Learning Adaptive Kernels for Statistical Independence Tests

Yixin Ren Fudan University
Yewei Xia Fudan University
Hao Zhang SIAT, CAS
Jihong Guan Tongji University
Shuigeng Zhou Fudan University

Abstract

We propose a novel framework for kernel-based statistical independence tests that enable adaptatively learning parameterized kernels to maximize test power. Our framework can effectively address the pitfall inherent in the existing signal-to-noise ratio criterion by modeling the change of the null distribution during the learning process. Based on the proposed framework, we design a new class of kernels that can adaptatively 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 our independence 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/renyixin666/HSIC-LK.git`.

1 Introduction

Given two variables X and Y, the statistical independence test determines whether the null hypothesis $\mathbb{P}_{XY} = \mathbb{P}_X\mathbb{P}_Y$ can be rejected. Traditional tests such as Pearson's correlation coefficient (Cohen et al., 2009) and Kendall's τ can only measure monotonic relations between low-dimensional variables. Many recent works have tried to capture more complex non-linear dependencies in higher dimensional spaces (Lyons, 2013; Liu et al., 2022; Chatterjee, 2021; Lopez-Paz et al., 2013; Zhang et al., 2012). These methods have been applied to various machine learning problems such as self-supervised learning (Li et al., 2021), feature selection (Camps-Valls et al., 2010) and causal discovery (Ren et al., 2023; Hoyer et al., 2008).

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) (Bach and Jordan, 2002) and the constrained covariance (COCO) (Gretton et al., 2005) 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) (Gretton et al., 2007) uses the squared Hilbert-Schmidt norm to detect dependence. The HSIC-based independence tests (Gretton et al., 2007; Zhang et al., 2018) 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 (Gretton et al., 2007) (e.g. the Gaussian kernel with median bandwidth) or use a randomized feature mapping (Zhang et al., 2018), thus have limited flexibility and cannot capture the differences in the distributions of complex structures. To solve this, some works (Jitkrittum et al., 2016; Liu et al., 2021) tried to learn the kernels/features to maximize the power of hypothesis tests. Jitkrittum et al. (2017) 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, Liu et al. (2020, 2021) learned the kernels using a criterion called the signal-to-noise ratio (Kübler et al., 2022) as the optimization objective in the two-sample test problem (Gretton et al., 2005). However, our study shows that this criterion may result in wrong solutions when learning kernels for independence tests

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

since the change of the null distribution is ignored. In this paper, 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: 1) We propose a novel framework for kernel-based statistical independence tests that enable adaptatively 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. 2) We further design a new class of kernels that can adaptatively 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. 3) 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. 4) We conduct extensive experiments on both synthetic and real data, which demonstrate the superiority of the proposed method.

Outline. The rest of the paper is organized as follows: Sec. 2 reviews HSIC-based statistical independence tests. Sec. 3 first presents the pitfall of the learning criterion with existing work and then introduces a novel framework to solve it. Furthermore, a class of importance-weighted kernels is designed. Sec. 4 gives the theoretical analyses and Sec. 5 is the performance evaluation. We conclude the paper in Sec. 6. All theoretical proofs are given in the Supplementary Material.

Related Work. 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 (Gretton et al., 2007) is recognized as one of the most powerful tests among them. In addition, some variants (Zhang et al., 2018) 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 (Székely et al., 2007; Székely and Rizzo, 2013) which utilizes characteristic functions to measure and test dependence. In fact, distance-based methods are equivalent to the HSIC with specific kernels (Sejdinovic et al., 2013). 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.

Learning kernels to maximize the power of the test has also been extensively studied in different applications (e.g. two-sample tests (Sutherland et al., 2016), independence tests (Albert et al., 2022), and goodness-of-fit tests (Schrab et al., 2022)), 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 (Li and Yuan, 2019) and learning deep kernels (Liu et al., 2020) 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 (Albert et al., 2022). 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). As for the second direction, due to the space limit we provide the results in the supplementary file.

2 Preliminaries

2.1 Statistical Independence Test

Let $\mathcal{X}, \mathcal{Y}$ be separable metric spaces, $\mathbb{P}_{XY}$ be Borel probability measure defined on $\mathcal{X} \times \mathcal{Y}$, $\mathbb{P}_X$ and $\mathbb{P}_Y$ be the respective marginal distributions on $\mathcal{X}$ and $\mathcal{Y}$. Given n independent and identically distributed (i.i.d) samples $Z := (X, Y) = \{(x_i, y_i)\}_{i=1}^n$ with distribution $\mathbb{P}_{XY}$, we wish to know whether $\mathbb{P}_{XY}$ can be factorized into $\mathbb{P}_X\mathbb{P}_Y$: does $\mathbb{P}_{XY} = \mathbb{P}_X\mathbb{P}_Y$ (i.e. $X \perp\!\!\!\perp Y$)?

The hypothesis testing framework is used, i.e.,

$$\mathcal{H}_0 : \mathbb{P}_{XY} = \mathbb{P}_X\mathbb{P}_Y \quad \text{versus} \quad \mathcal{H}_1 : \mathbb{P}_{XY} \neq \mathbb{P}_X\mathbb{P}_Y. \quad (1)$$

The independence hypothesis testing 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 is not greater than α.

Two types of errors may be generated during hypothesis testing. Type I error means the false rejection of $\mathcal{H}_0$, and Type II error indicates when $\mathcal{H}_0$ is wrong but not rejected. A good independence test requires that Type I error rate is upper bounded by α meanwhile Type II error is minimized (Zhang et al., 2012).

2.2 Hilbert-Schmidt Independence Criterion

We base our statistical independence tests on the Hilbert-Schmidt Independence Criterion (HSIC).

Definition 1. *(Gretton et al., 2007). Let $\mathcal{F}$ be an RKHS, with the kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$. Likewise, let $\mathcal{G}$ be a second RKHS on $\mathcal{Y}$ with kernel $l : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$. The Hilbert-Schmidt Independence Criterion between X and Y, denoted as HSIC(X, Y) is defined as*

$$\begin{aligned}\mathbf{E}\big[k(X,X')l(Y,Y')\big]+\mathbf{E}\big[k(X,X')\big]\mathbf{E}\big[l(Y,Y')\big] \\ - 2\mathbf{E}_{X'Y'}\big[\mathbf{E}_X k(X,X')\mathbf{E}_Y l(Y,Y')\big],\end{aligned} \tag{2}$$

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

For characteristic kernels (Gretton, 2015), the independence relationship ($X \perp\!\!\!\perp Y$) between X and Y can be judged by $\mathrm{HSIC}(X, Y) = 0$.

Given n i.i.d samples $Z = \{(x_i, y_i)\}_{i=1}^n$ with distribution $\mathbb{P}_{XY}$, an observation of $\mathrm{HSIC}(X, Y)$, denoted as $\mathrm{HSIC}_b(Z)$, can be given by

$$\frac{1}{n^2}\sum_{i,j} k_{ij}l_{ij} + \frac{1}{n^4}\sum_{i,j,q,r} k_{ij}l_{qr} - 2\frac{1}{n^3}\sum_{i,j,q} k_{ij}l_{iq}, \tag{3}$$

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 center matrix and $\mathbf{1}$ is an $n \times 1$ vector of ones.

2.3 Asymptotics of HSIC

The asymptotic distribution of the statistic under two hypotheses can be established by the following proposition (Gretton et al., 2007, Theorem 1, 2).

Proposition 1. (Asymptotics of $\mathrm{HSIC}_b(Z)$). *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$_b(Z)$ coverages in distribution as*

$$n\mathit{HSIC}_b(Z) \xrightarrow{d} \sum_{l=1}^{\infty} \lambda_l z_l^2, \tag{4}$$

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$_b(Z)$ converges in distribution to a Gaussian variable

$$n^{\frac{1}{2}}\Big(\mathit{HSIC}_b(Z) - \mathit{HSIC}(X,Y)\Big) \xrightarrow{d} \mathcal{N}(0, \sigma_u^2), \tag{5}$$

where the variance is given by

$$\sigma_u^2 = 16\big(\mathbf{E}_i(\mathbf{E}_{j,q,r}h_{ijqr})^2 - \mathit{HSIC}(X,Y)^2\big) \tag{6}$$

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

3 Learning Kernels

3.1 The Pitfall with 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

$$\mathbb{P}_{\mathcal{H}_1}\left(n\mathrm{HSIC}_b(Z) > r\right) \to \Phi\left(\frac{n\mathrm{HSIC}(X,Y) - r}{\sqrt{n}\sigma_u}\right), \tag{7}$$

where Φ is the standard normal CDF and r is the threshold, i.e., the $(1-\alpha)$-quantile of distribution of Eq. (4) 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 (Liu et al., 2020, 2021) in kernel learning for the two-sample test problem (Gretton et al., 2006), called signal-to-noise ratio (Kübler et al., 2022). Note that the criterion is theoretically biased as long as r in Eq. (7) 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., $k(x, x') := \exp\big(-\frac{\|x-x'\|^2}{2\omega_x^2}\big), l(y, y') := \exp\big(-\frac{\|y-y'\|^2}{2\omega_y^2}\big)$, where ω_x, ω_y are the width parameters. The estimate of signal-to-noise is taken as

$$J_{w/o} := \frac{n\mathrm{HSIC}_b(Z)}{\widehat{\sigma}_u(Z)}, \tag{8}$$

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

$$16 \cdot \Big(\frac{1}{n}\sum_i \big(\frac{1}{n^3}\sum_{j,q,r} h_{ijqr}\big)^2 - \big(\mathrm{HSIC}_b(Z)\big)^2\Big). \tag{9}$$

For fixed samples Z 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:

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

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

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

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 (Gretton et al., 2007):

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

$$\mathcal{E}_0 := 1 + \widehat{\|\mu_x\|^2}\widehat{\|\mu_y\|^2} - \widehat{\|\mu_y\|^2} - \widehat{\|\mu_x\|^2}, \tag{10}$$

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

$$\mathcal{V}_0 = \frac{2(n-4)(n-5)}{(n-1)^2(n-2)(n-3)}\mathbf{1}^T\big(\mathbf{B} - \mathit{diag}(\mathbf{B})\big)\mathbf{1}, \tag{11}$$

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.

The limit of these two moments can be calculated by

$$\begin{aligned}\mathcal{E}_0\big|_{\omega_x=0^+} &= \frac{1}{n-1}\mathrm{Tr}\mathbf{L_c},\\ \mathcal{V}_0\big|_{\omega_x=0^+} &= \frac{2(n-4)(n-5)}{n^2(n-1)^2(n-2)(n-3)}\sum_{i\neq j}(\mathbf{L_c})_{ij}^2.\end{aligned} \tag{12}$$

Since the variance is $\mathcal{O}(n^{-2})$ as in Eq. (12), 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

$$J_{w/,\mathcal{E}_0} := \frac{n\mathrm{HSIC}_b(Z) - \mathcal{E}_0}{\widehat{\sigma}_u(Z)}, \tag{13}$$

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.

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 the supplementary file.

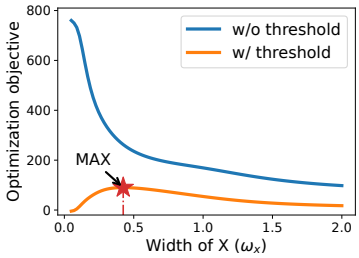

Figure 1: The values of optimization objective for different ω_x on the ISA dataset under the setting n=250, d=3, θ=π/10, ω_y=1.0. The "w/o threshold" line corresponds to Eq. (8) and the other to Eq. (16).

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.

3.2 Our 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 (Gretton et al., 2007), 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. (4). The first two moments of Eqs. (10), (11) are used to determine the two parameters, i.e.,

$$n\mathrm{HSIC}_b(Z) \sim \frac{x^{\gamma-1}e^{-x/\beta}}{\beta^\gamma\Gamma(\gamma)}, \gamma = \frac{\mathcal{E}_0^2}{\mathcal{V}_0},\ \beta = \frac{\mathcal{V}_0}{\mathcal{E}_0}, \tag{14}$$

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.,

$$\int_0^{\frac{\widehat{c_\alpha}}{\beta}} \frac{x^{\gamma-1}e^{-x}}{\Gamma(\gamma)}dx = 1-\alpha. \tag{15}$$

And the criterion can be obtained by [1]

$$J_{w/,\widehat{c_\alpha}} := \frac{n\mathrm{HSIC}_b(Z) - \widehat{c_\alpha}}{\widehat{\sigma}_u(Z)}. \tag{16}$$

[1] In practice, we add a small constant to the denominator as suggested in (Liu et al., 2020).

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 (Kingma and Ba, 2014). 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. (15). 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

$$\frac{n\mathrm{HSIC}_b(Z) - \widehat{c_\alpha}\big|_{\omega_*}}{\widehat{\sigma}_u(Z)} - \frac{\big(\partial_\beta \widehat{c_\alpha}\big|_{\omega_*}\big)\cdot\beta + \big(\partial_\gamma \widehat{c_\alpha}\big|_{\omega_*}\big)\cdot\gamma}{\widehat{\sigma}_u(Z)\big|_{\omega_*}} \tag{17}$$

by combining with an automatic differentiation framework such as PyTorch (Paszke et al., 2017) to estimate the gradient $\partial_\omega J_{w/,\widehat{c_\alpha}}$ at the point $\omega = \omega_*$.

Data split. Given samples Z, 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 (Kübler et al., 2020). Here we adapt the technique used in a variety of tests (Liu et al., 2020; Jitkrittum et al., 2017): 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 Alg. 1. As a pre-processing step, we split the data to training data Z^{tr} and test data Z^{te} (Line 1). The test contains two phases: 1) We learn the kernels with Adam optimizer using full batches on Z^{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 Z^{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.

3.3 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:

$$k(x, x') = \exp\big(-(x - x')\Sigma_x(x - x')\big), \tag{18}$$

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 (Gretton, 2015; Fukumizu et al., 2008). 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 (Williams and Rasmussen, 1995). 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,

$$k(x, x') := \prod_{i=1}^{d_x} \exp\Big(-\frac{w_i(x_i - x_i')^2}{2\omega_x^2}\Big), w_i \in (0, 1), \tag{19}$$

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. 5.2.2.

Algorithm 1 The learning and testing framework

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

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

4 Theoretical Analysis

We require some assumptions as follows:

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

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

(b) 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,Ω_1 is separately bounded by R_{ω_0}, R_{ω_1} respectively, i.e.,

$$\Omega_0\subseteq\Big\{\omega_0\big|\ \|\omega_0\|\le R_{\Omega_0}\Big\},\Omega_1\subseteq\Big\{\omega_1\big|\ \|\omega_1\|\le R_{\Omega_1}\Big\}.$$

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

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

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

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

with the nonnegative Lipschitz constant L_k,L_l.

The assumptions (a) (b) (c) 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. (4) and $r^{(n)}$ be the threshold with kernels of size n. Under Assumptions (a) to (c), then with probability at least $1-\delta$,*

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

The theorem extends the result in (Liu et al., 2020) 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) Let ω_0^*,ω_1^* be the kernel parameters after learning, Z^{te} be the testing samples of size m, then the probability of Type II error*

$$\mathbb{P}_{\mathcal{H}_1}\big(mHSIC_b(Z^{te})\le r^{(m)}|\omega_0^*,\omega_1^*\big)\sim\mathcal{O}(m^{-1/2}).\quad(20)$$

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 the supplementary file 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}.

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 Z of size n, the objective function we used in practice*

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

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 (Zou et al., 2019) in practice. Due to the space limit, we present proofs in the supplementary file.

5 Performance Evaluation

5.1 Compared Methods

We compare the following tests on several datasets.

We consider **the randomized dependence coefficient (RDC)** (Lopez-Paz et al., 2013). 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) (Zhang et al., 2018). 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) (Jitkrittum et al., 2017). 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 (Gretton et al., 2007). The original HSIC testing 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. 3.3.

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 (Jitkrittum et al., 2017). 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 the supplementary file.

5.1.1 Benchmark Datasets

5.2 Results on Synthetic Data

We consider the benchmarks from (Zhang et al., 2018; Jitkrittum et al., 2017) and the application on independent subspace analysis from (Gretton et al., 2007). We also conduct experiments on high-dimensional data based on 3Dshapes (Burgess and Kim, 2018).

We use the following three benchmarks:

Sine Dependency (SD). We begin with a nonlinear dependence model. Concretely,

$$X \sim \mathcal{N}_d(0, I_d), Y = 20\sin\big(4\pi(X_1^2 + X_2^2)\big) + Z, \quad (21)$$

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 nonlinear 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 p_{XY} be the probability density function on $\mathcal{X} \times \mathcal{Y} := [-\pi, \pi]^2$, i.e.,

$$(X, Y) \sim p_{xy}(x, y) \propto 1 + \sin(\omega x)\sin(\omega y), \quad (22)$$

where ω is frequency. Higher frequency makes the drawn samples more similar to those drawn from Uniform($[-\pi, \pi]^2$) (Sejdinovic et al., 2013), thus more difficult to detect dependency for small sample sizes.

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

$$X \sim \mathcal{N}_d(0, I_d), Y = |Z| \prod_{i=1}^{d} \operatorname{sgn}(X_i), \quad (23)$$

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

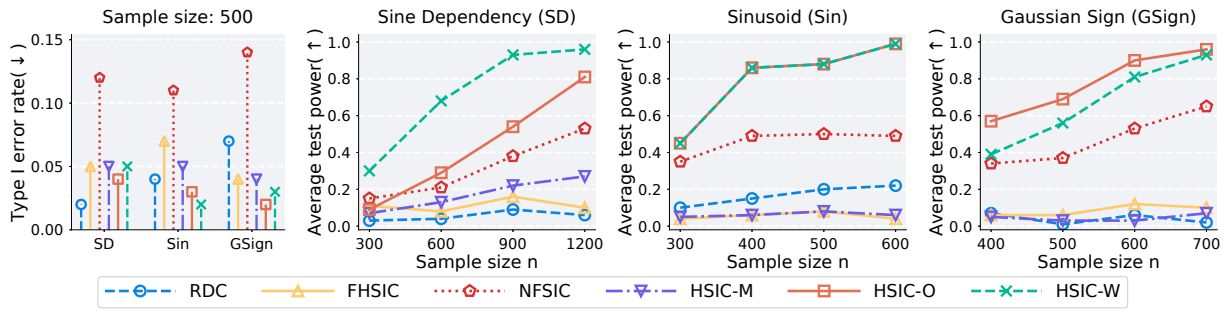

Figure 2: Left: Results of Type I error rate on the three benchmarks with sample size $n = 500$. The other three plots: the results of average test power on the three benchmarks.

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.

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.

5.2.1 Independence of Subspaces

One important application of independence testing is to determine the convergence of algorithms for independent component analysis (ICA) (Gretton et al., 2007), 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 (Theis, 2006). 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.

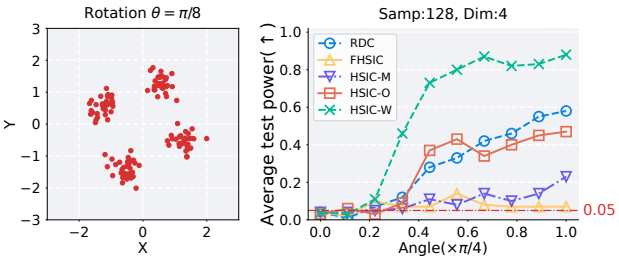

Figure 3: Left: Example dataset for $d = 1, n = 128$ and rotation angles $\theta = \pi/8$. Right: The average test power v.s. the rotation angle of each method.

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 $\leq$ 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.

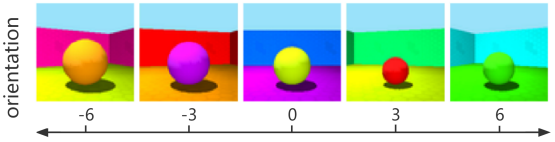

Figure 4: Examples of images generated by varying the orientation factor while fixing the object shape.

5.2.2 High-Dimensional Data

We consider a challenge setting on high-dimensional image data. 3DShapes (Burgess and Kim, 2018) 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.

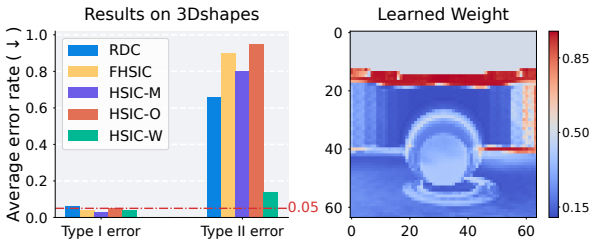

Figure 5: Left: The average rates of the two types of error on the 3Dshapes dataset. Right: The visualization of the learned weights of HSIC-W.

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.

5.3 Results on Real Data

As for real data testing, we consider the subset of the Million Song Data[2] (Bertin-Mahieux, 2011). 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 Alg. 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.

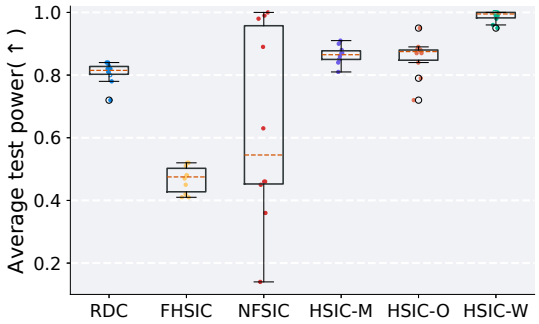

Figure 6: The average test power of 6 methods. The dashed line in each box is the mid-point.

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 Conclusion

In this paper, we propose a novel framework for kernel-based independence tests that enable adaptatively learning parameterized kernels to maximize 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.

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

Acknowledgements

This work was supported by National Key Research and Development Program of China (grant No. 2021YFC3340302). Jihong Guan was supported by National Key Research and Development Program of China (grant No. 2021YFC3300300). Hao Zhang was supported by Supported by the Strategic Priority Research Program of Chinese Academy of Sciences, (grant No. XDB38040200) and National Natural Science Foundation of China (grant No. 62073310, No. 62006051). The authors would like to thank the anonymous reviewers for their valuable comments.

Correspondence authors:
Prof. Shuigeng Zhou (sgzhou@fudan.edu.cn, School of Computer Science, Fudan University),
Dr. Hao Zhang (h.zhang10@siat.ac.cn, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences).

References

Albert, M., Laurent, B., Marrel, A., and Meynaoui, A. (2022). Adaptive test of independence based on hsic measures. *The Annals of Statistics*, 50(2):858–879.

Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48.

Bertin-Mahieux, T. (2011). YearPredictionMSD. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C50K61.

Burgess, C. and Kim, H. (2018). 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/.

Camps-Valls, G., Mooij, J., and Scholkopf, B. (2010). Remote sensing feature selection by kernel dependence measures. *IEEE Geoscience and Remote Sensing Letters*, 7(3):587–591.

Chatterjee, S. (2021). A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536):2009–2022.

Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Fukumizu, K., Gretton, A., Schölkopf, B., and Sriperumbudur, B. K. (2008). Characteristic kernels on groups and semigroups. *Advances in neural information processing systems*, 21.

Gretton, A. (2015). A simpler condition for consistency of a kernel independence test. *arXiv preprint arXiv:1501.06103*.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2006). A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. (2007). A kernel statistical test of independence. *Advances in neural information processing systems*, 20.

Gretton, A., Smola, A., Bousquet, O., Herbrich, R., Belitski, A., Augath, M., Murayama, Y., Pauls, J., Schölkopf, B., and Logothetis, N. (2005). Kernel constrained covariance for dependence measurement. In *International Workshop on Artificial Intelligence and Statistics*, pages 112–119. PMLR.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21.

Jitkrittum, W., Szabó, Z., Chwialkowski, K. P., and Gretton, A. (2016). Interpretable distribution features with maximum testing power. *Advances in Neural Information Processing Systems*, 29.

Jitkrittum, W., Szabó, Z., and Gretton, A. (2017). An adaptive test of independence with analytic kernel embeddings. In *International Conference on Machine Learning*, pages 1742–1751. PMLR.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kübler, J., Jitkrittum, W., Schölkopf, B., and Muandet, K. (2020). Learning kernel tests without data splitting. *Advances in Neural Information Processing Systems*, 33:6245–6255.

Kübler, J. M., Stimper, V., Buchholz, S., Muandet, K., and Schölkopf, B. (2022). Automl two-sample test. *Advances in Neural Information Processing Systems*, 35:15929–15941.

Li, T. and Yuan, M. (2019). On the optimality of gaussian kernel based nonparametric tests against smooth alternatives. *arXiv preprint arXiv:1909.03302*.

Li, Y., Pogodin, R., Sutherland, D. J., and Gretton, A. (2021). Self-supervised learning with kernel dependence maximization. *Advances in Neural Information Processing Systems*, 34:15543–15556.

Liu, F., Xu, W., Lu, J., and Sutherland, D. J. (2021). Meta two-sample testing: Learning kernels for testing with limited data. *Advances in Neural Information Processing Systems*, 34:5848–5860.

Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A., and Sutherland, D. J. (2020). Learning deep kernels for non-parametric two-sample tests. In *International conference on machine learning*, pages 6316–6326. PMLR.

Liu, L., Pal, S., and Harchaoui, Z. (2022). Entropy regularized optimal transport independence criterion. In *International Conference on Artificial Intelligence and Statistics*, pages 11247–11279. PMLR.

Lopez-Paz, D., Hennig, P., and Schölkopf, B. (2013). The randomized dependence coefficient. *Advances in neural information processing systems*, 26.

Lyons, R. (2013). Distance covariance in metric spaces.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch.

Ren, Y., Zhang, H., Xia, Y., Guan, J., and Zhou, S. (2023). Multi-level wavelet mapping correlation for statistical dependence measurement: methodology and performance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6499–6506.

Schrab, A., Kim, I., Guedj, B., and Gretton, A. (2022). Efficient aggregated kernel tests using incomplete u-statistics. *Advances in Neural Information Processing Systems*, 35:18793–18807.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The annals of statistics*, pages 2263–2291.

Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A., and Gretton, A. (2016). Generative models and model criticism via optimized maximum mean discrepancy. *arXiv preprint arXiv:1611.04488*.

Székely, G. J. and Rizzo, M. L. (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances.

Theis, F. (2006). Towards a general independent subspace analysis. *Advances in Neural Information Processing Systems*, 19.

Williams, C. and Rasmussen, C. (1995). Gaussian processes for regression. *Advances in neural information processing systems*, 8.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2012). Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*.

Zhang, Q., Filippi, S., Gretton, A., and Sejdinovic, D. (2018). Large-scale kernel methods for independence testing. *Statistics and Computing*, 28:113–130.

Zou, F., Shen, L., Jie, Z., Zhang, W., and Liu, W. (2019). A sufficient condition for convergences of adam and rmsprop. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11127–11135.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes]

   (b) The license information of the assets, if applicable. [Yes]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]

   (d) Information about consent from data providers/curators. [Yes]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Learning Adaptive Kernels for Statistical Independence Tests: Supplementary Materials

Appendices Organization

- Section A: List of Symbols and Notations.
- Section B and C: Preliminaries and Assumptions.
- Section D: Proof of Theorem 1.
- Section E: Proof of Proposition 3.
- Section F: Gamma Approximation of Threshold.
- Section G: Limit Behavior with Gaussian Kernels.
- Section H: Proof of Theorem 2.
- Section I: Time Complexity.
- Section J: Additional Experiment Results.

A List of Symbols and Notations

$\mathcal{O}$	big O notion
o	small O 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
$\mathbf{E}[X]$	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.

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 $\mathrm{HSIC}_b(Z)$). *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* $\mathcal{H}_0$, $HSIC_b(Z)$ *converges in distribution as*

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

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* $\mathcal{H}_1$, $HSIC_b(Z)$ *converges in distribution to a Gaussian variable*

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

where the variance is given by

$$\sigma_u^2 = 16\big(\mathbf{E}_i(\mathbf{E}_{j,q,r}h_{ijqr})^2 - HSIC(X,Y)^2\big) \tag{3}$$

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

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 $\mathrm{HSIC}_u(Z)$ under $\mathcal{H}_0$

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

$$\mathrm{HSIC}_u(Z) := \frac{1}{(n)_2}\sum_{(i,j)\in\mathbf{i}_2^n} k_{ij}l_{ij} + \frac{1}{(n)_4}\sum_{(i,j,q,r)\in\mathbf{i}_4^n} k_{ij}l_{qr} - 2\frac{1}{(n)_3}\sum_{(i,j,q)\in\mathbf{i}_3^n} k_{ij}l_{iq}, \tag{4}$$

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

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

The complete proof is given in (Gretton et al., 2007). We show only some of the key steps here. The biased estimate of $\mathrm{HSIC}(X,Y)$, denote as $\mathrm{HSIC}_b(Z)$, is a sum of three V-statistics

$$\mathrm{HSIC}_b(Z) := \frac{1}{n^2}\sum_{i,j}^{n} k_{ij}l_{ij} + \frac{1}{n^4}\sum_{i,j,q,r}^{n} k_{ij}l_{qr} - 2\frac{1}{n^3}\sum_{i,j,q}^{n} k_{ij}l_{iq}, \tag{5}$$

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

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

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

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

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

$$\mathbf{E}[\mathrm{HSIC}_b(Z)] = \frac{1}{n}\Big(\mathbf{E}_{xy}kl + \|\mu_x\|^2\|\mu_y\|^2 - \mathbf{E}_xk\|\mu_y\|^2 - \mathbf{E}_yl\|\mu_x\|^2\Big) + \mathcal{O}\big(n^{-2}\big) \tag{7}$$

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

$$\widehat{\|\mu_x\|^2} = \frac{1}{(n)_2}\sum_{(i,j)\in i_2^n} k_{ij},\ \widehat{\|\mu_y\|^2} = \frac{1}{(n)_2}\sum_{(i,j)\in \mathbf{i}_2^n} l_{ij}. \tag{8}$$

The obtained estimate

$$\mathbf{E}[n\mathrm{HSIC}_b(Z)] = 1 + \widehat{\|\mu_x\|^2}\widehat{\|\mu_y\|^2} - \widehat{\|\mu_y\|^2} - \widehat{\|\mu_x\|^2} \tag{9}$$

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 $\mathrm{HSIC}_u(Z)$ under $\mathcal{H}_0$

The complete proof is given in (Gretton et al., 2007). We show only some of the key steps here. According to (Serfling, 2009, Section 5.2.1), the variance of the U-statistic with the kernel can be calculated by

$$\mathbf{Var}[\mathrm{HSIC}_u(Z)] = \binom{n}{4}^{-1}\sum_{c=1}^{4}\binom{4}{c}\binom{n-4}{4-c}\zeta_c = \frac{4\binom{n-4}{3}}{\binom{n}{4}}\zeta_1 + \frac{6\binom{n-4}{2}}{\binom{n}{4}}\zeta_2 + \mathcal{O}(n^{-3}), \tag{10}$$

where we only need to consider the dominant term

$$\zeta_2 = \mathbf{E}_{i,j}\Big[(\mathbf{E}_{q,r}h_{ijqr})\Big]^2 - \underbrace{[\mathbf{E}\mathrm{HSIC}_u(Z)]^2}_{0 \text{ under } \mathcal{H}_0}. \tag{11}$$

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

$$\mathbf{E}_{q,r}h_{ijqr} = \frac{1}{6}(k_{ij} + \mathbf{E}_{xx'}k - \mathbf{E}_xk_i - \mathbf{E}_xk_j)(l_{ij} + \mathbf{E}_{yy'}l - \mathbf{E}_yl_i - \mathbf{E}_yl_j). \tag{12}$$

Combining with the results

$$\begin{aligned} \mathbf{E}_{ij}(k_{ij} + \mathbf{E}_{xx'}k - \mathbf{E}_xk_i - \mathbf{E}_xk_j)^2 &= \mathbf{E}_{ij}\langle\phi(x_i) - \mu_x, \phi(x_j) - \mu_x\rangle^2 \\ &= \mathbf{E}_{ij}\langle(\phi(x_i) - \mu_x)\otimes(\phi(x_i) - \mu_x), (\phi(x_j) - \mu_x)\otimes(\phi(x_j) - \mu_x)\rangle_{\mathrm{HS}} := \|C_{xx}\|^2, \end{aligned} \tag{13}$$

then the variance of the statistic is obtained by

$$\mathbf{Var}[\mathrm{HSIC}_u(Z)] = \frac{2(n-4)(n-5)}{(n)_4}\|C_{xx}\|_{\mathrm{HS}}^2\|C_{yy}\|_{\mathrm{HS}}^2 + \mathcal{O}(n^{-3}), \tag{14}$$

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

$$\frac{\mathbf{1}^T\big(\mathbf{B} - \mathrm{diag}(\mathbf{B})\big)\mathbf{1}}{n(n-1)}, \text{with } \mathbf{B} = \big((\mathbf{HKH})\odot(\mathbf{HLH})\big)^{\cdot 2}, \tag{15}$$

where $\odot$ is the entrywise matrix product and $()^{\cdot 2}$ is the entrywise matrix power. The estimate in Eq. (14) has a bias of $\mathcal{O}(n^{-3})$ and can be calculated within time cost $\mathcal{O}(n^2)$.

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

Since the additional terms of the bias vanish faster than Eq. (14), 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 $\mathrm{HSIC}_u(Z)$ and $\mathrm{HSIC}_b(Z)$

By definition of unbiased estimator $\mathrm{HSIC}_u(Z)$, we have $\mathbf{E}\mathrm{HSIC}_u(Z) = \mathrm{HSIC}(X,Y)$, i.e., the mean of $\mathbf{E}\mathrm{HSIC}_u(Z)$ is equal to the population mean $\mathrm{HSIC}(X,Y)$. And for the mean of $\mathrm{HSIC}_b(Z)$, the result is $\mathrm{HSIC}(X,Y)+\mathcal{O}(n^{-1})$ since the difference between $\mathrm{HSIC}_u(Z)$ and $\mathrm{HSIC}_b(Z)$ is $\mathcal{O}(n^{-1})$ according to Eq. (6).

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

Under $\mathcal{H}_1$, the term ζ_1 in Eq. (10) become positive. In this case, the variance becomes

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

In this paper, we denote

$$\sigma_u^2 := 16\Big(\mathbf{E}_i\big(\mathbf{E}_{j,q,r}h_{ijqr}\big)^2 - \mathrm{HSIC}(X,Y)^2\Big) \tag{17}$$

as the variance of $\sqrt{n}\mathrm{HSIC}_u(Z)$. The variance of $\sqrt{n}\mathrm{HSIC}_b(Z)$ are the same since the difference between them is given in Eq. (6) hence $\sqrt{n}(\mathrm{HSIC}_b(Z) - \mathrm{HSIC}_u(Z)) \sim \mathcal{O}(n^{-1/2})$. The estimator of Eq. (17) can be taken as

$$16 \cdot \Big(\frac{1}{n}\sum_i \big(\frac{1}{n^3}\sum_{j,q,r} h_{ijqr}\big)^2 - \big(\mathrm{HSIC}_b(Z)\big)^2\Big). \tag{18}$$

The terms in Eq. (18) 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

$$\begin{aligned}
\sum_{j,q,r} h_{ijqr} =& \sum_{j,q,r}\frac{1}{4!}\sum_{(t,u,v,w)}^{(i,j,q,r)} k_{tu}l_{tu} + k_{tu}l_{vw} - 2k_{tu}l_{tv} \\
=& \frac{1}{4!}\sum_{j,q,r}\sum_{(u,v,w)}^{(j,q,r)}(k_{iu}l_{iu} + k_{iu}l_{vw} - 2k_{iu}l_{iv}) + \frac{1}{4!}\sum_{j,q,r}\sum_{(t,v,w)}^{(j,q,r)}(k_{ti}l_{ti} + k_{ti}l_{vw} - 2k_{ti}l_{tv}) \\
&+ \frac{1}{4!}\sum_{j,q,r}\sum_{(t,u,w)}^{(j,q,r)}(k_{tu}l_{tu} + k_{tu}l_{iw} - 2k_{tu}l_{ti})] + \frac{1}{4!}\sum_{j,q,r}\sum_{(t,u,v)}^{(j,q,r)}(k_{tu}l_{tu} + k_{tu}l_{vi} - 2k_{tu}l_{tv}) \\
=& \frac{1}{4!}\sum_{j,q,r}\sum_{(u,v,w)}^{(j,q,r)}(2k_{iu}l_{iu} + 2k_{iu}l_{vw} - 2k_{iu}l_{iv}) - \frac{1}{4!}\sum_{j,q,r}\sum_{(t,v,w)}^{(j,q,r)}(2k_{ti}l_{tv}) \\
&+ \frac{1}{4!}\sum_{j,q,r}\sum_{(t,u,w)}^{(j,q,r)}(2k_{tu}l_{tu} + 2k_{tu}l_{iw} - 2k_{tu}l_{ti}) - \frac{1}{4!}\sum_{j,q,r}\sum_{(t,u,v)}^{(j,q,r)}(2k_{tu}l_{tv}) \\
=& \frac{1}{2}\Big[n^2(\mathbf{KL})_{i,i} + (\mathbf{K1})_i(\mathbf{1}^T\mathbf{L1}) - n[(\mathbf{K1})\odot(\mathbf{L1})]_i - n(\mathbf{KL1})_i \\
&+ n\mathrm{Tr}(\mathbf{KL}) + (\mathbf{L1})_i(\mathbf{1}^T\mathbf{K1}) - n(\mathbf{LK1})_i - (\mathbf{1}^T\mathbf{KL1})\Big],
\end{aligned} \tag{19}$$

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 $\mathbf{E}[nHSIC_b(Z)]$ can be given by*

$$\mathcal{E}_0 := 1 + \widehat{\|\mu_x\|^2}\widehat{\|\mu_y\|^2} - \widehat{\|\mu_y\|^2} - \widehat{\|\mu_x\|^2}, \tag{20}$$

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

$$\mathcal{V}_0 = \frac{2(n-4)(n-5)}{(n-1)^2(n-2)(n-3)}\mathbf{1}^T\big(\mathbf{B} - diag(\mathbf{B})\big)\mathbf{1}, \tag{21}$$

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.

C Assumptions

Below are some assumptions we required.

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

$$\sup_{\omega_0\in\Omega_0}\sup_{x\in\mathcal{X}} k_{\omega_0}(x,x) \le \nu, \quad \sup_{\omega_1\in\Omega_1}\sup_{y\in\mathcal{Y}} l_{\omega_1}(y,y) \le \nu. \tag{22}$$

For the kernels we use in practice (e.g. Gaussian kernels), $\nu = 1$.

(ii) The possible 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, Ω_1 is separately bounded by R_{ω_0}, R_{ω_1} respectively, i.e.,

$$\Omega_0 \subseteq \Big\{\omega_0\Big|\ \|\omega_0\| \le R_{\Omega_0}\Big\}, \Omega_1 \subseteq \Big\{\omega_1\Big|\ \|\omega_1\| \le R_{\Omega_1}\Big\}. \tag{23}$$

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

$$\big|k_{\omega_0}(x,x') - k_{\omega_0'}(x,x')\big| \le L_k \cdot \|\omega_0 - \omega_0'\| \tag{24}$$

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

$$\big|l_{\omega_1}(y,y') - l_{\omega_1'}(y,y')\big| \le L_l \cdot \|\omega_1 - \omega_1'\| \tag{25}$$

with the nonnegative Lipschitz constant L_k, L_l.

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

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 $\sigma_u \ge c > 0$ 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. (1) and $r^{(n)}$ be the threshold with kernels of size n. Under Assumptions (i) to (iii), then with probability at least $1-\delta$,*

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

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 $1-\delta$,*

$$\sup_{\omega_0,\omega_1}\left|\Delta^{(b)}_{\xi}(\omega_0,\omega_1)\right| \leq 6\nu^2\sqrt{\frac{2}{n}\log\frac{2}{\delta} + (D_0+D_1)\frac{\log n}{n}} + \frac{12\nu}{\sqrt{n}}(L_k\cdot R_{\Omega_0} + L_l\cdot R_{\Omega_1}) \tag{26}$$

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 Z'. The difference between

$$\hat{\xi} := \frac{1}{(n)_2}\sum_{(i,j)\in \mathbf{i}_2^n} k_{ij}l_{ij} + \frac{1}{(n)_4}\sum_{(i,j,q,r)\in \mathbf{i}_4^n} k_{ij}l_{qr} - \frac{2}{(n)_3}\sum_{(i,j,q)\in \mathbf{i}_3^n} k_{ij}l_{iq} \tag{27}$$

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

$$\begin{aligned}\left|\hat{\xi}-\hat{\xi}'\right| \leq & \frac{1}{(n)_2}\sum_{(i,j)\in \mathbf{i}_2^n, ij:1\in\{i,j\}}\left|k_{ij}l_{ij}-k'_{ij}l'_{ij}\right| + \frac{1}{(n)_4}\sum_{(i,j,q,r)\in \mathbf{i}_4^n, ijqr:1\in\{i,j,q,r\}}\left|k_{ij}l_{qr}-k'_{ij}l'_{qr}\right| \\ & + \frac{2}{(n)_3}\sum_{(i,j,q)\in \mathbf{i}_3^n, ijq:1\in\{i,j,q\}}\left|k_{ij}l_{iq}-k'_{ij}l'_{iq}\right| \leq \frac{(n-1)_1}{(n)_2}\nu^2 + \frac{2(n-1)_3}{(n)_4}\nu^2 + \frac{3(n-1)_2}{(n)_3}\nu^2 = \frac{12\nu^2}{n}.\end{aligned} \tag{28}$$

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 $1-\delta$,

$$\left|\Delta^{(u)}_{\xi}(\omega_0,\omega_1)\right| \leq 6\nu^2\sqrt{\frac{2}{n}\log\frac{2}{\delta}}. \tag{29}$$

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 (Vershynin, 2018, Proposition 4.2.12), by comparing the volumes, we have

$$\mathcal{N}(\Omega_0,r_0) \leq \left(\frac{2R_{\Omega_0}}{r_0}+1\right)^{D_0} \underbrace{\leq \left(\frac{3R_{\Omega_0}}{r_0}\right)^{D_0}}_{\text{when } r_0\leq R_{\Omega_0}}. \tag{30}$$

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 $\omega_0,\omega_0'\in\Omega_0$,

$$\begin{aligned}\left|\hat{\xi}^{(u)}_{\omega_0,\omega_1} - \hat{\xi}^{(u)}_{\omega_0',\omega_1}\right| \leq & \frac{\nu}{(n)_2}\sum_{(i,j)\in \mathbf{i}_2^n}|k^{\omega_0}_{ij}-k^{\omega_0'}_{ij}| + \frac{\nu}{(n)_4}\cdot \\ & \sum_{(i,j,q,r)\in \mathbf{i}_4^n}|k^{\omega_0}_{ij}-k^{\omega_0'}_{ij}| + \frac{2\nu}{(n)_3}\sum_{(i,j,q)\in \mathbf{i}_3^n}|k^{\omega_0}_{ij}-k^{\omega_0'}_{ij}| \leq 4\nu L_k\|\omega_0-\omega_0'\|,\end{aligned} \tag{31}$$

and using the property of unbiased estimate, then

$$\left|\xi_{\omega_0,\omega_1} - \xi_{\omega_0',\omega_1}\right| = \left|\mathbf{E}\hat{\xi}^{(u)}_{\omega_0,\omega_1} - \mathbf{E}\hat{\xi}^{(u)}_{\omega_0',\omega_1}\right| \leq \mathbf{E}\left[\left|\hat{\xi}^{(u)}_{\omega_0,\omega_1} - \hat{\xi}^{(u)}_{\omega_0',\omega_1}\right|\right] \leq 4\nu L_k\|\omega_0-\omega_0'\|. \tag{32}$$

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

$$\left|\xi_{\omega_0,\omega_1} - \xi_{\omega_0,\omega_1'}\right| \leq \left|\hat{\xi}^{(u)}_{\omega_0,\omega_1} - \hat{\xi}^{(u)}_{\omega_0',\omega_1}\right| \leq 4\nu L_l \|\omega_1 - \omega_1'\|. \tag{33}$$

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

$$\|\omega_{0,i} - \omega_0\| \leq r_0, \ \|\omega_{1,j} - \omega_1\| \leq r_1 \tag{34}$$

and also

$$\left|\Delta^{(u)}_\xi(\omega_0,\omega_1)\right| \leq \left|\Delta^{(u)}_\xi(\omega_{0,i},\omega_{1,j})\right| + 4\nu L_k \cdot r_0 + 4\nu L_l \cdot r_1. \tag{35}$$

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

$$\begin{aligned}\sup_{\omega_0,\omega_1} \left|\Delta^{(b)}_\xi(\omega_0,\omega_1)\right| \leq & \max_{ij} \left|\Delta^{(u)}_\xi(\omega_{0,i},\omega_{1,j})\right| + 4\nu L_k \cdot r_0 + 4\nu L_l \cdot r_1 \\ \leq & 6\nu^2 \sqrt{\frac{2}{n} \log \frac{2\mathcal{N}(\Omega_0, r_0)\mathcal{N}(\Omega_1, r_1)}{\delta}} + 4\nu L_k \cdot r_0 + 4\nu L_l \cdot r_1.\end{aligned} \tag{36}$$

We finish the proof of this part by combining Eq. (30) 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$_u(Z)$ and V-statistic estimator HSIC$_b(Z)$ is asymptotically bounded by $\mathcal{O}(n^{-1})$, that is,*

$$|HSIC_b(Z) - HSIC_u(Z)| \sim \mathcal{O}(n^{-1}) \tag{37}$$

Proof. By definition, we have

$$\begin{aligned}|\mathrm{HSIC}_b(Z) - \mathrm{HSIC}_u(Z)| \leq & \Big|\frac{1}{(n)_2} \sum_{(i,j)\in \mathbf{i}_2^n} k_{ij} l_{ij} - \frac{1}{n^2} \sum_{i,j} k_{ij} l_{ij}\Big| \\ & + \Big|\frac{1}{(n)_4} \sum_{(i,j,q,r)\in \mathbf{i}_4^n} k_{ij} l_{qr} - \frac{1}{n^4} \sum_{i,j,q,r} k_{ij} l_{qr}\Big| + 2 \cdot \Big|\frac{1}{(n)_3} \sum_{(i,j,q)\in \mathbf{i}_3^n} k_{ij} l_{iq} - \frac{1}{n^3} \sum_{i,j,q} k_{ij} l_{iq}\Big|.\end{aligned} \tag{38}$$

Take the first term in Eq. (38) as an example,

$$\Big|\frac{1}{(n)_2} \sum_{(i,j)\in \mathbf{i}_2^n} k_{ij} l_{ij} - \frac{1}{n^2} \sum_{i,j} k_{ij} l_{ij}\Big| = \Big|\Big(\frac{1}{(n)_2} - \frac{1}{n^2}\Big) \sum_{(i,j)\in \mathbf{i}_2^n} k_{ij} l_{ij} - \frac{1}{n^2} \sum_i k_{ii} l_{ii}\Big| \leq \frac{\nu^2}{n}, \tag{39}$$

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 HSIC$_b(Z)$ and HSIC$_u(Z)$ is

$$|\mathrm{HSIC}_b(Z) - \mathrm{HSIC}_u(Z)| \leq \frac{\nu^2}{n} + \frac{6\nu^2}{n} + 2 \cdot \frac{3\nu^2}{n} = \frac{13\nu^2}{n}. \tag{40}$$

Thus, the lemma is proved. □

D.2 Convergence results 2

As a start, we denote the random error

$$\Delta^{(b)}_\sigma(\omega_0,\omega_1) := \left|\widehat{\sigma}^2_u(Z) - \sigma^2_u\right|, \tag{41}$$

where the variance is

$$\sigma^2_u = 16\left(\mathbf{E}_i\big(\mathbf{E}_{j,q,r} h_{ijqr}\big)^2 - \mathrm{HSIC}(X,Y)^2\right) \tag{42}$$

and an estimate without regularization is

$$\widehat{\sigma}_u^2(Z) = 16 \cdot \left(\frac{1}{n} \sum_i \Big(\frac{1}{n^3} \sum_{j,q,r} h_{ijqr} \Big)^2 - \Big(\mathrm{HSIC}_b(Z) \Big)^2 \right), \tag{43}$$

where the term

$$h_{ijqr} = \frac{1}{4!} \sum_{(t,u,v,w)}^{(i,j,q,r)} k_{tu} l_{tu} + k_{tu} l_{vw} - 2 k_{uv} l_{tv} \tag{44}$$

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

$$\mathrm{HSIC}_b(Z) = \frac{1}{n^4} \sum_{i,j,q,r}^{n} h_{ijqr}. \tag{45}$$

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

$$\sup_{\omega_0, \omega_1} \big| \Delta_\sigma^{(u)}(\omega_0, \omega_1) \big| \leq 768 \nu^4 \sqrt{\frac{2}{n} \log \frac{2}{\delta} + (D_0 + D_1) \frac{\log n}{n}} + \frac{6272 \nu^4}{n} + \frac{1536 \nu^3}{\sqrt{n}} \Big(L_k \cdot R_{\Omega_0} + L_l \cdot R_{\Omega_1} \Big) \tag{46}$$

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

$$|h_{ijqr}| \leq \frac{1}{4!} \sum_{(t,u,v,w)}^{(i,j,q,r)} |k_{tu} l_{tu} + k_{tu} l_{vw} - 2 k_{uv} l_{tv}| \leq 2\nu^2, \tag{47}$$

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

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 Z'. We denote the counterpart of h_{ijqr} calculated on Z' as h'_{ijqr}. Now we require an upper bound on $|\widehat{\sigma}_u^2(Z) - \widehat{\sigma}_u^2(Z')|$. For the first term in Eq. (43), we have

$$\begin{aligned} &\Big| \frac{1}{n} \sum_i \Big(\frac{1}{n^3} \sum_{j,q,r} h_{ijqr} \Big)^2 - \frac{1}{n} \sum_i \Big(\frac{1}{n^3} \sum_{j,q,r} h'_{ijqr} \Big)^2 \Big| \\ \leq & \frac{1}{n} \sum_i \Big(\underbrace{\frac{1}{n^3} \sum_{j,q,r} \big| h_{ijqr} - h'_{ijqr} \big|}_{\leq 12\nu^2/n \text{ similar as Eq. (28)}} \Big) \cdot \Big(\underbrace{\frac{1}{n^3} \sum_{j,q,r} \big| h_{ijqr} + h'_{ijqr} \big|}_{\leq 4\nu^2 \text{ since } |h_{ijqr}| \leq 2} \Big) \leq \frac{48\nu^4}{n}. \end{aligned} \tag{48}$$

For the second term in Eq. (43), denote $\hat{\xi}_b = \mathrm{HSIC}_b(Z)$, $\hat{\xi}_b' = \mathrm{HSIC}_b(Z')$ (for this part only), we have

$$\Big| \big(\hat{\xi}_b \big)^2 - \big(\hat{\xi}_b' \big)^2 \Big| = \big| \hat{\xi}_b + \hat{\xi}_b' \big| \cdot \big| \hat{\xi}_b - \hat{\xi}_b' \big| \leq 4\nu^2 \cdot \frac{12\nu^2}{n} = \frac{48\nu^4}{n}, \tag{49}$$

since $\hat{\xi}_b, \hat{\xi}_b'$ is bounded in $[0, 2\nu]$ and the bound on difference term can be obtained by a similar analysis in Eq. (28). Therefore,

$$|\widehat{\sigma}_u^2(Z) - \widehat{\sigma}_u^2(Z')| \leq \frac{1536\nu^4}{n}. \tag{50}$$

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

$$|\widehat{\sigma}_u^2(Z) - \mathbf{E}[\widehat{\sigma}_u^2(Z)]| \leq 768 \nu^4 \sqrt{\frac{2}{n} \log \frac{2}{\delta}}. \tag{51}$$

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

$$\mathbf{E}[\widehat{\sigma}_u^2(Z)] = 16 \Big(\frac{1}{n^7} \sum_{ijqrj'q'r'} \mathbf{E}[h_{ijqr} h_{ij'q'r'}] - \frac{1}{n^8} \sum_{ijqri'j'q'r'} \mathbf{E}[h_{ijqr} h_{i'j'q'r'}] \Big) \tag{52}$$

and rewrite the second subterm with matching as

$$\sigma_u^2 = 16\Big(\frac{1}{n^7}\sum_{ijqrj'q'r'}\mathbf{E}[h_{1234}h_{1678}] - \frac{1}{n^8}\sum_{ijqri'j'q'r'}\mathbf{E}[h_{1234}h_{5678}]\Big), \tag{53}$$

where the number subscripts represent a specific set of values, then by calculating the number of non-zero terms and combining Eq. (47), we obtain that

$$|\mathbf{E}[\widehat{\sigma}_u^2(Z)] - \sigma_u^2| \le 16 \cdot \underbrace{\Big(2 - \frac{(n)_7}{n^7} - \frac{(n)_8}{n^8}\Big)}_{\text{by the number of non-zero terms}} \cdot 8\nu^4. \tag{54}$$

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

$$|\mathbf{E}[\widehat{\sigma}_u^2(Z)] - \sigma_u^2| \le 16 \cdot \frac{\binom{7}{2} + \binom{8}{2}}{n} \cdot 8\nu^4 = \frac{6272\nu^4}{n}. \tag{55}$$

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 proof the Lipschitz property of $\widehat{\sigma}_{u,\omega}^2(Z)$ and $\sigma_{u,\omega}^2$. One can check that for fixed a, b, c, d,

$$\big|k_{ab}^{(\omega)}l_{cd}^{(\omega)} - k_{ab}^{(\omega')}l_{cd}^{(\omega')}\big| \le \big|k_{ab}^{(\omega)}\big|\cdot\big|l_{cd}^{(\omega)} - l_{cd}^{(\omega')}\big| + \big|k_{ab}^{(\omega)} - k_{ab}^{(\omega')}\big|\cdot\big|l_{cd}^{(\omega')}\big| \le \nu\Big(L_k\|\omega_0 - \omega_0'\| + L_l\|\omega_1 - \omega_1'\|\Big). \tag{56}$$

by assumption (i) and (iii). Then combining with Eq. (44), for each term we can use the results of Eq. (56), then

$$|h_{ijqr}^{(\omega)} - h_{ijqr}^{(\omega')}| \le 4\nu\Big(L_k\|\omega_0 - \omega_0'\| + L_l\|\omega_1 - \omega_1'\|\Big). \tag{57}$$

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

$$\big|h_{ijqr}^{(\omega)}h_{abcd}^{(\omega)} - h_{ijqr}^{(\omega')}h_{abcd}^{(\omega')}\big| \le \big|h_{ijqr}^{(\omega)}\big|\cdot\big|h_{abcd}^{(\omega)} - h_{abcd}^{(\omega')}\big| + \big|h_{ijqr}^{(\omega)} - h_{ijqr}^{(\omega')}\big|\cdot\big|h_{abcd}^{(\omega')}\big| \le 16\nu^3\Big(L_k\|\omega_0 - \omega_0'\| + L_l\|\omega_1 - \omega_1'\|\Big) \tag{58}$$

As a result,

$$\begin{aligned}\big|\widehat{\sigma}_{u,\omega}^2(Z) - \widehat{\sigma}_{u,\omega'}^2(Z)\big| &\le \frac{16}{n^7}\sum_{ijqrbcd}\big|h_{ijqr}^{(\omega)}h_{ibcd}^{(\omega)} - h_{ijqr}^{(\omega')}h_{ibcd}^{(\omega')}\big| + \frac{16}{n^8}\sum_{ijqrabcd}\big|h_{ijqr}^{(\omega)}h_{abcd}^{(\omega)} - h_{ijqr}^{(\omega')}h_{abcd}^{(\omega')}\big| \\ &\le 512\nu^3\Big(L_k\|\omega_0 - \omega_0'\| + L_l\|\omega_1 - \omega_1'\|\Big).\end{aligned} \tag{59}$$

Again using a similar process, we can show that

$$\big|\sigma_{u,\omega}^2 - \sigma_{u,\omega'}^2\big| \le 512\nu^3\Big(L_k\|\omega_0 - \omega_0'\| + L_l\|\omega_1 - \omega_1'\|\Big) \tag{60}$$

Since we focus on the same parameter space as before, we take the same cover set we used in Eq. (30). Then, by combing the results of Eqs. (51), (55), (59) and (60), we have with probability at least $1 - \delta$,

$$\sup_{\omega_0,\omega_1}\big|\Delta_\sigma^{(u)}(\omega_0, \omega_1)\big| \le 768\nu^4\sqrt{\frac{2}{n}\log\frac{2\mathcal{N}(\Omega_0, r_0)\mathcal{N}(\Omega_1, r_1)}{\delta}} + \frac{6272\nu^4}{n} + 512\nu^3\Big(L_k \cdot r_0 + L_l \cdot r_1\Big). \tag{61}$$

We finish the proof of this part by combining Eq. (30) 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 $\sigma_u \geq c$ on $\overline{\Omega}_c := \overline{\Omega}_0 \times \overline{\Omega}_1$, according to the Eq. (61), when

$$n \geq N_0 = \left(\frac{2}{c}\right)^3 \cdot \left[768\nu^4\Big(\sqrt{2\log\frac{4}{\delta}} + \sqrt{D_0+D_1}\Big) + 1536\nu^3\big(L_k R_{\Omega_0} + L_l R_{\Omega_1}\big) + 6272\nu^4\right]^3. \tag{62}$$

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

$$\begin{aligned}\sup_{\omega_0,\omega_1}\left|\frac{\hat{\xi}-r^{(n)}/n}{\widehat{\sigma}} - \frac{\xi - r/n}{\sigma}\right| &\leq \sup_{\omega_0,\omega_1}\left(\left|\frac{\hat{\xi}-\xi}{\widehat{\sigma}}\right| + \left|\frac{r^{(n)}-r}{n\widehat{\sigma}}\right| + \left|\frac{\xi+r/n}{\sigma}\right|\cdot\left|\frac{\widehat{\sigma}-\sigma}{\widehat{\sigma}}\right|\right) \\ &\leq \sup_{\omega_0,\omega_1}\left(\frac{2}{c}\cdot\left|\hat{\xi}-\xi\right| + \frac{2}{cn}\cdot\left|r^{(n)}-r\right| + \frac{8(\nu+r/n)}{3c^3}\cdot\left|\widehat{\sigma}^2-\sigma^2\right|\right).\end{aligned} \tag{63}$$

Combining the results Eqs. (36) and (61), also according to the results (Korolyuk and Borovskich, 2013, Theorem 13) that shown that $|r^n - r| \sim o(n^{-1/2})$ and $\sup_{\omega_0,\omega_1} r < \infty$, thus we have

$$\sup_{\omega_0,\omega_1}\left|\frac{\hat{\xi}-r^{(n)}/n}{\widehat{\sigma}} - \frac{\xi - r/n}{\sigma}\right| = \mathcal{O}\left(\frac{1}{c^3}\left[\sqrt{\frac{2}{n}\log\frac{4}{\delta} + (D_0+D_1)\frac{\log n}{n}} + \frac{L_k R_{\Omega_0} + L_l R_{\Omega_1}}{\sqrt{n}}\right]\right) = \mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right), \tag{64}$$

with probability at least $1-\delta$.

E Proof of Proposition 3.

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

$$\mathbb{P}_{\mathcal{H}_1}\big(m\mathit{HSIC}_b(Z^{te}) \leq r^{(m)} | \omega_0^*, \omega_1^*\big) \sim \mathcal{O}(m^{-1/2}). \tag{65}$$

Proof. For simplify, we denote $\hat{\xi}_b$ as the biased estimator of $\mathrm{HSIC}_b(Z^{te})$ 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. (29), we can have a uniform bound for $r^{(m)}$ for $m\hat{\xi}_b$ that

$$r^{(m)} \leq 6\nu^2\sqrt{2\log\frac{2}{1-\alpha}}\sqrt{m} + 13\nu^2 \sim \mathcal{O}(m^{1/2}). \tag{66}$$

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

$$\begin{aligned}\mathbb{P}\Big(\text{Type II error}\Big) &= \mathbb{P}_{\mathcal{H}_1}\Big(m\hat{\xi}_b \leq r^{(m)} | \omega_0^*, \omega_1^*\Big) \leq \mathbb{P}_{\mathcal{H}_1}\Big(m\hat{\xi}_u \leq r^{(m)} + 13\nu^2 \big| \omega_0^*, \omega_1^*\Big) \\ &= \mathbb{P}_{\mathcal{H}_1}\left(\frac{\sqrt{m}(\hat{\xi}_u - \mathbf{E}\hat{\xi}_u)}{4\sigma^{1/2}} \leq \frac{r^{(m)}/\sqrt{m} - \sqrt{m}\mathbf{E}\hat{\xi}_u + 13\nu^2/\sqrt{m}}{4\sigma^{1/2}}\Big| \omega_0^*, \omega_1^*\right).\end{aligned} \tag{67}$$

According to the results shown in (Serfling, 2009, Section 5.5.1 Theorem B), and also $\sigma > 0$ under $\mathcal{H}_1$, we have

$$\begin{aligned}\mathbb{P}\Big(\text{Type II error}\Big) &\leq \Phi\left(\frac{r^{(m)}/\sqrt{m} - \sqrt{m}\mathbf{E}\hat{\xi}_u + 13\nu^2/\sqrt{m}}{4\sigma^{1/2}}\right) + \frac{C_1\nu_h}{\sigma^{3/2}}\frac{1}{\sqrt{m}} \\ &\leq \Phi\Big(C_2 - C_3\sqrt{m}\mathbf{E}\hat{\xi}_u + C_4/\sqrt{m}\Big) + C_5\frac{1}{\sqrt{m}}\end{aligned} \tag{68}$$

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

$$\Phi(x) = -\frac{e^{-x^2}}{2x\sqrt{\pi}}\left(1 + \sum_{n=1}^{\infty}(-1)^n \frac{1\cdot 3\cdot 5\cdots(2n-1)}{(2x^2)^n}\right), \tag{69}$$

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

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

$$n\mathrm{HSIC}_b(Z) \sim \frac{x^{\gamma-1}e^{-x/\beta}}{\beta^\gamma \Gamma(\gamma)}, \tag{70}$$

where the two parameters are

$$\gamma = \frac{(\mathbf{E}[\mathrm{HSIC}_b(Z)])^2}{\mathbf{Var}[\mathrm{HSIC}_b(Z)]},\ \beta = \frac{n\mathbf{Var}[\mathrm{HSIC}_b(Z)]}{\mathbf{E}[\mathrm{HSIC}_b(Z)]}. \tag{71}$$

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

$$\frac{1}{n}\mathrm{Tr}C_{xx}\mathrm{Tr}C_{yy} = \frac{1}{n}\Big(1 + \|\mu_x\|^2\|\mu_y\|^2 - \|\mu_x\|^2 - \|\mu_y\|^2\Big). \tag{72}$$

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

$$\widehat{\|\mu_x\|^2} = \frac{1}{(n)_2}\sum_{(i,j)\in i_2^n} k_{ij},\ \widehat{\|\mu_y\|^2} = \frac{1}{(n)_2}\sum_{(i,j)\in \mathbf{i}_2^n} l_{ij}. \tag{73}$$

Alternatively, a matrix expression

$$\mathbf{E}[\mathrm{HSIC}_b(Z)] = \frac{1}{n}\Big(\frac{1}{n-1}\mathrm{Tr}(\mathbf{HKH})\Big)\Big(\frac{1}{n-1}\mathrm{Tr}(\mathbf{HLH})\Big), \tag{74}$$

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

$$\mathbf{Var}[\mathrm{HSIC}_b(Z)] = \frac{2(n-4)(n-5)}{(n)_4}\|C_{xx}\|_{\mathrm{HS}}^2\|C_{yy}\|_{\mathrm{HS}}^2, \tag{75}$$

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

$$\frac{\mathbf{1}^T\big(\mathbf{B} - \mathrm{diag}(\mathbf{B})\big)\mathbf{1}}{n(n-1)}, \text{with } \mathbf{B} = \big((\mathbf{HKH})\odot(\mathbf{HLH})\big)^{\cdot 2}, \tag{76}$$

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

$$\mathcal{E}_0 = \Big(\frac{1}{n-1}\mathrm{Tr}\mathbf{K_c}\Big)\Big(\frac{1}{n-1}\mathrm{Tr}\mathbf{L_c}\Big), \mathcal{V}_0 = \frac{2(n-4)(n-5)}{(n-1)^2(n-2)(n-3)}\mathbf{1}^T\big(\mathbf{B} - \mathrm{diag}(\mathbf{B})\big)\mathbf{1}, \tag{77}$$

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

$$\mathcal{E}_0 = \mathbf{E}[n\mathrm{HSIC}_b(Z)], \mathcal{V}_0 = \mathbf{Var}[n\mathrm{HSIC}_b(Z)]. \tag{78}$$

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

$$\int_0^{\widehat{c_\alpha}} \frac{x^{\gamma-1}e^{-x/\beta}}{\beta^\gamma\Gamma(\gamma)}dx = \int_0^{\frac{\widehat{c_\alpha}}{\beta}} \frac{x^{\gamma-1}e^{-x}}{\Gamma(\gamma)}dx = 1-\alpha, \tag{79}$$

where Γ is the gamma function.

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

$$k(x, x') = \exp\big(-\eta \cdot \|x - x'\|^2\big), \tag{80}$$

where the parameter $\eta = 1/2\omega_x^2$. We also consider the case with fixed samples Z 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, $\mathcal{X} \subseteq \{x \in \mathbb{R}^d : \|x\| \leq R_{\mathcal{X}}/2\}$ 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. $(\mathbf{K}_c)_{ij}^2 > 0, (\mathbf{L}_c)_{ij}^2 > 0$ for all $i \neq j$ 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 $\|x_i - x_j\| \neq 0, \|y_i - y_j\| \neq 0$ for all $i \neq j$ 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

$$(\mathbf{K}_c)_{ij} = k_{ij} - \frac{1}{n}\sum_i k_{ij} - \frac{1}{n}\sum_j k_{ij} + \frac{1}{n^2}\sum_{ij} k_{ij}. \tag{81}$$

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

$$\mathbf{K}(\eta = 0^+) = \mathbf{1}\mathbf{1}^T, \mathbf{K}(\eta = +\infty) = \mathbf{I}. \tag{82}$$

and the limit of centering kernel matrix

$$\mathbf{K}_c(\eta = 0^+) = \mathbf{O}, \mathbf{K}_c(\eta = +\infty) = \mathbf{H}. \tag{83}$$

Then by substituting the results of Eqs. (82) and (83) gives the limit of two moments of null distribution

$$\begin{aligned} &\mathcal{E}_0(\eta = 0^+) = 0,\ \mathcal{E}_0(\eta = +\infty) = \Big(\frac{1}{n-1}\mathrm{Tr}\mathbf{L_c}\Big). \\ &\mathcal{V}_0(\eta = 0^+) = 0,\ \mathcal{V}_0(\eta = +\infty) = \frac{2(n-4)(n-5)}{n^2(n-1)^2(n-2)(n-3)}\sum_{i\neq j}(\mathbf{L_c})_{ij}^2. \end{aligned} \tag{84}$$

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

$$\mathcal{E}_1 := n\mathrm{HSIC}_b(Z) = \frac{1}{n}\mathrm{Tr}(\mathbf{HKHL}). \tag{85}$$

The limit is

$$\mathcal{E}_1(\eta = 0^+) = 0, \mathcal{E}_1(\eta = +\infty) = \frac{1}{n}\mathrm{Tr}(\mathbf{L_c}). \tag{86}$$

And the variance is

$$\mathcal{V}_1 := \widehat{\sigma}_u^2(Z) = 16 \cdot \left(\frac{1}{n}\sum_i \Big(\frac{1}{n^3}\sum_{j,q,r} h_{ijqr}\Big)^2 - \Big(\mathrm{HSIC}_b(Z)\Big)^2\right), \tag{87}$$

where the term

$$h_{ijqr} = \frac{1}{4!} \sum_{(t,u,v,w)}^{(i,j,q,r)} k_{tu}l_{tu} + k_{tu}l_{vw} - 2k_{uv}l_{tv} \tag{88}$$

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 $\mathrm{HSIC}_b(Z)$ can be expressed as

$$\mathrm{HSIC}_b(Z) = \frac{1}{n^4} \sum_{i,j,q,r}^{n} h_{ijqr}. \tag{89}$$

To compute the limit easily, we can also make use of its matrix expression of the term given in Eq. (19). Then

$$\left[\sum_{j,q,r} h_{ijqr}\right]\Big|_{\eta=0^+} = 0, \left[\sum_{j,q,r} h_{ijqr}\right]\Big|_{\eta=+\infty} = \frac{n^2(\mathbf{L}_c)_{ii} + n\mathrm{Tr}(\mathbf{L_c})}{2}. \tag{90}$$

After that, the limit of Eq. (87) can be calculated that

$$\mathcal{V}_1\big|_{\eta=0^+} = 0, \mathcal{V}_1\big|_{\eta=+\infty} = \frac{4}{n^2}\left[\frac{\mathrm{Tr}\big[(\mathbf{L_c})^{\cdot 2}\big]}{n} - \Big(\frac{\mathrm{Tr}(\mathbf{L_c})}{n}\Big)^2\right]. \tag{91}$$

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.

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 $J_\lambda(Z) \equiv 0$ 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 Z of size n, the objective function we used in practice*

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

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)

$$k(x, x') = \exp\big(-\eta \cdot \|x - x'\|^2\big) \tag{92}$$

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

$$\frac{\partial k(x,x')}{\partial \eta} = -\|x - x'\|^2 \exp\big(-\eta \cdot \|x - x'\|^2\big) \leq 0, \frac{\partial^2 k(x,x')}{\partial^2 \eta} = \|x - x'\|^4 \exp\big(-\eta \cdot \|x - x'\|^2\big) \geq 0. \tag{93}$$

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

$$0 \leq -\nabla_\eta k \leq R_{\mathcal{X}}^2, 0 \leq \nabla_\eta^2 k \leq R_{\mathcal{X}}^4. \tag{94}$$

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

$$0 \leq \nabla_\eta^2 k \leq -\nabla_\eta k \leq k \leq 1. \tag{95}$$

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. (79), 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

$$\mathcal{E} := \mathcal{E}_0 = \Big(\frac{1}{n-1}\mathrm{Tr}\mathbf{K_c}\Big)\Big(\frac{1}{n-1}\mathrm{Tr}\mathbf{L_c}\Big), \mathcal{V} := \frac{\mathbf{1}^T(\mathbf{B} - \mathrm{diag}(\mathbf{B}))\mathbf{1}}{n(n-1)} = \frac{(n-1)(n-2)(n-3)}{2n(n-4)(n-5)}\mathcal{V}_0 = C_0\mathcal{V}_0, \tag{96}$$

where $\mathcal{E}_0, \mathcal{V}_0$ are the two moments of null distribution defined in Eq. (84). 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

$$\begin{aligned}
\nabla_\eta \mathcal{E} &= \Big(\frac{1}{n-1}\mathrm{Tr}\nabla_\eta\mathbf{K_c}\Big)\Big(\frac{1}{n-1}\mathrm{Tr}\mathbf{L_c}\Big) = \Big(\frac{1}{n(n-1)}\mathbf{1}^T[-\nabla_\eta\mathbf{K}]\mathbf{1}\Big)\Big(\frac{1}{n-1}\mathrm{Tr}\mathbf{L_c}\Big) \geq 0 \\
\nabla_\eta^2 \mathcal{E} &= \Big(\frac{1}{n(n-1)}\mathbf{1}^T[-\nabla_\eta^2\mathbf{K}]\mathbf{1}\Big)\Big(\frac{1}{n-1}\mathrm{Tr}\mathbf{L_c}\Big) \leq 0.
\end{aligned} \tag{97}$$

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

$$|\nabla_\eta \mathcal{E}| \leq 1, \ |\nabla_\eta^2 \mathcal{E}| \leq 1, \tag{98}$$

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

$$\nabla_\eta \mathbf{B} = 2 \cdot \nabla_\eta\mathbf{K_c} \odot \mathbf{K_c} \odot (\mathbf{L_c})^{\cdot 2}, \nabla_\eta^2\mathbf{B} = 2 \cdot \Big(\nabla_\eta^2\mathbf{K_c} \odot \mathbf{K_c} + (\nabla_\eta\mathbf{K_c})^{\cdot 2}\Big) \odot (\mathbf{L_c})^{\cdot 2}. \tag{99}$$

Also, the following results can be given

$$\begin{aligned}
0 \leq \mathcal{V} &\leq \frac{1}{n(n-1)}\sum_{i\neq j}\mathbf{B}_{ij} \leq \frac{4}{n(n-1)}\sum_{i\neq j}|\mathbf{L}_c|_{ij}^2 \leq 16, \\
|\nabla_\eta\mathcal{V}| &\leq \frac{1}{n(n-1)}\sum_{i\neq j}|\nabla_\eta\mathbf{B}|_{ij} \leq \frac{8}{n(n-1)}\sum_{i\neq j}|\mathbf{L}_c|_{ij}^2 \leq 32, \\
|\nabla_\eta^2\mathcal{V}| &\leq \frac{1}{n(n-1)}\sum_{i\neq j}|\nabla_\eta^2\mathbf{B}|_{ij} \leq \frac{16}{n(n-1)}\sum_{i\neq j}|\mathbf{L}_c|_{ij}^2 \leq 64,
\end{aligned} \tag{100}$$

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

$$\mathcal{V} = \frac{1}{n(n-1)}\sum_{i\neq j}(\mathbf{K_c})_{ij}^2(\mathbf{L_c})_{ij}^2 \geq \frac{1}{n(n-1)}\frac{\big(\sum_{i\neq j}(\mathbf{K_c})_{ij}\big)^2}{\sum_{i\neq j}(\mathbf{L_c})_{ij}^{-2}}. \tag{101}$$

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

$$\mathcal{V} \geq \frac{1}{n(n-1)}\frac{(-\mathrm{Tr}\mathbf{K_c})^2}{\sum_{i\neq j}(\mathbf{L_c})_{ij}^{-2}} = \frac{n-1}{n}\frac{\mathcal{E}^2}{\Big(\frac{1}{n-1}\mathrm{Tr}\mathbf{L_c}\Big)^2}\frac{1}{\sum_{i\neq j}(\mathbf{L_c})_{ij}^{-2}}. \tag{102}$$

Since n and $\mathbf{L_c}$ is all fixed, we set the const in Eq. (102) 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 $C_1\big(C(\eta_{min})\big)^2 \leq \mathcal{V} \leq 16$.

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. (71) and Eq. (96). 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. (79), the threshold $\widehat{c_\alpha}$ is the solution of the following equation

$$\int_0^{\widehat{c_\alpha}} \frac{x^{\gamma-1}e^{-x/\beta}}{\beta^\gamma \Gamma(\gamma)} dx = \int_0^{\frac{\widehat{c_\alpha}}{\beta}} \frac{x^{\gamma-1}e^{-x}}{\Gamma(\gamma)} dx = 1 - \alpha, \tag{103}$$

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 inequation 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

$$\alpha = \mathbb{P}(T \geq \widehat{c_\alpha}) \leq \frac{\mathbf{E}e^{\lambda T}}{e^{\lambda \widehat{c_\alpha}}} = (1 - \beta\lambda)^{-\gamma} e^{-\lambda \widehat{c_\alpha}}, \tag{104}$$

which indicates that

$$\widehat{c_\alpha} \leq \min_{\lambda \in [0, \beta^{-1}]} \frac{-\log(\alpha) - \gamma \log(1 - \beta\lambda)}{\lambda}. \tag{105}$$

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

$$\widehat{c_\alpha} \leq 2 \cdot \Big(\frac{1}{\gamma} \log(\frac{1}{\alpha}) + \log 2\Big) \cdot \mathcal{E}. \tag{106}$$

And for the lower bound, we have

$$1 - \alpha = \int_0^{\frac{\widehat{c_\alpha}}{\beta}} \frac{x^{\gamma-1}e^{-x}}{\Gamma(\gamma)} dx < \int_0^{\frac{\widehat{c_\alpha}}{\beta}} \frac{x^{\gamma-1}}{\Gamma(\gamma)} dx = \frac{\widehat{c_\alpha}^\gamma}{\beta^\gamma \Gamma(\gamma+1)} \tag{107}$$

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

$$\widehat{c_\alpha} \geq \min_{\gamma \geq C_2} (1-\alpha)^{1/\gamma} \Gamma(\gamma+1)^{1/\gamma} \beta = (1-\alpha)^{1/C_2} \Gamma(C_2+1)^{1/C_2} \beta, \tag{108}$$

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

$$\int_0^{\frac{\widehat{c_\alpha}}{\beta}} \frac{x^{\gamma-1}e^{-x}}{\Gamma(\gamma)} dx = 1 - \alpha. \tag{109}$$

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,

$$\underbrace{g_\gamma\Big(\frac{\widehat{c_\alpha}}{\beta}\Big)}_{\bigstar} \cdot \nabla_\eta \Big(\frac{\widehat{c_\alpha}}{\beta}\Big) + \underbrace{\int_0^{\frac{\widehat{c_\alpha}}{\beta}} [\nabla_\eta g_\gamma](x) dx}_{\blacksquare} = 0. \tag{110}$$

For the first term in Eq. (110)

$$\bigstar = g_\gamma\Big(\frac{\widehat{c_\alpha}}{\beta}\Big) = \frac{\widehat{c_\alpha}^{\gamma-1} e^{-\widehat{c_\alpha}/\beta}}{\beta^{\gamma-1}\Gamma(\gamma)}, \tag{111}$$

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. (110)

$$\blacksquare = \big(\nabla_\eta \gamma\big) \cdot \underbrace{\left\{ \int_0^{\frac{\widehat{c_\alpha}}{\beta}} x^{\gamma-1} e^{-x} \left(\frac{\ln(x)}{\Gamma(\gamma)} + \nabla_\gamma \Big(\frac{1}{\Gamma(\gamma)} \Big) \right) dx \right\}}_{\spadesuit}, \tag{112}$$

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. $\nabla_\gamma \Big(\frac{1}{\Gamma(\gamma)} \Big)$ 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 $\left| \nabla_\eta \Big(\frac{\widehat{c_\alpha}}{\beta} \Big) \right|$, this futher 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. (110),

$$(\nabla_\eta \bigstar) \cdot \nabla_\eta \Big(\frac{\widehat{c_\alpha}}{\beta} \Big) + \bigstar \cdot \nabla_\eta^2 \Big(\frac{\widehat{c_\alpha}}{\beta} \Big) + \Big(\nabla_\eta \blacksquare \Big) = 0. \tag{113}$$

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

$$\begin{aligned} \nabla_\eta \bigstar = \nabla_\eta \left(\frac{\exp\left\{ (\gamma - 1) \log(\frac{\widehat{c_\alpha}}{\beta}) - \frac{\widehat{c_\alpha}}{\beta} \right\}}{\Gamma(\gamma)} \right) = \exp\left\{ (\gamma - 1) \log\Big(\frac{\widehat{c_\alpha}}{\beta} \Big) - \frac{\widehat{c_\alpha}}{\beta} \right\} \cdot \left\{ \nabla_\eta \Big(\frac{1}{\Gamma(\gamma)} \Big) + \frac{1}{\Gamma(\gamma)} \cdot \right. \\ \left. \Big[(\nabla_\eta \gamma) \log\Big(\frac{\widehat{c_\alpha}}{\beta} \Big) + (\gamma - 1) \Big(\frac{\widehat{c_\alpha}}{\beta} \Big)^{-1} \cdot \nabla_\eta \Big(\frac{\widehat{c_\alpha}}{\beta} \Big) - \nabla_\eta \Big(\frac{\widehat{c_\alpha}}{\beta} \Big) \Big] \right\} \end{aligned} \tag{114}$$

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

$$\nabla_\eta \blacksquare = \big(\nabla_\eta^2 \gamma \big) \cdot \spadesuit + \big(\nabla_\eta \gamma \big) \cdot \nabla_\eta \spadesuit, \tag{115}$$

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

$$\begin{aligned} \nabla_\eta \spadesuit =& \nabla_\eta \left\{ \int_0^{\frac{\widehat{c_\alpha}}{\beta}} x^{\gamma-1} e^{-x} \left(\frac{\ln(x)}{\Gamma(\gamma)} + \nabla_\gamma \Big(\frac{1}{\Gamma(\gamma)} \Big) \right) dx \right\} \\ =& \Big(\frac{\widehat{c_\alpha}}{\beta} \Big)^{\gamma-1} e^{-\frac{\widehat{c_\alpha}}{\beta}} \left(\frac{\ln\Big(\frac{\widehat{c_\alpha}}{\beta} \Big)}{\Gamma(\gamma)} + \nabla_\gamma \Big(\frac{1}{\Gamma(\gamma)} \Big) \right) \cdot \nabla_\eta \Big(\frac{\widehat{c_\alpha}}{\beta} \Big) + \\ & \left\{ \int_0^{\frac{\widehat{c_\alpha}}{\beta}} \frac{x^{\gamma-1}}{e^x} \left[\frac{\ln^2(x)}{\Gamma(\gamma)} + \nabla_\gamma \Big(\frac{2 \ln(x)}{\Gamma(\gamma)} \Big) + \nabla_\gamma^2 \Big(\frac{1}{\Gamma(\gamma)} \Big) \right] dx \right\} \cdot \nabla_\eta(\gamma). \end{aligned} \tag{116}$$

The terms in Eq. (116) all satisfy the bounded condition. As a result, we have proved that $\nabla_\eta \Big(\frac{\widehat{c_\alpha}}{\beta} \Big)$ is a Lipschitz continuous function. According to

$$\nabla_\eta \widehat{c_\alpha} = \beta \cdot \left(\nabla_\eta \Big(\frac{\widehat{c_\alpha}}{\beta} \Big) - \widehat{c_\alpha} \cdot \nabla_\eta \Big(\frac{1}{\beta} \Big) \right), \tag{117}$$

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 $n\text{HSIC}_b(Z)$ and $\widehat{\sigma}_u^2(Z) + \lambda$. We define

$$\mathcal{E}_1 := n\text{HSIC}_b(Z), \ \mathcal{V}_{1,\lambda} := \widehat{\sigma}_u^2(Z) + \lambda = \mathcal{V}_1 + \lambda, \tag{118}$$

where $\mathcal{V}_1$ is defined in Eq. (87). 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}.

$$|h_{ijqr}| \leq \frac{1}{4!} \sum_{(t,u,v,w)}^{(i,j,q,r)} |k_{tu}l_{tu} + k_{tu}l_{vw} - 2k_{uv}l_{tv}| \leq 2, \tag{119}$$

since for all i, j, k_{ij} and l_{ij} both take the value in $[0, 1]$. Combining with Eq. (89), we can quickly conclude that

$$0 \leq \mathcal{E}_1 \leq \frac{1}{n^3} \sum_{i,j,q,r}^{n} |h_{ijqr}| \leq 2n, \tag{120}$$

and the variance

$$0 < \lambda \leq \mathcal{V}_{1,\lambda} \leq 16 \cdot \left(\frac{1}{n} \sum_{i} \Big(\frac{1}{n^3} \sum_{j,q,r} h_{ijqr} \Big)^2 \right) + \lambda \leq 64 + \lambda. \tag{121}$$

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

$$\begin{aligned} \nabla_\eta h_{ijqr} &= \frac{1}{4!} \sum_{(t,u,v,w)}^{(i,j,q,r)} \nabla_\eta k_{tu}l_{tu} + \nabla_\eta k_{tu}l_{vw} - 2\nabla_\eta k_{uv}l_{tv}, \\ \nabla_\eta^2 h_{ijqr} &= \frac{1}{4!} \sum_{(t,u,v,w)}^{(i,j,q,r)} \nabla_\eta^2 k_{tu}l_{tu} + \nabla_\eta^2 k_{tu}l_{vw} - 2\nabla_\eta^2 k_{uv}l_{tv}. \end{aligned} \tag{122}$$

Hence the bounds can be obtained in a similar way,

$$|\nabla_\eta h_{ijqr}| \leq 2, |\nabla_\eta^2 h_{ijqr}| \leq 2, \tag{123}$$

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

$$\begin{aligned} &|\nabla_\eta \mathcal{V}_{1,\lambda}| \leq 32 \Big(\frac{1}{n} \sum_{i} \Big| \frac{1}{n^3} \sum_{j,q,r} h_{ijqr} \Big| \cdot \Big| \frac{1}{n^3} \sum_{j,q,r} \nabla_\eta h_{ijqr} \Big| + \Big| \frac{1}{n^4} \sum_{i,j,q,r}^{n} \nabla_\eta h_{ijqr} \Big| \cdot \Big| \frac{1}{n^4} \sum_{i,j,q,r}^{n} \nabla_\eta h_{ijqr} \Big| \Big) \leq 256 \\ &|\nabla_\eta^2 \mathcal{V}_{1,\lambda}| \leq 512. \end{aligned} \tag{124}$$

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

$$J_\lambda(Z) = \frac{n\mathrm{HSIC}_b(Z) - \widehat{c_\alpha}}{\widehat{\sigma}_{u,\lambda}(Z)} = \frac{\mathcal{E}_1 - \widehat{c_\alpha}}{\sqrt{\mathcal{V}_{1,\lambda}}} \tag{125}$$

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

I 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 $\mathcal{O}\big(n^2(d_x+d_y)\big)$ and computing the terms $\mathrm{HSIC}_b(Z)$, $\widehat{\sigma}_u^2(Z)$ and $\widehat{c_\alpha}$ costs $\mathcal{O}\big(n^2\big)$ as shown in Sec. B and by the definition of $\widehat{c_\alpha}$ given in Sec. F. Hence the total time complexity is $\mathcal{O}\big(Tn^2(d_x + d_y)\big)$, 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.

J Additional Experiment Results

J.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. 3 in the main paper). We first show the results of the experiments under more settings in Fig. 1. 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 $\frac{n\text{HSIC}(X,Y)}{\sqrt{n}\sigma_u} \sim \mathcal{O}(\sqrt{n})$ 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. 2. 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. 3. 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.

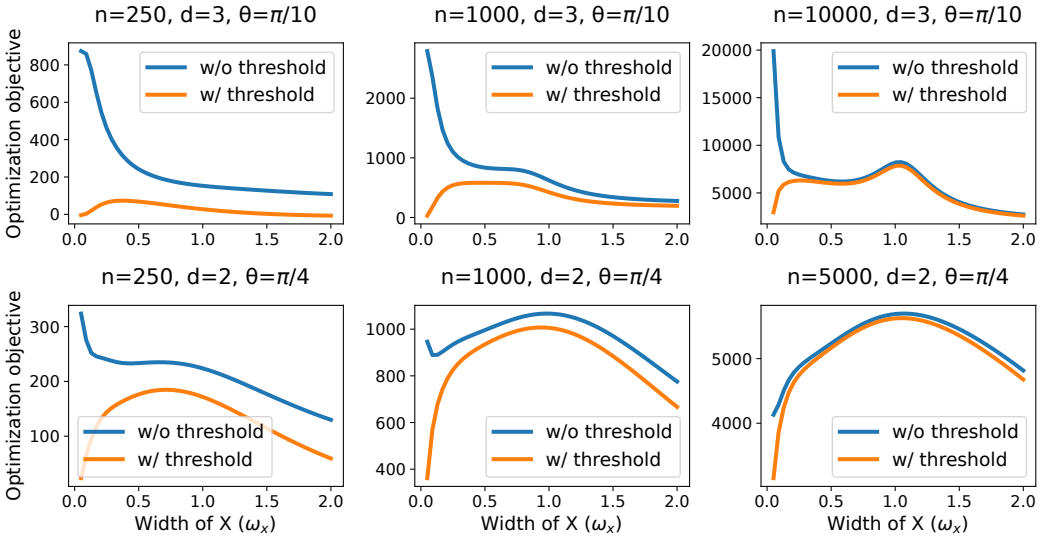

Figure 1: The values of optimization objective for different ω_x on the ISA dataset under more settings.

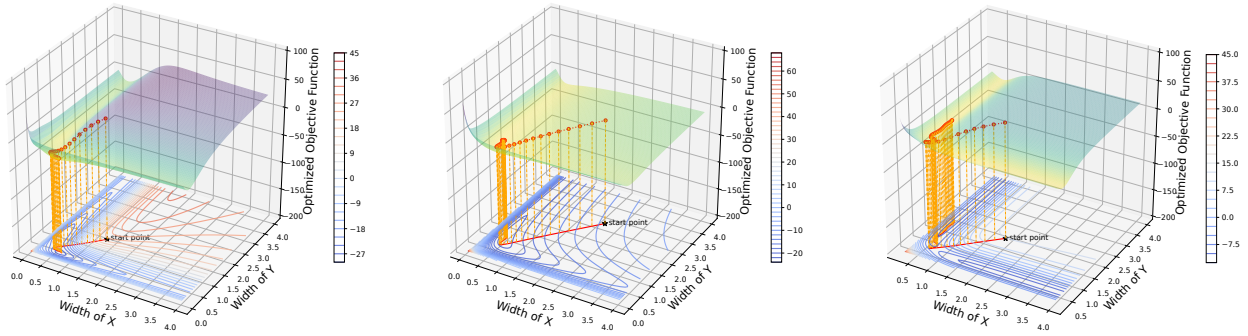

Figure 2: The visualization results of gradient descent process for HSIC-O on the ISA dataset. From left to right: 1) Gaussian kernel with learnable width. $n = 128$, $\theta = \pi/10$, $d = 2$. 2) Laplace kernel with learnable width. $n = 128$, $\theta = \pi/10$, $d = 2$. 3) Gaussian kernel with learnable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

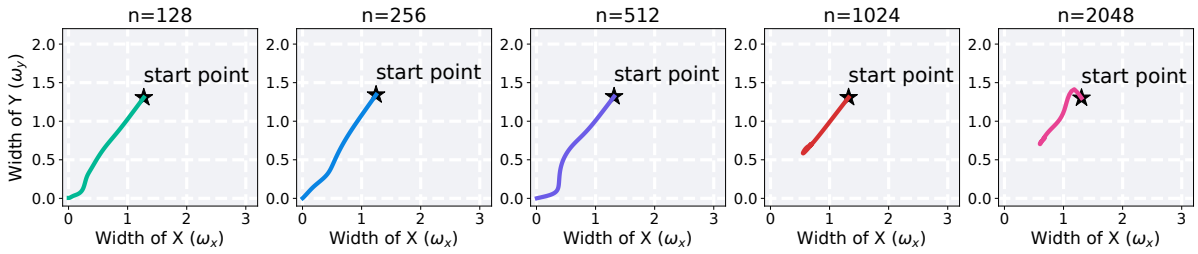

Figure 3: The visualization results of the kernel bandwidth optimization process when using the criterion (w.o. threshold). The number of iterations is sufficient to ensure convergence. Setup: The ISA dataset is used, d is fixed to 2 and $\theta = \pi/10$. The sample size n is changed from 128 to 2048.

J.2 Experiment Results with More Settings and More Compared Methods

In this section, we provide the experimental results under more settings in Fig. 4.

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, 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.

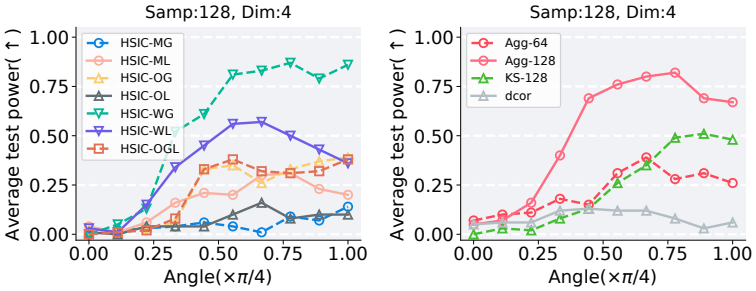

Figure 4: Results of experiments under the ISA dataset with more settings.

More compared methods. On the right, we compare with more methods. We consider the distance-based methods (Székely and Rizzo, 2013) called distance covariance (dcor) as well as the aggregated kernels tests (Schrab et al., 2022). 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. 4. 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.

J.3 Running Time.

The running time of each method is given in Fig. 5. 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.

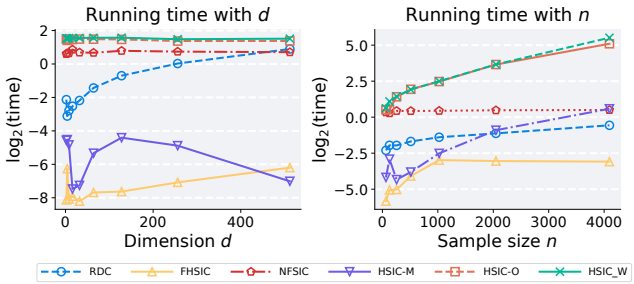

Figure 5: The practical running time of each method. Left: fix $n = 128$. Right: fix $d = 4$.

Potential future research directions.

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 (Schrab et al., 2022) 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.

References

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. (2007). A kernel statistical test of independence. *Advances in neural information processing systems*, 20.

Korolyuk, V. S. and Borovskich, Y. V. (2013). *Theory of U-statistics*, volume 273. Springer Science & Business Media.

Schrab, A., Kim, I., Guedj, B., and Gretton, A. (2022). Efficient aggregated kernel tests using incomplete u-statistics. *Advances in Neural Information Processing Systems*, 35:18793–18807.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*. John Wiley & Sons.

Székely, G. J. and Rizzo, M. L. (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.