Efficiently Learning Significant Fourier Feature Pairs for Statistical Independence Testing

Yixin Ren[1], **Yewei Xia**[1,4], **Hao Zhang**[3,*], **Jihong Guan**[2], **Shuigeng Zhou**[1,*]

[1]Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, Shanghai, China
[2]Department of Computer Science and Technology, Tongji University, Shanghai, China
[3]SIAT, Chinese Academy of Sciences, Shenzhen, China
[4]Machine Learning Department, MBZUAI, Abu Dhabi, UAE
{yxren21, ywxia23}@m.fudan.edu.cn, h.zhang10@siat.ac.cn
jhguan@tongji.edu.cn, sgzhou@fudan.edu.cn

Abstract

We propose a novel method to efficiently learn significant Fourier feature pairs for maximizing the power of Hilbert-Schmidt Independence Criterion (HSIC) based independence tests. We first reinterpret HSIC in the frequency domain, which reveals its limited discriminative power due to the inability to adapt to specific frequency-domain features under the current inflexible configuration. To remedy this shortcoming, we introduce a module of learnable Fourier features, thereby developing a new criterion. We then derive a finite sample estimate of the test power by modeling the behavior of the criterion, thus formulating an optimization objective for significant Fourier feature pairs learning. We show that this optimization objective can be computed in linear time (with respect to the sample size n), which ensures fast independence tests. We also prove the convergence property of the optimization objective and establish the consistency of the independence tests. Extensive empirical evaluation on both synthetic and real datasets validates our method's superiority in effectiveness and efficiency, particularly in handling high-dimensional data and dealing with large-scale scenarios.

1 Introduction

Testing for independence is a crucial and challenging task in machine learning and statistics, with wide-range applications in causal inference [16, 31], feature selection [6] and deep learning [23, 42]. Its primary objective is to determine whether two random variables, X and Y are independent, based on the observations of the underlying joint distribution $\mathbb{P}_{XY}$. While traditional independence tests, such as Pearson's correlation coefficient [9] and Kendall's τ, can only detect monotonic relationships between low-dimensional variables, more modern tests [26, 43, 7, 25, 27, 35, 19, 20] aim to deal with complex non-linear interactions in much more challenging higher-dimensional space [45, 29].

One class of nonlinear dependence measures [3, 15] aims to capture distributional characteristics using kernel embeddings [13], primarily derived from the cross-covariance operators in the reproducing kernel Hilbert space (RKHS). Among them, Hilbert-Schmidt Independence Criterion (HSIC) [14] is the most popular one. It utilizes the squared Hilbert-Schmidt norm to detect dependence and exhibits outstanding performance across various data contexts by choosing suitable kernels. On the other hand, some other fundamental nonlinear dependence measures employ characteristic functions

*Corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

to detect the smoothed discrepancy between the joint distribution and the product of marginals. By employing appropriate characteristic functions, the statistic [39, 40] computes the covariance between distances of variable pairs. It has been demonstrated that these distance-based methods are equivalent to HSIC with specific kernels [33]. However, all these measures suffer from the drawback of requiring quadratic time (w.r.t. the sample size n) to compute the feature covariance and necessitating fixed kernel or distance functions, rendering them impractical on large-scale datasets due to the unaffordable time cost and lacking flexibility in handling complex scenarios.

To address these challenges, a multitude of works grounded on these measures have emerged. Upon HSIC, [44] proposes some linear-time tests including a block-averaged statistic, a statistic with Nyström approximation, and one with finite-dimensional feature mappings using random Fourier features (RFF) [28]. For convenience, these tests are referred to as BHSIC, NyHSIC, and FHSIC, respectively. FHSIC and NyHSIC are observed to have a considerable advantage over BHSIC. However, a remaining drawback of these methods is that the features are not learnable. Therefore, these methods lack enough adaptability to complex settings, thus leading to performance degradation.

In addition to time efficiency, another research direction [1, 30] aims to make independence tests adaptive to better capturing distributional distinctions. These methods either select/combine appropriate kernels from a predefined set or learn parameterized kernels. Nonetheless, their criteria still inherit the quadratic time complexity of HSIC, thus cannot be readily applied to large-scale data.

Furthermore, some approaches [17, 32] try to address both challenges simultaneously. For instance, HSICAgg [32] suggests combining several kernels from a predefined set (e.g. kernels with different preset bandwidths) and aggregating the test results for improving performance. Additionally, an incomplete U-statistic of HSIC is proposed to ensure computational efficiency. Nevertheless, selecting from a predefined set of kernels imposes limitations on flexibility, and in cases where scaling optimization is required on each dimension, the number of kernel pairs escalates exponentially. Also, NFSIC [18] proposes to combine a time-efficient technique called analytic kernel embeddings [8, 17] and learn the important local distributional features. However, its learning objective is merely a lower bound of test power and demands a substantial number of samples to ensure accuracy.

In this paper, we propose a novel test method that flexibly learns distributional features while maintaining high efficiency. We first reinterpret HSIC from a frequency-domain perspective, then we point out its potential shortcomings with an elaborate example and indicate corresponding improvement directions. Finally an central optimization objective is derived by directly modeling test power, which can be computed in linear time while maximizing the test performance. Comparing with [30] that also addresses the kernel learning problem in independence testing with a time/space complexity of $\mathcal{O}(n^2)$, our criteria for learning are designed to have a complexity of $\mathcal{O}(n)$ for both space and time. Consequently, the whole test framework can efficiently handle large-scale data.

Contributions. In summary, the contributions of the work are as follows: 1) We propose a novel approach that efficiently learns significant Fourier feature pairs for maximizing the power of HSIC-based independence tests. 2) We design an optimization objective that can be computed in linear time, which is derived by directly modeling test power. 3) We theoretically establish the non-asymptotic convergence property of the optimization objective and demonstrate the consistency of our method. 4) We conduct extensive experiments on both synthetic and real data, showcasing its superiority in effectiveness and efficiency in handling high-dimensional data (e.g. image data) and addressing large-scale scenarios.

Outline. The rest of the paper is organized as follows: Sec. 2 reviews HSIC-based statistical independence tests. Sec. 3 reinterprets HSIC from a frequency-domain perspective, and explain its potential shortcomings with an elaborate example and indicate corresponding improvement directions. Sec. 4 designs an optimization objective by directly modeling test power, which can be computed in linear time. Sec. 5 presents the theoretical analysis and Sec. 6 evaluates the performance of the proposed method on synthetic and real dataset. We conclude the paper in Sec. 7.

2 Preliminaries and Notations

We begin by introducing notions and reviewing the hypothesis testing framework for independence tests. Let $\mathcal{X} \times \mathcal{Y}$ be separable metric space, typically $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$. $\mathbb{P}_{XY}$ denotes a Borel probability measure defined on $\mathcal{X} \times \mathcal{Y}$, while $\mathbb{P}_X$ and $\mathbb{P}_Y$ denote the respective marginal distributions. 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 aim to test whether X, Y are independent (i.e., $X \perp\!\!\!\perp Y$). This corresponds to a hypothesis testing problem formulated as $\mathcal{H}_0 : \mathbb{P}_{XY} = \mathbb{P}_X\mathbb{P}_Y$ versus $\mathcal{H}_1 : \mathbb{P}_{XY} \neq \mathbb{P}_X\mathbb{P}_Y$.

The testing procedure is as follows: First, define the statistic ρ and calculate its estimated value using the samples. Then, choose a significance level α (typically set to 0.05), which represents the probability that the sampling of ρ 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 occur in this procedure. Type I error occurs when $\mathcal{H}_0$ is falsely rejected, while Type II error happens when $\mathcal{H}_0$ is incorrect but not rejected. A good test [43] needs to control Type I error within α while maximizing the testing power ($1-$Type II error rate).

For independence tests, a commonly used statistic is HSIC, defined as follows:

Definition 1. *[14]. Let $\mathcal{F}$ be an RKHS with kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ and $\mathcal{G}$ be a second RKHS on $\mathcal{Y}$ with kernel $l : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$, the HSIC between X and Y, denoted as HSIC(X, Y) is defined as*

$$\mathbf{E}\big[k(X, X')l(Y, Y')\big]+\mathbf{E}\big[k(X, X')\big]\mathbf{E}\big[l(Y, Y')\big]-2\mathbf{E}_{X'Y'}\big[\mathbf{E}_X k(X, X')\mathbf{E}_Y l(Y, Y')\big], \quad (1)$$

where (X', Y') is a independent copy of (X, Y). An estimator of HSIC(X, Y) is given by

$$\textit{HSIC}_b(Z) := \frac{1}{n^2}\sum_{i,j} k_{ij}l_{ij} + \frac{1}{n^4}\sum_{i,j,q,r} k_{ij}l_{qr} - 2\frac{1}{n^3}\sum_{i,j,q} k_{ij}l_{iq} = \frac{1}{n^2}\textit{Tr}(\mathbf{KHLH}), \quad (2)$$

where $k_{ij} := k(x_i, x_j)$, $l_{ij} := l(y_i, y_j)$ are the entries of the $n \times n$ kernel matrices $\mathbf{K}$, $\mathbf{L}$ respectively, $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is the centering matrix and $\mathbf{1}$ is a vector of ones.

3 Revisiting HSIC from Frequency Domain Perspective

We denote $\mathcal{F}$ as the Fourier transform, and $\mathcal{F}^{-1}$ as its inverse. When the kernels k, l are translation-invariant, i.e., there exist functions ψ, ψ_k, ψ_l such that for all $(x, x') \in \mathcal{X} \times \mathcal{X}$ and $(y, y') \in \mathcal{Y} \times \mathcal{Y}$,

$$\psi(x - x', y - y') = \psi_k(x - x')\psi_l(y - y') = k(x, x')l(y, y'). \quad (3)$$

Then, according to the results of [36, Corollary 4], the HSIC with function ψ can be formulated as

$$\text{HSIC}(X, Y) = \int_{\mathbb{R}^{d_x}\times\mathbb{R}^{d_y}} \left|\phi_{\mathbb{P}_{XY}}(\omega) - \phi_{\mathbb{P}_X\mathbb{P}_Y}(\omega)\right|^2 (\mathcal{F}^{-1}\psi)(\omega)d\omega, \quad (4)$$

where $\omega = (\omega_x, \omega_y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, ω_x, ω_y are the frequencies of X and Y respectively, and

$$\phi_{\mathbb{P}_{XY}}(\omega) := \int e^{-i(\omega_x^T x+\omega_y^T y)}d\mathbb{P}_{XY},\ \phi_{\mathbb{P}_X\mathbb{P}_Y}(\omega) := \left(\int e^{-i\omega_x^T x}d\mathbb{P}_X\right)\left(\int e^{-i\omega_y^T y}d\mathbb{P}_Y\right) \quad (5)$$

are the characteristic functions of $\mathbb{P}_{XY}$ and $\mathbb{P}_X\mathbb{P}_Y$, respectively. Intuitively, Eq. (4) means that HSIC can be understood as the difference between the joint distribution and the product of the marginal distributions in the frequency domain, with different weights $(\mathcal{F}^{-1}\psi)(\omega)$ being attached to different frequencies, which are determined by the kernel function. When $\mathcal{F}^{-1}\psi$ is almost everywhere non-zero, it can be shown that the kernel is characteristic [36, 10]. The characteristic condition ensures that the criterion is discriminative for discrepancies at almost all frequencies. However, with inappropriate choices of $\mathcal{F}^{-1}\psi$, the differences may not be significant enough. We explain this with an example:

Example. Consider the Sinusoid model that $\mathcal{X} \times \mathcal{Y} := [-\pi, \pi]^2$ and $(X, Y) \sim p_{xy}(x, y) \propto 1 + \sin(\omega_0 x)\sin(\omega_0 y)$, where p_{xy} is the probability density function and ω_0 is a positive integer. Combining Eq. (5), we can calculate that $\phi_{\mathbb{P}_X\mathbb{P}_Y}(\omega) = \delta(\omega_x)\delta(\omega_y)$ and $\phi_{\mathbb{P}_{XY}}(\omega) = \delta(\omega_x)\delta(\omega_y) + [\delta(\omega_x + \omega_0) + \delta(\omega_x - \omega_0)][\delta(\omega_y + \omega_0) + \delta(\omega_y - \omega_0)]$, where δ is the Dirac delta function, thus the difference between them only relies on the frequency ω_0. When the Gaussian kernels with width $\sqrt{2}\lambda_x$ and $\sqrt{2}\lambda_y$ are used, i.e., $k(x, x') = \exp(-\|x - x'\|_2^2/(4\lambda_x^2)), l(y, y') = \exp(-\|y - y'\|_2^2/(4\lambda_y^2))$, then the inverse Fourier transform of ψ is $(\mathcal{F}^{-1}\psi)(\omega_x, \omega_y) = \pi^{-1}\lambda_x\lambda_y \exp(-(\lambda_x^2\omega_x^2 + \lambda_y^2\omega_y^2))$. Hence HSIC$(X, Y) = 4\pi^{-1}\lambda_x\lambda_y \exp(-(\lambda_x^2 + \lambda_y^2)\omega_0^2)$ whose maximum is taken at $\lambda_x^* = \lambda_y^* = 1/(\sqrt{2}\omega_0)$, indicating that the widths need to be adjusted to focus on some specific frequencies. If the common setting [14, 44] is adapted, which uses mid-widths (i.e., the median distance does

not change with ω_0 since the marginal distributions do not change with ω_0), then the criterion will exponentially decline to 0 as ω_0 increases. In contrast, the criterion using the adaptive optimization width $(1/\omega_0, 1/\omega_0)$ decreases at a rate of $\mathcal{O}(\omega_0^{-2})$, which is a considerable improvement.

This example illustrates the loss of the discriminatory power of the criterion when an inappropriate $\mathcal{F}^{-1}\psi$ is chosen. The discriminatory power of the criterion heavily impacts the sample size required for the test to obtain significant results in practice, and existing inflexible configurations may lead to inadequate test power in the presence of reasonably large sample sizes. Consequently, it is important to design learnable $\mathcal{F}^{-1}\psi$. To this end, we subsequently design a learnable objective and let it be optimized in a data-driven manner. Before this, we provide an approach to make the criterion be computed efficiently. This can be achieved by sampling in the frequency domain. Formally, a finite-dimensional approximation in the frequency domain of the integral in Eq. (4) is given as follows:

$$\mathrm{HSIC}_\omega(X,Y) := \frac{1}{D_x D_y}\sum_{i=1}^{D_x}\sum_{j=1}^{D_y}\left|\phi_{\mathbb{P}_{XY}}(\omega_{x;i},\omega_{y;j}) - \phi_{\mathbb{P}_X\mathbb{P}_Y}(\omega_{x;i},\omega_{y;j})\right|^2, \tag{6}$$

where $\{\omega_{x;i}\}_{i=1}^{D_x}, \{\omega_{y;j}\}_{j=1}^{D_y}$ are sampled independently with the measure $\mathcal{F}^{-1}\psi_k, \mathcal{F}^{-1}\psi_l$, respectively. Note that $\mathcal{F}^{-1}\psi$ is a product measure, i.e., $\mathcal{F}^{-1}\psi = (\mathcal{F}^{-1}\psi_k) \otimes (\mathcal{F}^{-1}\psi_l)$. This type of approximation is also called random Fourier features (RFF) [28] that had been applied to various kernel algorithms. We will incorporate this technique to efficiently perform computation later.

4 Learning Significant Fourier Feature Pairs

4.1 HSIC with Learnable Fourier Feature Pairs

To design $\mathcal{F}^{-1}\psi$, we need to make sure that $\mathrm{supp}(\mathcal{F}^{-1}\psi) = \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ to meet the characteristic condition and that its integral over the full space is 1 to ensure it is a probability measure. Also, for practical utility, $\mathcal{F}^{-1}\psi$ should embody a familiar probability density function, facilitating sampling procedures. Fortunately, a versatile array of options emerges through the judicious selection of kernels [2] with adjustable parameters. Take kernel k as an example, some commonly used kernels are listed in Tab. 1, and their inverse Fourier transforms are listed simultaneously. Additionally, to

Table 1: Some popular kernels (parameterized by σ, Σ) with corresponding density functions.

Kernel	$\psi_k(\Delta)$	$\mathcal{F}^{-1}\psi_k(\omega)$	$\mathcal{T}_{\theta_k}(x)$	$p_k(\omega)$
Gaussian	$e^{-\frac{\|\Delta\|_2^2}{2\sigma^2}}$	$(2\pi)^{-d_x/2}\sigma e^{-\sigma^2\|\omega\|_2^2/2}$	x/σ	$(2\pi)^{-d_x/2}e^{-\|\omega\|_2^2/2}$
Laplace	$e^{-\frac{\|\Delta\|_1}{\sigma}}$	$\sqrt{\frac{2}{\pi}}\prod_d \frac{\sigma}{\sigma^2+\omega_d^2}$	x/σ	$\sqrt{\frac{2}{\pi}}\prod_d \frac{1}{1+\omega_d^2}$
Mahalanobis	$e^{-\frac{1}{2}\Delta^T\Sigma^{-1}\Delta}$	$(2\pi)^{-d_x/2}\lvert\Sigma\rvert^{-1/2}e^{-\omega^T\Sigma^{-1}\omega/2}$	$\Sigma^{1/2}x$	$(2\pi)^{-d_x/2}e^{-\|\omega\|_2^2/2}$

be able to apply gradient-based optimization techniques, we invoke a method that disentangle the sampled objects and the learnable parameters. Specifically, we leverage a variable transform $\mathcal{T}_{\theta_k}$ (a bijection function parameterized with θ_k) to convert the probability measure $\mathcal{F}^{-1}\psi_k$ into a simple distribution (e.g. a standard Gaussian distribution) $p_k(\omega)$. Simultaneously, we relocate the learnable component onto X. Consequently, we can focus on learning parameterized transformations $\mathcal{T}_{\theta_k}$ and simplifying the computation by enabling sampling directly from $p_k(\omega)$.

Remark. The above scheme provides a broader form for designing. The mapping $\mathcal{T}_{\theta_k}$ can be viewed as a feature extractor, which makes it possible to flexibly combine models (e.g., neural network) thus incorporating deep kernel [24] into the framework. Also, it should be noted that the single kernel example can also be extended to multi-kernel setting [11] by executing the procedure for each kernel.

Next, we obtain the learnable independence criterion and utilize the sampling technique as in Eq. (6) to compute efficiently. Note that for simplicity, we take the same value for both D_x and D_y in Eq. (6) by default. By the definition, the kernel function can be expressed as

$$\psi_k\left(\mathcal{T}_{\theta_k}x - \mathcal{T}_{\theta_k}x'\right) = \mathcal{F}[\mathcal{F}^{-1}\psi_k(\omega)] = \int e^{-i\omega^T(\mathcal{T}_{\theta_k}x - \mathcal{T}_{\theta_k}x')}p_k(\omega)d\omega. \tag{7}$$

[2]The bounded, continuous, translation-invariant kernel satisfies the characteristic condition [12].

By applying the frequency sampling technique, we obtain the approximation as

$$\psi_k^{(\omega)}\left(\mathcal{T}_{\theta_k}x-\mathcal{T}_{\theta_k}x'\right):=\frac{2}{D}\sum_{j=1}^{D/2}e^{-i\omega_{k;j}^T(\mathcal{T}_{\theta_k}x-\mathcal{T}_{\theta_k}x')}=\frac{2}{D}\sum_{j=1}^{D/2}\cos\left(\omega_{k;j}^T(\mathcal{T}_{\theta_k}x-\mathcal{T}_{\theta_k}x')\right), \quad (8)$$

where $\{\omega_{k;j}\}_{j=1}^{D/2}$ are sampled independently with distribution $p_k(\omega)$ and the last equation is because the kernel function is real. To get a more computationally tractable form, we define

$$\Lambda_k(x):=\sqrt{\frac{2}{D}}\left[\cos\big(\omega_1^T\mathcal{T}_{\theta_k}x\big),\sin\big(\omega_1^T\mathcal{T}_{\theta_k}x\big),...,\cos\big(\omega_{D/2}^T\mathcal{T}_{\theta_k}x\big),\sin\big(\omega_{D/2}^T\mathcal{T}_{\theta_k}x\big)\right], \quad (9)$$

called learnable RFF of k then Eq. (8) becomes $\psi_k^{(\omega)}\left(\mathcal{T}_{\theta_k}x-\mathcal{T}_{\theta_k}x'\right)=\Lambda_k(x)\Lambda_k(x')^T$. The expression with a similar form is also given in [44], with the difference that we have added learnable parts. For Y, we define the corresponding symbols by substituting k for l and x for y. Also, for convenience, we default to keeping Y and X the same number of samples D from here on. Then the HSIC with learnable RFF pairs can be obtained by replacing k,l in Eq. (1) to $\psi_k^{(\omega)},\psi_l^{(\omega)}$. Also, the corresponding estimator with sample Z can be obtained by replacing $\mathbf{K},\mathbf{L}$ in Eq. (2) to the matrices $\mathbf{\Lambda}_X\mathbf{\Lambda}_X^T,\mathbf{\Lambda}_Y\mathbf{\Lambda}_Y^T$, where $\mathbf{\Lambda}_X:=[\Lambda_k(x_1);...;\Lambda_k(x_n)]_{n\times D}$ and so as define for $\mathbf{\Lambda}_Y$. As a result,

$$\text{HSIC}_\omega(Z):=\frac{1}{n^2}\text{Tr}(\mathbf{\Lambda}_X\mathbf{\Lambda}_X^T\mathbf{H}\mathbf{\Lambda}_Y\mathbf{\Lambda}_Y^T\mathbf{H})=\frac{1}{n^2}\text{Tr}(\mathbf{\Lambda}_X^T\mathbf{H}\mathbf{\Lambda}_Y\mathbf{\Lambda}_Y^T\mathbf{H}\mathbf{\Lambda}_X)=\frac{1}{n^2}\|\mathbf{\Lambda}_{Xc}^T\mathbf{\Lambda}_{Yc}\|_F^2, \quad (10)$$

where $\mathbf{\Lambda}_{Xc}:=\mathbf{H}\mathbf{\Lambda}_X,\mathbf{\Lambda}_{Yc}:=\mathbf{H}\mathbf{\Lambda}_Y$. The time complexity is analyzed as follows. Since the computation of the mapping $\mathcal{T}_{\theta_k}x$ depends on the specific design, here we default to analyzing the kernel case shown in Tab. 1. In this case, computing $\mathbf{\Lambda}_X,\mathbf{\Lambda}_Y$ requires $\mathcal{O}\big(nD(d_x+d_y)\big)$ time. Then calculate $\mathbf{\Lambda}_{Xc},\mathbf{\Lambda}_{Yc}$ cost $\mathcal{O}(nD)$. After that, calculate $\text{HSIC}_\omega(Z)$ cost $\mathcal{O}(nD^2)$. Hence, the overall time complexity is $\mathcal{O}\big(nD(d_x+d_y+D)\big)$, i.e. the running time is linear with n.

4.2 Linear-time Optimization Objective

Next, we model the behavior of $\text{HSIC}_\omega(Z)$ to obtain an optimization objective for maximizing the power of the test. By utilizing the property that $\text{HSIC}_\omega(Z)$ is a V-statistic, we can extend the results [14, Theorem 1, 2] for $\text{HSIC}_\omega(Z)$, as shown in the following proposition with the proof given in the Appendix. To simplify, we denote (x_i,y_i) as z_i to represent the i-th sample and denote $\psi_k^{(\omega)}\left(\mathcal{T}_{\theta_k}x_t-\mathcal{T}_{\theta_k}x_u\right)$ as $k_{tu}^{(\omega)}$ and $\psi_l^{(\omega)}\left(\mathcal{T}_{\theta_l}y_t-\mathcal{T}_{\theta_l}y_u\right)$ as $l_{tu}^{(\omega)}$.

Proposition 1 (Asymptotics). *Let $h_{ijqr}^{(\omega)}:=\frac{1}{4!}\sum_{(t,u,v,w)}^{(i,j,q,r)}k_{tu}^{(\omega)}l_{tu}^{(\omega)}+k_{tu}^{(\omega)}l_{vw}^{(\omega)}-2k_{uv}^{(\omega)}l_{tv}^{(\omega)}$, where the sum represents all ordered quadruples (t,u,v,w) drawn without replacement from (i,j,q,r). Then, Under the null hypothesis $\mathcal{H}_0$, $\text{HSIC}_\omega(Z)$ coverages in distribution to*

$$n\text{HSIC}_\omega(Z)\xrightarrow{d}\sum_{l=1}^{\infty}\lambda_l\chi_{1l}^2,\ \ \lambda_l g_l(z_j)=\int_{z_i,z_q,z_r}h_{ijqr}^{(\omega)}g_l(z_i)dF_{z_i,z_q,z_r}, \quad (11)$$

where $\chi_{11}^2,\chi_{12}^2,...$ are independent χ_1^2 variates and λ_l is the solution to the eigenvalue problem as in the right of Eq. (11). Also, under the alternative $\mathcal{H}_1$, $\text{HSIC}_\omega(Z)$ converges in distribution as

$$n^{\frac{1}{2}}\Big(\text{HSIC}_\omega(Z)-\mathbf{E}_Z\text{HSIC}_\omega(Z)\Big)\xrightarrow{d}\mathcal{N}(0,\sigma_\omega^2),\ \sigma_\omega^2:=16\Big[\mathbf{E}_i(\mathbf{E}_{j,q,r}h_{ijqr}^{(\omega)})^2-\big(\mathbf{E}_Z h_{ijqr}^{(\omega)}\big)^2\Big] \quad (12)$$

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

According to Proposition 1, the power of the test with HSIC_ω can be formulated by

$$\mathbb{P}_{\mathcal{H}_1}\left(n\text{HSIC}_\omega(Z)>r_\omega\right)\to\Phi\left(\frac{n\mathbf{E}_Z\text{HSIC}_\omega(Z)-r_\omega}{\sqrt{n}\sigma_\omega}\right), \quad (13)$$

where Φ is the standard normal CDF and r_ω is the threshold, i.e. $(1-\alpha)$-quantile of distribution given in Eq. (11) that exactly controls Type I error rate to the nominal level α. Hence, to maximize the power of the test, a natural criterion is $[n\mathbf{E}_Z\text{HSIC}_\omega(Z)-r_\omega]/(\sqrt{n}\sigma_\omega)$. Next, we provide its estimation which can be computed in linear time.

We first consider obtaining the estimator of the numerator part. For the term $\mathbf{E}_Z \text{HSIC}_\omega(Z)$, we can estimate it with $\text{HSIC}_\omega(Z)$ as in Eq. (10). The estimation of the threshold r_ω poses a challenge, primarily stemming from the lack of an explicit expression for the distribution of the infinite sum of chi-square variables. One avenue to address this challenge involves employing the permutation method [2, 38] to simulate the distribution under $\mathcal{H}_0$. However, this method necessitates a significant number of shuffles to accurately approximate the distribution. Furthermore, even with the implementation of parallel schemes, it incurs memory costs proportional to the number of permutations, rendering it impractical for resource-constrained scenarios. Here, we adopt a lightweight approach in practice, leveraging the gamma approximation as proposed by [14]. A gamma distribution is uniquely determined by its first and second-order moments. For these two moments, we present their corresponding linear-time estimators in Theorem 1. As a result, we can obtain the $(1-\alpha)$-quantile of the gamma distribution, denoted as $\widehat{c_\alpha}$, with estimated parameters $\gamma := \mathcal{E}_0^2/\mathcal{V}_0, \beta := \mathcal{V}_0/\mathcal{E}_0$ in linear time. Formally, with the term $\mathcal{E}_0$ and $\mathcal{V}_0$ defined in Theorem 1, $\widehat{c_\alpha}$ is calculated by

$$\mathcal{H}_0 : n\text{HSIC}_\omega(Z) \sim \frac{x^{\gamma-1}e^{-x/\beta}}{\beta^\gamma \Gamma(\gamma)}, \gamma = \frac{\mathcal{E}_0^2}{\mathcal{V}_0}, \ \beta = \frac{\mathcal{V}_0}{\mathcal{E}_0}, \quad \int_0^{\frac{\widehat{c_\alpha}}{\beta}} \frac{x^{\gamma-1}e^{-x}}{\Gamma(\gamma)} dx = 1-\alpha, \tag{14}$$

where $\Gamma(\cdot)$ is the gamma function. By combining the way to estimate the gradients of $\widehat{c_\alpha}$ [30], we enable it for gradient-based optimization with automatic differentiation framework. As a result, we obtain a linear-time differentiable estimator of the numerator part.

Theorem 1 (Linear-Time Estimators)**.** *Under $\mathcal{H}_0$, the estimators of mean and variance with bias of $\mathcal{O}(n^{-1})$ to $\mathbf{E}_Z[n\text{HSIC}_\omega(Z)]$ and $\mathbf{Var}_Z[n\text{HSIC}_\omega(Z)]$, denote as $\mathcal{E}_0$ and $\mathcal{V}_0$, respectively, are given by*

$$\mathcal{E}_0 := \frac{[\mathbf{1}^T \mathbf{\Lambda}_{Xc}^{\cdot 2} \mathbf{1}][\mathbf{1}^T \mathbf{\Lambda}_{Yc}^{\cdot 2} \mathbf{1}]}{(n-1)^2}, \mathcal{V}_0 := \frac{2n(n-4)(n-5)}{(n-1)(n-2)(n-3)} \frac{[\mathbf{1}^T (\mathbf{\Lambda}_{Xc}^T \mathbf{\Lambda}_{Xc})^{\cdot 2} \mathbf{1}][\mathbf{1}^T (\mathbf{\Lambda}_{Yc}^T \mathbf{\Lambda}_{Yc})^{\cdot 2} \mathbf{1}]}{n^4}, \tag{15}$$

where $()^{\cdot 2}$ is the entry-wise matrix power. Both $\mathcal{E}_0$ and $\mathcal{V}_0$ can be calculated in $\mathcal{O}(nD^2)$ time.

For the remain term σ_ω, we estimate it with $\widehat{\sigma}_\omega$ that $\widehat{\sigma}_\omega^2 := 16\big[\frac{1}{n}\sum_i (\frac{1}{n^3}\sum_{j,q,r} h_{ijqr}^{(\omega)})^2 - \text{HSIC}_\omega^2(Z)\big]$. To calculate $\sum_{j,q,r} h_{ijqr}^{(\omega)}$, the straightforward way is to compute each item $h_{ijqr}^{(\omega)}$, which requires total $\mathcal{O}(n^4)$ of computation. Here we provide a way to enable it to be calculated in linear time by obtaining a matrix expression. The main result is given by

$$\sum_{j,q,r} h_{ijqr}^{(\omega)} = \frac{1}{2}\left[n\mathbf{1}^T\mathbf{A}\mathbf{1} + n^2(\mathbf{A}\mathbf{1})_i + (\mathbf{1}^T\mathbf{C})\mathbf{B}_i + (\mathbf{1}^T\mathbf{B})\mathbf{C}_i - n\mathbf{E}_i - n\mathbf{F}_i - n\mathbf{D}_i - \mathbf{1}^T\mathbf{D}\right], \tag{16}$$

where the definition of variables $\mathbf{A}$ to $\mathbf{F}$ with the calculation cost are given in the Fig. 1 and the derivation of Eq. (16) is given in the Appendix. By checking the complexity of the remaining matrix operations in Eq. (16), all the elements with index i can be calculated in $\mathcal{O}(nD^2)$. Combining the results obtained before that $\text{HSIC}_\omega(Z)$ can also be calculated in $\mathcal{O}(nD^2)$, thus calculating the term $\widehat{\sigma}_\omega$ cost $\mathcal{O}(nD^2)$ time. As a result, we obtain the overall linear-time optimization objective $J := [\text{HSIC}_\omega(Z) - \widehat{c_\alpha}/n]/\widehat{\sigma}_\omega$, which is a clear contrast to the existing quadratic-time schemes [30].

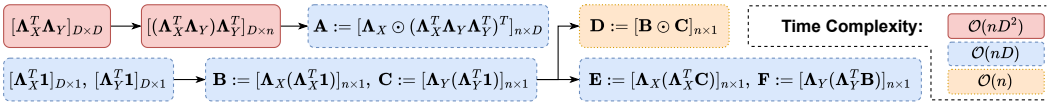

Figure 1: The diagram shows the definition of the quantities in Eq. (16), with styles representing the time complexity of the computational process in the current box. $\odot$: the element-wise product.

4.3 The Overall Learning Framework

After obtaining the differentiable optimization objective J, we can perform the training process end-to-end. In this process, the overfitting issues may happen especially with insufficient samples, which could influence both Type I and II errors. If we use the same sample for testing, the Type I errors may be uncontrollable [22] when the overfitting issues happen. To address this, we adopt the split scheme as in [24, 18] to allow our tests to maintain validity (controllable Type I errors). The split ratio is set to 0.5 to facilitate the balance between the two. Apart from controlling Type I errors, we still want to mitigate overfitting issues as much as possible in order to generalize the optimized

Fourier feature pairs on test data thus improving the power of our tests. To this end, we select smooth function classes to control the model complexity (e.g., as measured by the VC dimension), specifically in this paper, we consider the two classes in Tab. 1 and implement them for experiments later. One is the choice of Gaussian classes that optimize the global scale, and the other we consider Mahalanobis classes and set Σ to be $\text{diag}(\sigma_1, ..., \sigma_d)$ for optimization, which corresponds to optimizing the scale in each dimension (which allows to capture high-frequency signals such as the edge in the image). These smooth choices also bring the advantage of interpretability [30] and it is experimentally proven that this simple choice is already able to handle most of the cases in different settings.

More discussion about split strategy. Currently, there are two major classes of approaches for adaptive independence tests. One involves selecting kernels from a finite/countable set (discrete scenario) and the other involves performing kernel parameter searches in a continuous space (continuous scenario). For the former case, some methods [22, 32] control Type I errors by applying techniques from the selective inference literature without data splitting. However, these methods cannot be directly applied to a continuous scenario due to the uncountable set of kernels involved. To the best of our knowledge, both our scheme and existing methods [24, 18] rely on data splitting for the continuous case. Designing methods to control Type I errors in the continuous case without sample splitting remains a challenging and significant problem for future research.

Algorithm. Our algorithm is outlined in Alg. 1. As a pre-processing step, we split the data into the training data Z^{tr} and the testing data Z^{te} (Line 1). The test contains two phases: 1) We learn the Fourier feature pairs with Adam [21] optimizer using full batches on Z^{tr} (Lines 2-7). 2) With the learned Fourier feature pairs, we calculate the test statistic and threshold (Lines 8-10) to determine the independence (Lines 11) on Z^{te}. The overall time complexity is $\mathcal{O}\big(TnD(d_x + d_y + D)\big)$ and the space cost is $\mathcal{O}\big(n(d_x + d_y + D)\big)$ for storing the data as well as the Fourier feature pairs.

Algorithm 1 The learning and testing framework

Input: samples Z of X, Y, significance level α, the number of Fourier feature D.
Output: $X \perp\!\!\!\perp Y$ or $X \not\perp\!\!\!\perp Y$.

1: Split the data as $Z = Z^{tr} \cup Z^{te}$. Sampling $\{\omega_j\}_{j=1}^{D/2} = \{(\omega_{k;j}, \omega_{l;j})\}_{j=1}^{D/2}$ with $p(\omega)$.
2: $\triangleleft$ **Learning significant Fourier feature pairs on Z^{tr}.**
3: Initialize parameters θ_k, θ_l, set learning rate ϵ, and set iteration steps T.
4: **for** $t = 1, 2, ..., T$ **do**
5: Obtain learnable Fourier feature pairs $\mathbf{\Lambda}_X, \mathbf{\Lambda}_Y$ with parameters θ_k, θ_l and $\{\omega_j\}_{j=1}^{D/2}$.
6: Calculate criterion J with $\mathbf{\Lambda}_X, \mathbf{\Lambda}_Y$ then optimize J with $(\theta_k, \theta_l) \leftarrow (\theta_k, \theta_l) + \epsilon \nabla_{(\theta_k, \theta_l)} J$.
7: **end for**
8: After training, obtain optimized parameters θ_k^*, θ_l^*.
9: $\triangleleft$ **Testing with learned Fourier feature pairs on Z^{te}.**
10: Calculate the statistic $n^{te}\text{HSIC}_\omega(Z^{te})$, threshold $\widehat{c_\alpha}(Z^{te})$ with parameters θ_k^*, θ_l^* and $\{\omega_j\}_{j=1}^{D/2}$.
11: Return $X \not\perp\!\!\!\perp Y$ if $\widehat{c_\alpha}(Z^{te}) \leq n^{te}\text{HSIC}_\omega(Z^{te})$ holds, otherwise $X \perp\!\!\!\perp Y$.

5 Theoretical Results

We first give the uniform bound results over a ball in parameter space which guarantees the convergence of our optimizing objective thus ensuring its effectiveness in modeling test power.

Theorem 2 (Uniform Bound)**.** *Let θ_k, θ_l parameterize $\mathcal{T}_{\theta_k}, \mathcal{T}_{\theta_l}$ in Banach spaces of dimension d_k, d_l. And $\mathcal{T}_{\theta_k}, \mathcal{T}_{\theta_l}$ are Lipschitz to the parameters θ_k, θ_l with the non-negative constant L_k, L_l, respectively. Let Θ_c be a set of (θ_k, θ_l) for which $\sigma_\omega \geq c > 0$ with a positive constant c and $\|\theta_k\| \leq R_{\theta_k}, \|\theta_l\| \leq R_{\theta_l}$. Let r denote the threshold, i.e., $(1-\alpha)$-quantile for the distribution in Eq. (11) and $r^{(n)}$ be the threshold with sample size n. Let $\{(\omega_{k;j}, \omega_{l;j})\}_{j=1}^{D/2}$ be the samplings of frequency with the sampling number D. Also, we define $R_{\omega_k} := \sup_j \|\omega_{k;j}\|, R_{\omega_l} := \sup_j \|\omega_{l;j}\|, d_s := \max\{d_k, d_l\}$ and $\xi_\omega := \text{HSIC}_\omega(Z)$. Then with probability at least $1 - \delta$, we have*

$$\sup_{(\theta_k, \theta_l) \in \Theta_c} \left| \frac{\xi_\omega - r_\omega^{(n)}/n}{\widehat{\sigma}_\omega} - \frac{\mathbf{E}_Z \xi_\omega - r_\omega/n}{\sigma_\omega} \right| \sim \mathcal{O}\left(\left[\sqrt{\frac{1}{n} \log \frac{1}{\delta} + d_s \frac{\log n}{n}} + \frac{R_{\omega_k} L_k + R_{\omega_l} L_l}{\sqrt{n}} \right] \right).$$

Next, we show the consistency of the tests, i.e. the power of the test tends to 1 as the sample size increases. Let the U-statistic of $\text{HSIC}_\omega(Z^{te})$ be $\text{HSIC}_\omega^{(u)}(Z^{te})$, then we have the following results.

Theorem 3 (Consistency). *Let θ_k^*, θ_l^* be the parameters after learning, Z^{te} be the test samples of size m, when $\mathbf{E}_Z HSIC_\omega^{(u)}(Z^{te}) > 0$, then the probability of the Type II error*

$$\mathbb{P}\big(\textit{Type II error}\big) = \mathbb{P}_{\mathcal{H}_1}\left(mHSIC_\omega(Z^{te}) \leq r_\omega^{(m)} | \theta_k^*, \theta_l^*\right) \sim \mathcal{O}(m^{-1/2}). \tag{17}$$

Let the mapping functions with learned parameters θ_k^*, θ_l^* be $\mathcal{T}_{\theta_k^*}, \mathcal{T}_{\theta_l^*}$, and the corresponding range space be compact subsets of $\mathbb{R}^{d_{\mathcal{T}_x}}, \mathbb{R}^{d_{\mathcal{T}_y}}$, respectively. Also, the diameters of two range spaces are denoted by $diam(\mathcal{T}_{\theta_k^*}), diam(\mathcal{T}_{\theta_l^*})$, respectively. Let $\{(\omega_{k;j}, \omega_{l;j})\}_{j=1}^{D/2}$ be the frequency samplings with their second moment denoted by $\sigma_{\omega_k}^2 := \mathbf{E}_{p_k(\omega)}[\omega_{k;j}^T \omega_{k;j}]$, $\sigma_{\omega_l}^2 := \mathbf{E}_{p_l(\omega)}[\omega_{l;j}^T \omega_{l;j}]$. Additionally, we denote $\xi_u := HSIC(X, Y)$, then under $\mathcal{H}_1$, we have $\mathbf{E}_Z HSIC_\omega^{(u)}(Z^{te}) > 0$ with any constant probability when $D = \Omega\Big(\frac{d_{\mathcal{T}_x} + d_{\mathcal{T}_y}}{\xi_u^2} \log \frac{\sigma_{\omega_k} diam(\mathcal{T}_{\theta_k^*}) + \sigma_{\omega_l} diam(\mathcal{T}_{\theta_l^*})}{\xi_u}\Big)$.

This result can be understood in two parts. The first one is about consistency, i.e., the Type II error rate tends to 0 at the rate of $m^{-1/2}$ when condition $\mathbf{E}_Z \text{HSIC}_\omega^{(u)}(Z^{te}) > 0$ holds. The second part provides the condition when $\mathbf{E}_Z \text{HSIC}_\omega^{(u)}(Z^{te}) > 0$ holds, which requires sufficiently many frequency samplings. The theorem shows that the large value of the criterion $\text{HSIC}(X, Y)$ helps to reduce the required D. According to the results discussed in Sec. 3, there is an improvement in the criterion by finding the more significant features and thus helps to reduce the required D. To summarize, the significant features further help to guarantee the consistency of the test under the efficient requirements (smaller D). All proofs as well as additional results are given in the Appendix.

6 Performance Evaluation

We compare the following tests: distance-based statistic **dCor** [39], the original HSIC **QHSIC** [14], the copula-based method **RDC** [26], the three variants of HSIC **NyHSIC** [44], **FHSIC** [44], **BHSIC** [44] and **HSICAgg** [32], **NFSIC** [18] as introduced in Sec. 1. Among them, dCor and QHSIC are $\mathcal{O}(n^2)$ tests. RDC is calculated in $O(n \log n)$ time and the rest are $\mathcal{O}(n)$ tests. A detailed description of the comparing methods is given in the Appendix. For our methods, We provide two variants as mentioned in Sec. 4.3. We name the Gaussian class case as **LFHSIC-G**, and name the Mahalanobis class case (and set Σ as a diagonal matrix) **LFHSIC-M**. Additionally, for the comparative methods [30] that are relevant to us, due to their high time overhead and therefore inability to handle some settings of evaluation, we separately provide a comparison with our method under certain feasible experimental settings, the results are given in the Appendix.

Experimental setup. The significance level α is set to 0.05. We use Gaussian kernels for both X and Y in all kernel-based methods. And QHSIC, RDC, NyHSIC, FHSIC, BHSIC are all with the kernel width being set to the Euclidean distance median of the samples. The number of random features D for FHSIC, LFHSIC-G/M, the number of induced variables for NyHSIC, the block size for BHSIC as well as the number of sub-diagonals R for HSICAgg are all kept consistent as recommended in [44, 32] for fair evaluation. Parameter settings for the rest of the methods follow the defaults in the code. More details of the setups are given in the Appendix.

Evaluation protocol. We evaluate on four synthetic datasets [18, 30] and two real datasets [44, 30]. Synthetic datasets consist of Sine Dependency (SD), Sinusoid (Sin), Gaussian Sign (GSign), and independent subspace analysis (ISA) dataset [14]. On real data, we introduce high-dimensional image data and another music dataset to evaluate the capability of all methods in different data scenarios. Unless otherwise specified, we perform 100 repeated randomized experiments and report the average result of test power as default. More details of the generating process of each dataset and the details of the evaluation (including running time) are provided in the Appendix.

6.1 Results on Synthetic Datasets

Settings of SD, Sin, and GSign Dataset. The Sin data corresponds to the example in Sec. 3 that requires the method to focus on differences in specific frequencies. In SD, Y is dependent solely on the first two dimensions of X. In contrast, in GSign, Y is independent of any proper subset of X

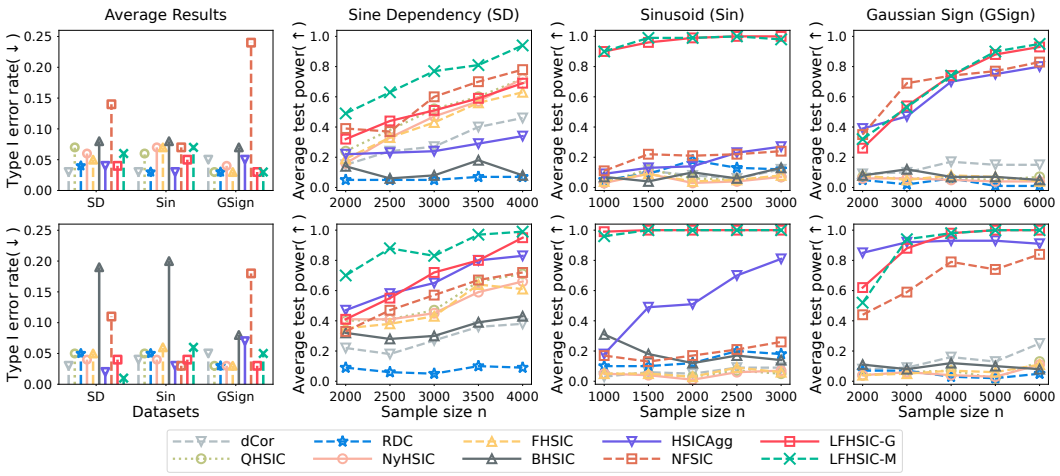

Figure 2: Top: ($D = 100$). Below: ($D = 500$). Left: The average Type I error rate on SD, Sin, and GSign datasets. The other three plots: The results of average test power on these three datasets.

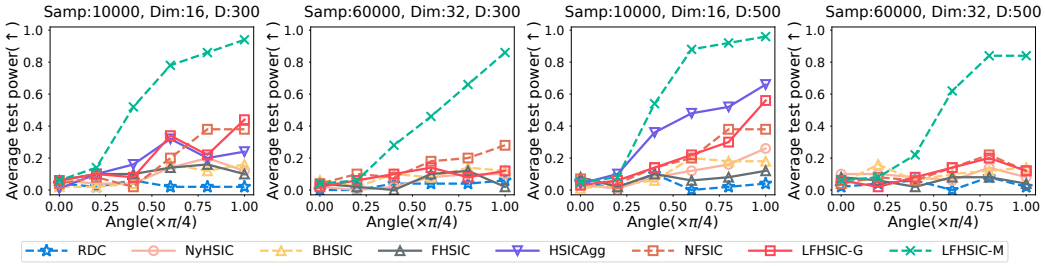

Figure 3: The average test power v.s. the rotation angle of each method on the ISA dataset.

but dependent on X as a whole. Therefore, it requires the method to learn important local/global features based on the characteristics of the data to improve the test power. For SD and GSign, we set the dimension of X as 4 and 5, respectively, and the dimension of Y is 1 for both. For Sin, we set the frequency parameter $\omega = 5$. For calculating the Type I error rate, we evaluate using samples ($n = 2000$) obtained by permutation for all three datasets.

Performance. The results for $D = 100$ and $D = 500$ are shown in Fig. 2. Except for NFSIC and BHSIC, all the other methods succeed in controlling the Type I error rate ≤ 0.05. LFHSIC-M/G, NFHSIC, and HSICAgg perform much better than other methods due to their ability to obtain more appropriate kernels/features for testing. LFHSIC-G/M performs on both settings of D and has a more significant advantage over the others when D is small, implying the optimization objective can still be successfully optimized and the criterion is still powerful under high-speed requirements. In addition, as the sample size increases the test power of LFHSIC-G/M is gradually converging to 1 in both settings, which corroborates the results of Theorem 3.

Settings of ISA Dataset (Large Scale). We set dimension (of both X, Y) and sample size as $d = 16, n = 10000$ and $d = 32, n = 60000$, then evaluate the average test power with angle parameter $\theta \in [0, \pi/4]$. Note that a larger angle signifies stronger dependency. The quadratic-time methods are not involved in the evaluation due to their inability to handle large-scale settings. For HSICAgg under the challenging setting $n = 60000, d = 32$, the memory space required for parallel implementation leads to memory overflow and hence the results are not given.

Performance. The results for $D = 300$ and $D = 500$ are shown in Fig. 3. The results obtained at $\theta = 0$ reflect the Type I error rate. All methods successfully control the Type I error rate ≤ 0.05. LFHSIC-M stably outperforms other methods significantly as the angle increases. Method (LFHSIC-G) that simply optimizes the global bandwidth performs worse as d increases, corroborating the need for more flexible kernel designs for more challenging tasks. Furthermore, comparing the

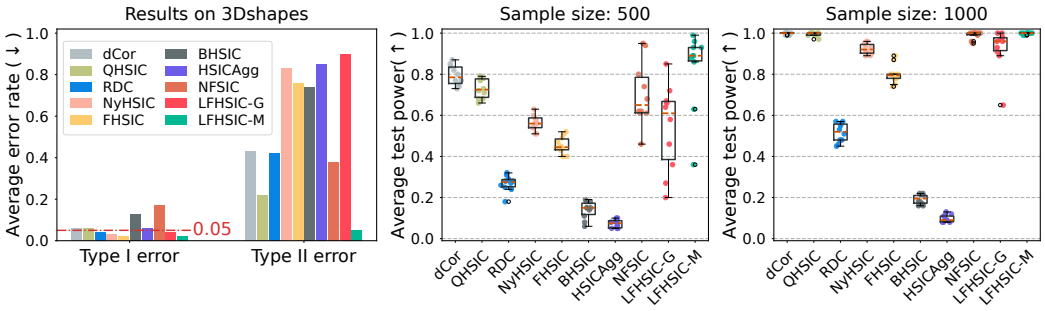

Figure 4: The results on two real data. Left: 3DShapes. Right two: MSD Dataset.

results under different settings of D, our method LFHSIC-M performs consistently well and exhibits progressively better performance as D increases.

6.2 Results on Real Data

Settings of Two Real Data. The first real dataset used is a high-dimensional image dataset 3Dshapes as in [30]. In our experiments, we vectorize image X to a vector with dimension $64\times64\times3 = 12,288$. The sample size is set as 128. We add standard Gaussian noise $\mathcal{N}(0,1)$ to the angle Y to make the setting more challenging. The Type I error rate is evaluated by the samples obtained by permutation. Besides, we consider the Million Song Data (MSD) as the second real dataset. The first dimension represents the year of release of each song and is referred to as variable Y. The remaining 90-dimensional features (e.g., mean timbre and timbre covariance) constitute variable X. We follow the recommended setting [44], i.e., disturbing each entry of the X with an independent Gaussian noise $\mathcal{N}(0,1000)$. For this dataset MSD, in order to fully utilize the data, we randomly select $n \in \{500, 1000\}$ samples as the training set and other n samples from the remaining data 100 times for the evaluation and obtain the average result. The above training and testing processes are repeated 10 times to evaluate the robustness of the optimization scheme.

Performance. The results of two real data with $D = 10$ are presented in Fig. 4. For the results on 3Dshapes (shown in the left of Fig. 4), all methods except BHSIC and NFSIC control the Type I error well. The linear-time test has relatively lower power compared to the quadratic-time test except for LFHSIC-M, proving that its more significant features obtained in high-dimensional scenarios enable it to achieve outstanding performance even in scenarios with high approximation requirements ($D = 10$). Similar conclusions can drawn from the MSD dataset (shown in the right of Fig. 4). Additionally, the results for NFSIC and LFHSIC-G/M with different sample sizes indicate increased robustness of the optimization as the sample size increases (reflected in the reduction of variance), and the more flexible design also contributes to this (comparing LFHSIC-M and LFHSIC-G), thus can be more effectively applied to real-world scenarios.

7 Conclusion

In this paper, we propose a novel method to efficiently learn significant Fourier feature pairs for maximizing the power of HSIC-based independence tests. By integrating a learnable Fourier feature module, we improve the flexibility of existing configurations and design a new criterion. The proposed linear-time optimization objective accurately models the power of the test and can be trained end-to-end in a data-driven manner, ensuring both effectiveness and efficiency. Both theoretical results and experimental results show the effectiveness of our proposed method. Future work includes further improving the sampling method in the frequency domain.

Acknowledgments and Disclosure of Funding

This work was supported by National Natural Science Foundation (NSFC) (62372116 and 62472415), and National Key Research and Development Program of China (2021YFC3340302 and 2021YFC3300304).

References

[1] 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.

[2] Arcones, M. A. and Gine, E. (1992). On the bootstrap of u and v statistics. *The Annals of Statistics*, pages 655–674.

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

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

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

[6] 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.

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

[8] Chwialkowski, K. P., Ramdas, A., Sejdinovic, D., and Gretton, A. (2015). Fast two-sample testing with analytic representations of probability measures. *Advances in Neural Information Processing Systems*, 28.

[9] 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.

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

[11] Gönen, M. and Alpaydın, E. (2011). Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] Kalinke, F. and Szabo, Z. (2024). The minimax rate of hsic estimation for translation-invariant kernels. *arXiv preprint arXiv:2403.07735*.

[20] Kim, I. and Schrab, A. (2023). Differentially private permutation tests: Applications to kernel methods. *arXiv preprint arXiv:2310.19043*.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

[27] Podkopaev, A. and Ramdas, A. (2024). Sequential predictive two-sample and independence testing. *Advances in Neural Information Processing Systems*, 36.

[28] Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20.

[29] Ramdas, A., Reddi, S. J., Póczos, B., Singh, A., and Wasserman, L. (2015). On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

[30] Ren, Y., Xia, Y., Zhang, H., Guan, J., and Zhou, S. (2024). Learning adaptive kernels for statistical independence tests. In *International Conference on Artificial Intelligence and Statistics*, pages 2494–2502. PMLR.

[31] 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(5), pages 6499–6506.

[32] 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.

[33] 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.

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

[35] Shekhar, S., Kim, I., and Ramdas, A. (2023). A permutation-free kernel independence test. *Journal of Machine Learning Research*, 24(369):1–68.

[36] Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561.

[37] Sutherland, D. J. and Schneider, J. (2015). On the error of random fourier features. *arXiv preprint arXiv:1506.02785*.

[38] 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*.

[39] Székely, G., Rizzo, M., and Bakirov, N. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6):2769–2794.

[40] 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.

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

[42] Wang, Z., Zhan, Z., Gong, Y., Shao, Y., Ioannidis, S., Wang, Y., and Dy, J. (2023). Dualhsic: Hsic-bottleneck and alignment for continual learning. In *International Conference on Machine Learning*, pages 36578–36592. PMLR.

[43] 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*.

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

[45] Zhang, T., Zhang, Y., and Zhou, T. (2024). Statistical insights into hsic in high dimensions. *Advances in Neural Information Processing Systems*, 36.

Appendix Organization

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 $\mathbb{R}$
$\mathbb{P}_X$	marginal distribution of X
$\mathbb{P}_{XY}$	joint distribution of X, Y
F_X	distribution function of X
$\mathbf{E}[X]$	expectation of X
$\text{Var}(X)$	variance of X
$X \perp\!\!\!\perp Y$	random variables X, Y are independent
$X \not\perp\!\!\!\perp Y$	random variables X, Y are not independent
$\mathbf{i}_r^n$	the set of all r-tuples drawn without replacement from the set $\{1, ..., n\}$
$\binom{n}{k}$	number of k-combinations of n elements
$(n)_k$	number of permutations, define as $\frac{n!}{(n-k)!}$
$\text{Tr}(\cdot)$	the trace of a square matrix
$\mathbf{1}$	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
$\xrightarrow{\text{d}}$	convergence in distribution
$\otimes$	the product symbol of measure
$\times$	the product symbol of topological space
$\mathcal{N}(\Theta, r)$	covering number with radii r for space Θ
$\mathcal{F}, \mathcal{F}^{-1}$	the Fourier transform, Fourier inverse transform

B Assumptions

The following are the assumptions required. We denote the parameter spaces of θ_k, θ_l as Θ_k, Θ_l.

(a) The mapping functions $\mathcal{T}_{\theta_k}, \mathcal{T}_{\theta_l}$ are Lipschitz to the parameters θ_k, θ_l, i.e. for all $x \in \mathcal{X}, y \in \mathcal{Y}$ and for all $\theta_k, \theta_k' \in \Theta_k, \theta_l, \theta_l' \in \Theta_l$,

$$\left\|\mathcal{T}_{\theta_k}(x) - \mathcal{T}_{\theta_k'}(x)\right\| \leq L_k \cdot \|\theta_k - \theta_k'\|, \ \left\|\mathcal{T}_{\theta_l}(y) - \mathcal{T}_{\theta_l'}(y)\right\| \leq L_l \cdot \|\theta_l - \theta_l'\| \tag{18}$$

with the nonnegative Lipschitz constant L_k, L_l.

(b) The range of the mapping functions $\mathcal{T}_{\theta_k}, \mathcal{T}_{\theta_l}$ are bounded.

(c) The parameters θ_0, θ_1 lie in Banach spaces of dimension d_k, d_l respectively. Also, the parameters θ_k, θ_l are bounded by $R_{\theta_k}, R_{\theta_l}$ respectively, i.e., $\|\theta_k\| \leq R_{\theta_k}, \|\theta_l\| \leq R_{\theta_l}$.

C Some Auxiliary Lemma

C.1 A Useful Expression

In this part, we give a useful expression for $h^{(\omega)}_{ijqr}$ for the subsequent proof. By the definition, we have $h^{(\omega)}_{ijqr} = \frac{1}{4!}\sum^{(i,j,q,r)}_{(t,u,v,w)} k^{(\omega)}_{tu} l^{(\omega)}_{tu} + k^{(\omega)}_{tu} l^{(\omega)}_{vw} - 2k^{(\omega)}_{tu} l^{(\omega)}_{tv}$. We simplify it by setting $t = i$, $u = i$, $v = i$ and $w = i$ in turn. Then we can show that $h^{(\omega)}_{ijqr}$ is equal to

$$\begin{aligned}
&\frac{1}{4!}\sum^{(j,q,r)}_{(u,v,w)} (k^{(\omega)}_{iu} l^{(\omega)}_{iu} + k^{(\omega)}_{iu} l^{(\omega)}_{vw} - 2k^{(\omega)}_{iu} l^{(\omega)}_{iv}) + \frac{1}{4!}\sum^{(j,q,r)}_{(t,v,w)} (k^{(\omega)}_{ti} l^{(\omega)}_{ti} + k^{(\omega)}_{ti} l^{(\omega)}_{vw} - 2k^{(\omega)}_{ti} l^{(\omega)}_{tv}) \\
+&\frac{1}{4!}\sum^{(j,q,r)}_{(t,u,w)} (k^{(\omega)}_{tu} l^{(\omega)}_{tu} + k^{(\omega)}_{tu} l^{(\omega)}_{iw} - 2k^{(\omega)}_{tu} l^{(\omega)}_{ti}) + \frac{1}{4!}\sum^{(j,q,r)}_{(t,u,v)} (k^{(\omega)}_{tu} l^{(\omega)}_{tu} + k^{(\omega)}_{tu} l^{(\omega)}_{vi} - 2k^{(\omega)}_{tu} l^{(\omega)}_{tv}).
\end{aligned} \tag{19}$$

By the definition, $k^{(\omega)}_{tu} := \psi^{(\omega)}_k(\mathcal{T}_{\theta_k} x_t - \mathcal{T}_{\theta_k} x_u) = \Lambda_X(x_t)\Lambda_X(x_u)^T$ is symmetric, i.e. $k^{(\omega)}_{tu} = k^{(\omega)}_{ut}$. And so as $l^{(\omega)}_{tu}$. Hence we can merge the identical items (marked with the same color). As a result,

$$\begin{aligned}
h^{(\omega)}_{ijqr} =&\frac{1}{4!}\sum^{(j,q,r)}_{(u,v,w)} (2k^{(\omega)}_{iu} l^{(\omega)}_{iu} + 2k^{(\omega)}_{iu} l^{(\omega)}_{vw} - 2k^{(\omega)}_{iu} l^{(\omega)}_{iv}) - \frac{1}{4!}\sum^{(j,q,r)}_{(t,v,w)} (2k^{(\omega)}_{ti} l^{(\omega)}_{tv}) \\
&+ \frac{1}{4!}\sum^{(j,q,r)}_{(t,u,w)} (2k^{(\omega)}_{tu} l^{(\omega)}_{tu} + 2k^{(\omega)}_{tu} l^{(\omega)}_{iw} - 2k^{(\omega)}_{tu} l^{(\omega)}_{ti}) - \frac{1}{4!}\sum^{(j,q,r)}_{(t,u,v)} (2k^{(\omega)}_{tu} l^{(\omega)}_{tv}).
\end{aligned} \tag{20}$$

We will use Eq. (20) many times in subsequent proofs.

C.2 Properties of Learnable Random Fourier Feature

Under the assumption (a), RFFs $\psi^{(\omega)}_k, \psi^{(\omega)}_l$ are Lipschitz to the parameters θ_k, θ_l. Formally,

Lemma 1. *(Lipschitz Property of Fourier Feature). Let $\mathcal{T}_{\theta_k}, \mathcal{T}_{\theta_l}$ be the mapping functions of X, Y that are Lipschitz to the parameters θ_k, θ_l with the non-negative constant L_k, L_l, respectively. Let $\{(\omega_{k;j}, \omega_{l;j})\}^{D/2}_{j=1}$ be the samplings of frequency with the sampling number D. Also, we define $R_{\omega_k} := \sup_j \|\omega_{k;j}\|, R_{\omega_l} := \sup_j \|\omega_{l;j}\|$, then for the RFFs $\psi^{(\omega)}_k, \psi^{(\omega)}_l$ with mapping functions $\mathcal{T}_{\theta_k}, \mathcal{T}_{\theta_l}$ and frequency samplings $\{(\omega_{k;j}, \omega_{l;j})\}^{D/2}_{j=1}$, for all $(x, x') \in \mathcal{X} \times \mathcal{X}, (y, y') \in \mathcal{Y} \times \mathcal{Y}$ and for all $\theta_k, \theta'_k \in \Theta_k, \theta_l, \theta'_l \in \Theta_l$, we have*

$$\begin{aligned}
\|\psi^{(\omega)}_k(\Delta_{x,x'}) - \psi^{(\omega)}_k(\Delta'_{x,x'})\| \le 2R_{\omega_k} L_k \cdot \|\theta_k - \theta'_k\|, \\
\|\psi^{(\omega)}_l(\Delta_{y,y'}) - \psi^{(\omega)}_l(\Delta'_{y,y'})\| \le 2R_{\omega_l} L_l \cdot \|\theta_l - \theta'_l\|,
\end{aligned} \tag{21}$$

where $\Delta_{x,x'} := \mathcal{T}_{\theta_k} x - \mathcal{T}_{\theta_k} x'$, $\Delta'_{x,x'} := \mathcal{T}_{\theta'_k} x - \mathcal{T}_{\theta'_k} x'$ and $\Delta_{y,y'}, \Delta'_{y,y'}$ are defined by analogy.

Proof. We prove the result for $\psi^{(\omega)}_k$ only since the proof for $\psi^{(\omega)}_l$ can be obtained in the same way. We start by recall the definition $\psi^{(\omega)}_k(\Delta_{x,x'}) := \frac{2}{D}\sum_{j=1}^{D/2} \cos(\omega^T_{k;j}\Delta_{x,x'})$. Then

$$\begin{aligned}
\|\psi^{(\omega)}_k(\Delta_{x,x'}) - \psi^{(\omega)}_k(\Delta'_{x,x'})\| &= \left\| \frac{2}{D}\sum_{j=1}^{D/2} \cos(\omega^T_{k;j}\Delta_{x,x'}) - \frac{2}{D}\sum_{j=1}^{D/2} \cos(\omega^T_{k;j}\Delta'_{x,x'}) \right\| \\
&\le \frac{2}{D}\sum_{j=1}^{D/2} \|\cos(\omega^T_{k;j}\Delta_{x,x'}) - \cos(\omega^T_{k;j}\Delta'_{x,x'})\|
\end{aligned} \tag{22}$$

Since the cosine function is bounded by 1, by the mean value theorem, for fixed j, we have

$$\|\cos(\omega^T_{k;j}\Delta_{x,x'}) - \cos(\omega^T_{k;j}\Delta'_{x,x'})\| \le |\omega^T_{k;j}\Delta_{x,x'} - \omega^T_{k;j}\Delta'_{x,x'}|. \tag{23}$$

Then according to the Cauchy–Schwarz inequality,

$$|\omega_{k;j}^T \Delta_{x,x'} - \omega_{k;j}^T \Delta'_{x,x'}| \leq \|\omega_{k;j}\| \cdot \|\Delta_{x,x'} - \Delta'_{x,x'}\|. \tag{24}$$

By the definition of $\Delta_{x,x'}$ and the Lipschitz property of the mapping functions $\mathcal{T}_{\theta_k}, \mathcal{T}_{\theta_l}$, we have

$$\begin{aligned} \|\Delta_{x,x'} - \Delta'_{x,x'}\| &= \|(\mathcal{T}_{\theta_k} x - \mathcal{T}_{\theta_k} x') - (\mathcal{T}_{\theta'_k} x - \mathcal{T}_{\theta'_k} x')\| \\ &\leq \|\mathcal{T}_{\theta_k} x - \mathcal{T}_{\theta'_k} x\| + \|\mathcal{T}_{\theta_k} x' - \mathcal{T}_{\theta'_k} x'\| \leq 2L_k \cdot \|\theta_k - \theta'_k\|. \end{aligned} \tag{25}$$

Combining the above results, we complete the proof. □

Under the assumption (b), we can obtain the uniform convergence property as follows.

Lemma 2. *(Uniform Convergence of Fourier Features). Let the mapping function of X, Y with parameters θ_k, θ_l be $\mathcal{T}_{\theta_k}, \mathcal{T}_{\theta_l}$, and the corresponding range space be a compact subset of $\mathbb{R}^{d_{\mathcal{T}_x}}, \mathbb{R}^{d_{\mathcal{T}_y}}$, respectively. Also, the diameter of two range spaces is denoted by diam$(\mathcal{T}_{\theta_k})$, diam$(\mathcal{T}_{\theta_l})$, respectively. Let $\{(\omega_{k;j}, \omega_{l;j})\}_{j=1}^{D/2}$ be the samplings of frequency with the sampling number D, then for the RFFs with mapping functions $\mathcal{T}_{\theta_k}, \mathcal{T}_{\theta_l}$ and frequency samplings $\{(\omega_{k;j}, \omega_{l;j})\}_{j=1}^{D/2}$, we have*

$$\begin{aligned} \mathbb{P}\left[\sup_{x,x' \in \mathcal{X}} |\Lambda_k(x)^T \Lambda_k(x') - k(x,x')| \geq \epsilon\right] &\leq 2^8 \left(\frac{\sigma_{\omega_k} diam(\mathcal{T}_{\theta_k})}{\epsilon}\right)^2 \exp\left(-\frac{D\epsilon^2}{4(d_{\mathcal{T}_x}+2)}\right), \\ \mathbb{P}\left[\sup_{y,y' \in \mathcal{Y}} |\Lambda_l(y)^T \Lambda_l(y') - l(y,y')| \geq \epsilon\right] &\leq 2^8 \left(\frac{\sigma_{\omega_l} diam(\mathcal{T}_{\theta_l})}{\epsilon}\right)^2 \exp\left(-\frac{D\epsilon^2}{4(d_{\mathcal{T}_y}+2)}\right), \end{aligned} \tag{26}$$

where the second moment of frequency samplings $\sigma_{\omega_k}^2 := \mathbf{E}_{p_k(\omega)}[\omega_{k;j}^T \omega_{k;j}]$, $\sigma_{\omega_l}^2 := \mathbf{E}_{p_l(\omega)}[\omega_{l;j}^T \omega_{l;j}]$.

Proof. Based on the derivation of RFFs in Sec. 4.1, We can view it as if the frequency sampling process is performed after the range space is obtained. Since the convergence bounds of the sampling process can be obtained directly through the results of [28, Claim 1], by replacing the input space in [28, Claim 1] to the range space here, then this part of the proof can be completed. □

Remark. Combining the technique in [37], the constants in bounds can be further improved.

C.3 Approximation Error Bound

Let $\text{HSIC}_\omega^{(u)}(Z)$, also denoted as $\xi_\omega^{(u)}$, be the U-statistic that corresponding to $\text{HSIC}_\omega(Z)$, i.e., $\text{HSIC}_\omega^{(u)}(Z) := \frac{1}{(n)_4} \sum_{(i,j,q,r) \in \mathbf{i}_4^n} h_{ijqr}^{(\omega)}$. The population value of $\text{HSIC}_\omega^{(u)}(Z)$ is given by $\mathbf{E}_Z \xi_\omega^{(u)}$ which can be viewed as the result obtained after a frequency sampling approximation on $\text{HSIC}(X, Y)$ is performed. The bound of approximation error is given by the following Lemma.

Lemma 3. *(Approximation Error Bound). For simplify, we denote $\Lambda_k(x)^T \Lambda_k(x'), \Lambda_l(y)^T \Lambda_l(y')$ as $k^{(\omega)}(x,x'), l^{(\omega)}(y,y')$, respectively. Then we have*

$$|\mathbf{E}_Z \xi_\omega^{(u)} - \mathit{HSIC}(X,Y)| \leq 4 \cdot \sup_{x,x' \in \mathcal{X}, y,y' \in \mathcal{Y}} |k^{(\omega)}(x,x') l^{(\omega)}(y,y') - k(x,x') l(y,y')|. \tag{27}$$

Proof. We first represent $\mathbf{E}_Z \xi_\omega^{(u)}$ in the form corresponding to Eq. (1), i.e.,

$$\begin{aligned} \mathbf{E}_Z \xi_\omega^{(u)} = \mathbf{E}_{XX'YY'}\left[k^{(\omega)}(X,X') l^{(\omega)}(Y,Y')\right] &+ \mathbf{E}_{XX'}\left[k^{(\omega)}(X,X')\right] \mathbf{E}_{YY'}\left[l^{(\omega)}(Y,Y')\right] \\ &- 2\mathbf{E}_{X'Y'}\left[\mathbf{E}_X k^{(\omega)}(X,X') \mathbf{E}_Y l^{(\omega)}(Y,Y')\right]. \end{aligned} \tag{28}$$

Taking one of the items as an example and comparing it to the corresponding item in Eq. (1),

$$\begin{aligned} &\left|\mathbf{E}_{X'Y'}\left[\mathbf{E}_X k^{(\omega)}(X,X') \mathbf{E}_Y l^{(\omega)}(Y,Y')\right] - \mathbf{E}_{X'Y'}\left[\mathbf{E}_X k(X,X') \mathbf{E}_Y l(Y,Y')\right]\right| \\ \leq &\int_{X',Y'} \int_X \int_Y |k^{(\omega)}(X,X') l^{(\omega)}(Y,Y') - k(X,X') l(Y,Y')| d\mathbb{P}_X d\mathbb{P}_Y d\mathbb{P}_{X'Y'} \\ \leq &\sup_{x,x' \in \mathcal{X}, y,y' \in \mathcal{Y}} |k^{(\omega)}(x,x') l^{(\omega)}(y,y') - k(x,x') l(y,y')|. \end{aligned} \tag{29}$$

The results can be obtained for the other terms in a similar way, which completes the proof. □

D Proof of Proposition 1

In this section, we give a proof of the Proposition 1. We first restate the Proposition 1 here.

Proposition 1 (Asymptotics). *Let* $h^{(\omega)}_{ijqr} := \frac{1}{4!}\sum^{(i,j,q,r)}_{(t,u,v,w)} k^{(\omega)}_{tu}l^{(\omega)}_{tu} + k^{(\omega)}_{tu}l^{(\omega)}_{vw} - 2k^{(\omega)}_{uv}l^{(\omega)}_{tv}$*, where the sum represents all ordered quadruples* (t,u,v,w) *drawn without replacement from* (i,j,q,r)*. Then, Under the null hypothesis* $\mathcal{H}_0$*,* $\text{HSIC}_\omega(Z)$ *coverages in distribution to*

$$n\text{HSIC}_\omega(Z) \xrightarrow{d} \sum_{l=1}^{\infty}\lambda_l\chi^2_{1l}, \;\; \lambda_l g_l(z_j) = \int_{z_i,z_q,z_r} h^{(\omega)}_{ijqr}g_l(z_i)dF_{z_i,z_q,z_r}, \tag{30}$$

where $\chi^2_{11},\chi^2_{12},...$ *are independent* χ^2_1 *variates and* λ_l *is the solution to the eigenvalue problem as in the right of Eq. (30). Also, under the alternative* $\mathcal{H}_1$*,* $\text{HSIC}_\omega(Z)$ *converges in distribution as*

$$n^{\frac{1}{2}}\Big(\text{HSIC}_\omega(Z) - \mathbf{E}_Z\text{HSIC}_\omega(Z)\Big) \xrightarrow{d} \mathcal{N}(0,\sigma^2_\omega),\;\; \sigma^2_\omega = 16\Big[\mathbf{E}_i(\mathbf{E}_{j,q,r}h^{(\omega)}_{ijqr})^2 - \big(\mathbf{E}_Z h^{(\omega)}_{ijqr}\big)^2\Big] \tag{31}$$

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

Proof. The proof is mainly based on [34, Chapter 5]. A proof for a similar result has been given in [14, Theorem 1,2]. The difference is that we consider the asymptotic distributions of $\text{HSIC}_\omega(Z)$ that used learnable RFF thus is a function of frequency samplings while they consider $\text{HSIC}_b(Z)$. Thus some steps need to be modified.

Step 1: we show that $\text{HSIC}_\omega(Z)$ is a V-statistic, this can be done since it can be expressed as $\text{HSIC}_\omega(Z) = \frac{1}{n^4}\sum_{i,j,q,r} h^{(\omega)}_{ijqr}$. To facilitate the conclusions in [34] for the U-statistic, we define the U-statistic $\text{HSIC}^{(u)}_\omega(Z) := \frac{1}{(n)_4}\sum_{(i,j,q,r)\in \mathbf{i}^n_4} h^{(\omega)}_{ijqr}$ that corresponds to $\text{HSIC}_\omega(Z)$.

Step 2: Then we prove the result under $\mathcal{H}_0$. To begin with, we first show that $\mathbf{E}_{j,q,r}h^{(\omega)}_{ijqr} = 0$ with $(i,j,q,r)\in\mathbf{i}^n_4$ under $\mathcal{H}_0$. According to Eq. (20), we can calculate $\mathbf{E}_{j,q,r}h^{(\omega)}_{ijqr}$ as

$$\begin{aligned}
\mathbf{E}_{j,q,r}h^{(\omega)}_{ijqr} =& \frac{2}{4!}\sum_{(u,v,w)}^{(j,q,r)}\mathbf{E}_{u,v,w}(k^{(\omega)}_{iu}l^{(\omega)}_{iu} + k^{(\omega)}_{iu}l^{(\omega)}_{vw} - k^{(\omega)}_{iu}l^{(\omega)}_{iv}) - \frac{2}{4!}\sum_{(t,v,w)}^{(j,q,r)}\mathbf{E}_{t,v,w}(k^{(\omega)}_{ti}l^{(\omega)}_{tv}) \\
&+ \frac{2}{4!}\sum_{(t,u,w)}^{(j,q,r)}\mathbf{E}_{t,u,w}(k^{(\omega)}_{tu}l^{(\omega)}_{tu} + k^{(\omega)}_{tu}l^{(\omega)}_{iw} - k^{(\omega)}_{tu}l^{(\omega)}_{ti}) - \frac{2}{4!}\sum_{(t,u,v)}^{(j,q,r)}\mathbf{E}_{t,u,v}(k^{(\omega)}_{tu}l^{(\omega)}_{tv}) \\
=& \frac{1}{2}\mathbf{E}^{i\neq u\neq v\neq w}_{u,v,w}(k^{(\omega)}_{iu}l^{(\omega)}_{iu} + k^{(\omega)}_{iu}l^{(\omega)}_{vw} - k^{(\omega)}_{iu}l^{(\omega)}_{iv}) - \frac{1}{2}\mathbf{E}^{i\neq t\neq v\neq w}_{t,v,w}(k^{(\omega)}_{ti}l^{(\omega)}_{tv}) \\
&+ \frac{1}{2}\mathbf{E}^{i\neq t\neq u\neq w}_{t,u,w}(k^{(\omega)}_{tu}l^{(\omega)}_{tu} + k^{(\omega)}_{tu}l^{(\omega)}_{iw} - k^{(\omega)}_{tu}l^{(\omega)}_{ti}) - \frac{1}{2}\mathbf{E}^{i\neq t\neq u\neq v}_{t,u,v}(k^{(\omega)}_{tu}l^{(\omega)}_{tv}),
\end{aligned} \tag{32}$$

where we define simplified notions $\mathbf{E}^{i\neq u\neq v\neq w}_{u,v,w}$ whose superscript indicates the restriction. For readability, we will require additional notations: $\mathbf{E}_x k^{(\omega)}_i := \mathbf{E}^{i\neq u}_u k^{(\omega)}_{iu}$ and $\mathbf{E}_x k^{(\omega)} := \mathbf{E}^{t\neq u}_{t,u}k^{(\omega)}_{tu}$ (the notations for Y is defined by analogy). Under $\mathcal{H}_0$, X and Y are independence. Hence

$$\begin{aligned}
2\mathbf{E}_{j,q,r}h^{(\omega)}_{ijqr} =& \mathbf{E}_x k^{(\omega)}_i\mathbf{E}_y l^{(\omega)}_i + \mathbf{E}_x k^{(\omega)}_i\mathbf{E}_y l^{(\omega)} - \mathbf{E}_x k^{(\omega)}_i\mathbf{E}_y l^{(\omega)}_i - \mathbf{E}_x k^{(\omega)}_i\mathbf{E}_y l^{(\omega)} \\
&+ \mathbf{E}_x k^{(\omega)}\mathbf{E}_y l^{(\omega)} + \mathbf{E}_x k^{(\omega)}\mathbf{E}_y l^{(\omega)}_i - \mathbf{E}_x k^{(\omega)}\mathbf{E}_y l^{(\omega)}_i - \mathbf{E}_x k^{(\omega)}\mathbf{E}_y l^{(\omega)} = 0.
\end{aligned} \tag{33}$$

Then combining the results [34, Section 5.5.2], we can prove Eq. (30).

Step 3: Next we prove the asymptotic distribution under $\mathcal{H}_1$. We only need to show that $|\text{HSIC}_\omega(Z) - \text{HSIC}^{(u)}_\omega(Z)| \sim \mathcal{O}(1/n)$. By the definition of $k^{(\omega)}_{tu}, l^{(\omega)}_{tu}$, we can check that $|k^{(\omega)}_{tu}| \leq 1, |l^{(\omega)}_{tu}| \leq 1$ for all t,u, thus $|h^{(\omega)}_{ijqr}| \leq 4$ for all i,j,q,r. Hence we have

$$|\text{HSIC}_\omega(Z) - \text{HSIC}^{(u)}_\omega(Z)| \leq \frac{n^4 - (n)_4}{n^4}\cdot 4 + \Big(\frac{1}{(n)_4} - \frac{1}{n^4}\Big)\cdot(n)_4\cdot 4 \sim \mathcal{O}(1/n). \tag{34}$$

Combining the results [34, Section 5.5.1], we can prove Eq. (31). □

E Proof of Theorem 1

In this section, we give a proof of the Theorem 1. We first restate the Theorem 1 here.

Theorem 1 (Linear-Time Estimators). *Under $\mathcal{H}_0$, the estimators of mean and variance with bias of $\mathcal{O}(n^{-1})$ to $\mathbf{E}_Z[n\text{HSIC}_\omega(Z)]$ and $\mathbf{Var}_Z[n\text{HSIC}_\omega(Z)]$, denote as $\mathcal{E}_0$ and $\mathcal{V}_0$, respectively, are given by*

$$\mathcal{E}_0 := \frac{[\mathbf{1}^T\mathbf{\Lambda}_{Xc}^{\cdot 2}\mathbf{1}][\mathbf{1}^T\mathbf{\Lambda}_{Yc}^{\cdot 2}\mathbf{1}]}{(n-1)^2}, \mathcal{V}_0 := \frac{2n(n-4)(n-5)}{(n-1)(n-2)(n-3)}\frac{[\mathbf{1}^T(\mathbf{\Lambda}_{Xc}^T\mathbf{\Lambda}_{Xc})^{\cdot 2}\mathbf{1}][\mathbf{1}^T(\mathbf{\Lambda}_{Yc}^T\mathbf{\Lambda}_{Yc})^{\cdot 2}\mathbf{1}]}{n^4}, \tag{35}$$

where $()^{\cdot 2}$ is the entrywise matrix power. Both $\mathcal{E}_0$ and $\mathcal{V}_0$ can be calculated in $\mathcal{O}(nD^2)$ time.

Proof. We first prove the part of the mean. Recall the definition of $\text{HSIC}_\omega(Z)$, we have $\text{HSIC}_\omega(Z) = \frac{1}{n^4}\sum_{i,j,q,r} h_{ijqr}^{(\omega)}$. Hence $\mathbf{E}_Z[\text{HSIC}_\omega(Z)] = \frac{1}{n^4}\sum_{i,j,q,r}\mathbf{E}_Z h_{ijqr}^{(\omega)}$. When $(i,j,q,r) \in \mathbf{i}_4^n$, then we can show that under $\mathcal{H}_0$, $\mathbf{E}_{i,j,q,r} h_{ijqr}^{(\omega)} = 0$ by performing the same analysis as in Eqs. (32) and (33). Then we consider the case where exactly two elements of i,j,q,r are the same, for a total of $6n(n-1)(n-2)$ terms. By the symmetry of $h_{ijqr}^{(\omega)}$, the expectation of these terms all take the same value, and here we take $h_{iiqr}^{(\omega)}$ as an example. According to Eq. (20), we can represent $h_{iiqr}^{(\omega)}$ as

$$\begin{aligned} h_{iiqr}^{(\omega)} =& \frac{2}{4!}\sum_{(u,v,w)}^{(i,q,r)}(k_{iu}^{(\omega)}l_{iu}^{(\omega)} + k_{iu}^{(\omega)}l_{vw}^{(\omega)} - k_{iu}^{(\omega)}l_{iv}^{(\omega)}) - \frac{2}{4!}\sum_{(t,v,w)}^{(i,q,r)}(k_{ti}^{(\omega)}l_{tv}^{(\omega)}) \\ &+ \frac{2}{4!}\sum_{(t,u,w)}^{(i,q,r)}(k_{tu}^{(\omega)}l_{tu}^{(\omega)} + k_{tu}^{(\omega)}l_{iw}^{(\omega)} - k_{tu}^{(\omega)}l_{ti}^{(\omega)}) - \frac{2}{4!}\sum_{(t,u,v)}^{(i,q,r)}(k_{tu}^{(\omega)}l_{tv}^{(\omega)}). \end{aligned} \tag{36}$$

Under $\mathcal{H}_0$, X and Y are independence. Take the expectation on both sides, we have

$$\begin{aligned} 12\mathbf{E}_{i,q,r}h_{iiqr}^{(\omega)} =& (2 + 4\mathbf{E}_x k^{(\omega)}\mathbf{E}_y l^{(\omega)}) + (2\mathbf{E}_y l^{(\omega)} + 4\mathbf{E}_x k^{(\omega)}\mathbf{E}_y l^{(\omega)}) \\ &- (2\mathbf{E}_x k^{(\omega)} + 2\mathbf{E}_y l^{(\omega)} + 2\mathbf{E}_x k^{(\omega)}\mathbf{E}_y l^{(\omega)}) - (2\mathbf{E}_y l^{(\omega)} + 4\mathbf{E}_x k^{(\omega)}\mathbf{E}_y l^{(\omega)}) \\ &+ (6\mathbf{E}_x k^{(\omega)}\mathbf{E}_y l^{(\omega)}) + (2\mathbf{E}_x k^{(\omega)} + 4\mathbf{E}_x k^{(\omega)}\mathbf{E}_y l^{(\omega)}) \\ &- (2\mathbf{E}_x k^{(\omega)} + 4\mathbf{E}_x k^{(\omega)}\mathbf{E}_y l^{(\omega)}) - (6\mathbf{E}_x k^{(\omega)}\mathbf{E}_y l^{(\omega)}) \\ =& 2(1 - \mathbf{E}_x k^{(\omega)} - \mathbf{E}_y l^{(\omega)} + \mathbf{E}_x k^{(\omega)}\mathbf{E}_y l^{(\omega)}), \end{aligned} \tag{37}$$

where we define additional notation $\mathbf{E}_x k^{(\omega)} := \mathbf{E}_{t,u}^{t\neq u} k_{tu}^{(\omega)}$ (the notation for Y is defined by analogy) and use $k_{tt}^{\omega} = l_{tt}^{\omega} = 1$. Hence in this case, the sum of the contributions of all terms to $\mathbf{E}_Z[n\text{HSIC}_\omega(Z)]$ is $(1 - \mathbf{E}_x k^{(\omega)} - \mathbf{E}_y l^{(\omega)} + \mathbf{E}_x k^{(\omega)}\mathbf{E}_y l^{(\omega)}) + \mathcal{O}(1/n)$. For the remaining terms, i.e., the case where at least three of i,j,q,r are equal, combined with the boundedness of $h_{ijqr}^{(\omega)}$, we can conclude that the sum of their contributions is $\mathcal{O}(1/n)$. As a result, we have shown that

$$\mathbf{E}_Z[n\text{HSIC}_\omega(Z)] = (1 - \mathbf{E}_x k^{(\omega)})(1 - \mathbf{E}_y l^{(\omega)}) + \mathcal{O}(n^{-1}). \tag{38}$$

The unbiased estimators of $\mathbf{E}_x k^{(\omega)}, \mathbf{E}_y l^{(\omega)}$ are given by $\mathbf{1}^T(\mathbf{\Lambda}_X\mathbf{\Lambda}_X^T - \mathbf{I}_n)\mathbf{1}, \mathbf{1}^T(\mathbf{\Lambda}_Y\mathbf{\Lambda}_Y^T - \mathbf{I}_n)\mathbf{1}$, respectively. Hence under $\mathcal{H}_0$, we obtain the estimator of mean with bias of $\mathcal{O}(n^{-1})$ as

$$\left[1 - \frac{\mathbf{1}^T(\mathbf{\Lambda}_X\mathbf{\Lambda}_X^T - \mathbf{I}_n)\mathbf{1}}{n(n-1)}\right]\left[1 - \frac{\mathbf{1}^T(\mathbf{\Lambda}_Y\mathbf{\Lambda}_Y^T - \mathbf{I}_n)\mathbf{1}}{n(n-1)}\right] = \frac{[\mathbf{1}^T\mathbf{\Lambda}_{Xc}^{\cdot 2}\mathbf{1}][\mathbf{1}^T\mathbf{\Lambda}_{Yc}^{\cdot 2}\mathbf{1}]}{(n-1)^2}. \tag{39}$$

Next, we prove the part of the variance. We start by calculating $\mathbf{Var}_Z[n\text{HSIC}_\omega^{(u)}(Z)]$, where the U-statistic $\text{HSIC}_\omega^{(u)}(Z) := \frac{1}{(n)_4}\sum_{(i,j,q,r)\in\mathbf{i}_4^n} h_{ijqr}^{(\omega)}$. According to the results [34, Section 5.2.1, Lemma A], we have

$$\mathbf{Var}[\text{HSIC}_\omega^{(u)}(Z)] = \binom{n}{4}^{-1}\sum_{c=1}^{4}\binom{4}{c}\binom{n-4}{4-c}\zeta_c = \frac{4\binom{n-4}{3}}{\binom{n}{4}}\zeta_1 + \frac{6\binom{n-4}{2}}{\binom{n}{4}}\zeta_2 + \mathcal{O}(n^{-3}), \tag{40}$$

where $\zeta_1 := \mathbf{E}_i\big(\mathbf{E}_{j,q,r}h^{(\omega)}_{ijqr}\big)^2$ and $\zeta_2 := \mathbf{E}_{i,j}\big(\mathbf{E}_{q,r}h^{(\omega)}_{ijqr}\big)^2$. Under $\mathcal{H}_0$, when $(i,j,q,r) \in \mathbf{i}^n_4$, we can show that $\mathbf{E}_{j,q,r}h^{(\omega)}_{ijqr} = 0$ by performing the same analysis as in Eqs. (32) and (33), thus $\zeta_1 = 0$. For calculating ζ_2, we mainly focus on the term $\mathbf{E}_{q,r}h^{(\omega)}_{ijqr}$. We use Eq. (20) again,

$$
\begin{aligned}
12h^{(\omega)}_{ijqr} = &\sum_{(u,v,w)}^{(j,q,r)} (k^{(\omega)}_{iu}l^{(\omega)}_{iu} + k^{(\omega)}_{iu}l^{(\omega)}_{vw} - k^{(\omega)}_{iu}l^{(\omega)}_{iv}) - \sum_{(t,v,w)}^{(j,q,r)} (k^{(\omega)}_{ti}l^{(\omega)}_{tv}) \\
&+ \sum_{(t,u,w)}^{(j,q,r)} (k^{(\omega)}_{tu}l^{(\omega)}_{tu} + k^{(\omega)}_{tu}l^{(\omega)}_{iw} - k^{(\omega)}_{tu}l^{(\omega)}_{ti}) - \sum_{(t,u,v)}^{(j,q,r)} (k^{(\omega)}_{tu}l^{(\omega)}_{tv}).
\end{aligned} \tag{41}
$$

Under $\mathcal{H}_0$, X and Y are independence. Take the expectation $\mathbf{E}_{q,r}$ on both sides, we have

$$
\begin{aligned}
12\mathbf{E}_{q,r}h^{(\omega)}_{ijqr} = &(2k^{(\omega)}_{ij}l^{(\omega)}_{ij} + 4\mathbf{E}_x k^{(\omega)}_i \mathbf{E}_y l^{(\omega)}_i) + (2k^{(\omega)}_{ij}\mathbf{E}_y l^{(\omega)} + 4\mathbf{E}_x k^{(\omega)}_i \mathbf{E}_y l^{(\omega)}_j) \\
&- (2k^{(\omega)}_{ij}\mathbf{E}_y l^{(\omega)}_i + 2\mathbf{E}_x k^{(\omega)}_i l^{(\omega)}_{ij} + 2\mathbf{E}_x k^{(\omega)}_i \mathbf{E}_y l^{(\omega)}_i) \\
&- (2k^{(\omega)}_{ij}\mathbf{E}_y l^{(\omega)}_j + 2\mathbf{E}_x k^{(\omega)}_i \mathbf{E}_y l^{(\omega)}_j + 2\mathbf{E}_x k^{(\omega)}_i \mathbf{E}_y l^{(\omega)}) \\
&+ (4\mathbf{E}_x k^{(\omega)}_j \mathbf{E}_y l^{(\omega)}_j + 2\mathbf{E}_x k^{(\omega)} \mathbf{E}_y l^{(\omega)}) + (4\mathbf{E}_x k^{(\omega)}_j \mathbf{E}_y l^{(\omega)}_i + 2\mathbf{E}_x k^{(\omega)} l^{(\omega)}_{ij}) \\
&- (2\mathbf{E}_x k^{(\omega)}_j \mathbf{E}_y l^{(\omega)}_{ij} + 2\mathbf{E}_x k^{(\omega)}_j \mathbf{E}_y l^{(\omega)}_i + 2\mathbf{E}_x k^{(\omega)} \mathbf{E}_y l^{(\omega)}_i) \\
&- (2\mathbf{E}_x k^{(\omega)}_j \mathbf{E}_y l^{(\omega)}_j + 2\mathbf{E}_x k^{(\omega)}_j \mathbf{E}_y l^{(\omega)} + 2\mathbf{E}_x k^{(\omega)} \mathbf{E}_y l^{(\omega)}_j) \\
= &\, 2(k^{(\omega)}_{ij} - \mathbf{E}_x k^{(\omega)}_i - \mathbf{E}_x k^{(\omega)}_j + \mathbf{E}_x k^{(\omega)})(l^{(\omega)}_{ij} - \mathbf{E}_y l^{(\omega)}_i - \mathbf{E}_y l^{(\omega)}_j + \mathbf{E}_y l^{(\omega)}),
\end{aligned} \tag{42}
$$

where we define additional notation $\mathbf{E}_x k^{(\omega)} := \mathbf{E}^{t\neq u}_{t,u} k^{(\omega)}_{tu}$, $\mathbf{E}_x k^{(\omega)}_j := \mathbf{E}^{j\neq u}_u k^{(\omega)}_{ju}$ (the notations for Y are defined by analogy). For simplify, we denote $k^{(\omega)}_{c;ij} := k^{(\omega)}_{ij} - \mathbf{E}_x k^{(\omega)}_i - \mathbf{E}_x k^{(\omega)}_j + \mathbf{E}_x k^{(\omega)}$ and $l^{(\omega)}_{c;ij} := l^{(\omega)}_{ij} - \mathbf{E}_y l^{(\omega)}_i - \mathbf{E}_y l^{(\omega)}_j + \mathbf{E}_y l^{(\omega)}$. Hence combining Eq. (40), we have

$$
\mathbf{Var}[n\mathrm{HSIC}^{(u)}_\omega(Z)] = \frac{2n(n-4)(n-5)}{(n-1)(n-2)(n-3)} \mathbf{E}_{i,j}(k^{(\omega)}_{c;ij})^2 \cdot \mathbf{E}_{i,j}(l^{(\omega)}_{c;ij})^2 + \mathcal{O}(n^{-1}). \tag{43}
$$

Since under $\mathcal{H}_0$, the bias lead by the difference terms between $\mathrm{HSIC}^{(u)}_\omega(Z)$ and $\mathrm{HSIC}_\omega(Z)$ vanish faster than Eq. (43), hence the variance of $\mathrm{HSIC}_\omega(Z)$ is identical. In the following part, we consider the empirical estimate of the leading term in Eq. (43). We estimate $k^{(\omega)}_{c;ij}$ with $(\mathbf{\Lambda}_{Xc}\mathbf{\Lambda}^T_{Xc})_{ij}$, then the estimation of $\mathbf{E}_{i,j}(k^{(\omega)}_{c;ij})^2$ is given by

$$
\frac{1}{n^2}\sum_{i,j}\big[(\Lambda_k(x_i) - \overline{\Lambda}_k)(\Lambda_k(x_j) - \overline{\Lambda}_k)^T\big]^2 = \frac{1}{n^2}\sum_{i,j}(\mathbf{\Lambda}_{Xc}\mathbf{\Lambda}^T_{Xc})^2_{ij} = \frac{\mathbf{1}^T(\mathbf{\Lambda}_{Xc}\mathbf{\Lambda}^T_{Xc})^{\cdot 2}\mathbf{1}}{n^2}, \tag{44}
$$

where we define notions $\overline{\Lambda}_k := \frac{1}{n}\sum_{u=1}^n \Lambda_k(x_u)$. Since computing the value of $\mathbf{1}^T(\mathbf{\Lambda}_{Xc}\mathbf{\Lambda}^T_{Xc})^{\cdot 2}\mathbf{1}$ requires $\mathcal{O}(n^2)$ time complexity, we transform it into a more computationally tractable form. We perform the following calculating as

$$
\mathbf{1}^T(\mathbf{\Lambda}_{Xc}\mathbf{\Lambda}^T_{Xc})^{\cdot 2}\mathbf{1} = \mathrm{Tr}(\mathbf{\Lambda}_{Xc}\mathbf{\Lambda}^T_{Xc}\mathbf{\Lambda}_{Xc}\mathbf{\Lambda}^T_{Xc}) = \mathrm{Tr}(\mathbf{\Lambda}^T_{Xc}\mathbf{\Lambda}_{Xc}\mathbf{\Lambda}^T_{Xc}\mathbf{\Lambda}_{Xc}) = \mathbf{1}^T(\mathbf{\Lambda}^T_{Xc}\mathbf{\Lambda}_{Xc})^{\cdot 2}\mathbf{1}. \tag{45}
$$

Recall the definition of $\mathbf{\Lambda}_{Xc} := [\mathbf{H}\mathbf{\Lambda}_X]_{n\times D}$ that can be calculated in $\mathcal{O}(nD)$ time, thus the term $[\mathbf{\Lambda}^T_{Xc}\mathbf{\Lambda}_{Xc}]_{D\times D}$ can be calculated in $\mathcal{O}(nD + nD^2)$ time. As a result, we obtain the estimator $[\mathbf{1}^T(\mathbf{\Lambda}^T_{Xc}\mathbf{\Lambda}_{Xc})^{\cdot 2}\mathbf{1}][\mathbf{1}^T(\mathbf{\Lambda}^T_{Yc}\mathbf{\Lambda}_{Yc})^{\cdot 2}\mathbf{1}]$ for the term $\mathbf{E}_{i,j}(k^{(\omega)}_{c;ij})^2 \cdot \mathbf{E}_{i,j}(l^{(\omega)}_{c;ij})^2$ that can be calculated in $\mathcal{O}(nD^2)$ time. The only thing left to do is to determine the bias of the estimator. For readable, we define $\widehat{k}^{(\omega)}_{ij} := k^{(\omega)}_{ij}$, $\widehat{k}^{(\omega)}_i := \frac{1}{n}\sum_u k^{(\omega)}_{iu}$ and $\widehat{k}^{(\omega)} := \frac{1}{n^2}\sum_{u,v} k^{(\omega)}_{uv}$, then by removing the terms with $i = j$, a estimate with difference $\mathcal{O}(n^{-1})$ to Eq. (44) is given by

$$
\frac{1}{n(n-1)}\sum_{i\neq j}\Big[\widehat{k}^{(\omega)}_{ij} - \widehat{k}^{(\omega)}_i - \widehat{k}^{(\omega)}_j + \widehat{k}^{(\omega)}\Big]^2. \tag{46}
$$

By comparing the difference between the expectation of Eq. (46) and $\mathbf{E}_{i,j}(k^{(\omega)}_{c;ij})^2$, we can show that this error is bound by $O(1/n)$. We illustrate this by taking one of the cross terms as an example and the other terms by analogy, as shown in the following,

$$\begin{aligned}\mathbf{E}\Big[\frac{1}{n(n-1)}\sum_{i\neq j}\widehat{k}^{(\omega)}_i\widehat{k}^{(\omega)}\Big] &= \frac{1}{n^3(n-1)}\mathbf{E}\Big[\sum_i\sum_{q,r}\sum_u k^{(\omega)}_{iu}k^{(\omega)}_{qr}\Big] \\ &= \frac{1}{(n)_4}\mathbf{E}\Big[\sum_{(i,q,r,u)\in \mathbf{i}^n_4} k^{(\omega)}_{iu}k^{(\omega)}_{qr}\Big] + \mathcal{O}(n^{-1}) = \mathbf{E}_x k^{(\omega)}_i \mathbf{E}_x k^{(\omega)} + \mathcal{O}(n^{-1}).\end{aligned} \tag{47}$$

Similarly, we can obtain the results for $\mathbf{E}_{i,j}(l^{(\omega)}_{c;ij})^2$. As a result, we have shown that $\mathcal{V}_0$ is a estimator of $\mathbf{Var}_Z[n\mathrm{HSIC}_\omega(Z)]$ with bias $\mathcal{O}(n^{-1})$ thus complete the whole proof. □

F Calculation of Eq. (16)

Here, we give the computational details of Eq. (16). We mark colors to indicate correspondences.

According to Eq. (20), we can calculate $\sum_{j,q,r} h^{(\omega)}_{ijqr}$ as

$$\begin{aligned}\sum_{j,q,r} h^{(\omega)}_{ijqr} = &\frac{1}{2}\sum_{u,v,w}(k^{(\omega)}_{iu}l^{(\omega)}_{iu} + k^{(\omega)}_{iu}l^{(\omega)}_{vw} - k^{(\omega)}_{iu}l^{(\omega)}_{iv}) - \frac{1}{2}\sum_{t,v,w}(k^{(\omega)}_{ti}l^{(\omega)}_{tv}) \\ &+ \frac{1}{2}\sum_{t,u,w}(k^{(\omega)}_{tu}l^{(\omega)}_{tu} + k^{(\omega)}_{tu}l^{(\omega)}_{iw} - k^{(\omega)}_{tu}l^{(\omega)}_{ti}) - \frac{1}{2}\sum_{t,u,v}(k^{(\omega)}_{tu}l^{(\omega)}_{tv}).\end{aligned} \tag{48}$$

We can further represent Eq. (48) in matrices form as

$$\begin{aligned}\sum_{j,q,r} h^{(\omega)}_{ijqr} = \frac{1}{2}\Big[&n^2(\mathbf{\Lambda}_X\mathbf{\Lambda}_X^T\mathbf{\Lambda}_Y\mathbf{\Lambda}_Y^T)_{i,i} + (\mathbf{\Lambda}_X\mathbf{\Lambda}_X^T\mathbf{1})_i(\mathbf{1}^T\mathbf{\Lambda}_Y\mathbf{\Lambda}_Y^T\mathbf{1}) - n[(\mathbf{\Lambda}_X\mathbf{\Lambda}_X^T\mathbf{1})\odot(\mathbf{\Lambda}_Y\mathbf{\Lambda}_Y^T\mathbf{1})]_i \\ &+ n\mathrm{Tr}(\mathbf{\Lambda}_X\mathbf{\Lambda}_X^T\mathbf{\Lambda}_Y\mathbf{\Lambda}_Y^T) + (\mathbf{\Lambda}_Y\mathbf{\Lambda}_Y^T\mathbf{1})_i(\mathbf{1}^T\mathbf{\Lambda}_X\mathbf{\Lambda}_X^T\mathbf{1}) - n(\mathbf{\Lambda}_Y\mathbf{\Lambda}_Y^T\mathbf{\Lambda}_X\mathbf{\Lambda}_X^T\mathbf{1})_i \\ &- n(\mathbf{\Lambda}_X\mathbf{\Lambda}_X^T\mathbf{\Lambda}_Y\mathbf{\Lambda}_Y^T\mathbf{1})_i - (\mathbf{1}^T\mathbf{\Lambda}_X\mathbf{\Lambda}_X^T\mathbf{\Lambda}_Y\mathbf{\Lambda}_Y^T\mathbf{1})\Big].\end{aligned} \tag{49}$$

Next, by variable substitution, we obtain the result as

$$\sum_{j,q,r} h^{(\omega)}_{ijqr} = \frac{1}{2}\left[n\mathbf{1}^T\mathbf{A}\mathbf{1} + n^2(\mathbf{A}\mathbf{1})_i + (\mathbf{1}^T\mathbf{C})\mathbf{B}_i + (\mathbf{1}^T\mathbf{B})\mathbf{C}_i - n\mathbf{E}_i - n\mathbf{F}_i - n\mathbf{D}_i - \mathbf{1}^T\mathbf{D}\right]. \tag{50}$$

where the definition of variables $\mathbf{A}$ to $\mathbf{F}$ with the calculation cost are given in the Fig. 1. For convenience, we re-show the diagram here for reference.

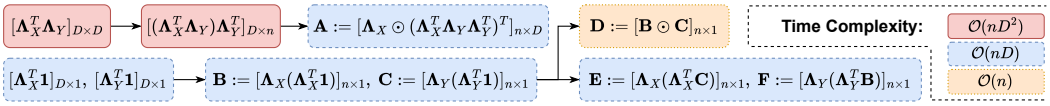

Figure 5: The diagram shows the definition of the quantities, with styles representing the time complexity of the computational process in the current box. $\odot$: the element-wise product.

The computational complexity of each step is illustrated in Fig. 5. We explain some steps here. As a start, recall that the size of $\mathbf{\Lambda}_X, \mathbf{\Lambda}_Y$ are both $n \times D$. Therefore a time complexity $\mathcal{O}(nD^2)$ is required to compute $\mathbf{\Lambda}_X^T\mathbf{\Lambda}_Y$ by matrix multiplication operation. Further multiplying the obtained $[\mathbf{\Lambda}_X^T\mathbf{\Lambda}_Y]_{D\times D}$ with $\mathbf{\Lambda}_Y^T$ requires $\mathcal{O}(nD^2)$ time complexity. Next, since both $\mathbf{\Lambda}_X$ and $(\mathbf{\Lambda}_X^T\mathbf{\Lambda}_Y\mathbf{\Lambda}_Y^T)^T$ are of size $n \times D$, the elemental product operation requires a time complexity of $\mathcal{O}(nD)$. In a similar way, we can check the time complexity for each remained step. After getting the variables $\mathbf{A}$ to $\mathbf{F}$, since $\mathbf{1}^T\mathbf{A}\mathbf{1}, \mathbf{A}\mathbf{1}$ all can be calculated in $\mathcal{O}(nD)$ and $\mathbf{1}^T\mathbf{B}, \mathbf{1}^T\mathbf{C}, \mathbf{1}^T\mathbf{D}$ all can be calculated in $\mathcal{O}(n)$, we conclude that the results with index i in Eq. (50) can be obtained in $\mathcal{O}(nD^2)$ time.

G Proof of Theorem 2

In this section, we give a proof of the Theorem 2. We first restate the Theorem 2 here.

Theorem 2 (Uniform Bound). *Let θ_k, θ_l parameterize $\mathcal{T}_{\theta_k}, \mathcal{T}_{\theta_l}$ in Banach spaces of dimension d_k, d_l. And $\mathcal{T}_{\theta_k}, \mathcal{T}_{\theta_l}$ are Lipschitz to the parameters θ_k, θ_l with the non-negative constant L_k, L_l, respectively. Let Θ_c be a set of (θ_k, θ_l) for which $\sigma_\omega \geq c > 0$ with a positive constant c and $\|\theta_k\| \leq R_{\theta_k}, \|\theta_l\| \leq R_{\theta_l}$. Let r denote the threshold, i.e., $(1-\alpha)$-quantile for the distribution in Eq. (11) and $r^{(n)}$ be the threshold with sample size n. Let $\{(\omega_{k;j}, \omega_{l;j})\}_{j=1}^{D/2}$ be the samplings of frequency with the sampling number D. Also, we define $R_{\omega_k} := \sup_j \|\omega_{k;j}\|, R_{\omega_l} := \sup_j \|\omega_{l;j}\|, d_s := \max\{d_k, d_l\}$ and $\xi_\omega :=$ HSIC$_\omega(Z)$. Then with probability at least $1-\delta$, we have*

$$\sup_{(\theta_k,\theta_l)\in\Theta_c}\left|\frac{\xi_\omega - r_\omega^{(n)}/n}{\widehat{\sigma}_\omega} - \frac{\mathbf{E}_Z\xi_\omega - r_\omega/n}{\sigma_\omega}\right| \sim \mathcal{O}\left(\left[\sqrt{\frac{1}{n}\log\frac{1}{\delta} + d_s\frac{\log n}{n}} + \frac{R_{\omega_k}L_k + R_{\omega_l}L_l}{\sqrt{n}}\right]\right).$$

Proof. We take a similar roadmap of proof as [30] and extend it to our optimization objective. The roadmap of the proof is as follows: we first obtain the convergence results (with sample size n) for each estimator with fixed parameters θ_k, θ_l, and then extend the results to the entire parameter space via ϵ-net arguments. We begin the proof of the first part, which is based on bounded differences inequality (McDiarmid's inequality) [41, Theorem 2.9.1].

Bound of $|\xi_\omega - \mathbf{E}_Z\xi_\omega|$. Recall the definition of $\xi_\omega := \text{HSIC}_\omega(Z) = \frac{1}{n^4}\sum_{i,j,q,r} h_{ijqr}^{(\omega)}$. By the definition of $k_{tu}^{(\omega)}, l_{tu}^{(\omega)}$, we have $|k_{tu}^{(\omega)}| \leq 1, |l_{tu}^{(\omega)}| \leq 1$ for all t, u, thus $|h_{ijqr}^{(\omega)}| \leq 4$ for all i, j, q, r. Now we begin by showing the bounded differences property of $h_{ijqr}^{(\omega)}$. Concretely, we replace the first sample $z_1 = (x_1, y_1)$ with $z_1' = (x_1', y_1')$ and keep the remaining samples the same. The obtained samples are named as Z'. Then the difference terms between $h_{ijqr}^{(\omega)}$ and the new substitution $\breve{h}_{ijqr}^{(\omega)}$ can only happen in the case that at least one of i, j, q, r is equal to 1. For the case that only one subscript is 1 (here take $i = 1$ for example), combining Eq. (20), we have

$$\Big|\sum_{j,q,r} h_{1jqr}^{(\omega)} - \sum_{j,q,r}\breve{h}_{1jqr}^{(\omega)}\Big| \leq \frac{2}{4!}(n-1)(n-2)(n-3)\cdot 6\cdot 16 = 8(n-1)(n-2)(n-3). \quad (51)$$

The whole contributes of remaining terms that at least two i, j, q, r are less than $\mathcal{O}(n^{-2})$, thus

$$\Big|\frac{1}{n^4}\sum_{i,j,q,r} h_{ijqr}^{(\omega)} - \frac{1}{n^4}\sum_{i,j,q,r}\breve{h}_{ijqr}^{(\omega)}\Big| \leq 4\cdot\Big|\frac{1}{n^4}\sum_{j,q,r} h_{1jqr}^{(\omega)} - \frac{1}{n^4}\sum_{j,q,r}\breve{h}_{1jqr}^{(\omega)}\Big| + \mathcal{O}(n^{-2}) = \mathcal{O}(n^{-1}). \quad (52)$$

Hence HSIC$_\omega(Z)$ satisfy the bounded differences property with $\mathcal{O}(n^{-1})$. Using McDiarmid's inequality, for fixed θ_0, θ_1, with probability at least $1-\delta$, there exist a universal constant C_1 such that

$$\big|\xi_\omega - \mathbf{E}_Z\xi_\omega\big| \leq C_1\sqrt{\frac{1}{n}\log\frac{2}{\delta}}. \quad (53)$$

Bound of $|r_\omega^{(n)} - r_\omega|$. As $r_\omega^{(n)}$ is the $(1-\alpha)$ of the distribution of $n\xi_\omega$ with sample size n under $\mathcal{H}_0$, according to the Eq. (53), when n is large enough, there exist a universal constant C_2 such that

$$|r_\omega^{(n)}|/n \leq C_1\sqrt{\frac{1}{n}\log\frac{2}{\alpha}} + |\mathbf{E}_Z\xi_\omega| \leq C_2\sqrt{\frac{1}{n}\log\frac{1}{\alpha}}, \quad (54)$$

where the last inequation is because $\mathbf{E}_Z\xi_\omega \sim \mathcal{O}(n^{-1})$ under $\mathcal{H}_0$ (see Theorem 1 for a detailed explanation). Hence $|r_\omega^{(n)}| \sim \mathcal{O}\big(\sqrt{n\log(1/\alpha)}\big)$. Also, by definition r_ω is a constant related to α.

Bound of $|\widehat{\sigma}_\omega^2 - \sigma_\omega^2|$. In this part, We first obtain the bound of $|\widehat{\sigma}_\omega^2 - \mathbf{E}_Z\widehat{\sigma}_\omega^2|$, then obtain the bound of $|\mathbf{E}_Z\widehat{\sigma}_\omega^2 - \sigma_\omega^2|$. As before we start by showing the bounded variance property of $\widehat{\sigma}_\omega^2$. We replace $z_1 = (x_1, y_1)$ with $z_1' = (x_1', y_1')$ and keep the remaining samples the same. The obtained samples are named as Z'. For readable, we denote $\widehat{\sigma}_\omega^2$ with sample Z, Z' as $\widehat{\sigma}_\omega^2(Z), \widehat{\sigma}_\omega^2(Z')$ respectively.

Recall the definition $\widehat{\sigma}_\omega^2 := 16\big[\frac{1}{n}\sum_i(\frac{1}{n^3}\sum_{j,q,r} h_{ijqr}^{(\omega)})^2 - \text{HSIC}_\omega^2(Z)\big]$. Since $|h_{ijqr}^{(\omega)}| \leq 4$, we have

$$\Big|\frac{1}{n}\sum_i\Big(\frac{1}{n^3}\sum_{j,q,r} h_{ijqr}^{(\omega)}\Big)^2 - \frac{1}{n}\sum_i\Big(\frac{1}{n^3}\sum_{j,q,r}\breve{h}_{ijqr}^{(\omega)}\Big)^2\Big| \leq \frac{8}{n^4}\sum_i\sum_{j,q,r}\big|h_{ijqr}^{(\omega)} - \breve{h}_{ijqr}^{(\omega)}\big|, \quad (55)$$

$$\Big|\Big(\frac{1}{n^4}\sum_{i,j,q,r} h_{ijqr}^{(\omega)}\Big)^2-\Big(\frac{1}{n^4}\sum_{i,j,q,r} \breve{h}_{ijqr}^{(\omega)}\Big)^2\Big|\leq \frac{8}{n^4}\sum_{i,j,q,r}\big|h_{ijqr}^{(\omega)}-\breve{h}_{ijqr}^{(\omega)}\big|. \tag{56}$$

Again, the difference terms between $h_{ijqr}^{(\omega)}$ and the new substitution $\breve{h}_{ijqr}^{(\omega)}$ can only happen in the case that at least one of i,j,q,r is equal to 1. Hence, Eqs. (55) and (56) are both $\mathcal{O}(n^{-1})$. As a result, $\widehat{\sigma}_\omega^2$ satisfy the bounded differences property with bound $\mathcal{O}(n^{-1})$. Using McDiarmid's inequality, with probability at least $1-\delta$, there exist a universal constant C_3 such that $\big|\widehat{\sigma}_\omega^2-\mathbf{E}_Z\widehat{\sigma}_\omega^2\big|\leq C_3\sqrt{\frac{1}{n}\log\frac{2}{\delta}}$. Next we obtain the bound of $|\mathbf{E}_Z\widehat{\sigma}_\omega^2-\sigma_\omega^2|$. We rewrite $\mathbf{E}_Z\widehat{\sigma}_\omega^2$ as

$$\mathbf{E}_Z\widehat{\sigma}_\omega^2=16\Big(\frac{1}{n^7}\sum_{ijqrj'q'r'}\mathbf{E}[h_{ijqr}^{(\omega)}h_{ij'q'r'}^{(\omega)}]-\frac{1}{n^8}\sum_{ijqri'j'q'r'}\mathbf{E}[h_{ijqr}^{(\omega)}h_{i'j'q'r'}^{(\omega)}]\Big). \tag{57}$$

By adding further restrictions that i,i',j,q,r,j',q',r' are all different, we can obtain the corresponding expression for σ_ω^2. Hence the difference between them can only happen when at least one subscript in i',j,q,r,j',q',r' is equal to i. Combining $|h_{ijqr}^{(\omega)}|\leq 4$, we have $|\mathbf{E}_Z\widehat{\sigma}_\omega^2-\sigma_\omega^2|\sim\mathcal{O}(n^{-1})$.

ϵ-net arguments. Next, we prove the second part with ϵ-net arguments. Take the parameter space Θ_k of θ_k as an example. We choose a cover with $\mathcal{N}(\Theta_k,r_k)$ points $\{p_i\}_{i=1}^{\mathcal{N}(\Theta_k,r_k)}$ such that for any point $p\in\Theta_k$, we have $\min_i\|p-p_i\|\leq r_k$. According to [41, Proposition 4.2.12], by comparing the volumes, we have $\mathcal{N}(\Theta_k,r_k)\leq(4R_{\Theta_k}/r_k)^{d_k}$. As for the parameter space Θ_l of θ_l, we can also obtain a cover with $\mathcal{N}(\Theta_l,r_l)$ points that $\mathcal{N}(\Theta_k,r_k)\leq(4R_{\Theta_l}/r_l)^{d_l}$. Here, we set $r_k=4R_{\Theta_k}/\sqrt{n}, r_l=4R_{\Theta_l}/\sqrt{n}$, thus $\mathcal{N}(\Theta_k,r_k)\leq(\sqrt{n})^{d_k}$,$\mathcal{N}(\Theta_l,r_l)\leq(\sqrt{n})^{d_l}$. Then combining the Lipschitz property as shown in Lemma 4, we have with probability at least $1-\delta$,

$$\begin{aligned}\sup_{(\theta_k,\theta_l)\in\Theta_c}\big|\xi_\omega-\mathbf{E}_Z\xi_\omega\big|&\leq C_1\sqrt{\frac{1}{n}\log\frac{2\mathcal{N}(\Theta_k,r_k)\mathcal{N}(\Theta_l,r_l)}{\delta}}+8R_{\omega_k}L_k\cdot r_k+8R_{\omega_l}L_lr_l\\&\leq C_1\sqrt{\frac{1}{n}\log\frac{2}{\delta}+(d_k+d_l)\frac{\log n}{2n}}+\frac{32R_{\omega_k}L_kR_{\Theta_k}}{\sqrt{n}}+\frac{32R_{\omega_l}L_lR_{\Theta_l}}{\sqrt{n}}.\end{aligned}$$

Hence when n is large enough, there exists a positive constant C_3, with probability at least $1-\delta$,

$$\sup_{(\theta_k,\theta_l)\in\Theta_c}\big|\xi_\omega-\mathbf{E}_Z\xi_\omega\big|\leq C_3\left[\sqrt{\frac{1}{n}\log\frac{1}{\delta}+(d_k+d_l)\frac{\log n}{n}}+\frac{R_{\omega_k}L_kR_{\Theta_k}}{\sqrt{n}}+\frac{R_{\omega_l}L_lR_{\Theta_l}}{\sqrt{n}}\right]. \tag{58}$$

Similar, when n is large enough, there exists a positive constant C_4, with probability at least $1-\delta$,

$$\sup_{(\theta_k,\theta_l)\in\Theta_c}\big|\widehat{\sigma}_\omega^2-\sigma_\omega^2\big|\leq C_4\left[\sqrt{\frac{1}{n}\log\frac{1}{\delta}+(d_k+d_l)\frac{\log n}{n}}+\frac{R_{\omega_k}L_kR_{\Theta_k}}{\sqrt{n}}+\frac{R_{\omega_l}L_lR_{\Theta_l}}{\sqrt{n}}\right]. \tag{59}$$

Overall Bound. Now we combine the previously obtained results. We have

$$\left|\frac{\xi_\omega-r_\omega^{(n)}/n}{\widehat{\sigma}_\omega}-\frac{\mathbf{E}_Z\xi_\omega-r_\omega/n}{\sigma_\omega}\right|\leq\left|\frac{\xi_\omega-\mathbf{E}_Z\xi_\omega}{\widehat{\sigma}_\omega}\right|+\left|\frac{r_\omega^{(n)}-r_\omega}{n\widehat{\sigma}_\omega}\right|+|\mathbf{E}_Z\xi_\omega-r_\omega/n|\cdot\left|\frac{1}{\widehat{\sigma}_\omega}-\frac{1}{\sigma_\omega}\right|.$$

Since on Θ_c, $\sigma_\omega\geq c$, according to Eq. (59), we can make $\widehat{\sigma}_\omega\geq c/2$ happen by assigning probability budget $\delta/2$ when n is large enough. Also, combining Eq. (54), $|\mathbf{E}_Z\xi_\omega|\leq 1$ and r_ω is a constant related to α, then with n large enough, there exist positive constants C_5,C_6,

$$\left|\frac{\xi_\omega-r_\omega^{(n)}/n}{\widehat{\sigma}_\omega}-\frac{\mathbf{E}_Z\xi_\omega-r_\omega/n}{\sigma_\omega}\right|\leq\frac{2}{c}\left|\xi_\omega-\mathbf{E}_Z\xi_\omega\right|+\frac{C_5}{c}\sqrt{\frac{1}{n}\log\frac{1}{\alpha}}+\frac{C_6}{c^3}\left|\widehat{\sigma}_\omega^2-\sigma_\omega^2\right|. \tag{60}$$

Note that we need to pay for probability budget $\delta/2$ for the above conclusion to hold. Then by taking the supremum on both sides in Eq. (60) and assigning the remained probability budget $\delta/2$ to Eqs. (58) and (59), we can show that when n is large enough, with probability at least $1-\delta$,

$$\begin{aligned}&\sup_{(\theta_k,\theta_l)\in\Theta_c}\left|\frac{\xi_\omega-r_\omega^{(n)}/n}{\widehat{\sigma}_\omega}-\frac{\mathbf{E}_Z\xi_\omega-r_\omega/n}{\sigma_\omega}\right|\\&\sim\mathcal{O}\left(\frac{1}{c^3}\left[\sqrt{\frac{1}{n}\log\frac{1}{\delta}+(d_k+d_l)\frac{\log n}{n}}+\frac{R_{\omega_k}L_kR_{\Theta_k}+R_{\omega_l}L_lR_{\Theta_l}}{\sqrt{n}}\right]\right)\end{aligned} \tag{61}$$

thus complete the proof. □

H Proof of Theorem 3

In this section, we give a proof of the Theorem 3. We first restate the Theorem 3 here.

Theorem 3 (Consistency)**.** *Let θ_k^*, θ_l^* be the parameters after learning, Z^{te} be the testing samples of size m, when $\mathbf{E}_Z HSIC_\omega^{(u)}(Z^{te}) > 0$, then the probability of Type II error*

$$\mathbb{P}\big(\textit{Type II error}\big) = \mathbb{P}_{\mathcal{H}_1}\big(m HSIC_\omega(Z^{te}) \leq r_\omega^{(m)} | \theta_k^*, \theta_l^*\big) \sim \mathcal{O}(m^{-1/2}). \tag{62}$$

Let the mapping functions with learned parameters θ_k^*, θ_l^* be $\mathcal{T}_{\theta_k^*}, \mathcal{T}_{\theta_l^*}$, and the corresponding range space be a compact subset of $\mathbb{R}^{d_{\mathcal{T}_x}}, \mathbb{R}^{d_{\mathcal{T}_y}}$, respectively. Also, the diameter of two range spaces is denoted by $diam(\mathcal{T}_{\theta_k^*}), diam(\mathcal{T}_{\theta_l^*})$, respectively. Let $\{(\omega_{k;j}, \omega_{l;j})\}_{j=1}^{D/2}$ be the frequency samplings with their second moment denoted by $\sigma_{\omega_k}^2 := \mathbf{E}_{p_k(\omega)}[\omega_{k;j}^T \omega_{k;j}]$, $\sigma_{\omega_l}^2 := \mathbf{E}_{p_l(\omega)}[\omega_{l;j}^T \omega_{l;j}]$. Additionally, we denote $\xi_u := HSIC(X, Y)$, then under $\mathcal{H}_1$, we have $\mathbf{E}_Z HSIC_\omega^{(u)}(Z^{te}) > 0$ with any constant probability when $D = \Omega\Big(\frac{d_{\mathcal{T}_x} + d_{\mathcal{T}_y}}{\xi_u^2} \log \frac{\sigma_{\omega_k} diam(\mathcal{T}_{\theta_k^*}) + \sigma_{\omega_l} diam(\mathcal{T}_{\theta_l^*})}{\xi_u}\Big)$.

Proof. The proof consists of two parts, we first give the rate of convergence of Type II error under condition $\mathbf{E}_Z \mathrm{HSIC}_\omega^{(u)}(Z^{te}) > 0$, and next for condition $\mathbf{E}_Z \mathrm{HSIC}_\omega^{(u)}(Z^{te}) > 0$ we give a lower bound on the number of frequency samplings required for it to hold. To simplify, we denote the U-statistic $\mathrm{HSIC}_\omega^{(u)}(Z^{te})$ as $\xi_\omega^{(u)}$ in this proof. We begin to prove the first part. With the learned parameters θ_k^*, θ_l^*, the probability of the Type II error is given by

$$\mathbb{P}\Big(\text{Type II error}\Big) = \mathbb{P}_{\mathcal{H}_1}\Big(m\xi_\omega \leq r_\omega^{(m)} | \theta_k^*, \theta_l^*\Big). \tag{63}$$

Combing the result of the difference between ξ_ω and $\xi_\omega^{(u)}$ as shown in Eq. (34), we have

$$\mathbb{P}_{\mathcal{H}_1}\Big(m\xi_\omega \leq r_\omega^{(m)} | \theta_k^*, \theta_l^*\Big) \leq \mathbb{P}_{\mathcal{H}_1}\left(m\xi_\omega^{(u)} \leq r_\omega^{(m)} + C_0 \big| \theta_k^*, \theta_l^*\right), \tag{64}$$

where C_0 is a positive constant. To apply the rate of convergence of the Central Limit Theorem, we rewrite the right equation in Eq. (64) as

$$\mathbb{P}_{\mathcal{H}_1}\left(\frac{\sqrt{m}(\xi_\omega^{(u)} - \mathbf{E}_Z \xi_\omega^{(u)})}{4\sigma_\omega^{1/2}} \leq \frac{r_\omega^{(m)}/\sqrt{m} - \sqrt{m}\mathbf{E}_Z \xi_\omega^{(u)} + C_0/\sqrt{m}}{4\sigma_\omega^{1/2}} \Big| \theta_k^*, \theta_l^*\right), \tag{65}$$

where the standard deviation (defined in Proposition 1) $\sigma_\omega > 0$ under $\mathcal{H}_1$. Then according to the results in [34, Section 5.5.1 Theorem B], there exist nonnegative constant C_1 such that

$$\mathbb{P}\Big(\text{Type II error}\Big) \leq \Phi\left(\frac{r_\omega^{(m)}/\sqrt{m} - \sqrt{m}\mathbf{E}_Z \xi_\omega^{(u)} + C_0/\sqrt{m}}{4\sigma_\omega^{1/2}}\right) + \frac{C_1 \nu_h}{\sigma_\omega^{3/2}} \frac{1}{\sqrt{m}} \tag{66}$$

where $\nu_h := \mathbf{E}_Z^{i \neq j \neq q \neq r} |h_{ijqr}^{(\omega)}|^3 < \infty$. When m is large enough, we further have

$$\mathbb{P}\Big(\text{Type II error}\Big) \leq \Phi\Big(C_2 - C_3\sqrt{m}\mathbf{E}_Z \xi_\omega^{(u)} + C_4/\sqrt{m}\Big) + C_5 \frac{1}{\sqrt{m}}. \tag{67}$$

where C_2, C_3, C_4 are positive constants and using $r^{(m)} \sim \mathcal{O}(m^{1/2})$ we prove in Eq. (54). Hence when $\mathbf{E}_Z \xi_\omega^{(u)} > 0$, the leading term $\sqrt{m}\mathbf{E}_Z \xi_\omega^{(u)}$ decrease as m increase. Further, to obtain the decrease rate when m is close to infinity, we consider the asymptotic expansion (when x is close to negative infinity) for the function $\Phi(x)$ as given by

$$\Phi(x) = -\frac{e^{-x^2}}{2x\sqrt{\pi}}\left(1 + \sum_{n=1}^{\infty}(-1)^n \frac{1 \cdot 3 \cdot 5 \cdots (2n-1)}{(2x^2)^n}\right), \tag{68}$$

thus $\Phi\Big(C_2 - C_3\sqrt{m}\mathbf{E}_Z \xi_\omega^{(u)} + C_4/\sqrt{m}\Big) \sim \mathcal{O}(m^{-1/2})$. As a result, the decreasing rate is at least $\mathcal{O}(m^{-1/2})$. We have so far completed the first part of the proof, and we next begin the second part of the proof, i.e. obtain the number of frequency samplings required for the condition $\mathbf{E}_Z \xi_\omega^{(u)} > 0$ to

hold. For simplify, we denote $\Lambda_k(x)^T\Lambda_k(x'), \Lambda_l(y)^T\Lambda_l(y')$ as $k^{(\omega)}(x,x'), l^{(\omega)}(y,y')$, respectively. Then according to Lemma 2, we have

$$\begin{aligned}
\mathbb{P}\left[\sup_{x,x'\in\mathcal{X}}|k^{(\omega)}(x,x')-k(x,x')|\geq\epsilon\right] &\leq 2^8\left(\frac{\sigma_{\omega_k}\mathrm{diam}(\mathcal{T}_{\theta_k^*})}{\epsilon}\right)^2\exp\left(-\frac{D\epsilon^2}{4(d_{\mathcal{T}_x}+2)}\right),\\
\mathbb{P}\left[\sup_{y,y'\in\mathcal{Y}}|l^{(\omega)}(y,y')-l(y,y')|\geq\epsilon\right] &\leq 2^8\left(\frac{\sigma_{\omega_l}\mathrm{diam}(\mathcal{T}_{\theta_l^*})}{\epsilon}\right)^2\exp\left(-\frac{D\epsilon^2}{4(d_{\mathcal{T}_y}+2)}\right).
\end{aligned} \tag{69}$$

Also, we denote the bounds in Eq. (69) as $\delta_x(\epsilon,D), \delta_y(\epsilon,D)$, respectively. Next we get the bound between $\mathbf{E}_Z\xi_\omega^{(u)}$ and $\mathrm{HSIC}(X,Y)$. According to Lemma 3, the bound is given by

$$|\mathbf{E}_Z\xi_\omega^{(u)}-\mathrm{HSIC}(X,Y)|\leq 4\cdot\sup_{x,x'\in\mathcal{X},y,y'\in\mathcal{Y}}|k^{(\omega)}(x,x')l^{(\omega)}(y,y')-k(x,x')l(y,y')|. \tag{70}$$

Since by the definition, for all $(x,x')\in\mathcal{X}\times\mathcal{X}, (y,y')\in\mathcal{Y}\times\mathcal{Y}$, we have $|k^{(\omega)}(x,x')|\leq 1$, $|l^{(\omega)}(y,y')|\leq 1$, $|k(x,x')|\leq 1$, $|l(y,y')|\leq 1$. Hence we have

$$|k^{(\omega)}(x,x')l^{(\omega)}(y,y')-k(x,x')l(y,y')|\leq|k^{(\omega)}(x,x')-k(x,x')|+|l^{(\omega)}(y,y')-l(y,y')|. \tag{71}$$

Combining the results of Eqs. (70) and (71), we obtain

$$|\mathbf{E}_Z\xi_\omega^{(u)}-\mathrm{HSIC}(X,Y)|\leq 4\cdot\sup_{x,x'\in\mathcal{X}}|k^{(\omega)}(x,x')-k(x,x')|+4\cdot\sup_{y,y'\in\mathcal{Y}}|l^{(\omega)}(y,y')-l(y,y')|. \tag{72}$$

Combining the results as shown in Eq. (69) and allocating the probability budget ϵ, we have

$$\mathbb{P}\left[\sup_{x,x'\in\mathcal{X},y,y'\in\mathcal{Y}}|\mathbf{E}_Z\xi_\omega^{(u)}-\mathrm{HSIC}(X,Y)|\geq\epsilon\right]\leq\delta_x(\epsilon/8,D)+\delta_y(\epsilon/8,D). \tag{73}$$

By setting $\epsilon=\xi_u/2$, and since $\xi_u>0$ under $\mathcal{H}_1$, we conclude that $\mathbf{E}_Z\xi_\omega^{(u)}>0$ holds with any constant probability when $D=\Omega\Big(\frac{d_{\mathcal{T}_x}+d_{\mathcal{T}_y}}{\xi_u^2}\log\frac{\sigma_{\omega_k}\mathrm{diam}(\mathcal{T}_{\theta_k^*})+\sigma_{\omega_l}\mathrm{diam}(\mathcal{T}_{\theta_l^*})}{\xi_u}\Big)$. □

In the proof of Theorem 3, we obtain the convergence bound of the statistic $\mathbf{E}_Z\xi_\omega^{(u)}$. Actually, the convergence result for its estimation can also be obtained, as shown in the following Corollary.

Corollary 1 (Approximation Error Bound of $\mathrm{HSIC}_\omega(Z^{te})$)**.** *Maintaining the same conditions and notions as in Theorem 3, we have the uniform convergence bound of $\mathit{HSIC}_\omega(Z^{te})$ as*

$$\mathbb{P}\left[\sup_{Z^{te}\in\mathcal{X}\times\mathcal{Y}}|\mathit{HSIC}_\omega(Z^{te})-\mathit{HSIC}_b(Z^{te})|\geq\epsilon\right]\leq\delta_x(\epsilon/8,D)+\delta_y(\epsilon/8,D). \tag{74}$$

Proof. By the definition, we have for all i,j,q,r, $|k_{ij}^{(\omega)}|\leq 1, |l_{ij}^{(\omega)}|\leq 1, |k_{ij}|\leq 1, |l_{ij}|\leq 1$, thus

$$|k_{ij}^{(\omega)}l_{qr}^{(\omega)}-k_{ij}l_{qr}|\leq|k_{ij}^{(\omega)}-k_{ij}||l_{qr}^{(\omega)}|+|k_{ij}||l_{qr}^{(\omega)}-l_{qr}|\leq|k_{ij}^{(\omega)}-k_{ij}|+|l_{qr}^{(\omega)}-l_{qr}|. \tag{75}$$

Then according to the results as shown in Eq. (69), we have for all i,j,q,r

$$\mathbb{P}\left[\sup_{x_i,x_j\in\mathcal{X},y_q,y_r\in\mathcal{Y}}|k_{ij}^{(\omega)}l_{qr}^{(\omega)}-k_{ij}l_{qr}|\geq\epsilon\right]\leq\delta_x(\epsilon/2,D)+\delta_y(\epsilon/2,D). \tag{76}$$

Recall the definition of $h_{ijqr}^{(\omega)}=\frac{1}{4!}\sum_{(t,u,v,w)}^{(i,j,q,r)}k_{tu}^{(\omega)}l_{tu}^{(\omega)}+k_{tu}^{(\omega)}l_{vw}^{(\omega)}-2k_{uv}^{(\omega)}l_{tv}^{(\omega)}$ and we further define the corresponding $h_{ijqr}=\frac{1}{4!}\sum_{(t,u,v,w)}^{(i,j,q,r)}k_{tu}l_{tu}+k_{tu}l_{vw}-2k_{uv}l_{tv}$, then for all i,j,q,r,

$$\mathbb{P}\left[\sup_{x_i,x_j\in\mathcal{X},y_q,y_r\in\mathcal{Y}}|h_{ijqr}^{(\omega)}-h_{ijqr}|\geq\epsilon\right]\leq\delta_x(\epsilon/8,D)+\delta_y(\epsilon/8,D). \tag{77}$$

After that, using the expressions that we obtained before, i.e., $\mathrm{HSIC}_\omega(Z^{te}):=\frac{1}{n^4}\sum_{i,j,q,r}h_{ijqr}^{(\omega)}$ and $\mathrm{HSIC}_b(Z^{te}):=\frac{1}{n^4}\sum_{i,j,q,r}h_{ijqr}$, we obtain the final bound that

$$\mathbb{P}\left[\sup_{Z^{te}\in\mathcal{X}\times\mathcal{Y}}|\mathrm{HSIC}_\omega(Z^{te})-\mathrm{HSIC}_b(Z^{te})|\geq\epsilon\right]\leq\delta_x(\epsilon/8,D)+\delta_y(\epsilon/8,D). \tag{78}$$

and thus complete the proof. □

I Smoothness of Optimization Objective

We first prove the Lipschitz property for some functions. For ease of reference, we re-list here the definitions of the terms that related to the optimization objective: $\xi_\omega := \text{HSIC}_\omega(Z) = \frac{1}{n^4}\sum_{i,j,q,r} h^{(\omega)}_{ijqr}$, $\widehat{\sigma}^2_\omega := 16\Big[\frac{1}{n}\sum_i(\frac{1}{n^3}\sum_{j,q,r} h^{(\omega)}_{ijqr})^2 - \text{HSIC}^2_\omega(Z)\Big]$ and $\sigma^2_\omega := 16\Big[\mathbf{E}_i(\mathbf{E}_{j,q,r}h^{(\omega)}_{ijqr})^2 - \big(\mathbf{E}_Z h^{(\omega)}_{ijqr}\big)^2\Big]$. The Lipschitz property of these terms are shown as follows.

Lemma 4 (Lipschitz Property of of $\xi_\omega, \mathbf{E}_Z\xi_\omega, \widehat{\sigma}^2_\omega, \sigma^2_\omega$)**.** *Maintaining the same conditions and notions as in Lemma 1, we have the following Lipschitz property*

$$\begin{aligned}
&|\xi_\omega(\theta_k,\theta_l) - \xi_\omega(\theta'_k,\theta'_l)| \le 8R_{\omega_k}L_k\cdot\|\theta_k-\theta'_k\| + 8R_{\omega_l}L_l\cdot\|\theta_l-\theta'_l\|,\\
&|\mathbf{E}_Z[\xi_\omega(\theta_k,\theta_l)] - \mathbf{E}_Z[\xi_\omega(\theta'_k,\theta'_l)]| \le 8R_{\omega_k}L_k\cdot\|\theta_k-\theta'_k\| + 8R_{\omega_l}L_l\cdot\|\theta_l-\theta'_l\|,\\
&|\widehat{\sigma}^2_\omega(\theta_k,\theta_l) - \widehat{\sigma}^2_\omega(\theta'_k,\theta'_l)| \le 1024R_{\omega_k}L_k\cdot\|\theta_k-\theta'_k\| + 1024R_{\omega_l}L_l\cdot\|\theta_l-\theta'_l\|,\\
&|\sigma^2_\omega(\theta_k,\theta_l) - \sigma^2_\omega(\theta'_k,\theta'_l)| \le 1024R_{\omega_k}L_k\cdot\|\theta_k-\theta'_k\| + 1024R_{\omega_l}L_l\cdot\|\theta_l-\theta'_l\|,
\end{aligned} \tag{79}$$

where we use the symbol $\xi_\omega(\theta_k,\theta_l)$ to denote ξ_ω with the parameter θ_k,θ_l and the others by analogy.

Proof. We start by obtaining the result of $h^{(\omega)}_{ijqr}$ for all i,j,q,r. Since for all i,j,q,r,

$$\big|k^{(\omega)}_{ij}(\theta_k)l^{(\omega)}_{qr}(\theta_l) - k^{(\omega)}_{ij}(\theta'_k)l^{(\omega)}_{qr}(\theta'_l)\big| \le \big|k^{(\omega)}_{ij}(\theta_k) - k^{(\omega)}_{ij}(\theta'_k)\big| + \big|l^{(\omega)}_{qr}(\theta_l) - l^{(\omega)}_{qr}(\theta'_l)\big|, \tag{80}$$

where the property $|k^{(\omega)}_{ij}| \le 1$ and $|l^{(\omega)}_{qr}| \le 1$ are used. According the Lemma 1, we have

$$\big|k^{(\omega)}_{ij}(\theta_k) - k^{(\omega)}_{ij}(\theta'_k)\big| \le 2R_{\omega_k}L_k\cdot\|\theta_k-\theta'_k\|,\ \big|l^{(\omega)}_{qr}(\theta_l) - l^{(\omega)}_{qr}(\theta'_l)\big| \le 2R_{\omega_l}L_l\cdot\|\theta_l-\theta'_l\|. \tag{81}$$

Combing the definition that $h^{(\omega)}_{ijqr} := \frac{1}{4!}\sum^{(i,j,q,r)}_{(t,u,v,w)} k^{(\omega)}_{tu}l^{(\omega)}_{tu} + k^{(\omega)}_{tu}l^{(\omega)}_{vw} - 2k^{(\omega)}_{uv}l^{(\omega)}_{tv}$, then

$$\big|h^{(\omega)}_{ijqr}(\theta_k,\theta_l) - h^{(\omega)}_{ijqr}(\theta'_k,\theta'_l)\big| \le 8R_{\omega_k}L_k\cdot\|\theta_k-\theta'_k\| + 8R_{\omega_l}L_l\cdot\|\theta_l-\theta'_l\|. \tag{82}$$

Also, combining the definition of $\xi_\omega := \text{HSIC}_\omega(Z) = \frac{1}{n^4}\sum_{i,j,q,r} h^{(\omega)}_{ijqr}$, we obtain

$$|\xi_\omega(\theta_k,\theta_l) - \xi_\omega(\theta'_k,\theta'_l)| \le 8R_{\omega_k}L_k\cdot\|\theta_k-\theta'_k\| + 8R_{\omega_l}L_l\cdot\|\theta_l-\theta'_l\|. \tag{83}$$

By using $|\mathbf{E}_Z[\xi_\omega(\theta_k,\theta_l)] - \mathbf{E}_Z[\xi_\omega(\theta'_k,\theta'_l)]| \le \mathbf{E}_Z[|\xi_\omega(\theta_k,\theta_l) - \xi_\omega(\theta'_k,\theta'_l)|]$, we have

$$|\mathbf{E}_Z[\xi_\omega(\theta_k,\theta_l)] - \mathbf{E}_Z[\xi_\omega(\theta'_k,\theta'_l)]| \le 8R_{\omega_k}L_k\cdot\|\theta_k-\theta'_k\| + 8R_{\omega_l}L_l\cdot\|\theta_l-\theta'_l\|. \tag{84}$$

For the results of $\widehat{\sigma}^2_\omega, \sigma^2_\omega$, we first proof the following results. For all i,j,q,r,i',j',q',r', we have

$$\begin{aligned}
&\big|h^{(\omega)}_{ijqr}(\theta_k,\theta_l)h^{(\omega)}_{i'j'q'r'}(\theta_k,\theta_l) - h^{(\omega)}_{ijqr}(\theta'_k,\theta'_l)h^{(\omega)}_{i'j'q'r'}(\theta'_k,\theta'_l)\big|\\
\le\ & 4\cdot\big|h^{(\omega)}_{ijqr}(\theta_k,\theta_l) - h^{(\omega)}_{ijqr}(\theta'_k,\theta'_l)\big| + 4\cdot|h^{(\omega)}_{i'j'q'r'}(\theta_k,\theta_l) - h^{(\omega)}_{i'j'q'r'}(\theta'_k,\theta'_l)\big|\\
\le\ & 64R_{\omega_k}L_k\cdot\|\theta_k-\theta'_k\| + 64R_{\omega_l}L_l\cdot\|\theta_l-\theta'_l\|,
\end{aligned} \tag{85}$$

where the first inequality holds due to property $|h^{(\omega)}_{ijqr}| \le 4$. Then we use the expression

$$\widehat{\sigma}^2_\omega = 16\Big[\frac{1}{n^7}\sum_{i,j,q,r,j',q',r'} h^{(\omega)}_{ijqr}h^{(\omega)}_{ij'q'r'} - \frac{1}{n^8}\sum_{i,j,q,r,i',j',q',r'} h^{(\omega)}_{ijqr}h^{(\omega)}_{i'j'q'r'}\Big] \tag{86}$$

and combine the results in Eq. (85). As a result, we obtain that

$$|\widehat{\sigma}^2_\omega(\theta_k,\theta_l) - \widehat{\sigma}^2_\omega(\theta'_k,\theta'_l)| \le 1024R_{\omega_k}L_k\cdot\|\theta_k-\theta'_k\| + 1024R_{\omega_l}L_l\cdot\|\theta_l-\theta'_l\|. \tag{87}$$

In a similar way, we can obtain the corresponding expression of σ^2_ω as

$$\sigma^2_\omega = 16\Big[\mathbf{E}_{i,j,q,r,j',q',r'}h^{(\omega)}_{ijqr}h^{(\omega)}_{ij'q'r'} - \mathbf{E}_{i,j,q,r,i',j',q',r'}h^{(\omega)}_{ijqr}h^{(\omega)}_{i'j'q'r'}\Big]. \tag{88}$$

Then we obtain a similar result as before, i.e.,

$$|\sigma^2_\omega(\theta_k,\theta_l) - \sigma^2_\omega(\theta'_k,\theta'_l)| \le 1024R_{\omega_k}L_k\cdot\|\theta_k-\theta'_k\| + 1024R_{\omega_l}L_l\cdot\|\theta_l-\theta'_l\| \tag{89}$$

which completes the proof. □

Also for the term associated with the estimated threshold (recall that it is computed from the first two moments), we obtain the following properties of $\mathcal{E}_0, \mathcal{V}_0$ as defined in Theorem 1.

Lemma 5 (Lipschitz Property of of $\mathcal{E}_0, \mathcal{V}_0$)**.** *Maintaining the same conditions and notions as in Lemma 4, we have the following Lipschitz property*

$$\begin{aligned}|\mathcal{E}_0(\theta_k,\theta_l)-\mathcal{E}_0(\theta'_k,\theta'_l)| &\le 2C_0R_{\omega_k}L_k\cdot\|\theta_k-\theta'_k\|+2C_0R_{\omega_l}L_l\cdot\|\theta_l-\theta'_l\|,\\ |\mathcal{V}_0(\theta_k,\theta_l)-\mathcal{V}_0(\theta'_k,\theta'_l)| &\le 128C_1R_{\omega_k}L_k\cdot\|\theta_k-\theta'_k\|+128C_1R_{\omega_l}L_l\cdot\|\theta_l-\theta'_l\|,\end{aligned} \tag{90}$$

where the constant $C_0(n) := \frac{n^2}{(n-1)^2}$ *and* $C_1(n) := \frac{n(n-4)(n-5)}{(n-1)(n-2)(n-3)}$.

Proof. The expression in Theorem 1, while easy to compute, is not suitable for this part of our proof. We begin by obtaining equivalent expressions for $\mathcal{E}_0, \mathcal{V}_0$. According to Eq. (39), we have

$$\begin{aligned}\mathcal{E}_0 &= \left[1-\frac{\mathbf{1}^T(\mathbf{\Lambda}_X\mathbf{\Lambda}_X^T-\mathbf{I}_n)\mathbf{1}}{n(n-1)}\right]\left[1-\frac{\mathbf{1}^T(\mathbf{\Lambda}_Y\mathbf{\Lambda}_Y^T-\mathbf{I}_n)\mathbf{1}}{n(n-1)}\right]\\ &= C_0\left[1-\frac{\mathbf{1}^T(\mathbf{\Lambda}_X\mathbf{\Lambda}_X^T)\mathbf{1}}{n^2}\right]\left[1-\frac{\mathbf{1}^T(\mathbf{\Lambda}_Y\mathbf{\Lambda}_Y^T)\mathbf{1}}{n^2}\right].\end{aligned} \tag{91}$$

And for $\mathcal{V}_0$, we use Eq. (45) and obtain

$$\mathcal{V}_0 = \frac{2C_1}{n^4}\big[\mathrm{Tr}(\mathbf{\Lambda}_X\mathbf{\Lambda}_X^T\mathbf{H}\mathbf{\Lambda}_X\mathbf{\Lambda}_X^T\mathbf{H})\big]\big[\mathrm{Tr}(\mathbf{\Lambda}_Y\mathbf{\Lambda}_Y^T\mathbf{H}\mathbf{\Lambda}_Y\mathbf{\Lambda}_Y^T\mathbf{H})\big]. \tag{92}$$

Also, we define $h^{(k)}_{ijqr} := \frac{1}{4!}\sum^{(i,j,q,r)}_{(t,u,v,w)} k^{(\omega)}_{tu}k^{(\omega)}_{tu} + k^{(\omega)}_{tu}k^{(\omega)}_{vw} - 2k^{(\omega)}_{uv}k^{(\omega)}_{tv}$ that corresponding to $h^{(\omega)}_{ijqr}$ and also define $h^{(l)}_{ijqr} := \frac{1}{4!}\sum^{(i,j,q,r)}_{(t,u,v,w)} l^{(\omega)}_{tu}l^{(\omega)}_{tu} + l^{(\omega)}_{tu}l^{(\omega)}_{vw} - 2l^{(\omega)}_{uv}l^{(\omega)}_{tv}$. Then we can further rewrite the term for X as $\frac{1}{n^2}\mathrm{Tr}(\mathbf{\Lambda}_X\mathbf{\Lambda}_X^T\mathbf{H}\mathbf{\Lambda}_X\mathbf{\Lambda}_X^T\mathbf{H}) = \frac{1}{n^4}\sum_{i,j,q,r} h^{(k)}_{ijqr}$ and for Y by analogy. Then the properties of $h^{(\omega)}_{ijqr}$ can also be obtained for $h^{(k)}_{ijqr}$ and $h^{(l)}_{ijqr}$, e.g., $|h^{(k)}_{ijqr}| \le 4, |h^{(l)}_{ijqr}| \le 4$ for all i,j,q,r. Next we start to prove the Lipschitz property of $\mathcal{E}_0, \mathcal{V}_0$. For $\mathcal{E}_0$, according to Eq. (81) and combining the results $0 \le \mathbf{1}^T(\mathbf{\Lambda}_X\mathbf{\Lambda}_X^T)\mathbf{1} \le n^2$ and that for Y, we can show that

$$\begin{aligned}|\mathcal{E}_0(\theta_k,\theta_l)-\mathcal{E}_0(\theta'_k,\theta'_l)| &\le \frac{C_0}{n^2}\sum_{i,j}\big|k^{(\omega)}_{ij}(\theta_k)-k^{(\omega)}_{ij}(\theta'_k)\big|+\frac{C_0}{n^2}\sum_{q,r}\big|l^{(\omega)}_{qr}(\theta_l)-l^{(\omega)}_{qr}(\theta'_l)\big|\\ &\le 2C_0R_{\omega_k}L_k\cdot\|\theta_k-\theta'_k\|+2C_0R_{\omega_l}L_l\cdot\|\theta_l-\theta'_l\|,\end{aligned} \tag{93}$$

where $\mathbf{1}^T(\mathbf{\Lambda}_X\mathbf{\Lambda}_X^T)\mathbf{1} = \sum_{i,j} k^{(\omega)}_{ij}$ by definition and so as for Y. And for $\mathcal{V}_0$, we can prove that

$$\begin{aligned}&|\mathcal{V}_0(\theta_k,\theta_l)-\mathcal{V}_0(\theta'_k,\theta'_l)|\\ \le\ &\frac{2C_1}{n^8}\sum_{i,j,q,r,i',j',q',r'}\big|h^{(k)}_{ijqr}(\theta_k,\theta_l)h^{(l)}_{i'j'q'r'}(\theta_k,\theta_l)-h^{(k)}_{ijqr}(\theta'_k,\theta'_l)h^{(l)}_{i'j'q'r'}(\theta'_k,\theta'_l)\big|\\ \le\ &128C_1R_{\omega_k}L_k\cdot\|\theta_k-\theta'_k\|+128C_1R_{\omega_l}L_l\cdot\|\theta_l-\theta'_l\|\end{aligned} \tag{94}$$

where the last inequation is obtained similar to Eq. (85), thus completes the proof. □

The following results extend the results in [30] to the more general case (we only restrict the mapping functions to satisfy the Lipschitz property, and thus include the Gaussian kernel case of their proof).

Theorem 4 (Smoothness of Optimization Objective)**.** *Let the sample of size* n *be* Z*, and with a small positive constant* c*, let the set of the parameters be* $\overline{\Theta}_c := \{(\theta_k,\theta_l)|\widehat{\sigma}_\omega \ge c, \mathcal{V}_0 \ge c, \mathcal{E}_0 \ge c\}$*, then there exist a nonnegative constant* L *such that* $\|\nabla_{(\theta_k,\theta_l)}J\| \le L$ *on* $\overline{\Theta}_c$*, where the optimization objective* $J := [\mathit{HSIC}_\omega(Z) - \widehat{c_\alpha}/n]/\widehat{\sigma}_\omega$ *is that we used in practice.*

Proof. According to Lemma 4, we have shown that $\mathrm{HSIC}_\omega(Z), \widehat{\sigma}_\omega$ both fits the Lipschitz condition. Also, according to Lemma 5, we have shown that $\mathcal{E}_0, \mathcal{V}_0$ are also Lipschitz with respect to θ_k, θ_l. Since the threshold $\widehat{c_\alpha}$ is completely determined by these two moments, combining the smoothness property of the mapping from two moments to thresholds as analyzed in [30, Theorem 2], we obtain that $\widehat{c_\alpha}$ is also Lipschitz with respect to θ_k, θ_l on $\overline{\Theta}_c$. As a result, we complete the entire proof based on the Lipschitz property of composite mappings. □

Remark. $\mathcal{E}_0$ and $\mathcal{V}_0$ are positive is almost satisfied in practice since according to the definition only $[\mathbf{1}^T\mathbf{\Lambda}_{Xc}^{\cdot 2}\mathbf{1}][\mathbf{1}^T\mathbf{\Lambda}_{Yc}^{\cdot 2}\mathbf{1}]$ and $[\mathbf{1}^T(\mathbf{\Lambda}_{Xc}^T\mathbf{\Lambda}_{Xc})^{\cdot 2}\mathbf{1}][\mathbf{1}^T(\mathbf{\Lambda}_{Yc}^T\mathbf{\Lambda}_{Yc})^{\cdot 2}\mathbf{1}]$ need to be greater than 0.

J Details of Experiment Setup

In this section, we give an introduction to the comparison methods in our experiments and provide the implementation details of each method.

J.1 Details of Comparison Methods

The methods of comparison used in the experiment are described below.

- dCor [39]: An independence test that is based on the distance covariance.
- QHSIC [14]: The original quadratic-time HSIC independence test.
- RDC [26]: The randomized dependence coefficient that measures the independence using the canonical correlation between a finite set of random features of the copula.
- NyHSIC [44]: A variant of HSIC that uses the Nyström method to approximate kernels.
- FHSIC [44]: A variant of HSIC that uses the random Fourier feature to approximate kernels.
- BHSIC [44]: A variant of HSIC with the block-based statistic.
- HSICAgg [32]: An aggregated kernel test with the incomplete statistic of HSIC.
- NFSIC [18]: A test uses the normalized version of the finite set independence criterion and chooses features on a hold-out validation set to optimize a lower bound on the test power.

Below are the **GitHub URLs** for each comparison method.

- dCor: `https://pypi.org/project/dcor`.
- QHSIC: `https://github.com/amber0309/HSIC/blob/master/HSIC.py`.
- RDC: `https://github.com/lopezpaz/randomized_dependence_coefficient`.
- NyHSIC: `https://github.com/oxcsml/kerpy/blob/master/independence_testing`.
- FHSIC: `https://github.com/oxcsml/kerpy/blob/master/independence_testing`.
- BHSIC: `https://github.com/oxcsml/kerpy/blob/master/independence_testing`.
- HSICAgg: `https://github.com/antoninschrab/mmdagg/tree/master/mmdagg`.
- NFSIC: `https://github.com/wittawatj/fsic-test/blob/master/fsic`.

Time Complexity. Among them, dCor and QHSIC are the tests of quadratic complexity with sample size n, i.e., $\mathcal{O}(n^2)$. RDC is calculated in $\mathcal{O}(n \log n)$ and the rest are linear-time tests, i.e., $\mathcal{O}(n)$.

Threshold. For dCor, QHSIC, RDC, NyHSIC, FHSIC, and BHSIC, 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 LFHSIC-G/M by gamma approximation.

Details of Setup. The number of random features for FHSIC, LFHSIC-G/M, the number of induced variables for NyHSIC, the block size for BHSIC and the number of sub-diagonals R for HSICAgg are all kept consistent as recommended in [44] for fair evaluation. Specifically, we set the number of random mappings in RDC to 20 to ensure compatibility with large-scale datasets. The test location parameter J of NFHSIC is set as default as 10, since it differs from other approximation methods that blindly increasing J may lead to a loss of power as shown in [18] and can significantly escalate time costs due to its cubic time complexity $\mathcal{O}(J^3)$. 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 LFHSIC-G/M, to be fair, we perform the same grid search on SD, Sin, and GSign datasets. In other experiments, we still use the median bandwidth as initialization for LFHSIC-G/M. Also, the maximum number of iterations for the optimization is set to 100 for NFSIC and LFHSIC-G/M. The default learning rate of the optimization of LFHSIC-G/M is set as 0.05 in all the experiments. As for HSICAgg, the default implementation of the predefined 25 pairs of bandwidths in its code is used. For synthetic data, we set the split ratio to 0.5 for NFSIC and LFHSIC-G/M, 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 MSD 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.

J.2 Details of Datasets

The details of the four synthetic datasets and two real datasets are described below.

- **Sine Dependency (SD):** In this model, X follows a d-dimension multivariate normal distribution $\mathcal{N}_d(0, I_d)$, and Y is defined as $20\sin\big(4\pi(X_1^2 + X_2^2)\big)+Z$, where X_i is the i-th dimension of X, and $Z \sim \mathcal{N}(0,1)$ represents independent noise. Notably, when $d > 2$, Y exhibits a nonlinear relationship solely with the first two dimensions of X.
- **Sinusoid (Sin):** This model introduces a localized alteration in the probability density function p_{xy} over $\mathcal{X} \times \mathcal{Y} := [-\pi, \pi]^2$, specified as $(X, Y) \sim p_{xy}(x, y) \propto 1 + \sin(\omega x)\sin(\omega y)$, where ω denotes the frequency. Increasing the frequency enhances the similarity between the sampled data and that drawn from Uniform($[-\pi, \pi]^2$), thereby augmenting the challenge of detecting dependency with limited sample sizes. An example visualization is shown on the left of Fig. 6.
- **Gaussian Sign (GSign):** In this model, X follows a d-dimension multivariate normal distribution $\mathcal{N}_d(0, I_d)$, and Y is expressed as $|Z|\prod_{i=1}^{d}\text{sgn}(X_i)$, where $\text{sgn}(\cdot)$ represents the sign function, X_i denotes the i-th dimension of X, and $Z \sim \mathcal{N}(0,1)$ is independent of X. The challenge lies in Y being independent of any proper subset of X but dependent on X as a whole, underscoring the importance of considering all dimensions of X simultaneously in independence testing.
- **ISA Dataset.** We construct the data through the following steps: First, we generate n i.i.d samples of two univariate random variables with a mixture of Gaussian model, i.e. $\frac{1}{2}\mathcal{N}(-1, 0.01) + \frac{1}{2}\mathcal{N}(1, 0.01)$. Second, we mix these random variables using a rotation matrix parameterized by an angle θ, which varies from 0 to $\pi/4$. A zero angle implies independence between the data, while a larger angle signifies stronger dependency. Third, we append noise with a distribution of $\mathcal{N}_{d-1}(0, I_{d-1})$ to each of the mixtures. Finally, we multiply 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 [14]. For the case $d = 1, \theta = \pi/10$, an example visualization is shown in the middle of Fig. 6.
- **3DShapes Dataset.** This dataset [5] comprises images depicting 3D scenes, complete with additional features like shadows and backgrounds. It encompasses six fundamental latent factors: floor hue, wall hue, object hue, object scale, object shape, and orientation, all adjustable to generate corresponding images. Orientation is treated as a dependency factor for independence testing, where we test the dependency between the image X and its orientation Y. To heighten the challenge, we maintain the object shape as a ball, minimizing the apparent orientation feature compared to other shapes like squares, while randomizing the remaining factors. An example visualization is given on the right side of Fig. 6.
- **Million Song Dataset.** The dataset, a subset of the Million Song Data[3] [4], comprises $515,345$ songs with 91-dimensional features. The first dimension represents the release year of each song, designated as variable Y, while the remaining features (e.g., timbre average and timbre covariance) form variable X. Our objective is to identify the dependency between X and Y.

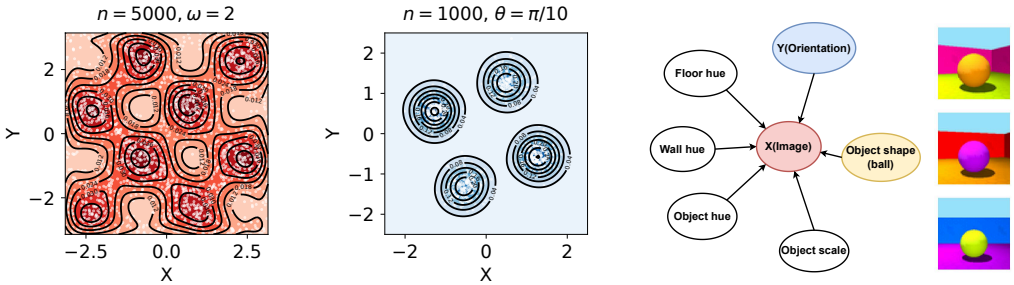

Figure 6: Examples of visualization of samples from different datasets. The two plots on the left correspond to the samples and their contour under Sin dataset ($n = 5000, w = 2$), and ISA dataset ($n = 1000, d = 1, \theta = \pi/10$), respectively. Right: a visualization of the causal diagram of the data generation process and some generated examples.

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

K Additional Experiment Results

In this section, we provide additional experimental results, mainly including the visualization results on the Sin synthetic dataset, the results with more comparing methods (as explained in the main paper, due to the high time overhead therefore do not participate in the evaluation of the main paper) as well as the running time of each method.

K.1 The visualization results on the Sin model

We provide the visualization results on the Sin synthetic dataset to illustrate the performance of our optimization objective for the example mentioned in the main paper. The results are shown in Fig. 7, where the setup follows our experiments ($n = 2000, \omega = 5, D = 100$). For visualization, the negative of our optimization objective J is shown. As can be seen, our optimization objective guides to letting the bandwidth adapt to improve the test power, and here we can see that regions with bandwidths around 0.2 (corresponding to the theoretical optimal solution in our main paper) indicate better power, thus corroborating the validity of our optimization objective. Also, notice that the landscape is smooth over a wide range as demonstrated in Theorem 4, and thus contributes to non-convex optimization.

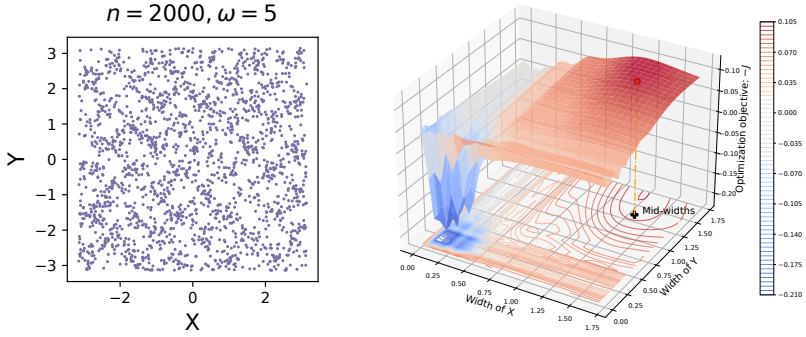

Figure 7: The visualization results on the Sin model. Left: the samples with $n = 2000, \omega = 5$. The visualization of the landscape of the negative of our optimization objective ($D = 100$).

K.2 Additional experimental comparisons

As mentioned in the main paper, the method[4] [30] is not involved in the comparison because it takes too much time to run in some settings. Here, we compare it with ours under some feasible settings to illustrate the improvement of our method on power-runtime trade-offs. The methods using Gaussian kernel with global width and Gaussian kernel with widths of each dimension (corresponding with ours) are employed, referred to QHSIC-O and QHSIC-W, respectively. For fairness, the same grid search procedure is employed as the initialization of the optimization. We perform the evaluation on the SD data and plot the results of the test power over time as shown in Fig. 8. Also, for our methods, we provide the results under the setting $D = 100$ and $D = 500$ as in the main paper. The experiments are conducted with the same equipment, specifically a 6-core CPU with a 3080 GPU.

Results. Our test consistently results in a better power-runtime tradeoff at different D settings. At $D = 100$, one test can be completed in less than a second when the sample size $n = 6000$. As D increases ($D = 500$), the number of samples required to achieve the same power decreases, but the increase in D leads to an overhead in runtime, and overall our test is still completed in a few seconds. In contrast, even though QHSIC-O/W requires fewer samples to reach the same power than our tests, the runtime rises rapidly as the sample size increases. When $n = 1000$, it already takes more than $10s$ to perform a test. When $n = 3000$, it needs nearly a minute to perform a test, which may greatly limit the practical application.

[4]The code is downloaded from `https://github.com/renyixin666/HSIC-LK`

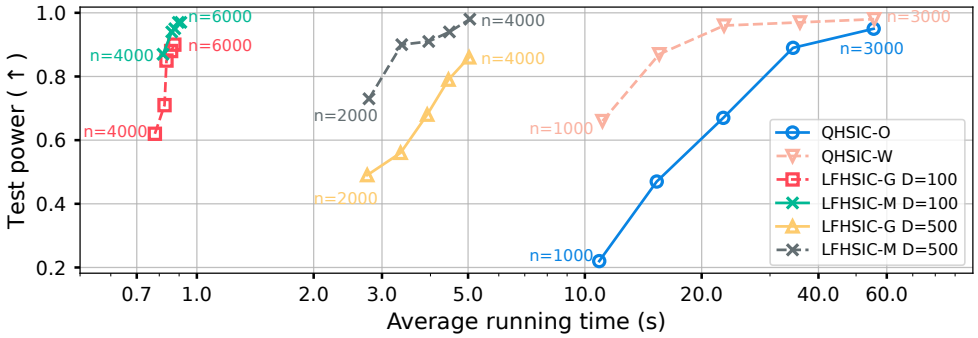

Figure 8: Time-power trade-off curves on the SD dataset.

K.3 Running Time

In this part, we evaluate the running time of each method on ISA datasets $d = 10$. We set $D = 100$ and plot the results of the running time versus sample size n as shown in Fig. 9. Shown on the left are the results of tests with $O(n)$ and $O(n \log n)$ time complexity. Since the quadratic time complexity test cannot handle large-scale inputs of 100,000 sample size (excessive runtime and large memory overhead to store the kernel matrix), we evaluate them separately on the right. The experiments are all conducted on the same equipment, specifically a 14-core CPU with a 4090 GPU.

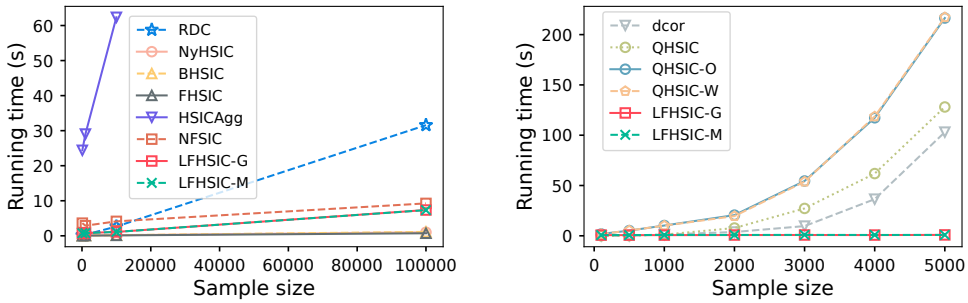

Figure 9: The running time curves with sample size n on the ISA dataset ($d = 10$).

Results. The experimental results on the left show that our method is faster than other methods with optimizable options (HSICAgg and NFSIC), and can complete a test within 10 seconds even with $100,000$ samples. Even though HSICAgg uses parallelism to optimize the computational efficiency of the scheme, the actual implementation of the parallelism is time-consuming, and hence leads to a high time overhead for a single practical test. For the results on the right, it can be seen that the quadratic complexity methods face a dramatic increase in time overhead as the sample size rises, and this is especially severe for the methods (QHSIC-O/W) that need to be optimized since the optimizing objective needs to perform multiple squared complexity operations. In contrast, our linear-time learning objective allows us to handle huge data samples very efficiently.

L Limitations and Broader Impacts

Limitations. According to the experimental results in the main paper as well as in the Appendix, no one method is better than the others in all settings, so it is important to choose several appropriate tests for real scenarios and summarize their results in order to obtain a more reliable conclusion.

Broader Impacts. This work proposes a novel framework for independence testing. The proposed linear-time optimization objective can be trained end-to-end in a data-driven manner, ensuring both effectiveness and efficiency in high-dimensional and large-scale scenarios. This could be beneficial for developing more reliable downstream algorithms in a variety of areas, including causal discovery, feature selection, and deep learning.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and precede the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: I've clearly stated the contribution.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See the Sec. L in the Appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: See Sec. 5 in the main paper. Also, the summarized assumptions and the proofs are provided in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the details to reproduce the main experimental results, please See Sec. J in Appendix, and we also provide the experimental data/code in the supplemental material.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the experimental data/code in the supplemental material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: we provide the details. See Sec. J in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experimental results are accompanied by statistical significance tests. See Sec. 6.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide sufficient information on the computer resources. See Sec. K in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: I read it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See the Sec. L in the Appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the methods used and list the URLs of the comparison methods.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.