu}{(n)_3}\sum_{(i,j,q)\in\mathbf{i}_3^n}|k_{ij}^{\omega_0}-k_{ij}^{\omega_0'}|\leq 4\nu L_k\|\omega_0-\omega_0'\|,\end{aligned}\tag{68}$$

and using the property of unbiased estimate, then

$$\left|\xi_{\omega_0,\omega_1}-\xi_{\omega_0',\omega_1}\right|=\left|\mathbb{E}\hat{\xi}^{(u)}_{\omega_0,\omega_1}-\mathbb{E}\hat{\xi}^{(u)}_{\omega_0',\omega_1}\right|\leq\mathbb{E}\left[\left|\hat{\xi}^{(u)}_{\omega_0,\omega_1}-\hat{\xi}^{(u)}_{\omega_0',\omega_1}\right|\right]\leq 4\nu L_k\|\omega_0-\omega_0'\|.\tag{69}$$

The above analysis also applies to $\omega_1,\omega_1'\in\Omega_1$ due to the symmetry, i.e.

$$\left|\xi_{\omega_0,\omega_1}-\xi_{\omega_0,\omega_1'}\right|\leq\left|\hat{\xi}^{(u)}_{\omega_0,\omega_1}-\hat{\xi}^{(u)}_{\omega_0',\omega_1}\right|\leq 4\nu L_l\|\omega_1-\omega_1'\|.\tag{70}$$

As a result, for any point $(\omega_0,\omega_1)\in\Omega_0\times\Omega_1$, we can find a point $(\omega_{0,i},\omega_{1,j})$ in the cover set such that

$$\|\omega_{0,i}-\omega_0\|\leq r_0,\ \|\omega_{1,j}-\omega_1\|\leq r_1\tag{71}$$

and also

$$\left|\Delta^{(u)}_{\xi}(\omega_0,\omega_1)\right|\leq\left|\Delta^{(u)}_{\xi}(\omega_{0,i},\omega_{1,j})\right|+4\nu L_k\cdot r_0+4\nu L_l\cdot r_1.\tag{72}$$

Combining with the result in the first part, we show that with probability at least $1-\delta$,

$$\begin{aligned}\sup_{\omega_0,\omega_1}\left|\Delta^{(b)}_{\xi}(\omega_0,\omega_1)\right|\leq&\max_{ij}\left|\Delta^{(u)}_{\xi}(\omega_{0,i},\omega_{1,j})\right|+4\nu L_k\cdot r_0+4\nu L_l\cdot r_1\\\leq&6\nu^2\sqrt{\frac{2}{n}\log\frac{2\mathcal{N}(\Omega_0,r_0)\mathcal{N}(\Omega_1,r_1)}{\delta}}+4\nu L_k\cdot r_0+4\nu L_l\cdot r_1.\end{aligned}\tag{73}$$

We finish the proof of this part by combining Eq. (67) and setting the radis $r_0=3R_{\Omega_0}/\sqrt{n}$ and $r_1=3R_{\Omega_1}/\sqrt{n}$. □

The analysis of the case of biased statistics we used in practice is the same as the unbiased case in our analysis. The reason is that they are bounded by a negligible gap against other dominant terms like $\mathcal{O}(\frac{1}{\sqrt{n}})$ in asymptotic analysis, which is shown by Lemma 2.

Lemma 2. *The bias term between the U-statistic estimator $HSIC_u(S)$ and V-statistic estimator $HSIC_b(S)$ is asymptotically bounded by $\mathcal{O}(n^{-1})$, that is,*

$$|HSIC_b(S)-HSIC_u(S)|\sim\mathcal{O}(n^{-1}).\tag{74}$$

Proof. By definition, we have

$$\begin{aligned}|\mathrm{HSIC}_b(S)-\mathrm{HSIC}_u(S)|\leq&\left|\frac{1}{(n)_2}\sum_{(i,j)\in\mathbf{i}_2^n}k_{ij}l_{ij}-\frac{1}{n^2}\sum_{i,j}k_{ij}l_{ij}\right|\\&+\left|\frac{1}{(n)_4}\sum_{(i,j,q,r)\in\mathbf{i}_4^n}k_{ij}l_{qr}-\frac{1}{n^4}\sum_{i,j,q,r}k_{ij}l_{qr}\right|+2\cdot\left|\frac{1}{(n)_3}\sum_{(i,j,q)\in\mathbf{i}_3^n}k_{ij}l_{iq}-\frac{1}{n^3}\sum_{i,j,q}k_{ij}l_{iq}\right|.\end{aligned}\tag{75}$$

Take the first term in Eq. (75) as an example,

$$\left|\frac{1}{(n)_2}\sum_{(i,j)\in\mathbf{i}_2^n}k_{ij}l_{ij}-\frac{1}{n^2}\sum_{i,j}k_{ij}l_{ij}\right|=\left|\left(\frac{1}{(n)_2}-\frac{1}{n^2}\right)\sum_{(i,j)\in\mathbf{i}_2^n}k_{ij}l_{ij}-\frac{1}{n^2}\sum_{i}k_{ii}l_{ii}\right|\leq\frac{\nu^2}{n},\tag{76}$$

since for all i,j, k_{ij},l_{ij} are in the range $[0,\nu]$. The same process can be applied to the second and third terms. Therefore, the total difference of $\mathrm{HSIC}_b(S)$ and $\mathrm{HSIC}_u(S)$ is

$$|\mathrm{HSIC}_b(S)-\mathrm{HSIC}_u(S)|\leq\frac{\nu^2}{n}+\frac{6\nu^2}{n}+2\cdot\frac{3\nu^2}{n}=\frac{13\nu^2}{n}.\tag{77}$$

Thus, the lemma is proved. □

D.2. Convergence results 2

As a start, we denote the random error

$$\Delta^{(b)}_{\sigma}(\omega_0,\omega_1):=\left|\widehat{\sigma}^2_u(S)-\sigma^2_u\right|,\tag{78}$$

where the variance is

$$\sigma_u^2 = 16\left(\mathbb{E}_i\left(\mathbb{E}_{j,q,r} h_{ijqr}\right)^2 - \mathrm{HSIC}(X,Y)^2\right) \tag{79}$$

and an estimate without regularization is

$$\widehat{\sigma}_u^2(S) = 16\cdot\left(\frac{1}{n}\sum_i\left(\frac{1}{n^3}\sum_{j,q,r} h_{ijqr}\right)^2 - \left(\mathrm{HSIC}_b(S)\right)^2\right), \tag{80}$$

where the term

$$h_{ijqr} = \frac{1}{4!}\sum_{(t,u,v,w)}^{(i,j,q,r)} k_{tu}l_{tu} + k_{tu}l_{vw} - 2k_{uv}l_{tv} \tag{81}$$

and the sum represents all ordered quadruples (t,u,v,w) drawn without replacement from (i,j,q,r). Also, the statistic $\mathrm{HSIC}_b(S)$ can be expressed as

$$\mathrm{HSIC}_b(S) = \frac{1}{n^4}\sum_{i,j,q,r}^{n} h_{ijqr}. \tag{82}$$

Lemma 3. *Under Assumptions (i) to (iii), then we have that with probability at least* $1-\delta$*,*

$$\sup_{\omega_0,\omega_1}\left|\Delta_\sigma^{(u)}(\omega_0,\omega_1)\right| \le 768v^4\sqrt{\frac{2}{n}\log\frac{2}{\delta} + (D_0+D_1)\frac{\log n}{n}} + \frac{6272v^4}{n} + \frac{1536v^3}{\sqrt{n}}\left(L_k\cdot R_{\Omega_0} + L_l\cdot R_{\Omega_1}\right). \tag{83}$$

Proof. First, we obtain the bound on h_{ijqr} with fixed i,j,q,r.

$$|h_{ijqr}| \le \frac{1}{4!}\sum_{(t,u,v,w)}^{(i,j,q,r)} |k_{tu}l_{tu} + k_{tu}l_{vw} - 2k_{uv}l_{tv}| \le 2v^2, \tag{84}$$

since for all i,j, k_{ij} and l_{ij} have range in $[0,v]$.

Suppose we change (x_1,y_1) to (x_1',y_1') and keep the remaining samples the same as before. The newly obtained samples are named as S'. We denote the counterpart of h_{ijqr} calculated on S' as h'_{ijqr}. Now we require an upper bound on $|\widehat{\sigma}_u^2(S) - \widehat{\sigma}_u^2(S')|$. For the first term in Eq. (80), we have

$$\begin{aligned}
&\left|\frac{1}{n}\sum_i\left(\frac{1}{n^3}\sum_{j,q,r} h_{ijqr}\right)^2 - \frac{1}{n}\sum_i\left(\frac{1}{n^3}\sum_{j,q,r} h'_{ijqr}\right)^2\right| \\
\le&\frac{1}{n}\sum_i\Bigg(\underbrace{\frac{1}{n^3}\sum_{j,q,r}\left|h_{ijqr} - h'_{ijqr}\right|}_{\le 12v^2/n \text{ similar as Eq. (65)}}\Bigg)\cdot\Bigg(\underbrace{\frac{1}{n^3}\sum_{j,q,r}\left|h_{ijqr} + h'_{ijqr}\right|}_{\le 4v^2 \text{ since } |h_{ijqr}|\le 2}\Bigg) \le \frac{48v^4}{n}.
\end{aligned} \tag{85}$$

For the second term in Eq. (80), denote $\hat{\xi}_b = \mathrm{HSIC}_b(S)$, $\hat{\xi}_b' = \mathrm{HSIC}_b(S')$ (for this part only), we have

$$\left|\left(\hat{\xi}_b\right)^2 - \left(\hat{\xi}_b'\right)^2\right| = \left|\hat{\xi}_b + \hat{\xi}_b'\right|\cdot\left|\hat{\xi}_b - \hat{\xi}_b'\right| \le 4v^2\cdot\frac{12v^2}{n} = \frac{48v^4}{n}, \tag{86}$$

since $\hat{\xi}_b, \hat{\xi}_b'$ is bounded in $[0,2v]$ and the bound on difference term can be obtained by a similar analysis in Eq. (65). Therefore,

$$|\widehat{\sigma}_u^2(S) - \widehat{\sigma}_u^2(S')| \le \frac{1536v^4}{n}. \tag{87}$$

Simply applying McDiamid's to $\widehat{\sigma}_u^2(S)$, we obtain that with probability at least $1-\delta$,

$$|\widehat{\sigma}_u^2(S) - \mathbb{E}[\widehat{\sigma}_u^2(S)]| \le 768v^4\sqrt{\frac{2}{n}\log\frac{2}{\delta}}. \tag{88}$$

Now we consider the bound of bias term $|\mathbb{E}[\widehat{\sigma}_u^2(S)] - \sigma_u^2|$. By definition, We can rewrite the first subterm as

$$\mathbb{E}[\widehat{\sigma}_u^2(S)] = 16\Big(\frac{1}{n^7}\sum_{ijqrj'q'r'}\mathbb{E}[h_{ijqr}h_{ij'q'r'}] - \frac{1}{n^8}\sum_{ijqri'j'q'r'}\mathbb{E}[h_{ijqr}h_{i'j'q'r'}]\Big) \tag{89}$$

and rewrite the second subterm with matching as

$$\sigma_u^2 = 16\Big(\frac{1}{n^7}\sum_{ijqrj'q'r'}\mathbb{E}[h_{1234}h_{1678}] - \frac{1}{n^8}\sum_{ijqri'j'q'r'}\mathbb{E}[h_{1234}h_{5678}]\Big), \tag{90}$$

where the number subscripts represent a specific set of values, then by calculating the number of non-zero terms and combining Eq. (84), we obtain that

$$|\mathbb{E}[\widehat{\sigma}_u^2(S)] - \sigma_u^2| \le 16 \cdot \underbrace{\left(2 - \frac{(n)_7}{n^7} - \frac{(n)_8}{n^8}\right)}_{\text{by the number of non-zero terms}} \cdot 8v^4. \tag{91}$$

Also, when $n > 8$, it results in

$$|\mathbb{E}[\widehat{\sigma}_u^2(S)] - \sigma_u^2| \le 16 \cdot \frac{\binom{7}{2} + \binom{8}{2}}{n} \cdot 8v^4 = \frac{6272v^4}{n}. \tag{92}$$

Next, we consider the case where ω_0, ω_1 changes. To simplify, in this part, we denote $\omega := (\omega_0, \omega_1)$ and $\omega' := (\omega_0', \omega_1')$. The superscript indicates the value taken under the corresponding parameter value, as an example, $k_{tu}^{(\omega)} = k_{tu}^{(\omega_0)}$ and $h_{ijqr}^{(\omega)} = h_{ijqr}^{(\omega_0,\omega_1)}$. Now we prove the Lipschitz property of $\widehat{\sigma}_{u,\omega}^2(S)$ and $\sigma_{u,\omega}^2$. One can check that for fixed a, b, c, d,

$$\left|k_{ab}^{(\omega)} l_{cd}^{(\omega)} - k_{ab}^{(\omega')} l_{cd}^{(\omega')}\right| \le \left|k_{ab}^{(\omega)}\right| \cdot \left|l_{cd}^{(\omega)} - l_{cd}^{(\omega')}\right| + \left|k_{ab}^{(\omega)} - k_{ab}^{(\omega')}\right| \cdot \left|l_{cd}^{(\omega')}\right| \le v\left(L_k\|\omega_0 - \omega_0'\| + L_l\|\omega_1 - \omega_1'\|\right), \tag{93}$$

by assumptions (i) and (iii). Then combining with Eq. (81), for each term we can use the results of Eq. (93), then

$$|h_{ijqr}^{(\omega)} - h_{ijqr}^{(\omega')}| \le 4v\left(L_k\|\omega_0 - \omega_0'\| + L_l\|\omega_1 - \omega_1'\|\right). \tag{94}$$

Also, since $|h_{ijqr}| \le 2v$,

$$\left|h_{ijqr}^{(\omega)} h_{abcd}^{(\omega)} - h_{ijqr}^{(\omega')} h_{abcd}^{(\omega')}\right| \le \left|h_{ijqr}^{(\omega)}\right| \cdot \left|h_{abcd}^{(\omega)} - h_{abcd}^{(\omega')}\right| + \left|h_{ijqr}^{(\omega)} - h_{ijqr}^{(\omega')}\right| \cdot \left|h_{abcd}^{(\omega')}\right| \le 16v^3\left(L_k\|\omega_0 - \omega_0'\| + L_l\|\omega_1 - \omega_1'\|\right). \tag{95}$$

As a result,

$$\begin{aligned}\left|\widehat{\sigma}_{u,\omega}^2(S) - \widehat{\sigma}_{u,\omega'}^2(S)\right| &\le \frac{16}{n^7} \sum_{ijqrbcd} \left|h_{ijqr}^{(\omega)} h_{ibcd}^{(\omega)} - h_{ijqr}^{(\omega')} h_{ibcd}^{(\omega')}\right| + \frac{16}{n^8} \sum_{ijqrabcd} \left|h_{ijqr}^{(\omega)} h_{abcd}^{(\omega)} - h_{ijqr}^{(\omega')} h_{abcd}^{(\omega')}\right| \\ &\le 512v^3\left(L_k\|\omega_0 - \omega_0'\| + L_l\|\omega_1 - \omega_1'\|\right).\end{aligned} \tag{96}$$

Again using a similar process, we can show that

$$\left|\sigma_{u,\omega}^2 - \sigma_{u,\omega'}^2\right| \le 512v^3\left(L_k\|\omega_0 - \omega_0'\| + L_l\|\omega_1 - \omega_1'\|\right). \tag{97}$$

Since we focus on the same parameter space as before, we take the same cover set we used in Eq. (67). Then, by combining the results of Eqs. (88), (92), (96), and (97), we have, with probability at least $1 - \delta$,

$$\sup_{\omega_0,\omega_1} \left|\Delta_\sigma^{(u)}(\omega_0, \omega_1)\right| \le 768v^4\sqrt{\frac{2}{n}\log\frac{2\mathcal{N}(\Omega_0, r_0)\mathcal{N}(\Omega_1, r_1)}{\delta}} + \frac{6272v^4}{n} + 512v^3\left(L_k \cdot r_0 + L_l \cdot r_1\right). \tag{98}$$

We finish the proof of this part by combining Eq. (67) and setting the radis $r_0 = 3R_{\Omega_0}/\sqrt{n}$ and $r_1 = 3R_{\Omega_1}/\sqrt{n}$. □

D.3. Convergence results 3

Recall that c is a small constant, we can set it to less than 1 for analysis. Since $\sigma_u \ge c$ on $\overline{\Omega}_c := \overline{\Omega}_0 \times \overline{\Omega}_1$, according to the Eq. (98), when

$$n \ge N_0 = \left(\frac{2}{c}\right)^3 \cdot \left[768v^4\left(\sqrt{2\log\frac{4}{\delta}} + \sqrt{D_0 + D_1}\right) + 1536v^3\left(L_k R_{\Omega_0} + L_l R_{\Omega_1}\right) + 6272v^4\right]^3, \tag{99}$$

we have $\widehat{\sigma}_b(S) \ge c/2$ with at least probability $1 - \delta/2$. And for simplicity, we denote $\widehat{\sigma}_u$ as σ, $\widehat{\sigma}_b(S)$ as $\widehat{\sigma}$, $\mathrm{HSIC}_b(S)$ as $\hat{\xi}$ and $\mathrm{HSIC}(X, Y)$ as ξ, then

$$\begin{aligned}\sup_{\omega_0,\omega_1} \left|\frac{\hat{\xi} - r^{(n)}/n}{\widehat{\sigma}} - \frac{\xi - r/n}{\sigma}\right| &\le \sup_{\omega_0,\omega_1} \left(\left|\frac{\hat{\xi} - \xi}{\widehat{\sigma}}\right| + \left|\frac{r^{(n)} - r}{n\widehat{\sigma}}\right| + \left|\frac{\xi + r/n}{\sigma}\right| \cdot \left|\frac{\widehat{\sigma} - \sigma}{\widehat{\sigma}}\right|\right) \\ &\le \sup_{\omega_0,\omega_1} \left(\frac{2}{c} \cdot \left|\hat{\xi} - \xi\right| + \frac{2}{cn} \cdot \left|r^{(n)} - r\right| + \frac{8(v + r/n)}{3c^3} \cdot \left|\widehat{\sigma}^2 - \sigma^2\right|\right).\end{aligned} \tag{100}$$

Combining the results Eqs. (73) and (98), also according to the results [19, Theorem 13] that shown that $|r^n/n - r/n| \sim o(n^{-1/2})$ and $\sup_{\omega_0,\omega_1} r < \infty$, thus we have

$$\sup_{\omega_0,\omega_1}\left|\frac{\hat{\xi}-r^{(n)}/n}{\hat{\sigma}}-\frac{\xi-r/n}{\sigma}\right|=\mathcal{O}\left(\frac{1}{c^3}\left[\sqrt{\frac{2}{n}\log\frac{4}{\delta}+(D_0+D_1)\frac{\log n}{n}}+\frac{L_k R_{\Omega_0}+L_l R_{\Omega_1}}{\sqrt{n}}\right]\right)=\mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right), \tag{101}$$

with probability at least $1-\delta$.

Appendix E. Proof of Proposition 3

Proposition 3 *(Consistency of Independence Test). Let ω_0^*,ω_1^* be the kernel parameters after learning, S^{te} be the testing samples of size m, then the probability of Type II error*

$$\mathbb{P}_{\mathcal{H}_1}\left(mHSIC_b(S^{te})\le r^{(m)}|\omega_0^*,\omega_1^*\right)\sim\mathcal{O}(m^{-1/2}). \tag{102}$$

Proof. For simplify, we denote $\hat{\xi}_b$ as the biased estimator of $\mathrm{HSIC}_b(S^{te})$ and $\hat{\xi}_u$ as the corresponding unbiased estimate. We first show a uniform bound of the threshold and then give the upper bound of the probability of Type II error.

Bound on the threshold. After the training process, let the fixed kernel bandwidths we find as ω_0^*,ω_1^*. For these fixed kernel bandwidths, according to the process in Eq. (66), we can have a uniform bound for $r^{(m)}$ for $m\hat{\xi}_b$ that

$$r^{(m)}\le 6v^2\sqrt{2\log\frac{2}{1-\alpha}}\sqrt{m}+13v^2\sim\mathcal{O}(m^{1/2}). \tag{103}$$

Decrease rate of Type II error. By definition, the probability of the Type II error is given by

$$\begin{aligned}\mathbb{P}\Big(\text{Type II error}\Big)&=\mathbb{P}_{\mathcal{H}_1}\left(m\hat{\xi}_b\le r^{(m)}|\omega_0^*,\omega_1^*\right)\le\mathbb{P}_{\mathcal{H}_1}\left(m\hat{\xi}_u\le r^{(m)}+13v^2\Big|\omega_0^*,\omega_1^*\right)\\&=\mathbb{P}_{\mathcal{H}_1}\left(\frac{\sqrt{m}(\hat{\xi}_u-\mathbb{E}\hat{\xi}_u)}{4\sigma^{1/2}}\le\frac{r^{(m)}/\sqrt{m}-\sqrt{m}\mathbb{E}\hat{\xi}_u+13v^2/\sqrt{m}}{4\sigma^{1/2}}\Big|\omega_0^*,\omega_1^*\right).\end{aligned} \tag{104}$$

According to the results shown in [37, Section 5.5.1 Theorem B], and also $\sigma>0$ under $\mathcal{H}_1$, we have

$$\begin{aligned}\mathbb{P}\Big(\text{Type II error}\Big)&\le\Phi\left(\frac{r^{(m)}/\sqrt{m}-\sqrt{m}\mathbb{E}\hat{\xi}_u+13v^2/\sqrt{m}}{4\sigma^{1/2}}\right)+\frac{C_1 v_h}{\sigma^{3/2}}\frac{1}{\sqrt{m}}\\&\le\Phi\Big(C_2-C_3\sqrt{m}\mathbb{E}\hat{\xi}_u+C_4/\sqrt{m}\Big)+C_5\frac{1}{\sqrt{m}}\end{aligned} \tag{105}$$

where $v_h:=\mathbb{E}|h_{1234}|^3<\infty$ and using $r^{(m)}\sim\mathcal{O}(m^{1/2})$ we prove before. Since under $\mathcal{H}_1$, $\xi>0$, hence $\mathbb{E}\hat{\xi}_u=\xi>0$. For the function $\Phi(x)$, we consider the asymptotic expansion when x is close to negative infinity, that is

$$\Phi(x)=-\frac{e^{-x^2}}{2x\sqrt{\pi}}\left(1+\sum_{n=1}^{\infty}(-1)^n\frac{1\cdot3\cdot5\cdots(2n-1)}{\left(2x^2\right)^n}\right), \tag{106}$$

then $\Phi\Big(C_2-C_3\sqrt{m}\mathbb{E}\hat{\xi}_u+C_4/\sqrt{m}\Big)\sim\mathcal{O}(m^{-1/2})$. As a result, the decreasing rate is at least $\mathcal{O}(m^{-1/2})$. □

Appendix F. Gamma approximation of threshold

We first restate the definition of $\widehat{c_\alpha}$. Under $\mathcal{H}_0$, we approximate the cumulative distribution function with the two-parameter gamma distribution

$$n\mathrm{HSIC}_b(S)\sim\frac{x^{\gamma-1}e^{-x/\beta}}{\beta^\gamma\Gamma(\gamma)}, \tag{107}$$

where the two parameters are

$$\gamma=\frac{(\mathbb{E}[\mathrm{HSIC}_b(S)])^2}{\mathbf{Var}[\mathrm{HSIC}_b(S)]},\ \beta=\frac{n\mathbf{Var}[\mathrm{HSIC}_b(S)]}{\mathbb{E}[\mathrm{HSIC}_b(S)]}. \tag{108}$$

Assume that $k_{ii}=l_{ii}=1$ (the Gaussian kernels satisfy this condition), then the mean of the statistic, denoted $\mathbb{E}[\mathrm{HSIC}_b(S)]$, is obtained as follows,

$$\frac{1}{n}\mathrm{Tr}C_{xx}\mathrm{Tr}C_{yy}=\frac{1}{n}\Big(1+\|\mu_x\|^2\|\mu_y\|^2-\|\mu_x\|^2-\|\mu_y\|^2\Big). \tag{109}$$

An empirical estimate can be obtained by replacing the term above with

$$\widehat{\|\mu_x\|^2}=\frac{1}{(n)_2}\sum_{(i,j)\in i_2^n}k_{ij},\ \widehat{\|\mu_y\|^2}=\frac{1}{(n)_2}\sum_{(i,j)\in i_2^n}l_{ij}. \tag{110}$$

Alternatively, a matrix expression

$$\mathbb{E}[\mathrm{HSIC}_b(S)] = \frac{1}{n}\Big(\frac{1}{n-1}\mathrm{Tr}(\mathbf{HKH})\Big)\Big(\frac{1}{n-1}\mathrm{Tr}(\mathbf{HLH})\Big), \tag{111}$$

where $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is the centering matrix. Also, the variance of the statistic is obtained by

$$\mathbf{Var}[\mathrm{HSIC}_b(S)] = \frac{2(n-4)(n-5)}{(n)_4}\|C_{xx}\|^2_{\mathrm{HS}}\|C_{yy}\|^2_{\mathrm{HS}}, \tag{112}$$

where $\|\cdot\|^2_{\mathrm{HS}}$ is the Hilbert-Schmidt norm. The empirical estimate of the product of HS norms $\|C_{xx}\|^2_{\mathrm{HS}}\|C_{yy}\|^2_{\mathrm{HS}}$ is

$$\frac{\mathbf{1}^T\big(\mathbf{B} - \mathrm{diag}(\mathbf{B})\big)\mathbf{1}}{n(n-1)}, \text{with } \mathbf{B} = \big((\mathbf{HKH}) \odot (\mathbf{HLH})\big)^{\cdot 2}, \tag{113}$$

where $\odot$ is the entrywise matrix product and $()^{\cdot 2}$ is the entrywise matrix power.

Next, we begin to discuss the properties of functions. For simplify, we denote $\mathbf{K_c} = \mathbf{HKH}$ and $\mathbf{L_c} = \mathbf{HLH}$. Also, we construct the following function to be analyzed

$$\mathcal{E}_0 = \Big(\frac{1}{n-1}\mathrm{Tr}\mathbf{K_c}\Big)\Big(\frac{1}{n-1}\mathrm{Tr}\mathbf{L_c}\Big), \mathcal{V}_0 = \frac{2(n-4)(n-5)}{(n-1)^2(n-2)(n-3)}\mathbf{1}^T\big(\mathbf{B} - \mathrm{diag}(\mathbf{B})\big)\mathbf{1}, \tag{114}$$

where $\mathbf{B} = (\mathbf{K_c} \odot \mathbf{L_c})^{\cdot 2}$, and one can check that

$$\mathcal{E}_0 = \mathbb{E}[n\mathrm{HSIC}_b(S)], \mathcal{V}_0 = \mathbf{Var}[n\mathrm{HSIC}_b(S)]. \tag{115}$$

The estimate of threshold $\hat{c_\alpha}$ with a confidence level of $1-\alpha$ is the solution of the following equation

$$\int_0^{\hat{c_\alpha}} \frac{x^{\gamma-1}e^{-x/\beta}}{\beta^\gamma\Gamma(\gamma)}dx = \int_0^{\frac{\hat{c_\alpha}}{\beta}} \frac{x^{\gamma-1}e^{-x}}{\Gamma(\gamma)}dx = 1-\alpha, \tag{116}$$

where Γ is the gamma function.

Appendix G. Limit behavior with Gaussian kernels

In this section, we study the limiting behavior (with respect to the bandwidth) of the statistics when the kernel function is taken to be a Gaussian function, in order to show the detailed calculations in the examples given in the main paper and to motivate Theorem 2. For analysis, we consider the Gaussian kernel with the form

$$k(x, x') = \exp\big(-\eta \cdot \|x - x'\|^2\big), \tag{117}$$

where the parameter $\eta = 1/2\omega_x^2$. We also consider the case with fixed samples S of sample size n and fixed width $\omega_y > 0$, then explore the behavior of the statistics of null and alternative distribution. In addition to the results at $\omega_x = 0$ (i.e., $\eta = +\infty$) shown in the main paper, we also show the results at $\eta = 0^+$. As a start, we begin by stating the following assumptions.

1. The domain $\mathcal{X}$ is Euclidean and bounded, $\mathcal{X} \subseteq \{x \in \mathbb{R}^d : \|x\| \le R_{\mathcal{X}}/2\}$ for some constant $R_{\mathcal{X}} < \infty$.
2. The non-diagonal elements of center matrices $\mathbf{K}_c, \mathbf{L}_c$ are not zero, i.e. $(\mathbf{K}_c)^2_{ij} > 0, (\mathbf{L}_c)^2_{ij} > 0$ for all $i \ne j$ when the kernel widths $(\omega_x, \omega_y) \in [\omega_{xl}, \omega_{xu}] \times [\omega_{yl}, \omega_{yu}]$ with given positive constants ω_{xl}, ω_{xu}, ω_{yl}, ω_{yu}.
3. The distributions of data are continuous. Hence $\|x_i - x_j\| \ne 0, \|y_i - y_j\| \ne 0$ for all $i \ne j$ almost surely.

Compared to assumption 3, which requires the data to have a continuous distribution, assumptions 1 and 2 are weaker, and we focus on analyzing assumption 2 due to it relates to both samples and bandwidth. Since by the definition $\mathbf{K}_c = \mathbf{HKH}$, hence

$$(\mathbf{K}_c)_{ij} = k_{ij} - \frac{1}{n}\sum_i k_{ij} - \frac{1}{n}\sum_j k_{ij} + \frac{1}{n^2}\sum_{ij} k_{ij}. \tag{118}$$

Intuitively, the value of $(\mathbf{K}_c)_{ij}$ is equal to itself minus the average of the row and column it is in and plus the average of all the elements. As a result, in practice, we hardly ever face a situation where it is strictly equal to 0 when the bandwidth is a positive constant. Under these assumptions, we now show the limit behavior.

We first get the limit of the kernel matrix

$$\mathbf{K}(\eta = 0^+) = \mathbf{1}\mathbf{1}^T, \mathbf{K}(\eta = +\infty) = \mathbf{I}, \tag{119}$$

and the limit of centering kernel matrix

$$\mathbf{K}_c(\eta = 0^+) = \mathbf{O}, \mathbf{K}_c(\eta = +\infty) = \mathbf{H}. \tag{120}$$

Then by substituting the results of Eqs. (119) and (120) gives the limit of two moments of null distribution

$$\begin{aligned}&\mathcal{E}_0(\eta=0^+)=0,\ \mathcal{E}_0(\eta=+\infty)=\Big(\frac{1}{n-1}\mathrm{Tr}\mathbf{L_c}\Big).\\&\mathcal{V}_0(\eta=0^+)=0,\ \mathcal{V}_0(\eta=+\infty)=\frac{2(n-4)(n-5)}{n^2(n-1)^2(n-2)(n-3)}\sum_{i\neq j}(\mathbf{L_c})_{ij}^2.\end{aligned}\tag{121}$$

Now we consider the limit behavior of the moments of alternative distribution. Recall the definitions, the mean of the distribution under $\mathcal{H}_1$ is given by

$$\mathcal{E}_1 := n\mathrm{HSIC}_b(S)=\frac{1}{n}\mathrm{Tr}(\mathbf{HKHL}).\tag{122}$$

The limit is

$$\mathcal{E}_1(\eta=0^+)=0,\mathcal{E}_1(\eta=+\infty)=\frac{1}{n}\mathrm{Tr}(\mathbf{L_c}).\tag{123}$$

And the variance is

$$\mathcal{V}_1 := \widehat{\sigma}_u^2(S)=16\cdot\left(\frac{1}{n}\sum_i\Big(\frac{1}{n^3}\sum_{j,q,r}h_{ijqr}\Big)^2-\Big(\mathrm{HSIC}_b(S)\Big)^2\right),\tag{124}$$

where the term

$$h_{ijqr}=\frac{1}{4!}\sum_{(t,u,v,w)}^{(i,j,q,r)}k_{tu}l_{tu}+k_{tu}l_{vw}-2k_{uv}l_{tv}\tag{125}$$

and the sum represents all ordered quadruples (t,u,v,w) drawn without replacement from (i,j,q,r). Also, we can show that the statistic $\mathrm{HSIC}_b(S)$ can be expressed as

$$\mathrm{HSIC}_b(S)=\frac{1}{n^4}\sum_{i,j,q,r}^{n}h_{ijqr}.\tag{126}$$

To compute the limit easily, we can also make use of its matrix expression of the term given in Eq. (60). Then

$$\left[\sum_{j,q,r}h_{ijqr}\right]\Big|_{\eta=0^+}=0,\left[\sum_{j,q,r}h_{ijqr}\right]\Big|_{\eta=+\infty}=\frac{n^2(\mathbf{L}_c)_{ii}+n\mathrm{Tr}(\mathbf{L_c})}{2}.\tag{127}$$

After that, the limit of Eq. (124) can be calculated that

$$\mathcal{V}_1\Big|_{\eta=0^+}=0,\mathcal{V}_1\Big|_{\eta=+\infty}=\frac{4}{n^2}\left[\frac{\mathrm{Tr}\big[(\mathbf{L_c})^{\cdot 2}\big]}{n}-\Big(\frac{\mathrm{Tr}(\mathbf{L_c})}{n}\Big)^2\right].\tag{128}$$

In summary, the variance $\mathcal{V}_0,\mathcal{V}_1$ are very small when η is taken to both 0 and $+\infty$ (when n is large), but the mean $\mathcal{E}_0,\mathcal{E}_1$ get constant value when $\eta=+\infty$. This further explains the overfitting problem in the main paper, i.e., disregarding the threshold at $\omega_x=0^+$ will result in very large values. In addition, the behavior of the variance in the limit encourages us to add a small regular constant to the denominator of the objective function to prevent numerical errors in practice. We will employ this in the next section and further show the result of the smoothness of the objective function.

Appendix H. Proof of Theorem 2

In this section, we give the proof of Theorem 2. In the beginning, we restate the Theorem 2 below. Our proof is based on assumptions 1 and 2, which as explained in the previous section almost hold in practice. For simplicity, we set $R_{\mathcal{X}}=1$ which can be achieved by normalization. In addition, we assume for x_i there exists $i\neq j,\|x_i-x_j\|>0$, i.e. all points x_i are prevented from taking the same value, and so are y_i. Note that this assumption is used to simplify the analysis but is not necessary due to the fact that if $x_i=x_j$ for all $i\neq j$, then $\mathbf{K}\equiv\mathbf{1}\mathbf{1}^T$ for all positive bandwidth thus $J_\lambda(S)\equiv 0$ for all positive bandwidth which indicates the smoothness.

Theorem 2 *(Smoothness of Objective Function). Let $k_{\omega_x},l_{\omega_y}$ be Gaussian kernels with bandwidth parameter ω_x,ω_y, for fixed samples S of size n, the objective function we used in practice*

$$J_\lambda(S):=\frac{nHSIC_b(S)-\widehat{c_\alpha}}{\sqrt{\widehat{\sigma}_u^2(S)+\lambda}},\ \lambda>0$$

satisfies the L-smoothing condition, i.e., its gradients of ω_x,ω_y are Lipschitz continuous on the compact domain $(\omega_x,\omega_y)\in[\omega_{xl},\omega_{xu}]\times[\omega_{yl},\omega_{yu}]$, for all positive constants ω_{xl}, ω_{xu}, ω_{yl}, ω_{yu}.

Proof. We consider the Gaussian kernel with the form (the conclusion remain the same as the form with ω_x)

$$k(x,x') = \exp\left(-\eta \cdot \|x-x'\|^2\right) \tag{129}$$

with parameter η for analysis and only show the results of smoothness on the bandwidth of x since the conclusions also hold for y due to symmetry. One can easily check the following

$$\frac{\partial k(x,x')}{\partial \eta} = -\|x-x'\|^2 \exp\left(-\eta \cdot \|x-x'\|^2\right) \leq 0, \frac{\partial^2 k(x,x')}{\partial^2 \eta} = \|x-x'\|^4 \exp\left(-\eta \cdot \|x-x'\|^2\right) \geq 0. \tag{130}$$

Hence for the case where the variable x has upper bounded norm $R_{\mathcal{X}}/2$ s.t. $\|x\| \leq R_{\mathcal{X}}/2$ for all $x \in \mathcal{X}$, then

$$0 \leq -\nabla_\eta k \leq R_{\mathcal{X}}^2, 0 \leq \nabla_\eta^2 k \leq R_{\mathcal{X}}^4. \tag{131}$$

Combining with the assumption 1 and $R_{\mathcal{X}} = 1$, we have

$$0 \leq \nabla_\eta^2 k \leq -\nabla_\eta k \leq k \leq 1. \tag{132}$$

This directly indicates that both kernel k and its derivative function are 1-Lipschitz continuous functions.

Smoothness of the threshold under $\mathcal{H}_0$. We now prove the smoothness of the threshold $\widehat{c_\alpha}$. Since $\widehat{c_\alpha}$ is defined as in Eq. (116), and it involves two variables β, γ, which are calculated by the first two moments of the null distribution, so we first prove the smoothness of these two moments. For analysis, we define

$$\mathcal{E} := \mathcal{E}_0 = \left(\frac{1}{n-1}\mathrm{Tr}\mathbf{K_c}\right)\left(\frac{1}{n-1}\mathrm{Tr}\mathbf{L_c}\right), \mathcal{V} := \frac{\mathbf{1}^T\left(\mathbf{B} - \mathrm{diag}(\mathbf{B})\right)\mathbf{1}}{n(n-1)} = \frac{(n-1)(n-2)(n-3)}{2n(n-4)(n-5)}\mathcal{V}_0 = C_0\mathcal{V}_0, \tag{133}$$

where $\mathcal{E}_0, \mathcal{V}_0$ are the two moments of null distribution defined in Eq. (121). Note that for fixed n, $\mathcal{E}, \mathcal{V}$ have the same smoothness property with $\mathcal{E}_0, \mathcal{V}_0$.

We show the smoothness of $\mathcal{E}$ by considering the higher-order gradients

$$\begin{aligned}
\nabla_\eta \mathcal{E} &= \left(\frac{1}{n-1}\mathrm{Tr}\nabla_\eta \mathbf{K_c}\right)\left(\frac{1}{n-1}\mathrm{Tr}\mathbf{L_c}\right) = \left(\frac{1}{n(n-1)}\mathbf{1}^T[-\nabla_\eta \mathbf{K}]\mathbf{1}\right)\left(\frac{1}{n-1}\mathrm{Tr}\mathbf{L_c}\right) \geq 0, \\
\nabla_\eta^2 \mathcal{E} &= \left(\frac{1}{n(n-1)}\mathbf{1}^T[-\nabla_\eta^2 \mathbf{K}]\mathbf{1}\right)\left(\frac{1}{n-1}\mathrm{Tr}\mathbf{L_c}\right) \leq 0.
\end{aligned} \tag{134}$$

Hence $\mathcal{E}$ is a monotonic non-decreasing function with η and

$$|\nabla_\eta \mathcal{E}| \leq 1, \ |\nabla_\eta^2 \mathcal{E}| \leq 1, \tag{135}$$

which indicates that both $\mathcal{E}$ and $\nabla_\eta \mathcal{E}$ are Lipschitz continuous functions.

And then we show the smoothness of $\mathcal{V}$. The higher-order gradients of $\mathbf{B}$ can be given by

$$\nabla_\eta \mathbf{B} = 2 \cdot \nabla_\eta \mathbf{K_c} \odot \mathbf{K_c} \odot (\mathbf{L_c})^{\cdot 2}, \nabla_\eta^2 \mathbf{B} = 2 \cdot \left(\nabla_\eta^2 \mathbf{K_c} \odot \mathbf{K_c} + (\nabla_\eta \mathbf{K_c})^{\cdot 2}\right) \odot (\mathbf{L_c})^{\cdot 2}. \tag{136}$$

Also, the following results can be given

$$\begin{aligned}
0 \leq \mathcal{V} &\leq \frac{1}{n(n-1)}\sum_{i\neq j}\mathbf{B}_{ij} \leq \frac{4}{n(n-1)}\sum_{i\neq j}|\mathbf{L}_c|_{ij}^2 \leq 16, \\
|\nabla_\eta \mathcal{V}| &\leq \frac{1}{n(n-1)}\sum_{i\neq j}|\nabla_\eta \mathbf{B}|_{ij} \leq \frac{8}{n(n-1)}\sum_{i\neq j}|\mathbf{L}_c|_{ij}^2 \leq 32, \\
|\nabla_\eta^2 \mathcal{V}| &\leq \frac{1}{n(n-1)}\sum_{i\neq j}|\nabla_\eta^2 \mathbf{B}|_{ij} \leq \frac{16}{n(n-1)}\sum_{i\neq j}|\mathbf{L}_c|_{ij}^2 \leq 64,
\end{aligned} \tag{137}$$

since the non-diagonal elements in $\mathbf{K_c}, \nabla_\eta \mathbf{K_c}, \nabla_\eta^2 \mathbf{K_c}$ all fall in the range $[-2, 2]$, and for $\mathbf{L_c}$ as well. Hence both $\mathcal{V}$ and $\nabla_\eta \mathcal{V}$ are Lipschitz continuous functions.

Under assumption 2, the following lower bound can be obtained by using Sedrakyan's inequality

$$\mathcal{V} = \frac{1}{n(n-1)}\sum_{i\neq j}(\mathbf{K_c})_{ij}^2(\mathbf{L_c})_{ij}^2 \geq \frac{1}{n(n-1)}\frac{\left(\sum_{i\neq j}(\mathbf{K_c})_{ij}\right)^2}{\sum_{i\neq j}(\mathbf{L_c})_{ij}^{-2}}. \tag{138}$$

Since the sum of the elements of the matrix $\mathbf{K_c} = \mathbf{HKH}$ is 0, we further have

$$\mathcal{V} \geq \frac{1}{n(n-1)}\frac{(-\mathrm{Tr}\mathbf{K_c})^2}{\sum_{i\neq j}(\mathbf{L_c})_{ij}^{-2}} = \frac{n-1}{n}\frac{\mathcal{E}^2}{\left(\frac{1}{n-1}\mathrm{Tr}\mathbf{L_c}\right)^2}\frac{1}{\sum_{i\neq j}(\mathbf{L_c})_{ij}^{-2}}. \tag{139}$$

Since n and $\mathbf{L_c}$ is all fixed, we set the const in Eq. (139) as C_1, then we have the bound between $\mathcal{V}$ and $\mathcal{E}$ as $\mathcal{V} \geq C_1\mathcal{E}^2$. If we restrict the minimum value of width to η_{min}, then according to the results that $\mathcal{E}$ is a monotonically increasing function and combining the assumption that we prevent all points of x_i from taking the same value and so as y_i, hence we have $0 < C(\eta_{min}) \leq \mathcal{E}$, where the const $C(\eta_{min}) := \mathcal{E}(\eta = \eta_{min})$. As a result, the variance has a bound $C_1\left(C(\eta_{min})\right)^2 \leq \mathcal{V} \leq 16$.

Since $\mathcal{E}, \mathcal{V}$ are all bounded in the convex set $[\eta_{min}, \eta_{max}]$ of the width, we can show that γ, β are have strictly positive (larger than a positive const) lower and upper bounds since $\gamma = \frac{\mathcal{E}_0^2}{\mathcal{V}_0}, \beta = \frac{n\mathcal{V}_0}{\mathcal{E}_0}$ defined in Eq. (108) and Eq. (133). Hence $\gamma, \beta, \gamma^{-1}, \beta^{-1}$ are all Lipschitz continuous functions in the domain $[\eta_{min}, \eta_{max}]$. In addition, we can further show that the gradients of $\gamma, \beta, \gamma^{-1}, \beta^{-1}$ are also Lipschitz continuous functions by proving that the second order derivative of them are all bounded.

Next, we show that $\widehat{c_\alpha}$ is smoothness. We restate the definition of $\widehat{c_\alpha}$ here. According to Eq. (116), the threshold $\widehat{c_\alpha}$ is the solution of the following equation

$$\int_0^{\widehat{c_\alpha}} \frac{x^{\gamma-1}e^{-x/\beta}}{\beta^\gamma\Gamma(\gamma)}dx = \int_0^{\frac{\widehat{c_\alpha}}{\beta}} \frac{x^{\gamma-1}e^{-x}}{\Gamma(\gamma)}dx = 1-\alpha, \tag{140}$$

where Γ is the gamma function. We now obtain the range of the threshold $\widehat{c_\alpha}$.

For the upper bound, we use the concentration in equation to bound the probability of the gamma distribution tail. Let T be the random variable with Gamma$(\gamma, 1/\beta)$ distribution, then for all $\lambda < \frac{1}{\beta}$, we have

$$\alpha = \mathbb{P}(T \geq \widehat{c_\alpha}) \leq \frac{\mathbb{E}e^{\lambda T}}{e^{\lambda\widehat{c_\alpha}}} = (1-\beta\lambda)^{-\gamma}e^{-\lambda\widehat{c_\alpha}}, \tag{141}$$

which indicates that

$$\widehat{c_\alpha} \leq \min_{\lambda\in[0,\beta^{-1}]} \frac{-\log(\alpha) - \gamma\log(1-\beta\lambda)}{\lambda}. \tag{142}$$

By heuristically setting $\lambda = \frac{1}{2\beta}$, we then obtain a upper bound

$$\widehat{c_\alpha} \leq 2\cdot\left(\frac{1}{\gamma}\log(\frac{1}{\alpha}) + \log 2\right)\cdot\mathcal{E}. \tag{143}$$

And for the lower bound, we have

$$1-\alpha = \int_0^{\frac{\widehat{c_\alpha}}{\beta}} \frac{x^{\gamma-1}e^{-x}}{\Gamma(\gamma)}dx < \int_0^{\frac{\widehat{c_\alpha}}{\beta}} \frac{x^{\gamma-1}}{\Gamma(\gamma)}dx = \frac{\widehat{c_\alpha}^{\gamma}}{\beta^\gamma\Gamma(\gamma+1)}, \tag{144}$$

which indicate that $\widehat{c_\alpha} > (1-\alpha)^{1/\gamma}\Gamma(\gamma+1)^{1/\gamma}\beta$. As a result, combining with $\gamma \geq \frac{C_0}{C_1} := C_2$ then

$$\widehat{c_\alpha} \geq \min_{\gamma\geq C_2}(1-\alpha)^{1/\gamma}\Gamma(\gamma+1)^{1/\gamma}\beta = (1-\alpha)^{1/C_2}\Gamma(C_2+1)^{1/C_2}\beta, \tag{145}$$

where the last equation uses the monotonically increasing properties of the function.

In summary, we show that the threshold is constrained by positive lower and upper bounds. Next, we further prove that the threshold satisfies the Lipschitz continuous condition. Here we start our study with $\frac{\widehat{c_\alpha}}{\beta}$ as the object in order to facilitate the calculation of the gradient since

$$\int_0^{\frac{\widehat{c_\alpha}}{\beta}} \frac{x^{\gamma-1}e^{-x}}{\Gamma(\gamma)}dx = 1-\alpha. \tag{146}$$

We denote the function $\frac{x^{\gamma-1}e^{-x}}{\Gamma(\gamma)}$ as $g_\gamma(x)$ and take the derivative on both sides. As a result,

$$\underbrace{g_\gamma\left(\frac{\widehat{c_\alpha}}{\beta}\right)}_{\bigstar}\cdot\nabla_\eta\left(\frac{\widehat{c_\alpha}}{\beta}\right) + \underbrace{\int_0^{\frac{\widehat{c_\alpha}}{\beta}} [\nabla_\eta g_\gamma](x)dx}_{\blacksquare} = 0. \tag{147}$$

For the first term in Eq. (147)

$$\bigstar = g_\gamma\left(\frac{\widehat{c_\alpha}}{\beta}\right) = \frac{\widehat{c_\alpha}^{\gamma-1}e^{-\widehat{c_\alpha}/\beta}}{\beta^{\gamma-1}\Gamma(\gamma)}, \tag{148}$$

according to the fact that both upper and lower bounds of $\beta, \gamma, \widehat{c_\alpha}, [1/\Gamma](\gamma)$ exist and the lower bound is greater than a positive constant, it is clear that the first term inherits the property. For the second term in Eq. (147)

$$\blacksquare = \left(\nabla_\eta \gamma\right) \cdot \underbrace{\left\{ \int_0^{\frac{\widehat{c_\alpha}}{\beta}} x^{\gamma-1} e^{-x} \left(\frac{\ln(x)}{\Gamma(\gamma)} + \nabla_\gamma \left(\frac{1}{\Gamma(\gamma)} \right) \right) dx \right\}}_{\spadesuit}, \tag{149}$$

it can be checked that the integral term is convergent because the integral is convergent at zero ($x = 0$) since $\gamma > 0$, and the upper limit of the integral is bounded. $\nabla_\gamma \left(\frac{1}{\Gamma(\gamma)} \right)$ is also bounded due to it being continuous on a closed interval of γ. And since we have proven that γ is a Lipschitz continuous function, $\nabla_\eta \gamma$ is also bounded. The above analysis of the two terms directly show that there is an upper bound C_3 of $\left| \nabla_\eta \left(\frac{\widehat{c_\alpha}}{\beta} \right) \right|$, this further shows that $\frac{\widehat{c_\alpha}}{\beta}$ is a Lipschitz continuous function.

We further consider the range of second-order derivatives of $\frac{\widehat{c_\alpha}}{\beta}$. We continue to derive both sides of Eq. (147),

$$(\nabla_\eta \bigstar) \cdot \nabla_\eta \left(\frac{\widehat{c_\alpha}}{\beta} \right) + \bigstar \cdot \nabla_\eta^2 \left(\frac{\widehat{c_\alpha}}{\beta} \right) + \left(\nabla_\eta \blacksquare \right) = 0. \tag{150}$$

As before, we show that the term $\nabla_\eta \bigstar, \nabla_\eta \blacksquare$ are bounded.

$$\begin{aligned} \nabla_\eta \bigstar = & \nabla_\eta \left(\frac{\exp\left\{ (\gamma - 1) \log(\frac{\widehat{c_\alpha}}{\beta}) - \frac{\widehat{c_\alpha}}{\beta} \right\}}{\Gamma(\gamma)} \right) = \exp\left\{ (\gamma - 1) \log\left(\frac{\widehat{c_\alpha}}{\beta} \right) - \frac{\widehat{c_\alpha}}{\beta} \right\} \\ & \cdot \left\{ \nabla_\eta \left(\frac{1}{\Gamma(\gamma)} \right) + \frac{1}{\Gamma(\gamma)} \cdot \left[(\nabla_\eta \gamma) \log\left(\frac{\widehat{c_\alpha}}{\beta} \right) + (\gamma - 1) \left(\frac{\widehat{c_\alpha}}{\beta} \right)^{-1} \cdot \nabla_\eta \left(\frac{\widehat{c_\alpha}}{\beta} \right) - \nabla_\eta \left(\frac{\widehat{c_\alpha}}{\beta} \right) \right] \right\}. \end{aligned} \tag{151}$$

By using a similar analysis, it is easy to verify that $\nabla_\eta \bigstar$ is bounded. And for the term $\nabla_\eta \blacksquare$, first we have

$$\nabla_\eta \blacksquare = \left(\nabla_\eta^2 \gamma \right) \cdot \spadesuit + \left(\nabla_\eta \gamma \right) \cdot \nabla_\eta \spadesuit, \tag{152}$$

and by the previous analysis $\nabla_\eta^2 \gamma, \spadesuit, \nabla_\eta \gamma$ are bounded, thus we only need to analyze $\nabla_\eta \spadesuit$.

$$\begin{aligned} \nabla_\eta \spadesuit = & \nabla_\eta \left\{ \int_0^{\frac{\widehat{c_\alpha}}{\beta}} x^{\gamma-1} e^{-x} \left(\frac{\ln(x)}{\Gamma(\gamma)} + \nabla_\gamma \left(\frac{1}{\Gamma(\gamma)} \right) \right) dx \right\} = \left(\frac{\widehat{c_\alpha}}{\beta} \right)^{\gamma-1} e^{-\frac{\widehat{c_\alpha}}{\beta}} \left(\frac{\ln\left(\frac{\widehat{c_\alpha}}{\beta} \right)}{\Gamma(\gamma)} + \nabla_\gamma \left(\frac{1}{\Gamma(\gamma)} \right) \right) \cdot \nabla_\eta \left(\frac{\widehat{c_\alpha}}{\beta} \right) \\ & + \left\{ \int_0^{\frac{\widehat{c_\alpha}}{\beta}} \frac{x^{\gamma-1}}{e^x} \left[\frac{\ln^2(x)}{\Gamma(\gamma)} + \nabla_\gamma \left(\frac{2 \ln(x)}{\Gamma(\gamma)} \right) + \nabla_\gamma^2 \left(\frac{1}{\Gamma(\gamma)} \right) \right] dx \right\} \cdot \nabla_\eta(\gamma). \end{aligned} \tag{153}$$

The terms in Eq. (153) all satisfy the bounded condition. As a result, we have proved that $\nabla_\eta \left(\frac{\widehat{c_\alpha}}{\beta} \right)$ is a Lipschitz continuous function. According to

$$\nabla_\eta \widehat{c_\alpha} = \beta \cdot \left(\nabla_\eta \left(\frac{\widehat{c_\alpha}}{\beta} \right) - \widehat{c_\alpha} \cdot \nabla_\eta \left(\frac{1}{\beta} \right) \right), \tag{154}$$

We can obtain the final conclusion that both $\widehat{c_\alpha}, \nabla_\eta \widehat{c_\alpha}$ satisfy the Lipschitz continuous condition.

Smoothness of the mean and variance under $\mathcal{H}_1$. Next we prove the smoothness of $n\text{HSIC}_b(S)$ and $\widehat{\sigma}_u^2(S) + \lambda$. We define

$$\mathcal{E}_1 := n\text{HSIC}_b(S), \ \mathcal{V}_{1,\lambda} := \widehat{\sigma}_u^2(S) + \lambda = \mathcal{V}_1 + \lambda, \tag{155}$$

where $\mathcal{V}_1$ is defined in Eq. (124). We first obtain the bound of $\mathcal{E}_1, \mathcal{V}_1$.

For fixed i, j, q, r, we can obtain the bound on h_{ijqr}.

$$|h_{ijqr}| \leq \frac{1}{4!} \sum_{(t,u,v,w)}^{(i,j,q,r)} |k_{tu} l_{tu} + k_{tu} l_{vw} - 2 k_{uv} l_{tv}| \leq 2, \tag{156}$$

since for all i, j, k_{ij} and l_{ij} both take the value in $[0, 1]$. Combining with Eq. (126), we can quickly conclude that

$$0 \le \mathcal{E}_1 \le \frac{1}{n^3} \sum_{i,j,q,r}^{n} |h_{ijqr}| \le 2n, \tag{157}$$

and the variance

$$0 < \lambda \le \mathcal{V}_{1,\lambda} \le 16 \cdot \left(\frac{1}{n} \sum_i \Big(\frac{1}{n^3} \sum_{j,q,r} h_{ijqr} \Big)^2 \right) + \lambda \le 64 + \lambda. \tag{158}$$

In addition, the range of the higher-order derivatives can be given by

$$\begin{aligned} \nabla_\eta h_{ijqr} &= \frac{1}{4!} \sum_{(t,u,v,w)}^{(i,j,q,r)} \nabla_\eta k_{tu} l_{tu} + \nabla_\eta k_{tu} l_{vw} - 2\nabla_\eta k_{uv} l_{tv}, \\ \nabla_\eta^2 h_{ijqr} &= \frac{1}{4!} \sum_{(t,u,v,w)}^{(i,j,q,r)} \nabla_\eta^2 k_{tu} l_{tu} + \nabla_\eta^2 k_{tu} l_{vw} - 2\nabla_\eta^2 k_{uv} l_{tv}. \end{aligned} \tag{159}$$

Hence the bounds can be obtained in a similar way,

$$|\nabla_\eta h_{ijqr}| \le 2, |\nabla_\eta^2 h_{ijqr}| \le 2, \tag{160}$$

since for all i, j, $[-\nabla_\eta k_{ij}]$ and $\nabla_\eta^2 k_{ij}$ are all bounded in $[0, 1]$.

As a result, we have

$$\begin{aligned} |\nabla_\eta \mathcal{V}_{1,\lambda}| &\le 32 \Big(\frac{1}{n} \sum_i \Big| \frac{1}{n^3} \sum_{j,q,r} h_{ijqr} \Big| \cdot \Big| \frac{1}{n^3} \sum_{j,q,r} \nabla_\eta h_{ijqr} \Big| + \Big| \frac{1}{n^4} \sum_{i,j,q,r}^{n} \nabla_\eta h_{ijqr} \Big| \cdot \Big| \frac{1}{n^4} \sum_{i,j,q,r}^{n} \nabla_\eta h_{ijqr} \Big| \Big) \le 256, \\ |\nabla_\eta^2 \mathcal{V}_{1,\lambda}| &\le 512. \end{aligned} \tag{161}$$

Therefore, we can conclude that both $\mathcal{E}_1, \mathcal{V}_{1,\lambda}$ satisfy the smoothness condition.

Smoothness of the optimization objective Finally we prove the smoothness of the objective function. According to the properties of the composite function, we can further prove that our optimization object

$$J_\lambda(S) = \frac{n\mathrm{HSIC}_b(S) - \widehat{c_\alpha}}{\widehat{\sigma}_{u,\lambda}(S)} = \frac{\mathcal{E}_1 - \widehat{c_\alpha}}{\sqrt{\mathcal{V}_{1,\lambda}}} \tag{162}$$

satisfies the Lipschitz smoothness condition with η when the width $\eta \in [\eta_{min}, \eta_{max}]$. □

Appendix I. Proof of Theorem 6

Theorem 6. (Type I and II error bounds for RCIT). *Under the assumptions (iv) to (vii), assume that n grow strictly faster than* $\omega\big(m^{\frac{2(\beta+p)}{\beta-1}}\big)$, *and let* $\breve{r}^{(m)}$ *be the threshold with kernels of size m that is obtained by permutation testing or wild bootstrapping, then with high probability, the bound for Type I error*

$$\mathbb{P}(\textit{Type I error}) = \mathbb{P}_{\mathcal{H}_0}\big(mHSIC_b(\widehat{R}^{(m)}_{X|Z}, \widehat{R}^{(m)}_{Y|Z}) > \breve{r}^{(m)}\big) \le \alpha + o(1) \tag{163}$$

and the probability for the Type II error

$$\mathbb{P}(\textit{Type II error}) = \mathbb{P}_{\mathcal{H}_1}\big(mHSIC_b(\widehat{R}^{(m)}_{X|Z}, \widehat{R}^{(m)}_{Y|Z}) \le \breve{r}^{(m)}\big) \sim \mathcal{O}(m^{-1/2}) \tag{164}$$

hold for any sufficiently large $m \ge 1$.

Proof. According to the results of Eq. (33), with high probability

$$\Big| \mathrm{HSIC}_b(\widehat{R}^{(m)}_{X|Z}, \widehat{R}^{(m)}_{Y|Z}) - \mathrm{HSIC}_b(R^{(m)}_{X|Z}, R^{(m)}_{Y|Z}) \Big| \sim \mathcal{O}\Big(n^{-\frac{\beta-1}{2(\beta+p)}} \Big), \tag{165}$$

thus if n grow strictly faster than $\omega\Big(m^{\frac{2(\beta+p)}{\beta-1}}\Big)$, we have with high probability

$$\Big| m\mathrm{HSIC}_b(\widehat{R}^{(m)}_{X|Z}, \widehat{R}^{(m)}_{Y|Z}) - m\mathrm{HSIC}_b(R^{(m)}_{X|Z}, R^{(m)}_{Y|Z}) \Big| \sim o(1). \tag{166}$$

And in the same way, we can show that $\Big|\breve{r}^{(m)} - \tilde{r}^{(m)}\Big| \sim o(1)$ with high probability. As a result, we obtain the result for the Type I error as

$$\begin{aligned}\mathbb{P}(\text{Type I error}) &= \mathbb{P}_{\mathcal{H}_0}\left(m\text{HSIC}_b(\widehat{R}^{(m)}_{X|Z}, \widehat{R}^{(m)}_{Y|Z}) > \breve{r}^{(m)}\right)\\ &\le \mathbb{P}_{\mathcal{H}_0}\left(m\text{HSIC}_b(R^{(m)}_{X|Z}, R^{(m)}_{Y|Z}) > \breve{r}^{(m)}\right) + o(1) \le \alpha + o(1).\end{aligned} \tag{167}$$

For the Type II error, since the term $\text{HSIC}_b(\widehat{R}^{(m)}_{X|Z}, \widehat{R}^{(m)}_{Y|Z})$ is still a V-statistic, thus the analysis can be done in the similar way as the proof of Proposition 3 in *Appendix* E, i.e., we can prove that with high probability

$$\begin{aligned}\mathbb{P}(\text{Type II error}) &= \mathbb{P}_{\mathcal{H}_1}\left(m\text{HSIC}_b(\widehat{R}^{(m)}_{X|Z}, \widehat{R}^{(m)}_{Y|Z}) \le \breve{r}^{(m)}\right)\\ &\le \mathbb{P}_{\mathcal{H}_1}\left(m\text{HSIC}_b(R^{(m)}_{X|Z}, R^{(m)}_{Y|Z}) \le \breve{r}^{(m)} + \left|m\text{HSIC}_b(\widehat{R}^{(m)}_{X|Z}, \widehat{R}^{(m)}_{Y|Z}) - m\text{HSIC}_b(R^{(m)}_{X|Z}, R^{(m)}_{Y|Z})\right|\right)\\ &\le \mathbb{P}_{\mathcal{H}_1}\left(m\text{HSIC}_b(R^{(m)}_{X|Z}, R^{(m)}_{Y|Z}) \le \breve{r}^{(m)} + C_h\right),\end{aligned} \tag{168}$$

where C_h is a constant. Since $\breve{r}^{(m)} \sim \mathcal{O}(m^{1/2})$, thus the influence of C_h can be ignored when m is large enough, the remain proof is the same as the proof of Proposition 3 in *Appendix* E. □

Appendix J. Time complexity

Let the sample size be n, the dimensions of X, Y be d_x, d_y, and then the time complexity is derived from the following steps. For each iteration of the train, computing the kernel matrix costs $\mathcal{O}\left(n^2(d_x + d_y)\right)$ and computing the terms $\text{HSIC}_b(\mathcal{S})$, $\widehat{\sigma}^2_u(\mathcal{S})$ and $\widehat{c_\alpha}$ costs $\mathcal{O}\left(n^2\right)$ as shown in *Appendix* B and by the definition of $\widehat{c_\alpha}$ given in *Appendix* F. Hence the total time complexity is $\mathcal{O}\left(Tn^2(d_x + d_y)\right)$, where T is the total number of iterations.

Remark. For the data with a large number of samples, we can train in a small batch way to reduce computational complexity. This paper uses full batch training since a small sample size is sufficient for good results.

Appendix K. Baseline methods and setup for real CI experiments

In this section, we present the details of baseline methods and implementation settings for real-world CI tests and causal discovery.

K.1. Baseline details

The methods of comparison used in the real-world experiments are described below.

- **WIT** [33]. A state-of-the-art method captures the non-linear dependencies between variables by measuring their correlation under different levels of wavelet mappings.
- **DARLING** [15]. Develop a neural residual independence test by continuous optimization algorithm to characterize independence as a set of zero correlations between one square-integrable function of $X(\phi(X))$ and Y, hereafter we denote it by NRIT.
- **KCIT** [55]. A kernel-based conditional independence test (KCI-test), by constructing an appropriate test statistic and deriving its asymptotic distribution under the null hypothesis of conditional independence.
- **SCIT** [53]. A Method measures conditional independence by comparing the similarity between $R_{X|Z}$ and $R_{Y|Z}$ with that between $R_{X|Z}$ and an independent copy R_r of $R_{Y|Z}$ in a high-dimensional space.
- **FRCIT** [49]. A simply but more effective way to test CI based on regression.
- **RDC** [26]. A state-of-the-art method based on the canonical correlation between a finite set of random Fourier features. We refer to as RDCPT hereafter.
- **PaCoT**. A way to compute Fisher's correlation coefficient [30] and is widely applied in testing linear independence and CI [6,48].

K.2. Baseline setup

To ensure a fair comparison with RDCPT, we follow the methodology outlined in [3] to implement RDC with a permutation test. For WIT, we set the number of permutations to $B = 50$, while for SCIT, we similarly set the permutation count to $k = 50$. For Darling, we initialize the network weights using a uniform distribution, which facilitates stable data propagation through non-linear activation functions such as ReLU [27], thereby mitigating potential issues related to gradient explosion or vanishing. For KCIT, our implementation is highly based on the *causal-learn* package,[6] with kernel bandwidths initialized using the median heuristic.

[6] https://causal-learn.readthedocs.io/en/latest/independence_tests_index/kci.html.

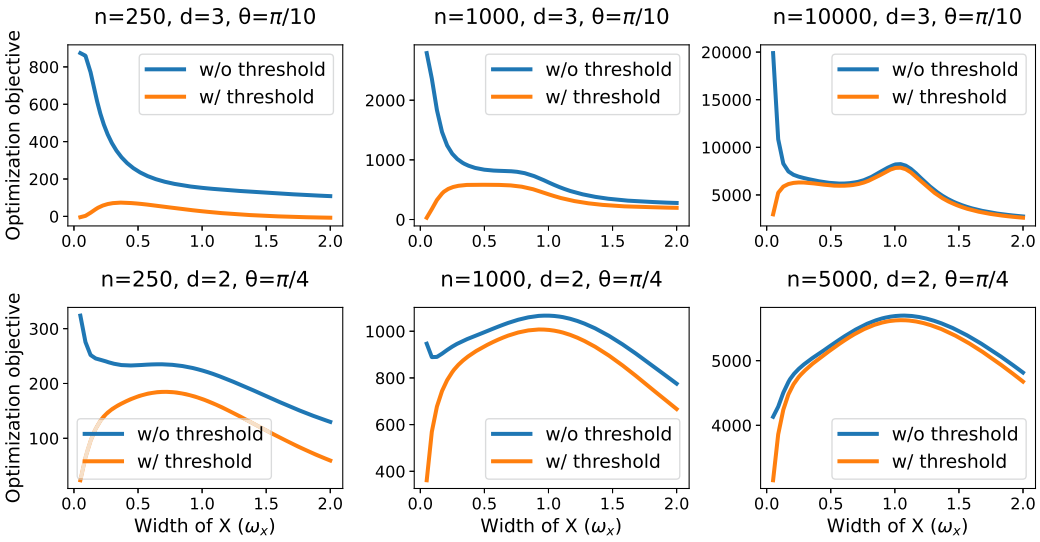

Fig. 10. The values of optimization objective for different ω_x on the ISA dataset under more settings.

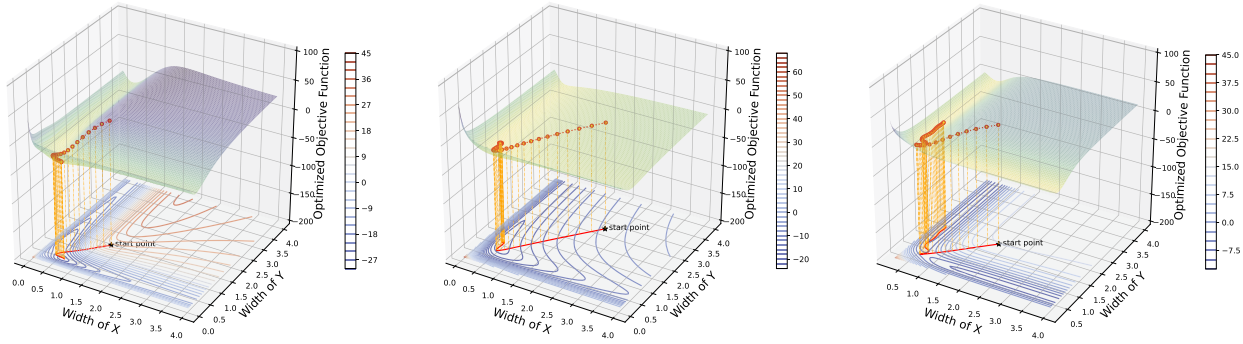

Fig. 11. The visualization results of gradient descent process for HSIC-O on the ISA dataset. From left to right: 1) Gaussian kernel with learnable width. $n = 128$, $\theta = \pi/10$, $d = 2$. 2) Laplace kernel with learnable width. $n = 128$, $\theta = \pi/10$, $d = 2$. 3) Gaussian kernel with learnable width. $n = 256$, $\theta = \pi/10$, $d = 4$.

Appendix L. Additional experiment results

L.1. More examples for the pitfall with the signal-to-noise ratio criterion

In this section, we present more experimental results as well as analysis under additional settings for the pitfall of existing methods (Sec. 4.1 in the main paper). We first show the results of the experiments under more settings in Fig. 10. From the above row, we can see that the overall difference between modeling thresholds or not diminish as the sample size increases (gradually closing over a wider range of kernel width). The next row has a simpler setup therefore the difference between the two is smaller. This corresponds to the theoretical explanation that as the sample size increases, the impact from the threshold gradually decreases compared to the signal-to-noise criterion. Formally, under the alternative hypothesis, the term $\frac{n\mathrm{HSIC}(X,Y)}{\sqrt{n}\sigma_u} \sim \mathcal{O}(\sqrt{n})$ while $\frac{r}{\sqrt{n}\sigma_u} \sim \mathcal{O}(\frac{1}{\sqrt{n}})$. We further give visualization results for more details of the optimization process. The results of our method (after modeling threshold) are given in Fig. 11. For different kernel and dimension settings, our method achieves a reliable optimization. This is also supported by our theoretical smoothness guarantee. For the case in which no threshold is applied, the results are presented in Fig. 12. It can be seen that even for the simpler case of $d = 2$, the optimization leads to the wrong solution (the solution with zero bandwidths), and the phenomenon has not been resolved until the sample size reaches 1024 that a converging solution is obtained for the first time. Overall, our method addresses the pitfalls of the original criterion, thus enabling it to deal with more challenging scenarios and greatly improving the stability of the optimization.

L.2. Experiment results with more settings and more compared methods

In this section, we provide the experimental results under more settings in Fig. 13.

More kernels. The results for more kernel settings are shown on the left. Among them, HSIC-MG, HSIC-ML, HSIC-OG, HSIC-OL, and HSIC-WG correspond to the results for the Gaussian kernel with median bandwidth, the Laplace kernel with median bandwidth,

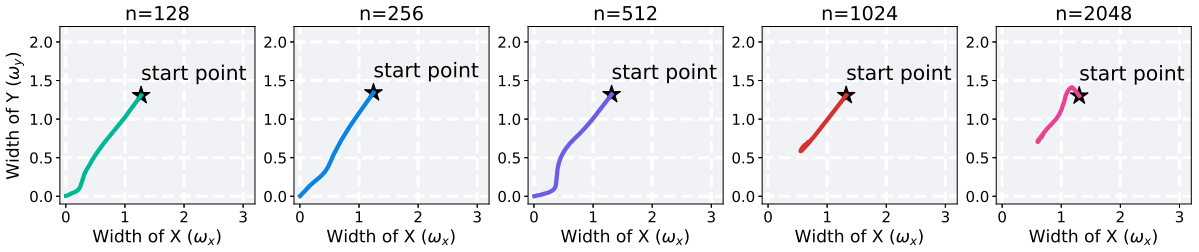

Fig. 12. The visualization results of the kernel bandwidth optimization process when using the criterion (w.o. threshold). The number of iterations is sufficient to ensure convergence. Setup: The ISA dataset is used, d is fixed to 2 and $\theta = \pi/10$. The sample size n is changed from 128 to 2048.

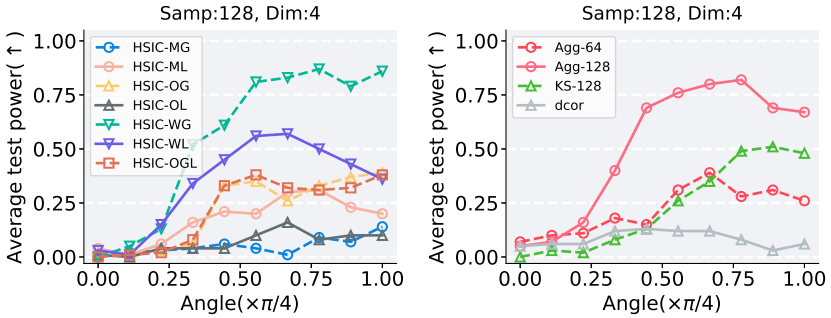

Fig. 13. Results of experiments under the ISA dataset with more settings.

the Gaussian kernel with optimized bandwidth, the Laplace kernel with optimized bandwidth, and the importance-weighted kernel, respectively. In addition, similar to the Gaussian version of the importance-weighted kernel, we also provide the kernel design for the Laplace case. Specifically, HSIC-WL corresponds to the result with the kernel $k(x,x') := \prod_{i=1}^{d_x} \exp\left(-\frac{w_i|x_i - x_i'|}{\omega_x}\right), w_i \in (0,1)$. For the combined kernel, an example (HSIC-OGL) is also given. Formally, the kernel has the following form $k = \pi_1 k_{OG} + \pi_2 k_{OL}$, where π_1, π_2 are the learnable combination coefficient and k_{OG}, k_{OL} are the Gaussian/Laplace kernels with learnable bandwidth parameters. From the results, we can see that the Gaussian kernel is better in general compared to Laplace in this setting of the ISA dataset. For example, HSIC-OG is better than HSIC-OL and HSIC-WG is better than HSIC-WL. But notice that HSIC-ML is superior to HSIC-MG due to better bandwidth initialization. Also note that optimized results are not always better than unoptimized (HSIC-ML is better than HSIC-OL), since only half of the sample is used for testing. Also, it is noted that the results of HSIC-OGL are similar to those of HSIC-OG since HSIC-OG achieves better results compared to HSIC-OL therefore the combined results tend to be closer to HSIC-OG. This also illustrates that our method can be applied to the scenario that learning the combination of kernels.

More compared methods. On the right, we compare with more methods. We consider the distance-based methods [42] called distance covariance (dcor) as well as the aggregated kernels tests [35]. To implement the aggregated kernels tests, we consider the kernel selection setting. Specifically, we first define the set of kernels, which contains a fixed form of kernels with different bandwidths. Here, we initialize the bandwidth as $\{2^i \omega_{mid}, i \in \{-5,-4,...,4,5\}\}$ for both kernels of X and Y. Correspondingly, we provide the results of our method under the kernel selection scenario as a comparison. We formulate the kernel with the form $k = \sum_{i=1}^{l} \pi_i k_i$, where $\{k_i, i = 1,2,...,l\}$ are the kernel defined as above and $\{\pi_i, i = 1,2,...,l\}$ are the learnable combination coefficients. For the method under the kernel selection setting with a sample size of 128, we show the results (Agg-128 and KS-128) in the right of Fig. 13. In this setup, Agg-128 performs better compared to KS-128, benefiting from its more adequate utilization of the sample size. As a comparison, we also provide the results of the aggregated kernels tests with a sample size of 64 (Agg-64). In this case the test sample size is the same and our method gives comparable results. In conclusion, even though our method can be applied to the kernel selection scenario, compared to the aggregation test, our scheme loses power due to the reduction of the sample size resulting from learning the kernel. This suggests that our method needs to compensate for the loss of sample size due to data splitting. Since this is beyond the scope of our paper, we treat it as an important direction for future work. For dcor, the test loses power due to its predefined distance function and therefore cannot handle this challenging setting flexibly.

L.3. Running time

The running time of each method is given in Fig. 14. The running time is consistent with our theoretical complexity, i.e., linear with dimension and proportional to the square of the samples. It also depends largely on the squared complexity of HSIC on which we are based. For high-dimensional settings, our method is competitive (for the case where n is relatively small and d is relatively large) and one test can be completed in a few seconds.

Potential future research directions.

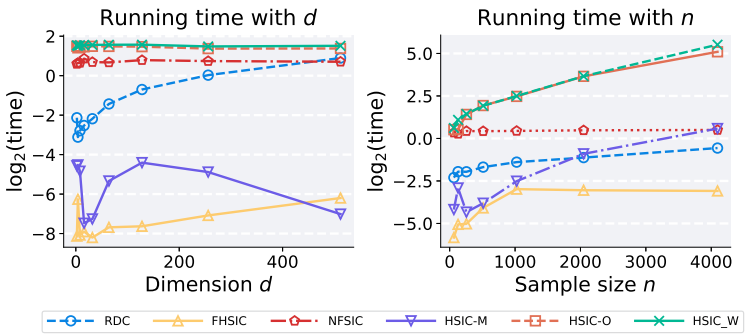

Fig. 14. The practical running time of each method. Left: fix $n = 128$. Right: fix $d = 4$.

Here, we discuss potential directions for future work.

- **Learning kernels without data splitting.** To the best of our knowledge, not only our schemes but also current existing methods that continuously learn the kernel rely on data splitting tend to hurt power performance. In the line of kernel selection implementation, some works such as [35] have been proposed to mitigate this problem. How to extend their methods to the scenario that continuous learning of kernel bandwidth parameters is an interesting direction.
- **Reducing computational costs.** As our method is based on the quadratic-time HSIC, it inherits the squared complexity concerning the sample size. To learn kernels in the case of large-scale kernel machines, kernel approximation methods such as random Fourier features as well as kernel thinning methods can be combined. We will further explore it in upcoming works.

Data availability

No data was used for the research described in the article.

References

[1] M. Albert, B. Laurent, A. Marrel, A. Meynaoui, Adaptive test of independence based on hsic measures, Ann. Stat. 50 (2022) 858–879.
[2] F.R. Bach, M.I. Jordan, Kernel independent component analysis, J. Mach. Learn. Res. 3 (2002) 1–48.
[3] A. Bellot, M.v.d. Schaar, Conditional independence testing using generative adversarial networks, in: Proceedings of the 33rd International Conference on Neural Information, Curran Associates Inc., 2019, pp. 2202–2211.
[4] T. Bertin-Mahieux, YearPredictionMSD, UCI Machine Learning Repository, 2011.
[5] C. Burgess, H. Kim, 3d shapes dataset, https://github.com/deepmind/3dshapes-dataset/, 2018.
[6] R. Cai, Z. Zhang, Z. Hao, Sada: a general framework to support robust causation discovery, in: International Conference on Machine Learning, PMLR, 2013, pp. 208–216.
[7] S. Fischer, I. Steinwart, Sobolev norm learning rates for regularized least-squares algorithms, J. Mach. Learn. Res. 21 (2020) 1–38.
[8] S.R. Flaxman, D.B. Neill, A.J. Smola, Gaussian processes for independence tests with non-iid data in causal inference, ACM Trans. Intell. Syst. Technol. 7 (2016) 22-1.
[9] K. Fukumizu, A. Gretton, B. Schölkopf, B.K. Sriperumbudur, Characteristic kernels on groups and semigroups, Adv. Neural Inf. Process. Syst. 21 (2008).
[10] A. Gretton, A simpler condition for consistency of a kernel independence test, preprint, arXiv:1501.06103, 2015.
[11] A. Gretton, K.M. Borgwardt, M. Rasch, B. Schölkopf, A.J. Smola, A kernel method for the two-sample-problem, in: Advances in Neural Information Processing Systems, 2006, pp. 513–520.
[12] A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, A. Smola, A kernel statistical test of independence, Adv. Neural Inf. Process. Syst. 20 (2007).
[13] A. Gretton, A. Smola, O. Bousquet, R. Herbrich, A. Belitski, M. Augath, Y. Murayama, J. Pauls, B. Schölkopf, N. Logothetis, Kernel constrained covariance for dependence measurement, in: International Workshop on Artificial Intelligence and Statistics, PMLR, 2005, pp. 112–119.
[14] M. Grosse-Wentrup, D. Janzing, M. Siegel, B. Schölkopf, Identification of causal relations in neuroimaging data with latent confounders: an instrumental variable approach, NeuroImage 125 (2016) 825–833.
[15] Y. He, P. Cui, Z. Shen, R. Xu, F. Liu, Y. Jiang, Daring: differentiable causal discovery with residual independence, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2021, pp. 596–605.
[16] W. Jitkrittum, Z. Szabó, K.P. Chwialkowski, A. Gretton, Interpretable distribution features with maximum testing power, Adv. Neural Inf. Process. Syst. 29 (2016).
[17] W. Jitkrittum, Z. Szabó, A. Gretton, An adaptive test of independence with analytic kernel embeddings, in: International Conference on Machine Learning, PMLR, 2017, pp. 1742–1751.
[18] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, preprint, arXiv:1412.6980, 2014.
[19] V.S. Korolyuk, Y.V. Borovskich, Theory of U-Statistics, vol. 273, Springer Science & Business Media, 2013.
[20] J. Kübler, W. Jitkrittum, B. Schölkopf, K. Muandet, Learning kernel tests without data splitting, Adv. Neural Inf. Process. Syst. 33 (2020) 6245–6255.
[21] J.M. Kübler, V. Stimper, S. Buchholz, K. Muandet, B. Schölkopf, Automl two-sample test, Adv. Neural Inf. Process. Syst. 35 (2022) 15929–15941.
[22] T. Li, M. Yuan, On the optimality of Gaussian kernel based nonparametric tests against smooth alternatives, preprint, arXiv:1909.03302, 2019.
[23] Z. Li, D. Meunier, M. Mollenhauer, A. Gretton, Optimal rates for regularized conditional mean embedding learning, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2022, pp. 4433–4445.
[24] F. Liu, W. Xu, J. Lu, D.J. Sutherland, Meta two-sample testing: learning kernels for testing with limited data, Adv. Neural Inf. Process. Syst. 34 (2021) 5848–5860.
[25] F. Liu, W. Xu, J. Lu, G. Zhang, A. Gretton, D.J. Sutherland, Learning deep kernels for non-parametric two-sample tests, in: International Conference on Machine Learning, PMLR, 2020, pp. 6316–6326.
[26] D. Lopez-Paz, P. Hennig, B. Schölkopf, The randomized dependence coefficient, Adv. Neural Inf. Process. Syst. 26 (2013).
[27] V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, in: Proceedings of the 27th International Conference on International Conference on Machine Learning, Omnipress, Madison, WI, USA, 2010, pp. 807–814.

[28] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch, in: Advances in Neural Information Processing Systems, 2017.
[29] J. Pearl, Causality, Cambridge University Press, 2009.
[30] S. Plata, A note on Fisher's correlation coefficient, Appl. Math. Lett. 19 (2006) 499–502.
[31] R. Pogodin, A. Schrab, Y. Li, D.J. Sutherland, A. Gretton, Practical kernel tests of conditional independence, arXiv:2402.13196, 2024.
[32] J.D. Ramsey, A scalable conditional independence test for nonlinear, non-gaussian data, preprint, arXiv:1401.5031, 2014.
[33] Y. Ren, H. Zhang, Y. Xia, J. Guan, S. Zhou, Multi-level wavelet mapping correlation for statistical dependence measurement: methodology and performance, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2023, pp. 6499–6506.
[34] K. Sachs, O. Perez, D. Pe'er, D.A. Lauffenburger, G.P. Nolan, Causal protein-signaling networks derived from multiparameter single-cell data, Science 308 (2005) 523–529.
[35] A. Schrab, I. Kim, B. Guedj, A. Gretton, Efficient aggregated kernel tests using incomplete u-statistics, Adv. Neural Inf. Process. Syst. 35 (2022) 18793–18807.
[36] D. Sejdinovic, B. Sriperumbudur, A. Gretton, K. Fukumizu, Equivalence of distance-based and rkhs-based statistics in hypothesis testing, Ann. Stat. (2013) 2263–2291.
[37] R.J. Serfling, Approximation Theorems of Mathematical Statistics, John Wiley & Sons, 2009.
[38] R.D. Shah, J. Peters, The hardness of conditional independence testing and the generalised covariance measure, Ann. Stat. 48 (2020).
[39] P. Spirtes, C. Glymour, R. Scheines, Causation, Prediction, and Search, MIT Press, 2001.
[40] D.J. Sutherland, H.Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, A. Gretton, Generative models and model criticism via optimized maximum mean discrepancy, preprint, arXiv:1611.04488, 2016.
[41] G. Szekely, M. Rizzo, N. Bakirov, Measuring and testing dependence by correlation of distances, Ann. Stat. 35 (2008).
[42] G.J. Székely, M.L. Rizzo, The distance correlation t-test of independence in high dimension, J. Multivar. Anal. 117 (2013) 193–213.
[43] F. Theis, Towards a general independent subspace analysis, Adv. Neural Inf. Process. Syst. 19 (2006).
[44] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc., Ser. B Stat. Methodol. 58 (1996) 267–288.
[45] R. Vershynin, High-Dimensional Probability: An Introduction with Applications in Data Science, vol. 47, Cambridge University Press, 2018.
[46] G.S. Watson, Linear least squares regression, Ann. Math. Stat. (1967) 1679–1699.
[47] C. Williams, C. Rasmussen, Gaussian processes for regression, Adv. Neural Inf. Process. Syst. 8 (1995).
[48] H. Zhang, Y. Ren, Y. Xia, S. Zhou, J. Guan, Towards effective causal partitioning by edge cutting of adjoint graph, IEEE Trans. Pattern Anal. Mach. Intell. 46 (2024) 10259–10271.
[49] H. Zhang, K. Zhang, S. Zhou, J. Guan, J. Zhang, Testing independence between linear combinations for causal discovery, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 6538–6546.
[50] H. Zhang, S. Zhou, J. Guan, Measuring conditional independence by independent residuals: theoretical results and application in causal discovery, in: AAAI Conference on Artificial Intelligence, 2018, pp. 2029–2036.
[51] H. Zhang, S. Zhou, J. Guan, J.L. Huan, Measuring conditional independence by independent residuals for causal discovery, ACM Trans. Intell. Syst. Technol. 10 (2019) 1–19.
[52] H. Zhang, S. Zhou, K. Zhang, J. Guan, Causal discovery using regression-based conditional independence tests, in: AAAI Conference on Artificial Intelligence, 2017, pp. 1250–1256.
[53] H. Zhang, S. Zhou, K. Zhang, J. Guan, Residual similarity based conditional independence test and its application in causal discovery, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 5942–5949.
[54] K. Zhang, A. Hyvärinen, On the identifiability of the post-nonlinear causal model, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, AUAI Press, 2009, pp. 647–655.
[55] K. Zhang, J. Peters, D. Janzing, B. Schölkopf, Kernel-based conditional independence test and application in causal discovery, preprint, arXiv:1202.3775, 2012.
[56] Q. Zhang, S. Filippi, A. Gretton, D. Sejdinovic, Large-scale kernel methods for independence testing, Stat. Comput. 28 (2018) 113–130.
[57] F. Zou, L. Shen, Z. Jie, W. Zhang, W. Liu, A sufficient condition for convergences of adam and rmsprop, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11127–11135.