Journal of Machine Learning Research 24 (2023) 1-68 Submitted 2/23; Revised 11/23 ; Published 11/23

A Permutation-Free Kernel Independence Test

Shubhanshu Shekhar SHUBHAN2@ANDREW.CMU.EDU
Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Ilmun Kim ILMUN@YONSEI.AC.KR
Department of Statistics and Data Science
Department of Applied Statistics
Yonsei University
Seodaemun-gu, Seoul, 03722, Republic of Korea

Aaditya Ramdas ARAMDAS@STAT.CMU.EDU
Department of Statistics and Data Science
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Editor: Jean-Philippe Vert

Abstract

In nonparametric independence testing, we observe i.i.d. data $\{(X_i, Y_i)\}_{i=1}^n$, where $X \in \mathcal{X}, Y \in \mathcal{Y}$ lie in any general spaces, and we wish to test the null that X is independent of Y. Modern test statistics such as the kernel Hilbert–Schmidt Independence Criterion (HSIC) and Distance Covariance (dCov) have intractable null distributions due to the degeneracy of the underlying U-statistics. Hence, in practice, one often resorts to using permutation testing, which provides a nonasymptotic guarantee at the expense of recalculating the quadratic-time statistics (say) a few hundred times. In this paper, we provide a simple but nontrivial modification of HSIC and dCov (called xHSIC and xdCov, pronounced "cross" HSIC/dCov) so that they have a limiting Gaussian distribution under the null, and thus do not require permutations. We show that our new tests, like the originals, are consistent against fixed alternatives, and minimax rate optimal against smooth local alternatives. Numerical simulations demonstrate that compared to the permutation tests, our variants have the same power within a constant factor, giving practitioners a new option for large problems or data-analysis pipelines where computation, not sample size, could be the bottleneck.

Keywords: independence testing, kernel-methods, permutation-free tests, Hilbert-Schmidt Independence Critrion (HSIC), Distance Covariance.

1. Introduction

We consider the following problem: given observations $\mathcal{D}_1^{2n} = \{(X_i, Y_i) : 1 \leq i \leq 2n\}$ drawn i.i.d. from a distribution P_{XY} on the observation space $\mathcal{X} \times \mathcal{Y}$, we wish to test whether X and Y are independent or not. Formally, this is stated as the following hypothesis testing problem:

$$H_0 : P_{XY} = P_X \times P_Y, \quad \text{versus} \quad H_1 : P_{XY} \neq P_X \times P_Y,$$

where P_X and P_Y denote the marginals of the joint distribution P_{XY}. For general observation spaces, a popular approach for independence testing is based on the Hilbert–Schmidt Independence Criterion (HSIC), first introduced by Gretton et al. (2005). As we describe formally in Theorem 1 in Section 2, the (population) HSIC of a joint distribution P_{XY} on $\mathcal{X} \times \mathcal{Y}$ is the sum of squared singular values (i.e., the Hilbert–Schmidt norm) of a cross-covariance operator defined using P_{XY} and positive definite kernels $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Given the data $\mathcal{D}_1^{2n}$, an unbiased empirical estimate of the population HSIC can be computed in quadratic time. Introducing the notation $k_{ij} \equiv k(X_i, X_j)$ and $\ell_{lm} \equiv \ell(Y_l, Y_m)$ for $1 \leq i, j, l, m \leq 2n$, and $(2n)_i \coloneqq \frac{(2n)!}{(2n-i)!}$, the empirical estimator of HSIC can be defined as:

$$\mathrm{HSIC}_n = \frac{1}{(2n)_2} \sum_{1 \leq i \neq j \leq 2n} k_{ij}\ell_{ij} + \frac{1}{(2n)_4} \sum_{\substack{1 \leq i,j,l,m \leq 2n \\ i,j,l,m \text{ distinct}}} k_{ij}\ell_{lm} - \frac{2}{(2n)_3} \sum_{\substack{1 \leq i,j,l \leq 2n \\ i,j,l \text{ distinct}}} k_{ij}\ell_{il}, \tag{1}$$

For characteristic kernels, Gretton et al. (2005) showed that the (population) HSIC is equal to zero if and only if $P_{XY} = P_X \times P_Y$, and thus it can serve as a measure of independence between P_X and P_Y. This can be then used to define an independence test that rejects the null when the statistic HSIC_n is larger than an appropriately chosen threshold. The choice of the threshold is crucial in ensuring that the test achieves large power under the alternative, while controlling the type-I error (at least asymptotically) at a specified level $\alpha \in (0, 1)$.

The empirical HSIC criterion introduced above is an instance of a degenerate U-statistic (Lee, 2019), and thus it has a complicated limiting null distribution — it is an infinite weighted linear combination of independent chi-squared random variables, the exact expression of which was derived by Gretton et al. (2007, Theorem 2). Due to the intractable nature of this null distribution, we cannot directly use this to calibrate the independence test based on the empirical HSIC statistic. Instead, in practice, the rejection threshold is often selected as the $(1 - \alpha)$-quantile after recomputing the statistic HSIC_n B times after permuting the indices of $(X_i)_{i=1}^{2n}$. Hence, this approach requires us to compute the quadratic time statistic HSIC_n a total of $B + 1$ times, which might make this method infeasible for larger n and B values.

To address the high computational cost of the permutation test, several alternatives such as tests based on deviation bounds between empirical and population HSIC, or using parametric approximations of the null distribution have been proposed. We discuss them in detail in Section 1.2. However, these existing approaches are either too conservative in practice or do not have theoretical guarantees on their performance.

In this paper, we propose a new and simple test that addresses the issues with existing kernel-based independence tests. In particular, we define a new unbiased empirical estimate of HSIC that has a tractable, standard normal, limiting null distribution under mild assumptions. We then use this statistic to define an independence test, that we call the cross-HSIC test, and show that it is consistent against arbitrary fixed alternatives and also achieves minimax rate-optimal power against smooth local alternatives.

1.1 Overview of Results

We propose a new statistic, based on the ideas of sample splitting and studentization, and use it to define a new test of independence. Given the observations $\mathcal{D}_1^{2n} = \{(X_i, Y_i) : 1 \leq i \leq 2n\}$, drawn i.i.d. from a joint distribution P_{XY} on $\mathcal{X} \times \mathcal{Y}$, we first split it into two equal parts, $\mathcal{D}_1^n = \{(X_i, Y_i) : 1 \leq i \leq n\}$ and $\mathcal{D}_{n+1}^{2n} = \{(X_i, Y_i) : n+1 \leq i \leq 2n\}$. Using these two splits, we construct two independent empirical estimates of the cross-covariance operator (denoted by f_1 and f_2), and define the statistic xHSIC_n as their Hilbert–Schmidt norm:

$$\begin{aligned} &\mathrm{xHSIC}_n = \langle f_1, f_2 \rangle_{HS} \;=\; \frac{1}{n(n-1)} \sum_{i \neq j} \langle h_{ij}, f_2 \rangle_{HS}, \\ &f_1 = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} h_{ij}, \quad \text{and} \quad f_2 = \frac{1}{n(n-1)} \sum_{n+1 \leq i \neq j \leq 2n} h_{ij}, \\ &h_{ij} \equiv h(Z_i, Z_j) = \frac{1}{2} \big\{ k(X_i, \cdot) - k(X_j, \cdot) \big\} \otimes \big\{ \ell(Y_i, \cdot) - \ell(Y_j, \cdot) \big\}. \end{aligned} \tag{2}$$

The final statistic, denoted by $\overline{\mathrm{x}}\mathrm{HSIC}_n$, is obtained by normalizing xHSIC_n with an empirical standard deviation term, stated in (6). Using this statistic, we define the cross-HSIC test, $\Psi = \mathbf{1}_{\overline{\mathrm{x}}\mathrm{HSIC}_n > z_{1-\alpha}}$, which rejects the null when $\overline{\mathrm{x}}\mathrm{HSIC}_n$ exceeds the $(1-\alpha)$-quantile of the standard normal distribution.

Our first set of results characterize the limiting null distribution of the $\overline{\mathrm{x}}\mathrm{HSIC}_n$ statistic, and justify the rejection threshold used in defining the cross-HSIC test. To motivate the more general results, we first consider the simple, but instructive, case of univariate observations (that is, $\mathcal{X} = \mathcal{Y} = \mathbb{R}$) and linear kernels $k(x, x') = xx'$ and $\ell(y, y') = yy'$ in Section 4. In this setting, we first show that for a fixed null distribution $P_{XY} = P_X \times P_Y$, the existence of finite second moment is sufficient for $\overline{\mathrm{x}}\mathrm{HSIC}_n$ to converge in distribution to $N(0, 1)$. Then we show that under stronger moment assumptions, the $\overline{\mathrm{x}}\mathrm{HSIC}_n$ statistic converges to $N(0, 1)$ uniformly over a composite class of null distributions. We then move to the case of general kernels and observation spaces, and derive analogous, but slightly more abstract, requirements for the asymptotic normality of the $\overline{\mathrm{x}}\mathrm{HSIC}_n$ statistic in Section 5. In particular, we identify sufficient conditions for the asymptotic normality of $\overline{\mathrm{x}}\mathrm{HSIC}_n$ for the case of fixed k, ℓ and P_{XY} in Theorem 6, and for the more general case where k_n, ℓ_n and $P_{XY,n}$ are all allowed to change with the sample-size n in Theorem 7. Overall, our results imply that our $\overline{\mathrm{x}}\mathrm{HSIC}_n$ statistic converges in distribution to a standard normal distribution, in most practically relevant scenarios, and thus the cross-HSIC test controls the type-I error asymptotically at the desired level α.

Having established the type-I error control achieved by our cross-HSIC test, we next analyze its power. Again, we first consider the case of univariate observations with linear kernels to gain some intuition in Section 4 where we show that the cross-HSIC test is consistent against any fixed alternative when P_X and P_Y are (linearly) correlated. We also consider the case of local alternatives and show that the cross-HSIC test can consistently detect alternatives separated by a $\Omega(1/\sqrt{n})$ boundary. Similar results also hold for the case of more general kernels. In particular, we first show in Theorem 9, that the cross-HSIC test with characteristic kernels k and ℓ is consistent against any fixed alternative.

Then in Theorem 10, we show that the cross-HSIC test instantiated with Gaussian kernels achieves minimax rate-optimal power against smooth local alternatives.

Finally, in Section 7, we consider a related class of statistics, called the distance-covariance (dCov). Using the equivalence between distance-based and kernel-based statistics (Sejdinovic et al., 2013), we introduce a new distance-based statistic, called the cross-dCov statistic. We then identify sufficient conditions for this statistic to have a standard normal limiting null distribution, and for it to be consistent against fixed alternatives.

1.2 Related Work

Nonparametric independence testing is a classical topic in statistics that still remains a subject of significant contemporary interest. A huge variety of methods have been developed recently for this problem, such as those based on ranks (Heller et al., 2013; Weihs et al., 2018; Deb and Sen, 2021; Shi et al., 2022), projections (Zhu et al., 2017), copula (Dette et al., 2013) and mutual information (Berrett and Samworth, 2019). This paper focuses on a particular class of kernel (and distance) based nonparametric tests, which are popular in machine learning and statistics. In the rest of this section, we present the details of the most closely related works to ours.

Nonasymptotic tail inequalities for HSIC. Gretton et al. (2005) introduced HSIC as a measure of dependence between two random variables, and obtained a high probability (nonasymptotic) deviation inequality between the empirical and population HSIC terms. The resulting test has finite sample validity, and is uniformly consistent against alternatives separated by a $\Omega(1/\sqrt{n})$ boundary (in terms of HSIC). However, this test is overly conservative in practice. To address this, Gretton et al. (2007) suggested to calibrate the test based on the null distribution of HSIC, and due to the intractability of the null distribution, they used parametric approximations of the null distribution to select the rejection threshold. This method, however, is a heuristic and does not have validity guarantees.

Modifications of the HSIC test statistic. Another class of permutation-free kernel independence tests take an approach similar to our paper, and modify the empirical HSIC statistic to make its null distribution tractable. Zhang et al. (2018) developed three such tests, using block-averaged HSIC, random Fourier features (RFFs) and the Nystrom approximation. The block-averaged HSIC statistic is obtained by partitioning the data $\mathcal{D}_1^{2n}$ into blocks of size b, and then computing the HSIC on these small blocks of data and taking their average; this takes $O(bn)$ time in total. The computational cost, then, varies from linear in n for constant block sizes, to quadratic when $b = \Omega(n)$. Furthermore, when $b = o(n)$, the block-averaged HSIC statistic has a limiting Gaussian distribution under the null, which can be used for calibrating the test. The other two methods (RFF and Nystrom) have a computational complexity of $\mathcal{O}(nb^2)$, where b denotes the number of features used. For the RFF method, the null distribution is a finite linear combination of chi-squared random variables, while in the Nystrom method, the null distribution is based on a low-rank approximation of the kernel matrices.

Jitkrittum et al. (2017) proposed a new measure of independence, called the finite set independence criterion (FSIC), that is computed as the average of the squared differences between the mean-embedding values of the joint distribution P_{XY}, and the product of marginals $P_X \times P_Y$, at J different locations in $\mathcal{X} \times \mathcal{Y}$. For the J locations drawn randomly from a continuous distribution, they showed that FSIC is equal to 0 if and only if X and Y are independent. Based on this, they constructed a Hotelling-type normalized FSIC statistic (denoted by $\widehat{\mathrm{NFSIC}}_n^2$) using J locations selected to optimize a lower bound on the power of the test. While the resulting linear-time test performed comparably to the quadratic-time permutation test in practice, there are several important factors that distinguish our work from theirs. Jitkrittum et al. (2017) analyze their test in the regime where the kernels (k and ℓ), and the number of features J are fixed; while the sample size n goes to infinity, and in particular, show that $\widehat{\mathrm{NFSIC}}_n^2$ has a limiting chi-squared distribution under the null. However, if these parameters (k, ℓ, and J) are also allowed to change with n, the null distribution of their statistic may change. Furthermore, the consistency of their test relies on an accurate estimation of the $J \times J$ covariance matrix, which is not possible when J also increases with n, and $\lim_{n\to\infty} J/n \approx 1$. Our test does not suffer from these issues, and we prove its consistency and type-I error control in more general settings.

For the specific case of Gaussian kernels, Li and Yuan (2019) analyzed an independence test based on a studentized version of the empirical HSIC statistic. In particular, for Gaussian kernels with scale parameter increasing at an appropriate rate with n, they showed that the studentized empirical HSIC statistic has a standard normal limiting null distribution, and the resulting independence test has minimax rate-optimal power against smooth local alternatives. In contrast, our studentized cross-HSIC statistic has a limiting null distribution for a much larger class of kernels, including unbounded kernels as well as kernels induced by semi-metrics (Sejdinovic et al., 2013). Furthermore, when specialized to Gaussian kernels, our test also achieves minimax rate-optimal power, matching the performance of the test studied by Li and Yuan (2019).

Other kernel-based independence tests. Deb et al. (2020) proposed a class of kernel-based nonparametric measures of association (called KMAc) between two random variables X and Y that satisfy the three desired criteria listed by Chatterjee (2021): they are equal to 0 for independent X and Y, and are equal to 1 when Y is a measurable function of X; they admit simple empirical estimates, and finally, they have simple asymptotic theory when X and Y are independent. These measures significantly generalize the univariate measure of association introduced by Chatterjee (2021). However, Deb et al. (2020) study the asymptotics of the empirical KMAc statistics only for the case of fixed kernels, and furthermore, they do not analyze the power of their independence tests against local alternatives.

Distance-Covariance tests. Székely et al. (2007) proposed a measure of dependence between two random variables, called the distance-Covariance (dCov) that is defined as the weighted L^2 norm between the characteristic function of the joint distribution, and the product of the marginal characteristic functions. For the case of Euclidean spaces, Székely et al. (2007) showed that the $(1-\alpha)$-quantile of a normalized version of the empirical dCov statistic can be upper bounded by that of a quadratic form of a Gaussian; thus providing a permutation-free test of independence. In practice, however, this test is quite conservative as the upper bound on the $(1-\alpha)$-quantile is tight only in the case of Bernoulli distributions. Lyons (2013) extended the definition and analysis of the dCov measure beyond Euclidean spaces, to arbitrary metric spaces. This was further generalized to the case of semi-metric spaces by Sejdinovic et al. (2013), who also derived a precise equivalence between dCov and kernel-based HSIC measures (we recall this in Theorem 11). In this paper, we use this equivalence to obtain new dCov based statistics, that also have a standard normal limiting distribution under the null.

High-dimensional asymptotics. A recent line of work has focused on exploring the limiting distribution of kernel and distance-based statistics in high dimensional regimes. For instance, Zhu et al. (2020) proved the asymptotic under the null for studentized versions of HSIC and unbiased dCov statistics in the regime where dimension and sample-size (n) both grow to infinity, but the growth of n is slower. Following them, Gao et al. (2021) derived a central limit theorem and also established explicit convergence rates for a rescaled dCov statistic under more general conditions (in particular, relaxing the growth requirement on n). Han and Shen (2021) obtained non-null central limit theorems for both kernel and distance-based statistics for the specific case of multivariate Gaussian observations. However, the asymptotic normality derivations in these papers rely critically on the eigenvalues of certain integral operators being bounded by positive constants both from above and below. These conditions cannot be verified in practice without additional prior information, thus reducing their practical applicability. Further, it is well known that such results cannot be true in low/moderate dimensional settings, where the limiting distribution is most certainly not Gaussian; it is a Gaussian chaos. In contrast, the asymptotic normality of our proposed cross-HSIC and cross-dCov statistics do not require these eigenvalue conditions, and are valid in all dimension regimes, with no restriction on the relationship between dimension and sample size.

Cross U-statistics. We note that the general design strategy used in this paper, based on sample-splitting and studentization, was first introduced by Kim and Ramdas (2023) for the case of one-sample U-statistics. The primary motivation of their work was to develop inference techniques that are *dimension-agnostic* — i.e., methods based on statistics whose asymptotic distribution remains the same, irrespective of the dimension regime. A key idea involved in their construction is to view sample-splitting as a tool for dimension reduction. That is, after splitting the dataset into two parts, they used the first split to learn a "witness function" $\hat{g}$ from some function class $\mathcal{G}$, and then obtained the cross-U statistic by taking the empirical mean of $\hat{g}$ over elements of the second split. This averaging can be viewed as projecting the points in the second split along the direction of $\hat{g}$. Shekhar et al. (2022) applied the ideas of Kim and Ramdas (2023) to the case of two-sample testing, to develop a permutation-free two-sample test based on kernel-MMD metric. In a similar vein, Kübler et al. (2022) proposed a two-sample test based on witness functions represented by machine learning (ML) models trained to optimize a weighted mean-squared loss. However, unlike Shekhar et al. (2022), the asymptotic validity of Kübler et al. (2022) assumes that the estimated witness function remains fixed as the sample size grows; thus the latter results do not apply if half (or a constant fraction of) the data is used to estimate the witness function, as would be necessary for minimax optimal power in theory and good performance in practice (comparable to the MMD and HSIC U-statistics). In this paper, we adapt and generalize the ideas from these papers, to develop a permutation-free kernel independence test. As we describe in Section 3.2, the analysis techniques developed by Shekhar et al. (2022) for their cross-MMD test cannot be directly applied to our setting, and hence we develop new methods for establishing the properties of our cross-HSIC test.

2. Hilbert–Schmidt Independence Criterion (HSIC)

As mentioned earlier, we assume that the observations (X, Y) lie in the space $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ may be different from $\mathcal{Y}$. Let $\mathcal{K}$ and $\mathcal{L}$ denote reproducing kernel Hilbert Spaces (RKHS) associated with positive-definite kernels $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, respectively. With $\phi(\cdot)$ and $\psi(\cdot)$ denoting the associated feature maps $x \mapsto k(x, \cdot)$ and $y \mapsto \ell(y, \cdot)$, we formally introduce the Hilbert–Schmidt independence criterion (HSIC).

Definition 1 *Let P_{XY} denote a probability distribution on $\mathcal{X} \times \mathcal{Y}$, and $\mathcal{H}$ and $\mathcal{G}$ denote RKHS associated with kernels $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, respectively. Then, the Hilbert–Schmidt Independence Criterion (HSIC) is defined as the Hilbert–Schmidt norm of the associated cross-covariance operator:*

$$\mathrm{HSIC}(P_{XY}, k, \ell) \coloneqq \|C_{XY}\|_{HS}^2, \quad \textit{where} \quad C_{XY} \coloneqq \mathbb{E}_{XY}\left[(\phi(X) - \mu) \otimes (\psi(Y) - \nu)\right].$$

Here $\otimes$ denotes the tensor product, and the terms μ and ν denote $\mathbb{E}_X[\phi(X)]$ and $\mathbb{E}_Y[\psi(Y)]$ respectively.

As shown by Gretton et al. (2005, Lemma 1), $\mathrm{HSIC}(P_{XY}, k, \ell)$ can be expressed as follows, in terms of the kernel functions, with (X, Y) and (X', Y') denoting two independent draws from the distribution P_{XY}:

$$\mathrm{HSIC}(P_{XY}, k, \ell) = \mathbb{E}_{XX'YY'}[k(X, X')\ell(Y, Y')] + \mathbb{E}_{XX'}[k(X, X')]\mathbb{E}_{YY'}[\ell(Y, Y')] - 2\mathbb{E}_{XY}[\mu(X)\nu(Y)].$$

Given data $\mathcal{D}_1^{2n} = \{(X_i, Y_i) : 1 \le i \le 2n\}$ consisting of $2n$ independent draws from P_{XY}, the empirical estimate stated in (1) is constructed using the above expression for HSIC. Under the null, the statistic HSIC_n is a degenerate one-sample U-statistic, and thus, for fixed kernels k and ℓ, its asymptotic null distribution is an infinite weighted combination of independent χ^2 random variables, as shown by Gretton et al. (2007, Theorem 2), where the weights depend on the distribution P_{XY}. Due to the intractable nature of this distribution, practical independence tests based on the HSIC_n statistic are usually calibrated using the permutation distribution, that leads to a significant increase in computation.

The HSIC metric is also known to be equal to the squared MMD distance between P_{XY} and the product of marginals, $P_X \times P_Y$, using the product kernel $K((x, y), (x', y')) = k(x, x')\ell(y, y')$. With the notation $\mu = \mathbb{E}_{P_X}[k(X, \cdot)]$, $\nu = \mathbb{E}_{P_Y}[\ell(Y, \cdot)]$ and $\omega = \mathbb{E}_{P_{XY}}[k(X, \cdot)\ell(Y, \cdot)]$, we can write $\mathrm{HSIC}(P_{XY}, k, \ell)$ as

$$\mathrm{HSIC}(P_{XY}, k, \ell) = \mathrm{MMD}^2(P_{XY}, P_X \times P_Y) = \langle \omega - \mu \times \nu,\ \omega - \mu \times \nu \rangle_{k \times \ell}, \tag{3}$$

where $\langle \cdot, \cdot \rangle_{k \times \ell}$ denotes the inner product in the RKHS associated with the product kernel $k \times \ell$. It is known that this RKHS is isometrically isomorphic to the tensor product space $\mathcal{H}_k \otimes \mathcal{H}_\ell$ endowed with the Hilbert–Schmidt norm (Steinwart and Christmann, 2008, Lemma 4.6 and Appendix A.5.2). We will use the above formulation in the next section to propose a new statistic with a tractable asymptotic null distribution.

3. The Cross-HSIC Test

We now propose a new statistic, called cross-HSIC, that has a standard Gaussian limiting null distribution under mild conditions on the kernels and the distribution P_{XY}. The construction of this new statistic relies on two key ideas of sample splitting and studentization.

Given $2n$ independent draws from the joint distribution P_{XY}, denoted by $\mathcal{D}_1^{2n} = \{(X_1, X_2), \ldots, (X_{2n}, Y_{2n})\}$, the studentized cross-HSIC statistic, $\overline{\mathrm{x}}\mathrm{HSIC}_n$, is defined in two steps:

- First, we split $\mathcal{D}_1^{2n}$ into two equal parts, denoted by $\mathcal{D}_1^n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ and $\mathcal{D}_{n+1}^{2n} = \{(X_{n+1}, Y_{n+1}), \ldots, (X_{2n}, Y_{2n})\}$. With Z_i denoting the paired observations (X_i, Y_i), we introduce the following terms:

$$f_1 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} h_{ij}, \quad \text{and} \quad f_2 = \frac{1}{n(n-1)} \sum_{t=n+1}^{2n} \sum_{\substack{u=n+1 \\ n \neq t}}^{2n} h_{tu},$$

  where we now denote $h_{ij} \equiv h(Z_i, Z_j) = \frac{1}{2}\{k(X_i, \cdot) - k(X_j, \cdot)\} \times \{\ell(Y_i, \cdot) - \ell(Y_j, \cdot)\}$, introduced earlier in (2), as an element of the RKHS $\mathcal{H}_{k \times \ell}$. Using these terms, we define the cross-HSIC statistic as

$$\mathrm{xHSIC}_n := \langle f_1, f_2 \rangle_{k \times \ell} = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \langle h_{ij}, f_2 \rangle_{k \times \ell} \, . \tag{4}$$

  It is easy to check that, for any $i \neq j$, we have $\mathbb{E}[h_{ij}] = \omega - \mu \times \nu$. Hence, f_1 and f_2 are two independent unbiased estimates of $\omega - \mu \times \nu$, which implies that xHSIC_n is an unbiased estimate of the HSIC metric. As the expression in (4) indicates, conditioned on the second-half of the data (i.e., $\mathcal{D}_{n+1}^{2n}$), xHSIC_n is a one-sample U-statistic.

- Our final statistic, denoted by $\overline{\mathrm{x}}\mathrm{HSIC}_n$, is obtained by normalizing the cross-statistic xHSIC_n with the empirical standard deviation as follows:

$$\overline{\mathrm{x}}\mathrm{HSIC}_n := \frac{\sqrt{n}\,\mathrm{xHSIC}_n}{s_n}, \quad \text{where} \tag{5}$$

$$s_n^2 = \frac{4(n-1)}{(n-2)^2} \sum_{i=1}^{n} \left(\frac{1}{n-1} \sum_{1 \leq j \neq i \leq n} \langle h_{ij}, f_2 \rangle_{k \times \ell} - \mathrm{xHSIC}_n \right)^2 . \tag{6}$$

  We note that $n^{-1} s_n^2$ is the jackknife estimator of the variance of $\mathbb{E}[h(Z_1, Z_2) | Z_1]$ also considered in Jing et al. (2000).

In Section 5, we show that under certain assumptions on the kernel, the $\overline{\mathrm{x}}\mathrm{HSIC}_n$ statistic has a standard normal limiting distribution under the null. This suggests the following natural level-α test of independence:

$$\Psi \equiv \Psi\left(\mathcal{D}_1^{2n}\right) = \mathbb{1}_{\overline{\mathrm{x}}\mathrm{HSIC}_n > z_{1-\alpha}},$$

where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of the $N(0,1)$ distribution.

3.1 Quadratic-time Computation of $\overline{\mathrm{x}}\mathrm{HSIC}_n$

A naive computation of the $\overline{\mathrm{x}}\mathrm{HSIC}_n$ statistic has a $\mathcal{O}(n^4)$ computational complexity, that is infeasible for all but very small problems. However, a more careful look at the terms involved in defining $\overline{\mathrm{x}}\mathrm{HSIC}_n$ indicates that it can actually be computed in quadratic time.

Theorem 2 *The $\overline{\mathrm{x}}\mathrm{HSIC}_n$ statistic, introduced in* (5)*, can be computed in $\mathcal{O}(n^2)$ time.*

Proof outline of Theorem 2. It suffices to prove that, both, the numerator (xHSIC_n) and the denominator (s_n) in the statistic $\overline{\mathrm{x}}\mathrm{HSIC}_n$ can be computed in quadratic time. First, we consider the numerator, and observe that it can be decomposed as follows:

$$\begin{aligned}
&\mathrm{xHSIC}_n = T_1 - T_2 - T_3 + T_4, \quad \text{where} \\
&T_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=n+1}^{2n} k(X_i, X_j)\ell(Y_i, Y_j), \\
&T_2 = \frac{1}{n^2(n-1)}\sum_{i=1}^{n}\sum_{n+1\le j_1\neq j_2\le 2n} k(X_i, X_{j_1})\ell(Y_i, Y_{j_2}), \\
&T_3 = \frac{1}{n^2(n-1)}\sum_{i=n+1}^{2n}\sum_{1\le j_1\neq j_2\le n} k(X_i, X_{j_1})\ell(Y_i, Y_{j_2}), \quad \text{and} \\
&T_4 = \frac{1}{n^2(n-1)^2}\sum_{1\le i_1\neq i_2\le n}\ \sum_{n+1\le j_1\neq j_2\le 2n} k(X_{i_1}, X_{j_1})\ell(Y_{i_2}, Y_{j_2}).
\end{aligned}$$

The above expressions indicate that a direct computation of T_1, T_2, T_3 and T_4 incur $\mathcal{O}(n^2)$, $\mathcal{O}(n^3)$, $\mathcal{O}(n^3)$ and $\mathcal{O}(n^4)$ cost respectively, making the overall computational cost $\mathcal{O}(n^4)$. In Theorem 16, we show how the terms T_2 and T_3 can be computed in quadratic time, and in Theorem 17 we show how the term T_4 can be computed in quadratic time. These two results together imply that the numerator of $\overline{\mathrm{x}}\mathrm{HSIC}_n$ has an overall quadratic complexity. Finally, we consider the denominator s_n, and first note that

$$s_n^2 = \frac{4(n-1)}{(n-2)^2}\left[\frac{1}{(n-1)^2}\sum_{i=1}^{n}\left(\sum_{j=1}^{n,j\neq i}\langle h_{ij}, f_2\rangle_{k\times\ell}\right)^2 - n\mathrm{xHSIC}_n^2\right]. \tag{7}$$

We complete the proof by showing that the first term inside the square brackets above can also be computed in quadratic time. The details are in Theorem 18.

3.2 Connections to xMMD_n^2

The kernel-MMD distance between two probability distributions (on the same observation space) is a widely used metric in the two-sample testing literature (Gretton et al., 2012). With $\mathcal{X} = \mathcal{Y}$ and $k = \ell$, the MMD distance between P_X and P_Y is defined as $\mathrm{MMD}(P_X, P_Y) := \langle \mu - \nu, \mu - \nu\rangle_k^{1/2}$. The usual empirical estimates of $\mathrm{MMD}^2(P_X, P_Y)$ based on observations are known to have a complex asymptotic null distribution (similar to HSIC). To address this, Shekhar et al. (2022) proposed the xMMD_n^2 statistic, based on splitting the observations $\mathcal{D}_1^{2n} = \{(X_i, Y_i) : i \in [2n]\}$ into two (usually equal) splits: $\mathcal{D}_1^n = \{(X_i, Y_i) : i \in [n]\}$ and $\mathcal{D}_{n+1}^{2n} = \{(X_i, Y_i) : n+1 \leq i \leq 2n\}$. In particular, let $(\widehat{\mu}_1, \widehat{\nu}_1)$ and $(\widehat{\mu}_2, \widehat{\nu}_2)$ denote the empirical estimates of μ and ν based on $\mathcal{D}_1^n$ and $\mathcal{D}_{n+1}^{2n}$ respectively. Then, the xMMD_n^2 is defined as

$$\mathrm{xMMD}_n^2 \equiv \mathrm{xMMD}_n^2(\mathcal{D}_1^{2n}) = \langle \widehat{\mu}_1 - \widehat{\nu}_1,\, \widehat{\mu}_2 - \widehat{\nu}_2 \rangle_k.$$

Shekhar et al. (2022) showed that the asymptotic distribution of a studentized version of xMMD_n^2 under the null, that is with $P_X = P_Y$, is $N(0,1)$ under mild conditions.

As we mentioned earlier in Section 3, the HSIC metric can be interpreted as the kernel-MMD distance, with the product kernel $k \times \ell$, between the joint distribution P_{XY} and the product of marginals $P_X \times P_Y$. Hence, similar to the definition of xMMD_n^2, the xHSIC_n statistic based on $\mathcal{D}_1^{2n}$, can be rewritten as

$$\mathrm{xHSIC}_n \equiv \mathrm{xHSIC}_n(\mathcal{D}_1^{2n}) = \langle \widehat{\mu}_1 \times \widehat{\nu}_1 - \widehat{\omega}_1,\, \widehat{\mu}_2 \times \widehat{\nu}_2 - \widehat{\omega}_2 \rangle_{k \times \ell}, \tag{8}$$

where $(\widehat{\omega}_1, \widehat{\mu}_1, \widehat{\nu}_1)$ and $(\widehat{\omega}_2, \widehat{\mu}_2, \widehat{\nu}_2)$ denote the empirical estimates of (ω, μ, ν) based on $\mathcal{D}_1^n$ and $\mathcal{D}_{n+1}^{2n}$ respectively.

Given the similarity in the definitions of xHSIC_n and xMMD_n^2, it might appear that the theoretical guarantees of xMMD_n^2 also carry over directly to the case of xHSIC_n. However, there is a subtle issue that prevents this. For analyzing xMMD_n^2, Shekhar et al. (2022) rely strongly on the fact that $\widehat{\mu}_1$ and $\widehat{\nu}_1$ are independent, and thus, conditioned on the second half of data, the terms $\langle \widehat{\mu}_1, \widehat{\mu}_2 - \widehat{\nu}_2 \rangle_k$ and $\langle \widehat{\nu}_1, \widehat{\mu}_2 - \widehat{\nu}_2 \rangle_k$ can be analyzed separately, and shown to converge (after appropriate normalization) to conditionally independent Gaussian distributions. The final distribution of the studentized xMMD_n^2 statistic is then obtained by using the fact that the sum of two Gaussian distributions is also Gaussian.

With xHSIC_n as defined in (8), however, the terms $\widehat{\omega}_1$ and $\widehat{\mu}_1 \times \widehat{\nu}_1$ are not independent. Hence, the techniques used for analyzing xMMD_n^2 do not directly apply to our case, and we develop a different approach for analyzing the studentized version of xHSIC_n statistic in the next two sections.

Remark 3 *It is well-known that the MMD is an integral probability metric associated with a unit ball in a RKHS (Gretton et al., 2012, equation (2)). In view of the relationship* (3)*, the square root of the HSIC is also an integral probability metric between P_{XY} and $P_X \times P_Y$. Specifically, we have the identity*

$$\sqrt{\mathrm{HSIC}(P_{XY}, k, \ell)} = \sup_{\|f\|_{k \times \ell} \leq 1} \langle f,\, \omega - \mu \times \nu \rangle_{k \times \ell} = \langle f^\star,\, \omega - \mu \times \nu \rangle_{k \times \ell},$$

where $f^\star := (\omega - \mu \times \nu)/\|\omega - \mu \times \nu\|_{k \times \ell}$ is a witness function achieving the supremum. This alternative expression offers a different viewpoint on our approach: we first estimate the witness function $f^\star$ that maximizes the problem signal using $\mathcal{D}_1^n$. We then apply this estimated witness function to project the dataset $\mathcal{D}_{n+1}^{2n}$ and aggregate the results by averaging them. This perspective aligns with Kim and Ramdas (2023, Section 1.2) that interprets sample splitting as a tool for dimensionality reduction.

4. Warmup: Testing Linear Dependence in One Dimension

The theoretical properties of our test, under both null and alternative, are subtle and nontrivial. Thus, to build intuition gently for later generalizations in the paper, we first briefly analyze our cross-HSIC test in the simplest testing: testing linear dependence of real-valued random variables X and Y.

Formally, this section studies the following (much simpler) problem: given $\mathcal{D}_1^{2n}$ drawn according to a joint distribution P_{XY} over $\mathbb{R}^2$, suppose we want to test whether $P_{XY} = P_X \times P_Y$ or that $\mathrm{Cov}(X, Y)$ is non-zero. Of course, there are many methods for this setting and we do not recommend using ours in particular; the only reason to focus only on our statistic is in order to get a handle of the more complex theoretical analysis that follows in the general case.

To formally describe our results in this setting, we set the observation spaces $\mathcal{X}$ and $\mathcal{Y}$ to $\mathbb{R}$, and assume that the kernels k and ℓ are both linear kernels: i.e., $k(x, x') = xx'$ and $\ell(y, y') = yy'$. In this case, the cross-HSIC statistic can be written as follows:

$$\mathrm{xHSIC}_n = \left(\frac{1}{n}\sum_{i=1}^{n} X_iY_i - \frac{1}{n(n-1)}\sum_{1\le i\neq j\le n} X_iY_j\right) \times \underbrace{\left(\frac{1}{n}\sum_{t=n+1}^{2n} X_tY_t - \frac{1}{n(n-1)}\sum_{n+1\le t\neq u\le 2n} X_tY_u\right)}_{=f_2}.$$

To develop intuition for validity of our procedure, we remark that xHSIC_n is a non-degenerate U-statistic conditional on f_2. It is well-known that a non-degenerate U-statistic is asymptotically Gaussian (Lee, 2019) under mild moment conditions, and thus one can expect that xHSIC_n is also asymptotically Gaussian conditional on f_2. Indeed, after studentization, we can isolate the randomness from f_2 as the sign of f_2, and our final statistic with linear kernels k and ℓ can be approximated as

$$\overline{\mathrm{x}}\mathrm{HSIC}_n = \mathrm{sign}(f_2) \times \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{(X_i - \mathbb{E}[X])(Y_i - \mathbb{E}[Y])}{\sqrt{\mathbb{V}\{(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\}}} + o_P(1) \quad \text{under the null.}$$

Notably, $\mathrm{sign}(f_2)$ is independent of the other terms, and converges to the Rademacher distribution. Using this fact, along with the central limit theorem, we prove that $\overline{\mathrm{x}}\mathrm{HSIC}_n$ is asymptotically $N(0, 1)$ under the null. Similar statistics have also been analyzed in other works, such as Kim and Ramdas (2023) and Lundborg et al. (2022). In the next theorem, we make this heuristic explanation rigorous and investigate the asymptotic behavior of $\mathbb{P}(\mathrm{xHSIC}_n > z_{1-\alpha})$ under both null and alternative. Below and throughout this section, we often omit the dependence of X and Y on P_{XY}, and simply write P_{XY} as P.

Theorem 4 *Suppose $\mathcal{D}_1^{2n} = \{(X_i, Y_i) \in \mathbb{R}^2 : 1 \le i \le 2n\}$ is drawn i.i.d. from a distribution P. Consider $\overline{\mathrm{x}}\mathrm{HSIC}_n$ computed with linear kernels k and ℓ based on $\mathcal{D}_1^{2n}$. Introduce the terms $\tilde{X} = X - \mathbb{E}[X]$, $\tilde{Y} = Y - \mathbb{E}[Y]$, and $\rho = \mathbb{E}[\tilde{X}\tilde{Y}]/\mathbb{V}^{1/2}[\tilde{X}\tilde{Y}]$. Then, for any fixed $\alpha \in (0, 1)$, we have the following:*

(a) (Pointwise asymptotic) Suppose P is fixed, and $\mathbb{E}[\tilde{X}^2\tilde{Y}^2], \mathbb{E}[\tilde{X}^2], \mathbb{E}[\tilde{Y}^2] \in (0, \infty)$. Then we have

$$\lim_{n\to\infty}\left|\mathbb{P}\big(\overline{\mathrm{x}}\mathrm{HSIC}_n > z_{1-\alpha}\big) - \Phi(-\sqrt{n}\rho)\Phi(z_\alpha - \sqrt{n}\rho) - \Phi(\sqrt{n}\rho)\Phi(z_\alpha + \sqrt{n}\rho)\right| = 0.$$

(b) *(Uniform asymptotic) Let $\mathcal{P}_n$ be a family of distributions of (X, Y), potentially changing with n, such that*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_n} \frac{\max\left\{\mathbb{E}_P[\tilde{X}^4\tilde{Y}^4],\, \mathbb{E}_P[\tilde{X}^4]\cdot\mathbb{E}_P[\tilde{Y}^4]\right\}}{n\mathbb{V}_P^2[\tilde{X}\tilde{Y}]} = 0. \tag{9}$$

Then we have

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_n} \left|\mathbb{P}_P\big(\overline{\mathrm{x}}\mathrm{HSIC}_n > z_{1-\alpha}\big) - \Phi(-\sqrt{n}\rho_n)\Phi(z_\alpha - \sqrt{n}\rho_n) - \Phi(\sqrt{n}\rho_n)\Phi(z_\alpha + \sqrt{n}\rho_n)\right| = 0.$$

As described earlier, the proof of Theorem 4 relies on the special structure of the linear kernel in the case of real-valued observations. Nevertheless, in Section 5, we will obtain similar results for general kernels, using significantly different proof techniques.

Part (a) of Theorem 4 implies that the cross-HSIC test is pointwise asymptotically of level α and it is consistent in power against any fixed alternative under the finite second moment condition. Part (b) provides a stronger uniform approximation under condition (9), involving fourth and second moments of $\tilde{X}$ and $\tilde{Y}$. This uniform result in part (b) directly allows us to state the uniform type-I error and power guarantees of the cross-HSIC test. We formally record this result in Theorem 5 below, showing that the cross-HSIC test controls type-I error at level-α uniformly over a large class of null distributions, and it is also consistent against local alternatives separated by a $\Omega(1/\sqrt{n})$ detection boundary.

Corollary 5 *Consider the same settings as in part (b) of Theorem 4, and recall that $\mathcal{P}_n$ is the family of distributions satisfying condition* (9)*. Then we have the following uniform guarantees:*

(a) *(Uniform validity) Let $\mathcal{P}_n^{(0)} = \{P_{XY} \in \mathcal{P}_n : P_{XY} = P_X \times P_Y\}$. The cross-HSIC test controls the asymptotic type-I error at level α uniformly over $\mathcal{P}_n^{(0)}$, i.e., $\lim_{n\to\infty}\sup_{P\in\mathcal{P}_n^{(0)}} \mathbb{E}_P[\Psi(\mathcal{D}_1^{2n})] = \alpha$.*

(b) *(Uniform consistency) Let $\mathcal{P}_n^{(1)}$ be a subset of $\mathcal{P}_n$ such that $\sqrt{n}|\rho_n| \to \infty$. The type-II error of the cross-HSIC test converges to zero uniformly over $\mathcal{P}_n^{(1)}$, i.e., $\lim_{n\to\infty}\sup_{P\in\mathcal{P}_n^{(1)}} \mathbb{E}_P[1 - \Psi(\mathcal{D}_1^{2n})] = 0$.*

We note again that condition (9) for $\mathcal{P}_n$ involves the fourth and second moments of X and Y. These moment condition can be relaxed to the finite $2 + \delta$ moment condition (see e.g., Appendix B.2), but with a more involved analysis. Nevertheless, we can show that condition (9) is not too strong, establishing an example where the asymptotic normality of $\overline{\mathrm{x}}\mathrm{HSIC}_n$ is no longer guaranteed without such conditions.

Example 1 *Suppose that $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ are i.i.d. random variables following a Bernoulli distribution with parameter p_n. Assume that $np_n^2 = \lambda > 0$ and λ is not an integer for all n. In this scenario, the limiting null distribution of $\overline{\mathrm{x}}\mathrm{HSIC}_n$ is not Gaussian and it instead satisfies*

$$\lim_{n\to\infty} \mathbb{P}(\overline{\mathrm{x}}\mathrm{HSIC}_n \leq 0) = \mathbb{P}\big(\mathrm{sign}(V') \times V \leq 0\big),$$

where V, V' are i.i.d. centered Poisson random variables with parameter λ.

We first remark that depending on the value of λ, the limiting probability in Example 1 is far from 1/2, which should be the case if $\overline{\mathrm{x}}\mathrm{HSIC}_n$ is asymptotically Gaussian. It is also worth mentioning that $n^{-1}\mathbb{E}_P[\tilde{X}^4\tilde{Y}^4]\mathbb{V}_P^{-2}(\tilde{X}\tilde{Y}) \gtrsim \lambda^{-1}$ under the conditions of Example 1. Thus the condition for $\mathcal{P}_n$ is violated. Lastly, we note that λ is assumed to be a non-integer for technical reasons and it may be removed with more effort. The detailed derivation of Example 1 can be found in Appendix B.3.

In the next section, we will see that we can obtain similar results with general observation spaces, controlling the type-I error of cross-HSIC test uniformly over composite null classes and establishing its consistency against local alternatives.

5. Asymptotic Gaussian Distribution under Null

We now generalize the results from previous section, derived for linear kernels and real-valued observations, to hold for general kernels and observation spaces. In particular, we first consider the case where the kernels k and ℓ, as well as the distribution P_{XY}, are fixed with n in Theorem 6, and then consider the most general setting with n-varying kernels and distributions in Theorem 7.

First, we introduce the "centered" kernels $\widetilde{k}$ and $\widetilde{\ell}$, defined as

$$\widetilde{k}(x,\cdot) = k(x,\cdot) - \mu, \quad \text{and} \quad \widetilde{\ell}(y,\cdot) := \ell(y,\cdot) - \nu.$$

Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and introduce the kernel $\widetilde{g} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, defined as

$$\widetilde{g}(z,z') = \widetilde{g}\big((x,y),(x',y')\big) = \langle \widetilde{k}(x,\cdot), \widetilde{k}(x',\cdot)\rangle_k \langle \widetilde{\ell}(y,\cdot), \widetilde{\ell}(y',\cdot)\rangle_\ell \tag{10}$$

Note that $\widetilde{g}$ is a symmetric function that is also square integrable if the kernel $k \times \ell$ is square integrable, when $(X,Y) \sim P_{XY,n}$ for some $n \geq 1$. Under this assumption, it admits the following orthonormal expansion:

$$\widetilde{g}(z,z') = \sum_{i=1}^{\infty} \lambda_{i,n} e_{i,n}(z) e_{i,n}(z'), \tag{11}$$

where $\{(\lambda_{i,n}, e_{i,n}) : i \geq 1\}$ form the orthonormal sequences of eigenvalue-eigenfunction pairs of the Hilbert–Schmidt operator associated with the product kernel $K = k \times \ell$; that is, $e \mapsto \int_{\mathcal{Z}} e(z)K(z,\cdot)dP_{XY,n}(z)$, for any $e \in L^2(P_{XY,n})$.

Our first main result of this section shows that if $\widetilde{g}(Z_1,Z_2)$ has a finite second moment under the null, then $\overline{\mathrm{x}}\mathrm{HSIC}_n$ converges in distribution to $N(0,1)$.

Theorem 6 *Suppose* $\overline{\mathrm{x}}\mathrm{HSIC}_n$ *is computed with fixed kernels* k *and* ℓ*, with observations* $\mathcal{D}_1^{2n}$ *drawn i.i.d. from a fixed distribution* P_{XY}*. If* k, ℓ *and* P_{XY} *satisfy the condition* $\mathbb{E}[\widetilde{g}(Z_1,Z_2)^2] \in (0,\infty)$ *with* $Z_1, Z_2 \sim P_{XY} = P_X \times P_Y$*, then we have* $\overline{\mathrm{x}}\mathrm{HSIC}_n \xrightarrow{d} N(0,1)$*.*

The finite second moment assumption on $\widetilde{g}$ is commonly used in the literature for studying the limiting distribution of HSIC statistic (e.g., Gretton et al., 2007) using the asymptotic theory of degenerate U- or V-statistics (e.g., Serfling, 2009). In Theorem 6, we show that under the same condition, our $\overline{\mathrm{x}}\mathrm{HSIC}_n$ attains an asymptotic normal limiting distribution. However, this type of fixed-asymptotic analysis excludes more dynamic and arguably more interesting settings where k, ℓ, P_{XY} may change with the sample size. Therefore we now present a more general condition in Assumption 1, which requires higher moment conditions than the finite-second moment of $\widetilde{g}$.

Assumption 1 *Let $\{P_{XY,n} : n \geq 1\}$ denote a sequence of probability distributions on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, with $P_{XY,n} = P_{X,n} \times P_{Y,n}$, and let $\{K_n = k_n \times \ell_n : n \geq 1\}$ denote a sequence of positive definite kernels on $\mathcal{Z}$. With $Z_{1,n}, Z_{2,n}, Z_{3,n}$ denoting three independent draws from $P_{XY,n}$, we assume that*

$$\lim_{n \to \infty} \frac{\mathbb{E}[\widetilde{g}(Z_{1,n}, Z_{2,n})^4]n^{-1} + \mathbb{E}[\widetilde{g}(Z_{1,n}, Z_{2,n})^2 \widetilde{g}(Z_{1,n}, Z_{3,n})^2]}{n\mathbb{E}[\widetilde{g}(Z_{1,n}, Z_{2,n})^2]^2} = 0.$$

where we recall that $\widetilde{g}$ was introduced in (10).

Our next result which establishes the asymptotic normality of the $\overline{\mathrm{x}}\mathrm{HSIC}_n$ statistic for n-varying kernels and distributions.

Theorem 7 *Suppose the $\overline{\mathrm{x}}\mathrm{HSIC}_n$ is computed with kernels $\{k_n, \ell_n : n \geq 1\}$, and let $\mathcal{P}_n^{(0)}$ denote the set of distributions $P_{XY,n} = P_{X,n} \times P_{Y,n}$ satisfying Assumption 1, for each $n \geq 1$. Then, the $\overline{\mathrm{x}}\mathrm{HSIC}_n$ statistic computed with kernels k_n and ℓ_n converges in distribution to $N(0,1)$ uniformly over $\mathcal{P}_n^{(0)}$.*

The proof of this statement is given in Appendix D, and it proceeds by verifying that under Assumption 1, the conditions required for the Berry–Esseen theorem for studentized U-statistics derived by Jing et al. (2000) are satisfied.

Remark 8 *While we have presented all the results of this section under the assumption that the two splits, $\mathcal{D}_1^n$ and $\mathcal{D}_{n+1}^{2n}$, are drawn i.i.d. from the same distribution P_{XY}, a closer inspection of the proof of these results indicates that the asymptotic normality of the $\overline{\mathrm{x}}\mathrm{HSIC}_n$ statistic (and the $\overline{\mathcal{V}}_n^2$ statistic, introduced later in Section 7) holds even when the two splits are only independent, and not identically distributed. Thus, the techniques developed in this paper are also applicable in more general scenarios; such as when $\mathcal{D}_1^n$ and $\mathcal{D}_{n+1}^{2n}$ are obtained by separately processing independent outputs of some common source.*

6. Power of the cross-HSIC Test

The results of the previous section establish the limiting standard normal distribution of the $\overline{\mathrm{x}}\mathrm{HSIC}_n$ statistic under general conditions, which in turn, implies that the cross-HSIC test controls type-I error at the desired level α asymptotically. In this section, we analyze the power of our test under the alternative. In particular, we first consider the case of an arbitrary fixed alternative, and prove the consistency of the cross-HSIC test with characteristic kernels in this case. Then, we consider the case of smooth local alternatives, and show that the cross-HSIC test with Gaussian kernels has minimax rate-optimal power. The proof of both these results can be inferred from a more general result (Theorem 29) that establishes sufficient conditions for the consistency of our cross-HSIC test, which we state and prove in Appendix E.1.

Consistency against fixed alternatives. First, we show that under some mild moment assumptions on the kernels, our cross-HSIC test is consistent against an arbitrary fixed alternative distribution $P_{XY} \neq P_X \times P_Y$.

Theorem 9 *Let P_{XY} be a distribution such that $\mathrm{HSIC}(P_{XY}, \mathcal{K}, \mathcal{L}) > 0$, where $\mathcal{K}$ and $\mathcal{L}$ denote the RKHS associated with characteristic kernels k and ℓ. Then, if $\mathbb{E}[k(X,X)\ell(Y,Y)] + \mathbb{E}[k(X,X)]\mathbb{E}[\ell(Y,Y)] < \infty$, the cross-HSIC test is consistent; that is, $\lim_{n\to\infty} \mathbb{P}_{P_{XY}}(\Psi = 1) = 1$.*

The details of obtaining this result from the more general result of Theorem 29 mentioned above are given in Appendix E.2. We now study the case of local alternatives, where the alternative distributions are allowed to change with n.

Consistency against smooth local alternatives. In our next result, we show that when constructed using Gaussian kernels, the cross-HSIC test is also minimax rate optimal against local alternatives with smooth density functions that are separated in L^2 sense. In particular, we set $\mathcal{X} = \mathbb{R}^{d_1}$ and $\mathcal{Y} = \mathbb{R}^{d_2}$, with $d \coloneqq d_1 + d_2$, and assume that the distributions P_{XY}, P_X and P_Y have smooth marginals p_{XY}, p_X and p_Y respectively. Furthermore, we assume that there exist constants M, M_X and M_Y such that $p_X \in \mathcal{W}^{\beta,2}(M_X)$, $p_Y \in \mathcal{W}^{\beta,2}(M_Y)$ and $M = M_X \times M_Y$. Here, $\mathcal{W}^{\beta,2}(M)$ denotes the ball with radius M in the fractional Sobolev space of order $\beta > 0$ (e.g., Li and Yuan, 2019). We can then define the null class of distributions, $\mathcal{P}_n^{(0)}$, and the Δ_n-separated alternative class of distributions, $\mathcal{P}_n^{(1)}$, as follows for all $n \geq 1$

$$\begin{aligned}
&\mathcal{P}_n^{(0)} = \{p_{XY} = p_X \times p_Y : p_X \in \mathcal{W}^{\beta,2}(M_X),\ p_Y \in \mathcal{W}^{\beta,2}(M_Y)\}, \quad \text{and} \\
&\mathcal{P}_n^{(1)} = \{p_{XY} : p_X \in \mathcal{W}^{\beta,2}(M_X),\ p_Y \in \mathcal{W}^{\beta,2}(M_Y),\ \text{and } \|p - p_X \times p_Y\|_{L_2} \geq \Delta_n\}.
\end{aligned}$$

For the independence testing problem described above, we will show that the cross-HSIC test, when instantiated using Gaussian kernels with appropriate scale factors, is minimax near-optimal. In particular, we define $k_n(x,x') = \exp\left(-c_n\|x - x'\|^2\right)$ and $\ell_n(y,y') = \exp\left(-c_n\|y - y'\|^2\right)$, where we have overloaded the term $\|\cdot\|$ to represent the Euclidean norm on both $\mathcal{X}$ and $\mathcal{Y}$.

Theorem 10 *Suppose $\mathcal{X} = \mathbb{R}^{d_1}$ and $\mathcal{Y} = \mathbb{R}^{d_2}$ with $d = d_1 + d_2$, and let $\{\Delta_n : n \geq 1\}$ is a non-negative sequence with $\lim_{n\to\infty} \Delta_n n^{2\beta/(d+4\beta)} = \infty$. Suppose the cross-HSIC test is instantiated with Gaussian kernels, $k_n(x,x') = \exp\left(-c_n\|x - x'\|^2\right)$ and $\ell_n(y,y') = \exp\left(-c_n\|y - y'\|^2\right)$, with $c_n \asymp n^{4/(d+4\beta)}$. Then, we have the following:*

$$\lim_{n\to\infty} \sup_{P_{XY,n} \in \mathcal{P}_n^{(0)}} \mathbb{E}[\Psi] = \alpha, \quad \textit{and} \quad \lim_{n\to\infty} \inf_{P_{XY,n} \in \mathcal{P}_n^{(1)}} \mathbb{E}[\Psi] = 1.$$

The proof of this statement is given in Appendix E.3. The first part of this result implies that the cross-HSIC test controls the type-I error at the specified level $\alpha \in (0,1)$ uniformly over the entire class of null distributions with smooth densities, $\mathcal{P}_n^{(0)}$. The second part of the result implies that our cross-HSIC test has a detection boundary of the order $\mathcal{O}\left(n^{-2\beta/(d+4\beta)}\right)$ in terms of the L^2-distance. As shown by Li and Yuan (2019), this rate cannot be improved in the worst case, thus establishing the minimax rate-optimality of our test. More formally, Li and Yuan (2019) showed that if $\lim_{n \to \infty} \Delta_n n^{2\beta/(d+4\beta)} < \infty$, then there exists an $\alpha \in (0,1)$ for which there exists no independence test that is consistent against such local alternatives (separated by Δ_n).

7. The Cross Distance Covariance Test (xdCov)

A popular alternative to kernel based method for measuring the dependence between two distributions is the distance-covariance metric introduced by Székely et al. (2007). When, $\mathcal{X} = \mathbb{R}^{d_1}$ and $\mathcal{Y} = \mathbb{R}^{d_2}$ for $d_1, d_2 \geq 1$, the distance-covariance associated with the joint distribution P_{XY} is

$$\begin{aligned} \mathcal{V}^2(P_{XY}) = \mathbb{E}_{X,X',Y,Y'}[\|X - X'\|\|Y - Y'\|] + \mathbb{E}_{X,X'}[\|X - X'\|]\mathbb{E}_{Y,Y'}\|Y - Y'\|] \\ - 2\mathbb{E}_{X,Y}[\mathbb{E}_{X'}[\|X - X'\|]\mathbb{E}_{Y'}[\|Y - Y'\|]]. \end{aligned} \tag{12}$$

In the above display, we overload the notation $\|\cdot\|$ to represent the Euclidean norm on both $\mathcal{X}$ and $\mathcal{Y}$. The distance-covariance metric has the property that it is equal to 0 if and only if X and Y are independent; that is, $P_{XY} = P_X \times P_Y$. This measure was extended beyond Euclidean spaces to general metric spaces by Lyons (2013), and further generalized to semi-metric spaces (i.e., distance-measures that do not satisfy the triangle inequality) by Sejdinovic et al. (2013). In particular, if $(\mathcal{X}, \rho_{\mathcal{X}})$ and $(\mathcal{Y}, \rho_{\mathcal{Y}})$ are semi-metric spaces, we can define the corresponding distance-covariance in a manner analogous to (12), as

$$\begin{aligned} \mathcal{V}^2(P_{XY}, \rho_{\mathcal{X}}, \rho_{\mathcal{Y}}) = \mathbb{E}_{X,X',Y,Y'}[\rho_{\mathcal{X}}(X,X')\rho_{\mathcal{Y}}(Y,Y')] + \mathbb{E}_{X,X'}[\rho_{\mathcal{X}}(X,X')]\mathbb{E}_{Y,Y'}[\rho_{\mathcal{Y}}(Y,Y')] \\ - 2\mathbb{E}_{X,Y}[\mathbb{E}_{X'}[\rho_{\mathcal{X}}(X,X')]\mathbb{E}_{Y'}[\rho_{\mathcal{Y}}(Y,Y')]]. \end{aligned}$$

Sejdinovic et al. (2013) showed an equivalence between the above distance-covariance metric and the HSIC computed with with the so-called *distance kernels* that we recall next.

Fact 11 *Define the distance kernels $k_{\mathcal{X}}$ and $\ell_{\mathcal{Y}}$ as*

$$\begin{aligned} k_{\mathcal{X}}(x,x') &\coloneqq \frac{1}{2}\left(\rho_{\mathcal{X}}(x,x_0) + \rho_{\mathcal{X}}(x',x_0) - 2\rho_{\mathcal{X}}(x,x')\right), \\ \text{and} \quad \ell_{\mathcal{Y}}(y,y') &\coloneqq \frac{1}{2}\left(\rho_{\mathcal{Y}}(y,y_0) + \rho_{\mathcal{Y}}(y',y_0) - 2\rho_{\mathcal{Y}}(y,y')\right), \end{aligned}$$

where x_0 and y_0 are arbitrary elements of $\mathcal{X}$ and $\mathcal{Y}$ respectively. Then, we have $\mathcal{V}^2(P_{XY}, \rho_{\mathcal{X}}, \rho_{\mathcal{Y}}) = \mathrm{HSIC}(P_{XY}, k_{\mathcal{X}}, \ell_{\mathcal{Y}})$.

Using this equivalence, we can define a cross-distance-covariance statistic similar to the cross-HSIC statistic of Section 3 by using sample-splitting and studentization.

Definition 12 *Given observations $\mathcal{D}_1^{2n}$ drawn from a distribution P_{XY}, we can define the cross-distance-covariance statistic, denoted by $\mathcal{V}_n^2$, as follows:*

$$\mathcal{V}_n^2 = \left\langle \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} h_{ij},\ \frac{1}{n(n-1)} \sum_{t=n+1}^{2n} \sum_{\substack{u=n+1 \\ u \neq t}}^{2n} h_{tu} \right\rangle \quad \text{where}$$

$$2h_{ij} = a_{ii} + a_{jj} - a_{ij} - a_{ji}, \quad \text{and} \quad a_{ij} = k_{\mathcal{X}}(X_i, \cdot)\ell_{\mathcal{Y}}(Y_j, \cdot).$$

We can then define a studentized version of cross-distance-covariance statistic, denoted by $\overline{\mathcal{V}}_n^2$, by normalizing $\mathcal{V}_n^2$ with $s_n/\sqrt{n}$, with s_n defined similar to (6).

Having defined the studentized cross-dCov statistic ($\overline{\mathcal{V}}_n^2$), we can now characterize its limiting null distribution by exploiting its equivalence to the studentized cross-HSIC statistic, and using the results derived in Section 5.

Corollary 13 *For a fixed distribution $P_{XY} = P_X \times P_Y$, if there exist a pair of points $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$ such that $\mathbb{E}[\rho_{\mathcal{X}}(X, x_0)^2]\mathbb{E}[\rho_{\mathcal{Y}}(Y, y_0)^2] < \infty$, then $\overline{\mathcal{V}}_n^2 \xrightarrow{d} N(0,1)$.*

More generally, suppose the sequence of distributions $\{P_{XY,n} : n \geq 1\}$ and the distance kernels $\{(k_{\mathcal{X},n}, \ell_{\mathcal{Y},n}) : n \geq 1\}$ and satisfy Assumption 1. Then, the statistic $\overline{\mathcal{V}}_n^2$ converges in distribution to $N(0,1)$ uniformly over the composite class of null distributions satisfying Assumption 1.

The above asymptotic normality of the $\overline{\mathcal{V}}_n^2$ statistic under the null suggests the definition of an independence test (Ψ^ρ), based on the cross-distance-covariance statistic that rejects the null when $\overline{\mathcal{V}}_n^2$ exceeds $z_{1-\alpha}$, the $1-\alpha$ quantile of the standard normal distribution. That is, $\Psi^\rho = \mathbf{1}_{\overline{\mathcal{V}}_n^2 > z_{1-\alpha}}$. Using the analogous results for the $\overline{\mathrm{x}}\mathrm{HSIC}_n$ statistic, we can obtain sufficient conditions for Ψ^ρ to be consistent.

Corollary 14 *For a fixed alternative distribution, $P_{XY} \neq P_X \times P_Y$, if there exist a pair of points $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$, such that $\mathbb{E}[\rho_{\mathcal{X}}(X, x_0)^2 \rho_{\mathcal{Y}}(Y, y_0)^2] + \mathbb{E}[\rho_{\mathcal{X}}(X, x_0)^2]\,\mathbb{E}[\rho_{\mathcal{Y}}(Y, y_0)^2] < \infty$, then the test Ψ^ρ is consistent.*

More generally, for a sequence of local alternatives $\{P_{XY,n} : n \geq 1\}$, if the distance kernels $k_{\mathcal{X}}$ and $\ell_{\mathcal{Y}}$ satisfy the conditions in Theorem 29 in Appendix E.1, the test Ψ^ρ is consistent.

Remark 15 *Székely et al. (2007) proposed a permutation-free independence test based on the dCov statistic, which relied on the fact that asymptotically, a suitably normalized variant of dCov is stochastically dominated by a quadratic form of a centered standard normal random variable (Székely et al., 2007, Theorem 6). However, as noted by Székely et al. (2007), and empirically verified by Sejdinovic et al. (2013), the resulting independence test can be extremely conservative in practice. We also illustrate this through an example in Figure 1. Furthermore, the validity of this approach when the distributions and distance-measures can change with the sample-size has not been established. Our proposed cross-dCov test addresses both these issues.*

8. Experiments

In this section, we experimentally validate the theoretical results presented in the previous sections. The code for reproducing these results is available in the repository: `https://github.com/sshekhar17/PermFreeHSIC`.

8.1 Null Distribution

Sufficiency of finite second moment. In the first experiment, we verify the claim of Theorem 6 that states that finite second moment of the kernel is sufficient for the asymptotic normality of the $\overline{\mathrm{x}}\mathrm{HSIC}_n$ statistic under the null. In particular, we use linear kernels

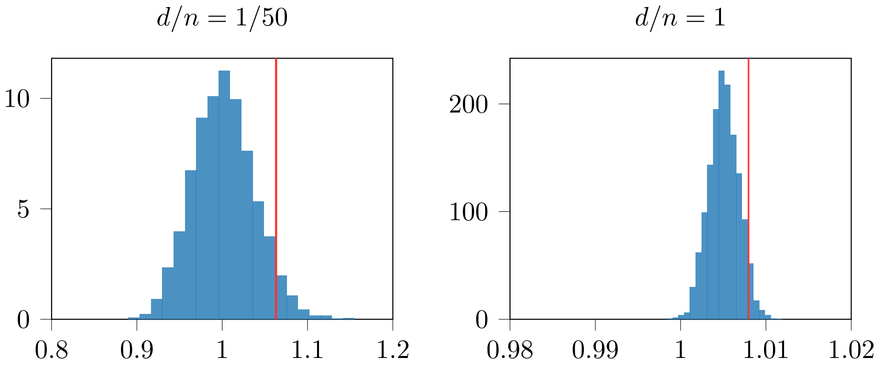

Figure 1: Plot of the empirical null distribution (over 5000 trials) of the normalized dcov statistic from Székely et al. (2007, Theorem 6) in the low ($d/n = 1/50$) and high ($d/n = 1$) dimensional settings. The red vertical line shows the true (empirical) $(1-\alpha)$-quantile of the null distribution for $\alpha = 0.05$. In both instances, this value is significantly smaller than $\left(\Phi^{-1}(1-\alpha/2)\right)^2 \approx 3.84$ — the value suggested by Székely et al. (2007), illustrating the highly conservative nature of their test.

k and ℓ, and consider the case where P_X and P_Y are distribution in $\mathbb{R}^d$; with each component drawn independently from a t-distribution with `dof` degrees of freedom. Recall that such distributions have finite moments of order up to `dof` $- 1$. The null distribution of the $\overline{\text{x}}\text{HSIC}_n$ statistic on such distributions with `dof` $\in \{1, 2, 3\}$ are shown in Figure 2. For `dof` $= 1$ and `dof` $= 3$, the null distribution appears to be clearly non-Gaussian and Gaussian respectively; with `dof` $= 2$ representing an intermediate state.

Effect of kernels and dimension regimes. In the next experiment, we verify that the $\overline{\text{x}}\text{HSIC}_n$ statistic has a limiting null distribution for different choices of kernels (Gaussian vs Rational Quadratic) and in different dimension regimes ($d/n = 3/4$ vs $d/n = 1/20$). The observations are drawn from independent multivariate Gaussian distributions with unit covariance matrix. The results, plotted in Figure 3, show that as expected, the null distribution of $\overline{\text{x}}\text{HSIC}_n$ approaches the standard normal distribution, even for relatively small sample sizes (all the plots have $n = 200$).

Control of type-I error. The previous experiment shows that visually, the null distribution of $\overline{\text{x}}\text{HSIC}_n$ statistic approaches the standard normal distribution even at relatively small n values. We now show in Figure 4 that this also translates into tight control over the type-I error at the desired level α (= 0.05 in the plots) of the resulting cross-HSIC test based on the $\overline{\text{x}}\text{HSIC}_n$ statistic.

8.2 Power

We now empirically compare the power of our cross-HSIC test with the original HSIC permutation test for various kernels and different dimension regimes. We limit our baselines to the HSIC permutation test, since it matches our stated objective of developing a kernel-based independence test that achieves a more favorable computational-vs-statistical efficiency trade-

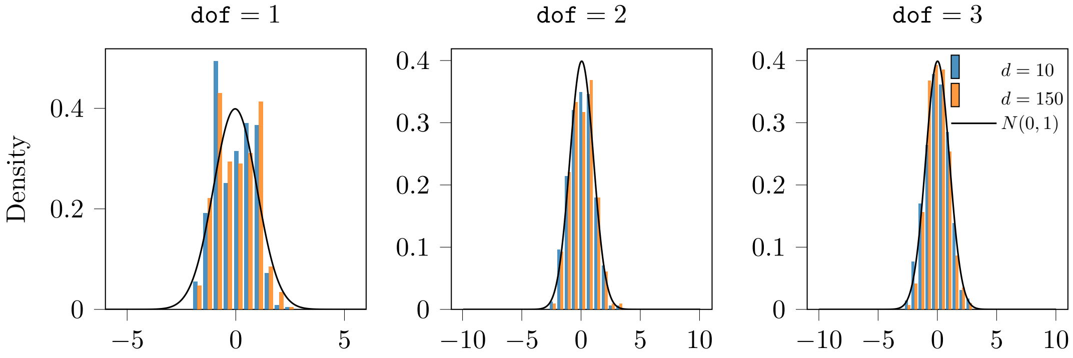

Figure 2: The figures show the null distribution of $\overline{\mathrm{x}}\mathrm{HSIC}_n$ with $n = 500$, $d \in \{10, 150\}$ using 500 trials. Both P_X and P_Y consist of d independent components drawn from t-distributions with degrees of freedom dof $\in \{1, 2, 3\}$; the data has finite variance only in the third plot. Thus, the plots above indicate that the existence of finite second moment is sufficient for the asymptotic normality of the $\overline{\mathrm{x}}\mathrm{HSIC}_n$ statistic, and appear somewhat necessary as well (the second plot may be close to Gaussian, but the first plot is far from it).

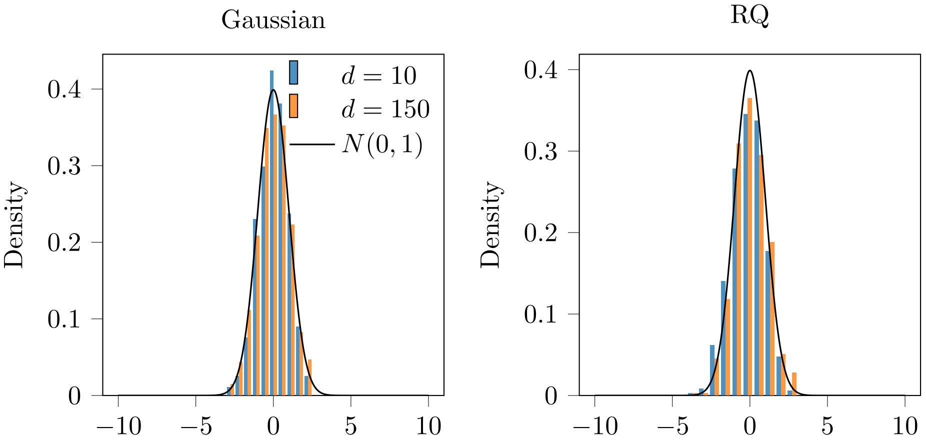

Figure 3: Plots of the null distribution of the $\overline{\mathrm{x}}\mathrm{HSIC}_n$ statistic for two commonly used kernels, Gaussian and Rational-Quadratic (RQ), under two dimension regimes each, with $d/n \in \{3/4, 1/20\}$, with $n = 200$.

off than the HSIC permutation test. In previous sections, we proved that our cross-HSIC is the first test that is permutation-free (hence computationally efficient), equally valid in different dimension regimes, while also retaining the minimax rate-optimality against smooth alternatives. We now benchmark the empirical performance of our test against the HSIC permutation test, while omitting comparisons to other methods. This is because, the power achieved by our test in all experiments is within a constant factor of that of the HSIC per-

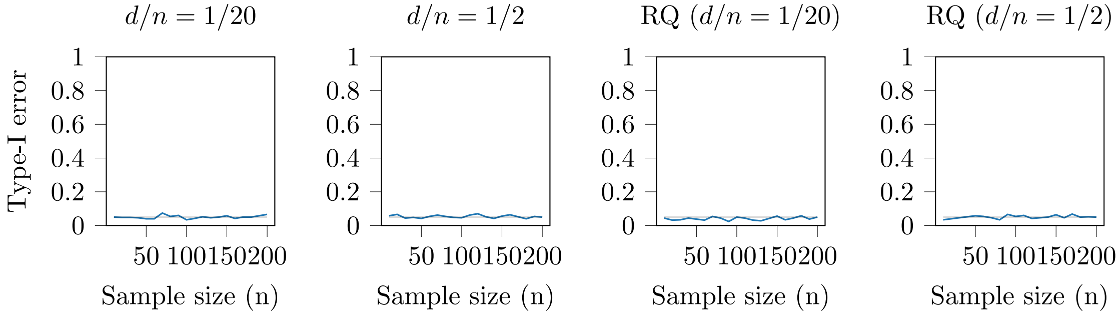

Figure 4: The figures show the variation of the type-I error of the cross-HSIC test under two different dimension regimes: $d/n \in \{1/2, 1/20\}$, and for two commonly used kernels: Gaussian (the first two plots) and Rational Quadratic (RQ). In all cases, we see that the type-I error is controlled at level $\alpha = 0.05$ for $n \geq 100$.

mutation test. Hence, all the power comparisons of existing tests with HSIC permutation tests are immediately also applicable to our test, with this correction.

In Figure 5, we plot the power curves of our cross-HSIC test and the HSIC permutation test for two kernels (Gaussian and Rational-Quadratic), for different levels of dependence (measured by ϵ). For this experiment, we set P_X to a multivariate Gaussian distribution in d dimensions with identity covariance matrix; and then generated Y as

$$Y = \epsilon \times X^b + (1-\epsilon) \times (X')^b, \quad \text{for } b > 0. \tag{13}$$

In the above display, the exponentiation is done component-wise and X' is an independent copy of X. As shown in Figure 5, our cross-HSIC test is slightly less powerful than the HSIC permutation test across different scenarios. This power loss can be attributed to a less efficient use of the data due to sample splitting. Nevertheless, we believe that this marginal decrease in power is a reasonable trade-off for the computational advantages gained by avoiding the permutation procedure.

Finally, in Figure 6, we show the improved power-vs-computation trade-off achieved by our cross-HSIC test, in comparison to the HSIC permutation test. In particular, note that to reach the same power, the running time required by our cross-HSIC test is approximately two orders of magnitude lower than the permutation test.

9. Conclusion and Future Work

In this paper, we proposed a variant of the HSIC statistic, called the cross-HSIC statistic, that is constructed using the ideas of sample-splitting and studentization. Under very mild conditions, we showed that this statistic has a standard normal limiting null distribution. Based on this result, we proposed a simple permutation-free independence test, that rejects the null when the studentized cross-HSIC statistic exceeds $z_{1-\alpha}$, the $(1-\alpha)$-quantile of the standard normal distribution. We present a thorough theoretical analysis of the performance of this test, and empirically validate the theoretical predictions on some experiments with synthetic data.

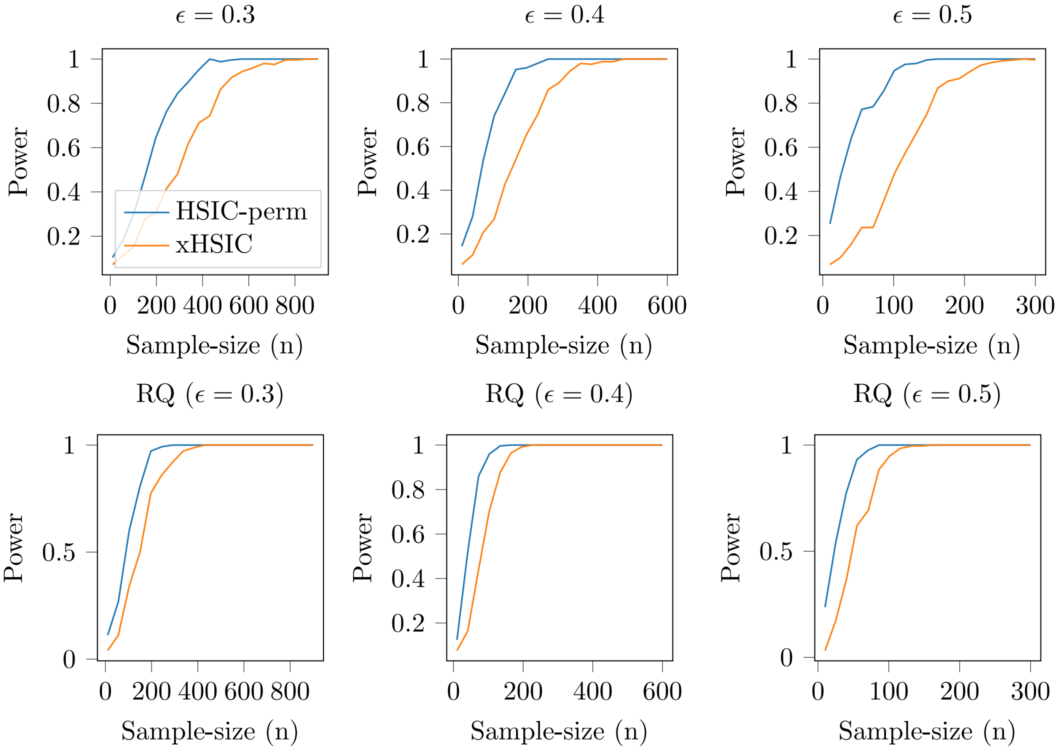

Figure 5: The figures in the top row show the power curves for HSIC permutation test, and cross-HSIC test with Gaussian kernels, while the bottom row corresponds to the same tests with Rational-Quadratic (RQ) kernel. In all figures, we have $d = 10$ and $b = 2$, while ϵ is set to $0.3, 0.4$, and 0.5 in the three columns. Recall that ϵ and b correspond to the expression in (13).

Our work opens up several interesting directions for future work, as we discuss below:

- **Robust cross-HSIC test.** The cross-HSIC statistic that we proposed can be thought of as the studentized sample covariance of projected feature maps onto one-dimensional spaces. The association between these projected feature maps can be measured through other methods as well such as Kendall's rank coefficient and Spearman's rank coefficient. Given that these rank-based approaches are more robust to outliers than the sample covariance, it would be interesting to develop robust cross-HSIC and investigate its theoretical guarantees and empirical performance.

- **Minimaxity beyond Gaussian kernels.** Another interesting future direction would involve extending the minimax result in Theorem 10 to non-Gaussian kernels, particularly the characteristic translation invariant kernels discussed in Schrab et al. (2023). We believe such a result is true, but would require significantly more technical effort to prove using rather recently developed tools, so we reserve it for future work.

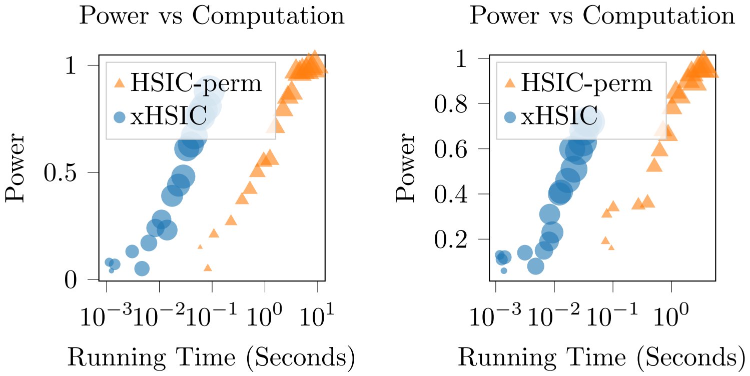

Figure 6: The figure shows the power versus running time curves for our cross-HSIC statistic and the HSIC permutation-test (with 150 permutations) on two different problems (with $b = 2$ and $\epsilon \in \{0.30, 0.35\}$. The size of the marker in the figures is proportional to the sample-size used for estimating the power. As indicated by the figures, if sample-size is not an issue, the cross-HSIC test can achieve the same power at a significantly lower running time (or computational cost) as compared to the permutation-test.

- **Alternatives to HSIC.** While our focus in this paper was on designing kernel- and distance-based dimension-agnostic independence tests, the same design principles (with appropriate modifications) might also be useful in other scenarios. For example, when dealing with complex data-types such as images or text, it would be interesting to explore designing independence tests based on features learned via ML models, following Kübler et al. (2022). Another interesting direction is to construct a linear-time dimension-agnostic independence test using a "cross-FSIC" statistic, by combining our ideas with those of Jitkrittum et al. (2017). Such a test would be particularly useful when working with large datasets or with limited computational resources. Although both these extensions are certainly feasible, they lie outside the scope of this work, and we leave them as interesting questions for future work.

- **Conditional independence testing.** Zhang et al. (2011) propose a kernel-based test for conditional independence (CI) building on a plug-in HSIC estimator. This method is typically calibrated by a Monte Carlo approach as their limiting distribution is intractable under the null. We believe that the use of the cross HSIC can alleviate their calibration issue, yielding a more reliable and computationally efficient CI test. Extending our framework to CI testing, and providing a rigorous justification would be an interesting direction for future work.

- **Testing independence of random processes.** In many practical applications, the observations are drawn from some time series or stochastic process; thus violating the i.i.d. assumption required in our analysis. To enable to applicability of our ideas in such applications, we need to establish the limiting distribution of cross-HSIC in an appropriate non-i.i.d. setting, such as under mixing conditions, as considered by

Chwialkowski and Gretton (2014). Some of the techniques discussed by Peña et al. (2009, Chapter 8 and 15) might be a good starting point for deriving such a result.

Acknowledgements

IK acknowledges support from the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2022R1A4A1033384), and the Korea government (MSIT) RS-2023-00211073.

References

T. B. Berrett and R. J. Samworth. Nonparametric independence testing via mutual information. *Biometrika*, 106(3):547–566, 2019.

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

K. Chwialkowski and A. Gretton. A kernel independence test for random processes. In *International Conference on Machine Learning*, pages 1422–1430. PMLR, 2014.

N. Deb and B. Sen. Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Journal of the American Statistical Association*, 118(541):192–207, 2021.

N. Deb, P. Ghosal, and B. Sen. Measuring association on topological spaces using kernels and geometric graphs. *arXiv preprint arXiv:2010.01768*, 2020.

H. Dette, K. F. Siburg, and P. A. Stoimenov. A Copula-Based Non-parametric Measure of Regression Dependence. *Scandinavian Journal of Statistics*, 40(1):21–41, 2013.

R. Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

L. Gao, Y. Fan, J. Lv, and Q.-M. Shao. Asymptotic distributions of high-dimensional distance correlation inference. *Annals of Statistics*, 49(4):1999, 2021.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *International Conference on Algorithmic Learning Theory*, pages 63–77. Springer, 2005.

A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. *Advances in Neural Information Processing Systems*, 20, 2007.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Q. Han and Y. Shen. Generalized kernel distance covariance in high dimensions: non-null CLTs and power universality. *arXiv preprint arXiv:2106.07725*, 2021.

R. Heller, Y. Heller, and M. Gorfine. A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2):503–510, 2013.

W. Hoeffding. The Strong Law of Large Numbers for U-statistics. Technical report, North Carolina State University. Dept. of Statistics, 1961.

B.-Y. Jing, Q. Wang, and L. Zhao. The Berry-Esséen bound for studentized statistics. *The Annals of Probability*, 28(1):511–535, 2000.

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

I. Kim and A. Ramdas. Dimension-agnostic inference using cross u-statistics. *Bernoulli*, 2023.

J. M. Kübler, V. Stimper, S. Buchholz, K. Muandet, and B. Schölkopf. AutoML Two-Sample Test. *Advances in Neural Information Processing Systems*, 35:15929–15941, 2022.

A. J. Lee. *U-statistics: Theory and Practice.* Routledge, 2019.

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

A. R. Lundborg, I. Kim, R. D. Shah, and R. J. Samworth. The Projected Covariance Measure for assumption-lean variable significance testing. *arXiv preprint arXiv:2211.02039*, 2022.

R. Lyons. Distance covariance in metric spaces. *The Annals of Probability*, 41(5):3284–3305, 2013.

V. H. Peña, T. L. Lai, and Q.-M. Shao. *Self-normalized processes: Limit theory and Statistical Applications.* Springer, 2009.

A. Schrab, I. Kim, M. Albert, B. Laurent, B. Guedj, and A. Gretton. MMD Aggregated Two-Sample Test. *Journal of Machine Learning Research (JMLR)*, 24, 2023.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5): 2263–2291, 2013.

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

R. D. Shah and J. Peters. The Hardness of Conditional Independence Testing and the Generalised Covariance Measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

S. Shekhar, I. Kim, and A. Ramdas. A permutation-free kernel two-sample test. *Proceedings of the Thirty-sixth Conference Neural Information Processing Systems*, 2022.

H. Shi, M. Drton, and F. Han. Distribution-free consistent independence tests via center-outward ranks and signs. *Journal of the American Statistical Association*, 117(537): 395–410, 2022.

I. Steinwart and A. Christmann. *Support vector machines.* Springer Science & Business Media, 2008.

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

L. Weihs, M. Drton, and N. Meinshausen. Symmetric rank covariances: a generalized framework for nonparametric measures of dependence. *Biometrika*, 105(3):547–562, 2018.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-Based Conditional Independence Test and Application in Causal Discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, page 804–813, 2011.

Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130, 2018.

C. Zhu, X. Zhange, S. Yao, and X. Shao. Distance-based and rkhs-based dependence metrics in high dimension. *The Annals of Statistics*, 48(6):3366–3394, 2020.

L. Zhu, K. Xu, R. Li, and W. Zhong. Projection correlation between two random vectors. *Biometrika*, 104(4):829–843, 2017.

Appendix A. Quadratic Computational Complexity of $\overline{\text{x}}\text{HSIC}_n$ (Theorem 2)

We begin by introducing the necessary notation for proving this result. First let K and L denote $2n \times 2n$ matrices, defined as

$$[K]_{ij} = \begin{cases} 0 & \text{if } 1 \leq i, j \leq n, \text{ or } n+1 \leq i, j \leq 2n \\ k(X_i, X_j) & \text{otherwise.} \end{cases}$$

and

$$[L]_{ij} = \begin{cases} 0 & \text{if } 1 \leq i, j \leq n, \text{ or } n+1 \leq i, j \leq 2n \\ \ell(X_i, X_j) & \text{otherwise.} \end{cases}$$

Also introduce the following two vectors in $\mathbb{R}^{2n}$.

$$\mathbf{1}_u = (\underbrace{1, \ldots, 1}_{n \text{ terms}}, 0, \ldots, 0) \quad \text{and} \quad \mathbf{1}_l = (\underbrace{0, \ldots, 0}_{n \text{ terms}}, 1, \ldots, 1).$$

Finally, note that we will use $\circ$ to denote the elementwise product of two matrices.

Proof of Theorem 2. To show the quadratic complexity of the cross-HSIC statistic, we it suffices to show that T_2, T_3, and T_4 can each be computed in quadratic time.

Lemma 16 *T_2 and T_3 satisfy the following:*

$$T_2 = \frac{1}{n^2(n-1)} \left(\mathbf{1}_l K L \mathbf{1}_l - \frac{1}{2} \text{tr}\,(KL) \right), \quad \textit{and} \quad T_3 = \frac{1}{n^2(n-1)} \left(\mathbf{1}_u K L \mathbf{1}_u - \frac{1}{2} \text{tr}\,(KL) \right),$$

where tr$(\cdot)$ denote the trace of a matrix.

Proof We show the details of the calculation for the term T_2 (the result for T_3 follows the exact same steps). In particular, we have the following:

$$\begin{aligned} n^2(n-1)T_2 &= \sum_{i=1}^{n} \sum_{n+1 \leq j_1 \neq j_2 \leq 2n} k(X_i, X_{j_1}) \ell(Y_i, Y_{j_2}) \\ &= \sum_{i=1}^{n} \sum_{j_1=n+1}^{2n} \sum_{j_2=n+1}^{2n} k(X_i, X_{j_1}) \ell(Y_i, Y_{j_2}) - \sum_{i=1}^{n} \sum_{j=n+1}^{2n} k(X_i, X_j) \ell(Y_i, Y_j) \\ &= \mathbf{1}_l^T K L \mathbf{1}_l - \frac{1}{2} tr(KL). \end{aligned}$$

Both the matrix multiplications involved in the last expression can be done in $\mathcal{O}(n^2)$ time, implying the quadratic complexity of T_2. ∎

Lemma 17 *The term T_4 satisfies*

$$T_4 = \frac{1}{n^2(n-1)^2} \left[(\mathbf{1}_l^T K \mathbf{1}_u)(\mathbf{1}_l^T L \mathbf{1}_u) - \mathbf{1}_l^T K L \mathbf{1}_l - \mathbf{1}_u^T K L \mathbf{1}_u + \frac{1}{2} \text{tr}(KL) \right].$$

Proof We again proceed by expanding the summation defining T_4.

$$\begin{aligned}
n^2(n-1)^2 T_4 &= \sum_{1\leq i_1 \neq i_2 \leq n} \sum_{n+1 \leq j_1 \neq j_2 \leq 2n} k(X_{i_1}, X_{j_1})\ell(Y_{i_2}, Y_{j_2}) \\
&= \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{j_1=n+1}^{2n}\sum_{j_2=n+1}^{2n} k(X_{i_1}, X_{j_1})\ell(Y_{i_2}, Y_{j_2}) - \sum_{i=1}^{n}\sum_{j_1=n+1}^{2n}\sum_{j_2=n+1}^{2n} k(X_i, X_{j_1})\ell(Y_i, Y_{j_2}) \\
&- \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{j=n+1}^{2n} k(X_{i_1}, X_j)\ell(Y_{i_2}, Y_j) + \sum_{i=1}^{n}\sum_{j=n+1}^{2n} k(X_i, X_j)\ell(Y_i, Y_j). \\
&:= T_{4,1} - T_{4,2} - T_{4,3} + T_{4,4}.
\end{aligned}$$

The first term, $T_{4,1}$, can be factored into a product of two terms, each of which can be computed in quadratic time, as follows:

$$T_{4,1} = \sum_{i_1, j_1} k(X_{i_1}, X_{j_1}) \sum_{i_2, j_2} \ell(Y_{i_2}, Y_{j_2}) = \left(\mathbf{1}_l^T K \mathbf{1}_u\right)\left(\mathbf{1}_l^T L \mathbf{1}_u\right).$$

The next three terms were evaluated in Theorem 16, as

$$T_{4,2} = \mathbf{1}_l^T KL\mathbf{1}_l, \quad T_{4,3} = \mathbf{1}_u^T KL\mathbf{1}_u, \quad \text{and} \quad T_{4,4} = \frac{1}{2}tr\left(KL\right).$$

Together, the above expressions imply the required result. ■

Thus the previous two lemmas imply that the statistic xHSIC$_n$ can be computed in quadratic time. To establish the quadratic computational complexity of $\overline{\text{x}}$HSIC$_n$, it remains to show that s_n^2 can also be computed in quadratic time. We now give an outline of this result, omitting some long but standard calculations.

Lemma 18 *The variance term s_n^2 can be computed in quadratic time.*

Proof outline of Theorem 18. In the first step, we expand the expression of s_n^2 to get the following:

$$\begin{aligned}
s_n^2 &= \frac{4(n-1)}{(n-2)^2}\sum_{i=1}^{n}\left(\frac{1}{n-1}\sum_{j=1}^{j\neq i,n}\langle h_{ij}, f_2\rangle_{k\times\ell} - \text{xHSIC}_n\right)^2 \\
&= \frac{4(n-1)}{(n-2)^2}\sum_{i=1}^{n}\left(\left(\frac{1}{n-1}\sum_{j=1}^{j\neq i,n}\langle h_{ij}, f_2\rangle_{k\times\ell}\right)^2 + \text{xHSIC}_n^2 - 2\text{xHSIC}_n\left(\frac{1}{n-1}\sum_{j=1}^{j\neq i,n}\langle h_{ij}, f_2\rangle_{k\times\ell}\right)\right) \\
&= \frac{4(n-1)}{(n-2)^2(n-1)^2}\sum_{i=1}^{n}\left(\frac{1}{n-1}\sum_{j=1}^{j\neq i,n}\langle h_{ij}, f_2\rangle_{k\times\ell}\right)^2 - \frac{4n(n-1)}{(n-2)^2}\text{xHSIC}_n^2.
\end{aligned}$$

We have already proved that xHSIC$_n$ can be computed with quadratic cost. Excluding the leading factors, the first term can written as

$$\sum_{i=1}^{n}\left(\frac{1}{n-1}\sum_{j=1}^{j\neq i,n}\langle h_{ij}, f_2\rangle_{k\times\ell}\right)^2 = \sum_{i=1}^{n} w_i^2, \quad \text{where} \quad w_i = \frac{1}{n-1}\sum_{j=1}^{j\neq i,n}\langle h_{ij}, f_2\rangle_{k\times\ell}.$$

Thus to complete the proof, it suffices to show that each $\boldsymbol{w} = (w_1, \ldots, w_n)^\top$ can be computed in quadratic time. By expanding the terms involved in the definition of $\boldsymbol{w}$, it can be verified that (with I_n denoting the $n \times n$ identity matrix, and $\mathbf{1}$ denoting the all ones vector of appropriate dimension):

$$\begin{aligned}
&\boldsymbol{w} = I_n \widetilde{\boldsymbol{w}}, \quad \text{where} \\
&2(n-1)\widetilde{\boldsymbol{w}} = n(K \circ L)\mathbf{1} + \frac{1}{2} tr(KL)\mathbf{1} - (KL + LK)\mathbf{1} - (K\mathbf{1}_l) \circ (L\mathbf{1}_l) - \frac{1}{2n}\mathbf{1}_l^T KL\mathbf{1}_l\mathbf{1} \\
&\quad + \frac{1}{2n}\left((\mathbf{1}^T L\mathbf{1})K\mathbf{1}_l + (\mathbf{1}^T K\mathbf{1})L\mathbf{1}_l\right).
\end{aligned}$$

Note that every term involved in the definition of $\widetilde{\boldsymbol{w}}$, and hence in the definition of $\boldsymbol{w}$, can be computed in quadratic time. This in turn, implies that the variance term s_n^2 can also be computed in quadratic time as required.

Appendix B. Testing Linear Dependence (Section 4)

In this appendix, we collect proofs of the results on linear kernels.

B.1 Proof of Theorem 4

Both results in Theorem 4 rely on some simplified representation of the $\overline{\mathrm{x}}\mathrm{HSIC}_n$ statistic for linear kernels in one dimension. We present the details of the common steps here, before moving on to the specifics of the proof of part (a) in Appendix B.1.1 and of part (b) in Appendix B.1.2.

Since xHSIC_n is location-invariant, we may assume $\mathbb{E}[X] = \mathbb{E}[Y] = 0$ without loss of generality, and we assume this throughout the proof. We first note that xHSIC_n can be written as:

$$\begin{aligned}
\mathrm{xHSIC}_n &= f_2 \cdot \left(\frac{1}{n}\sum_{i=1}^n X_iY_i - \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} X_iY_j \right) \\
&= f_2 \cdot \left(\frac{1}{n}\sum_{i=1}^n X_iY_i - \frac{1}{n(n-1)} \left\{ \left(\sum_{i=1}^n X_iY_i \right)^2 - \sum_{i=1}^n X_i^2Y_i^2 \right\} \right). \qquad (14)
\end{aligned}$$

On the other hand, the squared denominator of the studentized statistic is

$$s_n^2 = f_2^2 \cdot (\mathrm{I}_n - \mathrm{II}_n),$$

where

$$\mathrm{I}_n = \frac{1}{(n-1)(n-2)^2} \sum_{i=1}^n \left(\sum_{j=1}^{n, j \ne i} (X_i - X_j)(Y_i - Y_j) \right)^2 \quad \text{and} \qquad (15)$$

$$\mathrm{II}_n = \frac{4n(n-1)}{(n-2)^2} \left(\frac{1}{n}\sum_{i=1}^n X_iY_i - \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} X_iY_j \right)^2. \qquad (16)$$

More specifically, recall from (7) that s_n^2 can be expressed as

$$s_n^2 \; = \; \frac{4}{(n-1)(n-2)^2} \sum_{i=1}^{n} \left(\sum_{j=1}^{n,j\neq i} \langle h(Z_i, Z_j), f_2 \rangle \right)^2 - \frac{4n(n-1)}{(n-2)^2} \mathrm{xHSIC}_n^2.$$

Since $\langle h(Z_i, Z_j), f_2 \rangle = \frac{1}{2}(X_i - X_j)(Y_i - Y_j) \cdot f_2$ for the linear kernel, we see that

$$\frac{4}{(n-1)(n-2)^2} \sum_{i=1}^{n} \left(\sum_{j=1}^{n,j\neq i} \langle h(Z_i, Z_j), f_2 \rangle \right)^2 = f_2^2 \cdot \mathrm{I}_n.$$

Similarly, using expression (14), we see that

$$\frac{4n(n-1)}{(n-2)^2} \mathrm{xHSIC}_n^2 \; = \; f_2^2 \cdot \mathrm{II}_n.$$

Combining the above two expressions yields $s_n^2 = f_2^2 \cdot (\mathrm{I}_n - \mathrm{II}_n)$, and consequently we have

$$\overline{\mathrm{x}}\mathrm{HSIC}_n = \frac{\sqrt{n}\mathrm{xHSIC}_n}{s_n} = \mathrm{sign}(f_2) \times \frac{\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^n X_i Y_i - \frac{1}{n(n-1)} \sum_{1\leq i\neq j \leq n} X_i Y_j \Big)}{\sqrt{\mathrm{I}_n - \mathrm{II}_n}}. \tag{17}$$

B.1.1 Proof of part (a)

Having presented a simplified representation of the $\overline{\mathrm{x}}\mathrm{HSIC}_n$ with linear kernels, let us begin proving the pointwise asymptotic result in part (a) of Theorem 4. Under the finite second moment condition in the theorem statement, notice that

$$\begin{aligned} \mathbb{E}\bigg[\bigg(\frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} X_i Y_j \bigg)^2 \bigg] \; &\lesssim \; \frac{1}{n^2(n-1)^2} \sum_{1\leq i\neq j \leq n} \big\{ \mathbb{E}[X_i^2 Y_i^2] + \mathbb{E}[X_i Y_i] \mathbb{E}[X_j Y_j] \big\} \\ &\lesssim \; \frac{\mathbb{E}[X^2 Y^2]}{n^2} \to 0, \end{aligned}$$

where the last inequality uses Jensen's inequality. Thus an application of Markov's inequality verifies that $\frac{1}{n(n-1)} \sum_{1\leq i \neq j \leq n} X_i Y_j = O_P(n^{-1}) = o_P(1)$. We also know by the weak law of large numbers that $\frac{1}{n}\sum_{i=1}^n X_i Y_i - \mathbb{E}[XY] = o_P(1)$ and as a result, we have

$$\mathrm{II}_n = \frac{4n(n-1)}{(n-2)^2} \Big(\mathbb{E}[XY] + o_P(1) \Big)^2 = 4\{\mathbb{E}[XY]\}^2 + o_P(1).$$

To control I_n, note that

$$\begin{aligned} &(n-1)(n-2)^2 \mathrm{I}_n \\ = \;& \sum_{i=1}^{n} \left(\sum_{j=1}^{n,j\neq i} (X_i - X_j)(Y_i - Y_j) \right)^2 \\ = \;& \sum_{1\leq i\neq j\leq n} (X_i - X_j)^2 (Y_i - Y_j)^2 + \sum_{\substack{1\leq i,j,q \leq n \\ i,j,q \text{ distinct}}} (X_i - X_j)(Y_i - Y_j)(X_i - X_q)(Y_i - Y_q). \end{aligned}$$

We then apply the law of large numbers for U-statistics (Hoeffding, 1961) for each term above, and observe

$$A_n := \frac{1}{(n-1)(n-2)^2} \sum_{1 \leq i \neq j \leq n} (X_i - X_j)^2 (Y_i - Y_j)^2 \xrightarrow{p} 0 \quad \text{and} \tag{18}$$

$$B_n := \frac{1}{(n-1)(n-2)^2} \sum_{\substack{1 \leq i,j,q \leq n \\ i,j,q \text{ distinct}}} (X_i - X_j)(Y_i - Y_j)(X_i - X_q)(Y_i - Y_q) \xrightarrow{p} \mathbb{V}[XY] + 4\{\mathbb{E}[XY]\}^2, \tag{19}$$

which hold under the finite second moment assumption. This further implies that $\mathrm{I}_n - \mathrm{II}_n = \mathbb{V}[XY] + o_P(1)$.

As verified before, it holds that $\frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} X_i Y_j = O_P(n^{-1})$ and thus Slutsky's theorem together with the central limit theorem gives

$$\frac{\sqrt{n}\big(\frac{1}{n}\sum_{i=1}^n \{X_i Y_i - \mathbb{E}[XY]\} - \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} X_i Y_j\big)}{\sqrt{\mathbb{V}[XY]}} \xrightarrow{d} N(0,1).$$

Moreover since $\mathrm{I}_n - \mathrm{II}_n = \mathbb{V}[XY] + o_P(1)$, the continuous mapping theorem along with Slutsky's theorem establishes

$$\frac{\sqrt{n}\big(\frac{1}{n}\sum_{i=1}^n \{X_i Y_i - \mathbb{E}[XY]\} - \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} X_i Y_j\big)}{\sqrt{\mathrm{I}_n - \mathrm{II}_n}} \xrightarrow{d} N(0,1).$$

By symmetry with f_1 and recalling $\rho = \mathbb{E}[XY]/\mathbb{V}^{1/2}[XY]$, it can be seen that f_2 satisfies

$$\begin{aligned} \mathbb{P}\big(\mathrm{sign}(f_2) > 0\big) &= \mathbb{P}\left(\frac{\frac{1}{\sqrt{n}}\sum_{i=n+1}^{2n} \{X_i Y_i - \mathbb{E}[XY]\} - \frac{1}{\sqrt{n}(n-1)} \sum_{n+1 \leq i \neq j \leq 2n} X_i Y_j}{\sqrt{\mathbb{V}[XY]}} > -\sqrt{n}\rho \right) \\ &= \Phi\big(\sqrt{n}\rho\big) + o(1). \end{aligned}$$

Similarly, observe that

$$\mathbb{P}\big(\mathrm{sign}(f_2) < 0\big) = \Phi\big(-\sqrt{n}\rho\big) + o(1) \quad \text{and} \quad \mathbb{P}\big(\mathrm{sign}(f_2) = 0\big) = o(1).$$

Combining all the pieces together,

$$\begin{aligned} \mathbb{P}(\overline{\mathrm{x}}\mathrm{HSIC}_n > z_{1-\alpha}) &= \mathbb{P}\big(\mathrm{sign}(f_2) > 0\big)\mathbb{P}\left(\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n \{X_i Y_i - \mathbb{E}[XY]\}}{\sqrt{\mathbb{V}[XY]}} > z_{1-\alpha} - \sqrt{n}\rho \right) \\ &\quad + \mathbb{P}\big(\mathrm{sign}(f_2) \leq 0\big)\mathbb{P}\left(-\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n \{X_i Y_i - \mathbb{E}[XY]\}}{\sqrt{\mathbb{V}[XY]}} > z_{1-\alpha} + \sqrt{n}\rho \right) + o(1) \\ &= \Phi(-\sqrt{n}\rho)\Phi(z_\alpha - \sqrt{n}\rho) + \Phi(\sqrt{n}\rho)\Phi(z_\alpha + \sqrt{n}\rho) + o(1). \end{aligned}$$

This proves the first part of Theorem 4.

B.1.2 Proof of part (b)

In this subsection, we make the asymptotic guarantee hold uniformly over the family of distributions satisfying condition (9). Let r_n be a sequence of positive numbers. Throughout this subsection, we say that a random sequence $X_n = o_{\mathcal{P}}(r_n)$ for a family of distributions $\mathcal{P}$ if $\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \mathbb{P}_P(r_n^{-1}|X_n| > \epsilon) = 0$ for all $\epsilon > 0$.

We first collect several lemmas that will be used in the main body of the proof. The proofs of these results can be found in Appendix B.4.

Lemma 19 *For $\mathcal{P}_n$ satisfying condition* (9)*, we have*

$$\frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n} \frac{X_iY_j}{\sqrt{\mathbb{V}_P[XY]}} = o_{\mathcal{P}_n}(n^{-3/2}).$$

Lemma 20 *For $\mathcal{P}_n$ satisfying condition* (9) *and any fixed $\varepsilon \in (0, 1/2]$, we have*

$$\frac{1}{n}\sum_{i=1}^{n} \frac{X_iY_i - \mathbb{E}_P[XY]}{\sqrt{\mathbb{V}_P[XY]}} = o_{\mathcal{P}_n}(n^{-1/2+\varepsilon}).$$

Lemma 21 *For $\mathcal{P}_n$ satisfying condition* (9)*, we have*

$$\frac{1}{(n-1)(n-2)^2} \sum_{1\leq i\neq j\leq n} \frac{(X_i - X_j)^2(Y_i - Y_j)^2}{\mathbb{V}_P[XY]} = o_{\mathcal{P}_n}(1).$$

Lemma 22 *For $\mathcal{P}_n$ satisfying condition* (9)*, we have*

$$\begin{aligned}&\frac{1}{(n-1)(n-2)^2} \sum_{\substack{1\leq i,j,q\leq n\\ i,j,q \text{ distinct}}} \frac{(X_i - X_j)(Y_i - Y_j)(X_i - X_q)(Y_i - Y_q)}{\mathbb{V}_P[XY]}\\ =&1 + 4\left\{\frac{\mathbb{E}_P[XY]}{\sqrt{\mathbb{V}_P[XY]}}\right\}^2 + o_{\mathcal{P}_n}(1).\end{aligned}$$

Lemma 23 *Let V_n be a random sequence defined as*

$$V_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{X_iY_i - \mathbb{E}_P[XY]}{\sqrt{\mathbb{V}_P[XY]}}\right\}.$$

For $\mathcal{P}_n$ satisfying condition (9)*, we have the uniform central limit theorem as*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_n} \sup_{t\in\mathbb{R}} |\mathbb{P}_P(V_n \leq t) - \Phi(t)| = 0.$$

The following lemma corresponds to Lemma 20 of Shah and Peters (2020) with a slight modification for our purpose.

Lemma 24 *Let $\mathcal{P}$ be a family of distributions determining the laws of random sequences V_n, W_n and R_n. Suppose that*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{t\in\mathbb{R}} |\mathbb{P}_P(V_n \le t) - \Phi(t)| = 0.$$

Then we have the following.

(a) If $R_n = o_{\mathcal{P}}(1)$, we have

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{t\in\mathbb{R}} |\mathbb{P}_P(V_n + R_n \le t) - \Phi(t)| = 0.$$

(b) If $W_n = 1 + o_{\mathcal{P}}(1)$, we have

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{t\in\mathbb{R}} |\mathbb{P}_P(V_n/W_n \le t) - \Phi(t)| = 0.$$

(c) Fix $t \in \mathbb{R}$. If $W_n = 1 + o_{\mathcal{P}}(1)$, we have

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{a\in\mathbb{R}} |\mathbb{P}_P\big((V_n - a)/W_n \le t\big) - \Phi(t + a)| = 0.$$

Main body of the proof of part (b). Moving to the main proof of part (b), Theorem 19 and Theorem 20 imply that by letting $\varepsilon = 1/8 \in (0, 1/2)$,

$$\frac{\mathrm{II}_n}{\sqrt{\mathbb{V}_P[XY]}} = 4\{1 + o(1)\} \cdot \left\{ \frac{\mathbb{E}_P[XY]}{\sqrt{\mathbb{V}_P[XY]}} + o_{\mathcal{P}_n}(n^{-1/2+\varepsilon}) \right\}^2 = 4\left\{ \frac{\mathbb{E}_P[XY]}{\sqrt{\mathbb{V}_P[XY]}} \right\}^2 + o_{\mathcal{P}_n}(1).$$

This follows since

$$\sup_{P\in\mathcal{P}_n} n^{-3/8} \left| \frac{\mathbb{E}_P[XY]}{\sqrt{\mathbb{V}_P[XY]}} \right| \le \sup_{P\in\mathcal{P}_n} \frac{n^{1/4}}{n^{3/8}} \left\{ \frac{\mathbb{E}_P[X^4Y^4]}{n\mathbb{V}_P^2[XY]} \right\}^{1/4} = o(1),$$

under condition (9), and consequently

$$\frac{\mathbb{E}_P[XY]}{\sqrt{\mathbb{V}_P[XY]}} \cdot o_{\mathcal{P}_n}(n^{-1/2+1/8}) = o_{\mathcal{P}_n}(1).$$

In addition, by Theorem 21 and Theorem 22,

$$\frac{\mathrm{I}_n}{\sqrt{\mathbb{V}_P[XY]}} = 1 + 4\left\{ \frac{\mathbb{E}_P[XY]}{\sqrt{\mathbb{V}_P[XY]}} \right\}^2 + o_{\mathcal{P}_n}(1),$$

and combining the pieces yields

$$\frac{\mathrm{I}_n - \mathrm{II}_n}{\mathbb{V}_P[XY]} = 1 + o_{\mathcal{P}_n}(1).$$

Moreover the uniform continuous mapping theorem (Lundborg et al., 2022, Lemma 15) shows that

$$\sqrt{\frac{\mathrm{I}_n - \mathrm{II}_n}{\mathbb{V}_P[XY]}} = 1 + o_{\mathcal{P}_n}(1). \tag{20}$$

Now let us write

$$V_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{X_i Y_i - \mathbb{E}_P[XY]}{\sqrt{\mathbb{V}_P[XY]}} \right\}, \quad R_n := -\frac{1}{\sqrt{n}(n-1)} \sum_{1 \leq i \neq j \leq n} \frac{X_i Y_j}{\sqrt{\mathbb{V}_P[XY]}},$$

$$a_n := \sqrt{n}\rho_n = \frac{\sqrt{n}\mathbb{E}_P[XY]}{\sqrt{\mathbb{V}_P[XY]}} \quad \text{and} \quad W_n := \sqrt{\frac{\mathrm{I}_n - \mathrm{II}_n}{\mathbb{V}_P[XY]}},$$

and recall that $W_n = 1 + o_{\mathcal{P}_n}(1)$ as in (20) and $R_n = o_{\mathcal{P}_n}(1)$ by Theorem 19. Then Theorem 24 combined with Theorem 23 and Theorem 24 proves that

$$\lim_{n\to\infty} \sup_{P \in \mathcal{P}_n} \big| \mathbb{P}_P\big((V_n + R_n + a_n)/W_n > z_{1-\alpha}\big) - \Phi(z_\alpha + a_n)\big| = 0 \quad \text{and} \tag{21}$$

$$\lim_{n\to\infty} \sup_{P \in \mathcal{P}_n} \big| \mathbb{P}_P\big((V_n + R_n + a_n)/W_n < -z_{1-\alpha}\big) - \Phi(z_\alpha - a_n)\big| = 0.$$

We also note by Theorem 23 that

$$\lim_{n\to\infty} \sup_{P \in \mathcal{P}_n} \big| \mathbb{P}_P\big(f_2 > 0\big) - \Phi(a_n)\big| = 0, \quad \lim_{n\to\infty} \sup_{P \in \mathcal{P}_n} \big| \mathbb{P}_P\big(f_2 < 0\big) - \Phi(-a_n)\big| = 0 \tag{22}$$

$$\text{and} \quad \lim_{n\to\infty} \sup_{P \in \mathcal{P}_n} \mathbb{P}_P\big(f_2 = 0\big) = 0.$$

Finally, by an alternative expression of $\overline{\mathrm{x}}\mathrm{HSIC}_n$ given in (17), that is

$$\overline{\mathrm{x}}\mathrm{HSIC}_n = \mathrm{sign}(f_2) \times \frac{V_n + R_n + \sqrt{n}\rho_n}{W_n},$$

and by using the triangle inequality, we have

$$\begin{aligned} & \lim_{n\to\infty} \sup_{P \in \mathcal{P}_n} \big| \mathbb{P}_P\big(\overline{\mathrm{x}}\mathrm{HSIC}_n > z_{1-\alpha}\big) - \Phi(-\sqrt{n}\rho_n)\Phi(z_\alpha - \sqrt{n}\rho_n) - \Phi(\sqrt{n}\rho_n)\Phi(z_\alpha + \sqrt{n}\rho_n)\big| \\ \leq \; & \lim_{n\to\infty} \sup_{P \in \mathcal{P}_n} \big| \mathbb{P}_P(f_2 > 0)\mathbb{P}_P\big((V_n + R_n + \sqrt{n}\rho_n)/W_n > z_{1-\alpha}\big) - \Phi(\sqrt{n}\rho_n)\Phi(z_\alpha + \sqrt{n}\rho_n)\big| \\ & + \lim_{n\to\infty} \sup_{P \in \mathcal{P}_n} \big| \mathbb{P}_P(f_2 < 0)\mathbb{P}_P\big((V_n + R_n + \sqrt{n}\rho_n)/W_n < -z_{1-\alpha}\big) - \Phi(-\sqrt{n}\rho_n)\Phi(z_\alpha - \sqrt{n}\rho_n)\big| \\ & + \lim_{n\to\infty} \sup_{P \in \mathcal{P}_n} \mathbb{P}_P\big(f_2 = 0\big). \end{aligned}$$

The upper bound can be shown to be zero by the preliminary approximations (21) and (22). Therefore the desired result follows and we complete the proof of Theorem 4.

B.2 Uniform asymptotic normality with finite $2+\delta$ moments

For a fixed $\delta \in (0,2]$, consider a family of distributions $\mathcal{P}_{n,\delta}$ such that

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_{n,\delta}} \frac{\mathbb{E}_P\big[|X-\mathbb{E}_P[X]|^{2+\delta}\big]}{n^{\delta/4}\mathbb{V}_P[X]^{1+\delta/2}} = 0 \quad \text{and} \quad \lim_{n\to\infty} \sup_{P\in\mathcal{P}_{n,\delta}} \frac{\mathbb{E}_P\big[|Y-\mathbb{E}_P[Y]|^{2+\delta}\big]}{n^{\delta/4}\mathbb{V}_P[Y]^{1+\delta/2}} = 0.$$

In this subsection, we show that $\overline{\mathrm{x}}\mathrm{HSIC}_n$ is asymptotically $N(0,1)$ uniformly over the class $\mathcal{P}_{n,\delta}^{(0)}$ given as

$$\mathcal{P}_{n,\delta}^{(0)} = \big\{P_{XY} \in \mathcal{P}_{n,\delta} : P_{XY} = P_X \times P_Y\big\}.$$

In particular, we will show that

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_{n,\delta}^{(0)}} \sup_{t\in\mathbb{R}} \big|\mathbb{P}_P(\overline{\mathrm{x}}\mathrm{HSIC}_n \le t) - \Phi(t)\big| = 0. \tag{23}$$

Notice that any $P \in \mathcal{P}_{n,\delta}^{(0)}$ satisfies

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_{n,\delta}^{(0)}} \frac{\mathbb{E}_P\big[|X-\mathbb{E}_P(X)|^{2+\delta}\cdot|Y-\mathbb{E}_P(Y)|^{2+\delta}\big]}{n^{\frac{\delta}{2}}\big\{\mathbb{V}_P\big[\big(X-\mathbb{E}_P(X)\big)\big(Y-\mathbb{E}_P(Y)\big)\big]\big\}^{2+\delta}} = 0. \tag{24}$$

Since $\overline{\mathrm{x}}\mathrm{HSIC}_n$ is both location-invariant and scale-invariant, we may assume that $\mathbb{E}_P[X] = \mathbb{E}_P[Y] = 0$ and $\mathbb{V}_P[X] = \mathbb{V}_P[Y] = 1$ without loss generality (note that we have $\mathbb{V}_P[XY] = \mathbb{V}_P[X]\mathbb{V}_P[Y] = 1$ under the null when X and Y are centered). We therefore assume that X and Y are standardized throughout this proof.

For a triangular array of random variables with a distribution $P \in \mathcal{P}_{n,\delta}^{(0)}$, the uniform Lyapunov central limit theorem verifies that

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_{n,\delta}^{(0)}} \sup_{t\in\mathbb{R}} \left|\mathbb{P}_P\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_iY_i \le t\right) - \Phi(t)\right| = 0,$$

which can be shown similarly as Theorem 23. We also note that $\lim_{n\to\infty}\sup_{P\in\mathcal{P}_{n,\delta}^{(0)}} \mathbb{E}_P[|X^2Y^2|^{1+\delta/2}]n^{-\delta/2} = 0$ by condition (24). Thus Lemma 17 of Lundborg et al. (2022) yields

$$\frac{1}{n}\sum_{i=1}^n X_i^2Y_i^2 - 1 = o_{\mathcal{P}_{n,\delta}^{(0)}}(1),$$

which implies that $\frac{1}{n^{3/2}}\sum_{i=1}^n X_i^2Y_i^2 = o_{\mathcal{P}_{n,\delta}^{(0)}}(1)$. Therefore an application of Theorem 24 yields that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n X_iY_i - \frac{1}{n(n-1)}\sum_{1\le i\neq j\le n} X_iY_j\right)$$

is asymptotically $N(0,1)$ uniformly over $\mathcal{P}_{n,\delta}^{(0)}$. Following the same logic, we also have $\sqrt{n}f_2$ is asymptotically $N(0,1)$ uniformly over $\mathcal{P}_{n,\delta}^{(0)}$.

Now let us turn to the terms I_n and II_n where I_n and II_n are recalled in (15) and (16), respectively. Note that the second term satisfies

$$\mathrm{II}_n = O(1) \cdot f_1^2.$$

We already showed that $\sqrt{n}f_1$ is asymptotically $N(0,1)$ over $\mathcal{P}_{n,\delta}^{(0)}$ and thus $\mathrm{II}_n = o_{\mathcal{P}_{n,\delta}^{(0)}}(1)$. For the first term, note that

$$\begin{aligned}
\mathrm{I}_n \;&=\; \{1+o(1)\} \cdot \frac{1}{n^3}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n (X_i - X_j)(Y_i - Y_j)(X_i - X_k)(Y_i - Y_k) \\
&=\; \{1+o(1)\} \cdot \Bigg\{ \frac{1}{n}\sum_{i=1}^n X_i^2Y_i^2 + O(1)\bigg(\frac{1}{n}\sum_{i=1}^n X_i^2 Y_i\bigg)\bigg(\frac{1}{n}\sum_{j=1}^n Y_j\bigg) \\
&\qquad + O(1)\bigg(\frac{1}{n}\sum_{i=1}^n X_i Y_i^2\bigg)\bigg(\frac{1}{n}\sum_{j=1}^n X_j\bigg) + O(1)\bigg(\frac{1}{n}\sum_{i=1}^n X_i^2\bigg)\bigg(\frac{1}{n}\sum_{j=1}^n Y_j\bigg)^2 \\
&\qquad + O(1)\bigg(\frac{1}{n}\sum_{i=1}^n Y_i^2\bigg)\bigg(\frac{1}{n}\sum_{j=1}^n X_j\bigg)^2 \Bigg\}.
\end{aligned}$$

Under the condition on $\mathcal{P}_{n,\delta}^{(0)}$, it can be checked that

$$\sup_{P\in\mathcal{P}_{n,\delta}^{(0)}} \mathbb{E}_P[|X|^{1+\delta/2}]n^{-\delta/2} = o(1), \quad \sup_{P\in\mathcal{P}_{n,\delta}^{(0)}} \mathbb{E}_P[|Y|^{1+\delta/2}]n^{-\delta/2} = o(1), \tag{25}$$

$$\sup_{P\in\mathcal{P}_{n,\delta}^{(0)}} \mathbb{E}_P[|X^2Y|^{1+\delta/2}]n^{-\delta/2} = o(1) \quad \text{and} \quad \sup_{P\in\mathcal{P}_{n,\delta}^{(0)}} \mathbb{E}_P[|XY^2|^{1+\delta/2}]n^{-\delta/2} = o(1). \tag{26}$$

For instance, by Cauchy–Schwarz inequality, we see that

$$\sup_{P\in\mathcal{P}_{n,\delta}^{(0)}} \mathbb{E}_P[|X^2Y|^{1+\delta/2}]n^{-\delta/2} \le \sup_{P\in\mathcal{P}_{n,\delta}^{(0)}} \sqrt{\mathbb{E}_P[|X^2Y^2|^{1+\delta/2}]n^{-\delta/2}}\sqrt{\mathbb{E}_P[|X^2|^{1+\delta/2}]n^{-\delta/2}} \to 0. \tag{27}$$

Thus Lemma 17 of Lundborg et al. (2022) yields

$$\frac{1}{n}\sum_{i=1}^n X_i^2 Y_i = o_{\mathcal{P}_{n,\delta}^{(0)}}(1), \; \frac{1}{n}\sum_{i=1}^n X_i Y_i^2 = o_{\mathcal{P}_{n,\delta}^{(0)}}(1), \tag{28}$$

$$\frac{1}{n}\sum_{i=1}^n X_i = o_{\mathcal{P}_{n,\delta}^{(0)}}(1) \;\text{ and }\; \frac{1}{n}\sum_{i=1}^n Y_i = o_{\mathcal{P}_{n,\delta}^{(0)}}(1). \tag{29}$$

These approximations verify that $\mathrm{I}_n = 1 + o_{\mathcal{P}_{n,\delta}^{(0)}}(1)$, and thus $\mathrm{I}_n - \mathrm{II}_n = 1 + o_{\mathcal{P}_{n,\delta}^{(0)}}(1)$.

Having all the ingredients, we may follow the proof of part (b) of Theorem 4 and show the uniform normality result (23) as desired.

B.3 Details of Example 1

Based on our previous discussion in (17), $\overline{\mathrm{x}}\mathrm{HSIC}_n$ with linear kernels can be expressed as

$$\overline{\mathrm{x}}\mathrm{HSIC}_n = \mathrm{sign}(f_2)\frac{\sqrt{n}f_1}{\sqrt{\mathrm{I}_n - \mathrm{II}_n}},$$

where I_n and II_n can be found in (15) and (16), respectively. Moreover, since $\mathrm{I}_n - \mathrm{II}_n \geq 0$, we have

$$\mathbb{P}(\overline{\mathrm{x}}\mathrm{HSIC}_n \leq 0) = \mathbb{P}\big(\mathrm{sign}(f_2) \cdot nf_1 \leq 0\big).$$

To approximate this probability to the target probability in Example 1, note that

$$nf_1 = \sum_{i=1}^{n}(X_i - p_n)(Y_i - p_n) - \frac{1}{n-1}\sum_{1\leq i\neq j\leq n}(X_i - p_n)(Y_j - p_n)$$

by the location-invariance property. We also note that since

$$\mathbb{V}\bigg(\frac{1}{n-1}\sum_{1\leq i\neq j\leq n}(X_i - p_n)(Y_j - p_n)\bigg) = \frac{n}{n-1}p_n^2(1-p_n)^2 \to 0,$$

an application of Chebyshev's inequality ensures that

$$nf_1 = \sum_{i=1}^{n}(X_i - p_n)(Y_i - p_n) + o_P(1).$$

A similar argument shows that $p_n\sum_{i=1}^{n}(X_i - p_n) = o_P(1)$ and $p_n\sum_{i=1}^{n}(Y_i - p_n) = o_P(1)$. Therefore

$$nf_1 = \sum_{i=1}^{n}X_iY_i - np_n^2 + o_P(1).$$

Since X_iY_is are i.i.d. Bernoulli random variables with parameter p_n^2 satisfying $np_n^2 = \lambda > 0$, Poisson limit theorem (Theorem 3.6.1 of Durrett, 2019) together with Slutsky's theorem yields

$$nf_1 \xrightarrow{d} \mathrm{Poisson}(\lambda) - \lambda.$$

Similarly it also follows that $\mathrm{sign}(f_2) = \mathrm{sign}(nf_2) \xrightarrow{d} \mathrm{sign}(\mathrm{Poisson}(\lambda) - \lambda)$ by the continuous mapping theorem. Strictly speaking, the sign function is discontinuous at $x = 0$, and thus the continuous mapping theorem does not directly apply. However the same conclusion follows since $\mathbb{P}(\mathrm{Poisson}(\lambda) - \lambda = 0) = 0$, i.e., the set of discontinuity points has zero probability, as λ is a non-integer by our assumption. Combining the pieces and observing that f_1 and f_2 are independent, we therefore conclude that $\mathrm{sign}(f_2) \cdot nf_1 \xrightarrow{d} \mathrm{sign}(V') \times V$ where V, V' are i.i.d. centered Poisson random variables. Lastly, the distribution of $\mathrm{sign}(V') \times V$ is continuous at 0 since λ is not an integer. Hence the definition of convergence in distribution implies the desired result.

B.4 Proof of Auxiliary Lemmas

This subsection collects the proofs of lemmas used for part (b) of Theorem 4.

B.4.1 Proof of Theorem 19

Similarly as before in Appendix B.1.1, it can be shown that

$$\begin{aligned}\mathbb{E}_P\left[\left(\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\frac{X_iY_j}{\sqrt{\mathbb{V}_P[XY]}}\right)^2\right] &\lesssim \frac{1}{n^2(n-1)^2}\sum_{1\le i\neq j\le n}\left\{\frac{\mathbb{E}_P[X_i^2Y_i^2]}{\mathbb{V}_P[XY]}+\frac{\mathbb{E}_P[X_iY_i]\mathbb{E}_P[X_jY_j]}{\mathbb{V}_P[XY]}\right\}\\ &\lesssim \frac{\mathbb{E}_P[X^2Y^2]}{n^2\mathbb{V}_P[XY]}\lesssim\frac{1}{n^{3/2}}\sqrt{\frac{\mathbb{E}_P[X^4Y^4]}{n\mathbb{V}_P^2[XY]}}.\end{aligned}$$

Thus Markov's inequality yields

$$\sup_{P\in\mathcal{P}_n}\mathbb{P}_n\left(n^{3/4}\left|\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\frac{X_iY_j}{\sqrt{\mathbb{V}_P[XY]}}\right|\ge t\right)\lesssim\frac{1}{t^2}\sqrt{\sup_{P\in\mathcal{P}_n}\frac{\mathbb{E}_P[X^4Y^4]}{n\mathbb{V}_P^2[XY]}}\to 0,$$

where the last convergence result holds under condition (9). Hence the result follows.

B.4.2 Proof of Theorem 20

The result follows by Chebyshev's inequality using

$$\sup_{P\in\mathcal{P}_n}\mathbb{E}_P\left[\left(\frac{1}{n}\sum_{i=1}^n\frac{X_iY_i-\mathbb{E}_P[XY]}{\sqrt{\mathbb{V}_P[XY]}}\right)^2\right]=\frac{1}{n}.$$

B.4.3 Proof of Theorem 21

Note that condition (9) guarantees that

$$\begin{aligned}&\sup_{P\in\mathcal{P}_n}\mathbb{E}_P\left[\frac{1}{(n-1)(n-2)^2}\sum_{1\le i\neq j\le n}\frac{(X_i-X_j)^2(Y_i-Y_j)^2}{\mathbb{V}_P[XY]}\right]\\ \lesssim\ &\sup_{P\in\mathcal{P}_n}\frac{1}{\sqrt{n}}\sqrt{\frac{\max\left\{\mathbb{E}_P[X^4Y^4],\,\mathbb{E}_P[X^4]\mathbb{E}_P[Y^4]\right\}}{n\mathbb{V}_P^2[XY]}}=o(1),\end{aligned}$$

where the last convergence result holds under condition (9). Hence the result follows by Markov's inequality.

B.4.4 Proof of Theorem 22

We first note by the law of total expectation that

$$\mathbb{E}_P[(X_i-X_j)(Y_i-Y_j)(X_i-X_q)(Y_i-Y_q)]=\mathbb{V}_P[XY]+4\{\mathbb{E}_P[XY]\}^2.$$

Let

$$U_n=\frac{1}{n(n-1)(n-2)}\sum_{\substack{1\le i,j,q\le n\\ i,j,q\text{ distinct}}}\frac{(X_i-X_j)(Y_i-Y_j)(X_i-X_q)(Y_i-Y_q)}{\mathbb{V}_P[XY]},$$

and notice that U_n is a U-statistic. Then using an upper bound for the variance of U-statistics (Lee, 2019, Chapter 1.3),

$$\begin{aligned}
\mathbb{V}_P[U_n] &\lesssim \frac{1}{n}\mathbb{V}_P\left[\frac{(X_1-X_2)(Y_1-Y_2)(X_1-X_3)(Y_1-Y_3)}{\mathbb{V}_P[XY]}\right] \\
&\lesssim \frac{1}{n}\mathbb{E}_P\left[\frac{(X_1-X_2)^2(Y_1-Y_2)^2(X_1-X_3)^2(Y_1-Y_3)^2}{\mathbb{V}_P^2[XY]}\right] \\
&\lesssim \frac{\max\left\{\mathbb{E}_P[X^4Y^4],\, \mathbb{E}_P[X^4]\mathbb{E}_P[Y^4]\right\}}{n\mathbb{V}_P^2[XY]}.
\end{aligned}$$

Therefore Chebyshev's inequality yields $U_n = 1+4\{\mathbb{E}_P[XY]/\sqrt{\mathbb{V}_P[XY]}\}^2+o_{\mathcal{P}_n}(1)$. Finally, since $\frac{n}{n-2} = 1+O(n^{-1})$,

$$\begin{aligned}
&\frac{1}{(n-1)(n-2)^2}\sum_{\substack{1\leq i,j,q\leq n\\ i,j,q \text{ distinct}}} \frac{(X_i-X_j)(Y_i-Y_j)(X_i-X_q)(Y_i-Y_q)}{\mathbb{V}_P[XY]} \\
= \ &\frac{n}{n-2}\left[1+4\left\{\frac{\mathbb{E}_P[XY]}{\sqrt{\mathbb{V}_P[XY]}}\right\}^2\right]+o_{\mathcal{P}_n}(1) = 1+4\left\{\frac{\mathbb{E}_P[XY]}{\sqrt{\mathbb{V}_P[XY]}}\right\}^2+o_{\mathcal{P}_n}(1).
\end{aligned}$$

B.4.5 Proof of Theorem 23

This proof follows the exact same lines of the proof of Shah and Peters (2020, Lemma 18) combined with the Lyapunov central limit theorem for triangular arrays. Hence we omit the details.

B.4.6 Proof of Theorem 24

We only need to prove part (c) as the proofs of part (a) and part (b) are given in Lemma 20 of Shah and Peters (2020). For a given $\epsilon > 0$ and a fixed $t \in \mathbb{R}$, choose $0 < \delta \leq \min\left\{\frac{1}{2}, \frac{\sqrt{2\pi}}{3|t|}\epsilon\right\}$ and N such that for all $n \geq N$,

$$\sup_{P\in\mathcal{P}}\sup_{t\in\mathbb{R}}|\mathbb{P}_P(V_n\leq t)-\Phi(t)| < \epsilon/3 \quad \text{and} \quad \sup_{P\in\mathcal{P}}\mathbb{P}_P(|W_n-1|>\delta) < \epsilon/3. \tag{30}$$

Then for all $n \geq N$, for all $P \in \mathcal{P}$ and for all $a \in \mathbb{R}$,

$$\begin{aligned}
&\mathbb{P}_P\big((V_n-a)/W_n\leq t\big)-\Phi(t+a) \\
\leq \ &\mathbb{P}_P\big(V_n\leq t(1+\delta)+a\big)-\Phi(t+a)+\mathbb{P}_P(|W_n-1|>\delta) \\
\leq \ &\mathbb{P}_P\big(V_n\leq t(1+\delta)+a\big)-\Phi\big(t(1+\delta)+a\big)+\Phi\big(t(1+\delta)+a\big)-\Phi(t+a)+\epsilon/3 \\
\overset{\text{(i)}}{\leq} \ &\frac{2\epsilon}{3}+\big|\Phi\big(t(1+\delta)+a\big)-\Phi(t+a)\big| \\
\overset{\text{(ii)}}{\leq} \ &\epsilon,
\end{aligned}$$

where step (i) uses condition (30), and step (ii) uses the fact that Φ is $1/\sqrt{2\pi}$-Lipschitz and our condition on δ. An analogous argument may show that

$$\mathbb{P}_P\big((V_n - a)/W_n \le t\big) - \Phi(t+a) \ge -\epsilon,$$

for all $n \ge N$, for all $P \in \mathcal{P}$ and for all $a \in \mathbb{R}$. Therefore the desired statement holds.

Appendix C. Pointwise Asymptotic Null Distribution (Theorem 6)

In this section, we prove that $\overline{\text{x}}\text{HSIC}_n$ is asymptotically $N(0,1)$ under the null, provided that (i) the kernels k and ℓ are fixed in n, (ii) the data generating distribution P_{XY} is fixed in n, and (iii) $0 < \mathbb{E}[\widetilde{g}^2(Z_1,Z_2)] < \infty$ for Z_1, Z_2 independent draws from P_{XY}. We first present an outline of the proof in Appendix C.1, breaking the overall argument into three steps, and then present the details of these steps in the next three subsections.

Remark 25 (moment condition for $\overline{\text{x}}$dCov) *Suppose that we use the Euclidean distance kernel in Fact 11 with $x_0 = x'$ and $y_0 = y'$. Then the kernels become $k(x,x') = -\frac{1}{2}\|x-x'\|$ and $\ell(y,y') = -\frac{1}{2}\|y-y'\|$. Due to the non-linearity of the Euclidean norm, it is difficult to obtain an explicit form of $\mathbb{E}[\widetilde{g}^2(Z_1,Z_2)]$ associated with the Euclidean distance kernel. Nevertheless, Jensen's inequality gives a sufficient condition for $\mathbb{E}[\widetilde{g}^2(Z_1,Z_2)] < \infty$. First note that*

$$\begin{aligned}\widetilde{g}\big((x,x'),(y,y')\big) \;&=\; \frac{1}{4}\big\{\|x-x'\| - \mathbb{E}[\|X-x'\|] - \mathbb{E}[\|x-X'\|] + \mathbb{E}[\|X-X'\|]\big\} \\ &\quad\times\big\{\|y-y'\| - \mathbb{E}[\|Y-y'\|] - \mathbb{E}[\|y-Y'\|] + \mathbb{E}[\|Y-Y'\|]\big\}.\end{aligned}$$

Then Jensen's inequality (more specifically, $\{\mathbb{E}[\|X-X'\|]\}^2 \le \mathbb{E}[\|X-X'\|^2]$) along with independence of X and Y yields

$$\mathbb{E}[\widetilde{g}^2(Z_1,Z_2)] \lesssim \mathbb{E}[\|X-X'\|^2]\mathbb{E}[\|Y-Y'\|^2].$$

Therefore, $\overline{\text{x}}$dCov is asymptotically $N(0,1)$ given that $\mathbb{E}[\|X-X'\|^2] < \infty$ and $\mathbb{E}[\|Y-Y'\|^2] < \infty$.

C.1 Outline of the proof

Throughout the proof, we work with the centered features $\widetilde{\phi}(\cdot) := \phi(\cdot) - \mu$ and $\widetilde{\psi}(\cdot) := \psi(\cdot) - \nu$, and use $\langle\cdot,\cdot\rangle$ without the subscript to denote the RKHS inner product $\langle\cdot,\cdot\rangle_{k\times\ell}$. This can be done without loss of generality due to the following simple observation:

$$\begin{aligned}h(Z_i,Z_j) = h_{ij} &= \frac{1}{2}\big\{\phi(X_i) - \phi(X_j)\big\}\big\{\psi(Y_i) - \psi(Y_j)\big\} \\ &= \frac{1}{2}\big\{\phi(X_i) - \mu + \mu - \phi(X_j)\big\}\big\{\psi(Y_i) - \nu + \nu - \psi(Y_j)\big\} \\ &= \frac{1}{2}\big\{\widetilde{\phi}(X_i) - \widetilde{\phi}(X_j)\big\}\big\{\widetilde{\psi}(Y_i) - \widetilde{\psi}(Y_j)\big\}.\end{aligned}$$

Having this observation in mind, the main technical ingredient of the proof is the orthonormal expansion of $\widetilde{g}$ in equation (11). In fact, we can express this orthonormal expansion as the product of the orthonormal expansions of $\widetilde{k}$ and $\widetilde{\ell}$, respectively. This leads to

$$\begin{aligned}\widetilde{g}(z,z') &= \widetilde{g}((x,y),(x',y')) = \widetilde{k}(x,x')\widetilde{\ell}(y,y') \\ &= \left\{\sum_{k=1}^{\infty}\lambda_{X,k}e_{X,k}(x)e_{X,k}(x')\right\}\left\{\sum_{k'=1}^{\infty}\lambda_{Y,k'}e_{Y,k'}(y)e_{Y,k'}(y')\right\} \\ &= \sum_{k=1}^{\infty}\sum_{k'=1}^{\infty}\lambda_{X,k}\lambda_{Y,k'}e_{X,k}(x)e_{X,k}(x')e_{Y,k'}(y)e_{Y,k'}(y') \\ &= \sum_{k=1}^{\infty}\lambda_k e_k(z)e_k(z').\end{aligned}$$

In the proof, one of the main challenges is to handle the infinite sum associated with eigenvalues and eigenfunctions. When the sum is finite, then the usual fixed-dimensional law of large numbers and multivariate central limit theorem establish the desired result in a straightforward manner. However, when dealing with an infinite sum, we need extra care, and we bypass this technical difficulty by leveraging the truncation argument used in the asymptotic analysis of degenerate U-statistics (e.g. Serfling, 2009).

We break the proof into several pieces for readability.

- Step 1: In the first step, we prove

$$n\mathrm{xHSIC}_n = \sum_{k=1}^{\infty}\lambda_k\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}e_k(Z_i)\right)\left(\frac{1}{\sqrt{n}}\sum_{t=n+1}^{2n}e_k(Z_t)\right) + o_P(1).$$

- Step 2: In the second step, we prove

$$(n-1)s_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{k=1}^{\infty}\lambda_k e_k(Z_i)\left(\frac{1}{\sqrt{n}}\sum_{t=n+1}^{2n}e_k(Z_t)\right)\right\}^2 + o_P(1).$$

- Step 3: In the final step, we prove the bivariate central limit theorem

$$\begin{pmatrix}\sum_{k=1}^{\infty}\lambda_k\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}e_k(Z_i)\right)\left(\frac{1}{\sqrt{n}}\sum_{t=n+1}^{2n}e_k(Z_t)\right) \\ \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{k=1}^{\infty}\lambda_k e_k(Z_i)\left(\frac{1}{\sqrt{n}}\sum_{t=n+1}^{2n}e_k(Z_t)\right)\right\}^2\end{pmatrix} \xrightarrow{d} \begin{pmatrix}\sum_{k=1}^{\infty}\lambda_k W_k\widetilde{W}_k \\ \sum_{k=1}^{\infty}\lambda_k^2\widetilde{W}_k^2\end{pmatrix},$$

where $W_1, \widetilde{W}_1, W_2, \widetilde{W}_2, \ldots$ are i.i.d. $N(0,1)$. Then Slutsky's theorem along with the continuous mapping theorem proves that

$$\begin{aligned}
\overline{\mathrm{x}}\mathrm{HSIC}_n &= \frac{n\mathrm{xHSIC}_n}{\sqrt{(n-1)s_n^2}} \frac{\sqrt{n-1}}{\sqrt{n}} \\
&= \frac{\sum_{k=1}^{\infty} \lambda_k \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} e_k(Z_i) \right) \left(\frac{1}{\sqrt{n}} \sum_{t=n+1}^{2n} e_k(Z_t) \right) + o_P(1)}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{k=1}^{\infty} \lambda_k e_k(Z_i) \left(\frac{1}{\sqrt{n}} \sum_{t=n+1}^{2n} e_k(Z_t) \right) \right\}^2 + o_P(1)}} \{1 + o_P(1)\} \\
&\xrightarrow{d} \frac{\sum_{k=1}^{\infty} \lambda_k W_k \widetilde{W}_k}{\sqrt{\sum_{k=1}^{\infty} \lambda_k^2 \widetilde{W}_k^2}} \overset{d}{=} N(0,1).
\end{aligned}$$

In the subsequent subsections, we present detailed proofs of the above results in each step.

C.2 Proof of Step 1 (Numerator)

We start by decomposing xHSIC_n as

$$\begin{aligned}
\mathrm{xHSIC}_n &= \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \langle h_{ij}, f_2 \rangle \\
&= \underbrace{\frac{1}{n} \sum_{i=1}^{n} \langle \widetilde{\phi}(X_i) \widetilde{\psi}(Y_i), f_2 \rangle}_{\mathrm{xHSIC}_n^{(1)}} - \underbrace{\frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \langle \widetilde{\phi}(X_i) \widetilde{\psi}(Y_j), f_2 \rangle}_{\mathrm{xHSIC}_n^{(2)}}.
\end{aligned}$$

Notice that $\mathrm{xHSIC}_n^{(2)}$ is a degenerate U-statistic of order 2 whose kernel satisfies

$$\mathbb{E}\big[\langle \widetilde{\phi}(X_i) \widetilde{\psi}(Y_j), f_2 \rangle | f_2, Z_i \big] = \mathbb{E}\big[\langle \widetilde{\phi}(X_i) \widetilde{\psi}(Y_j), f_2 \rangle | f_2, Z_j \big] = 0.$$

Under the null, $\mathbb{E}[\mathrm{xHSIC}_n^{(2)} | f_2] = 0$ and

$$\mathbb{E}\Big[\big\{ \mathrm{xHSIC}_n^{(2)} \big\}^2 | f_2 \Big] = O\bigg(\frac{1}{n^2} \bigg) \mathbb{E}\big[\langle \widetilde{\phi}(X_1) \widetilde{\psi}(Y_1), f_2 \rangle^2 | f_2 \big].$$

Moreover,

$$\begin{aligned}
\langle \widetilde{\phi}(X_1) \widetilde{\psi}(Y_1), f_2 \rangle^2 &= \bigg(\frac{1}{n(n-1)} \sum_{n+1 \leq t \neq u \leq 2n} \langle \widetilde{\phi}(X_1) \widetilde{\psi}(Y_1), h(Z_t, Z_u) \rangle \bigg)^2 \\
&\lesssim \bigg(\frac{1}{n} \sum_{t=n+1}^{2n} \langle \widetilde{\phi}(X_1) \widetilde{\psi}(Y_1), \widetilde{\phi}(X_t) \widetilde{\psi}(Y_t) \rangle \bigg)^2 \\
&\quad + \bigg(\frac{1}{n(n-1)} \sum_{n+1 \leq t \neq u \leq 2n} \langle \widetilde{\phi}(X_t) \widetilde{\psi}(Y_u), \widetilde{\phi}(X_1) \widetilde{\psi}(Y_1) \rangle \bigg)^2,
\end{aligned}$$

and therefore its expectation is bounded by

$$\mathbb{E}\big[\langle \widetilde{\phi}(X_1)\widetilde{\psi}(Y_1), f_2\rangle^2\big] \lesssim \frac{1}{n}\mathbb{E}\big[\langle \widetilde{\phi}(X_1)\widetilde{\psi}(Y_1), \widetilde{\phi}(X_2)\widetilde{\psi}(Y_2)\rangle^2\big] = \frac{1}{n}\mathbb{E}\big[\widetilde{g}(Z_1,Z_2)^2\big]. \qquad (31)$$

Hence, for $t \geq 0$, Chebyshev's inequality yields

$$\begin{aligned}\mathbb{E}\big[\mathbb{P}\big(|\mathrm{xHSIC}_n^{(2)}| \geq t | f_2\big)\big] &\lesssim \frac{1}{t^2 n^2}\mathbb{E}\Big[\mathbb{E}\big[\langle \widetilde{\phi}(X_1)\widetilde{\psi}(Y_1), f_2\rangle^2 | f_2\big]\Big] \\ &\lesssim \frac{1}{t^2 n^3}\mathbb{E}\big[\widetilde{g}(Z_1,Z_2)^2\big],\end{aligned}$$

which concludes that $\mathrm{xHSIC}_n^{(2)} = O_P(n^{-3/2}) = o_P(n^{-1})$.

Next we study $\mathrm{xHSIC}_n^{(1)}$, which equals

$$\begin{aligned}\mathrm{xHSIC}_n^{(1)} &= \frac{1}{n}\sum_{i=1}^{n}\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_2\rangle \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{t=n+1}^{2n}\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), \widetilde{\phi}(X_t)\widetilde{\psi}(Y_t)\rangle - \frac{1}{n^2(n-1)}\sum_{i=1}^{n}\sum_{n+1\leq t\neq u\leq 2n}\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), \widetilde{\phi}(X_t)\widetilde{\psi}(Y_u)\rangle.\end{aligned}$$

The second term above has zero expectation and its variance is bounded above by

$$O\bigg(\frac{1}{n^3}\mathbb{E}\big[\widetilde{g}(Z_1,Z_2)^2\big]\bigg).$$

As a result, we have

$$\begin{aligned}\mathrm{xHSIC}_n^{(1)} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{t=n+1}^{2n}\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), \widetilde{\phi}(X_t)\widetilde{\psi}(Y_t)\rangle + O_P(n^{-3/2}) \\ &= \sum_{k=1}^{\infty}\lambda_k\bigg(\frac{1}{n}\sum_{i=1}^{n}e_k(Z_i)\bigg)\bigg(\frac{1}{n}\sum_{t=n+1}^{2n}e_k(Z_t)\bigg) + O_P(n^{-3/2}).\end{aligned}$$

Putting all together, we see

$$n\mathrm{xHSIC}_n = \sum_{k=1}^{\infty}\lambda_k\bigg(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}e_k(Z_i)\bigg)\bigg(\frac{1}{\sqrt{n}}\sum_{t=n+1}^{2n}e_k(Z_t)\bigg) + o_P(1),$$

as claimed in Step 1.

C.3 Proof of Step 2 (Denominator)

In this step, we would like to show $(n-1)s_n^2$ approximates

$$(n-1)s_n^2 = \frac{1}{n}\sum_{i=1}^{n}\bigg\{\sum_{k=1}^{\infty}\lambda_k e_k(Z_i)\bigg(\frac{1}{\sqrt{n}}\sum_{t=n+1}^{2n}e_k(Z_t)\bigg)\bigg\}^2 + o_P(1).$$

This part is much more challenging to prove than Step 1 partly due to the complexity of s_n^2. To simplify the problem a bit, we first prove that $\mathrm{xHSIC}_n^2 = O_P(n^{-2})$. Indeed, from the previous results in Step 1, for this claim to hold, it suffices to prove

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{t=n+1}^{2n}\langle\widetilde{\phi}(X_i)\widetilde{\psi}(Y_i),\widetilde{\phi}(X_t)\widetilde{\psi}(Y_t)\rangle = O_P(n^{-1}).$$

This follows by the Chebyshev's argument as before given that it has the expectation zero and the variance bounded by $O(n^{-2}\mathbb{E}\big[\widetilde{g}(Z_1,Z_2)^2\big])$. Consequently, we have $\mathrm{xHSIC}_n^2 = O_P(n^{-2})$, and thus

$$\begin{aligned}
s_n^2 &= \frac{4(n-1)}{(n-2)^2}\left[\frac{1}{(n-1)^2}\sum_{i=1}^{n}\left(\sum_{j=1}^{n,j\neq i}\langle h(Z_i,Z_j),f_2\rangle\right)^2 - n\mathrm{xHSIC}_n^2\right] \\
&= \frac{4(n-1)}{(n-1)^2(n-2)^2}\sum_{i=1}^{n}\left(\sum_{j=1}^{n,j\neq i}\langle h(Z_i,Z_j),f_2\rangle\right)^2 + O_P(n^{-2}).
\end{aligned}$$

The first term above further approximates

$$\begin{aligned}
&\frac{4(n-1)}{(n-1)^2(n-2)^2}\sum_{i=1}^{n}\left(\sum_{j=1}^{n,j\neq i}\langle h(Z_i,Z_j),f_2\rangle\right)^2 \\
= \;&\frac{4(n-1)}{(n-1)^2(n-2)^2}\sum_{1\leq i\neq j\leq n}\langle h(Z_i,Z_j),f_2\rangle^2 + \frac{4(n-1)}{(n-1)^2(n-2)^2}\sum_{\substack{1\leq i,j,q\leq n\\ i,j,q\text{ distinct}}}\langle h(Z_i,Z_j),f_2\rangle\langle h(Z_i,Z_q),f_2\rangle \\
= \;&\frac{4}{(n-1)(n-2)^2}\sum_{\substack{1\leq i,j,q\leq n\\ i,j,q\text{ distinct}}}\langle h(Z_i,Z_j),f_2\rangle\langle h(Z_i,Z_q),f_2\rangle + O_P(n^{-2}),
\end{aligned}$$

where the last step uses Markov's inequality together with

$$\begin{aligned}
\mathbb{E}\big[\langle h(Z_i,Z_j),f_2\rangle^2\big] &\lesssim \mathbb{E}\left[\left\langle\widetilde{\phi}(X_i)\widetilde{\psi}(Y_i),f_2\right\rangle^2\right] + \mathbb{E}\left[\left\langle\widetilde{\phi}(X_i)\widetilde{\psi}(Y_j),f_2\right\rangle^2\right] \\
&\quad + \mathbb{E}\left[\left\langle\widetilde{\phi}(X_j)\widetilde{\psi}(Y_j),f_2\right\rangle^2\right] + \mathbb{E}\left[\left\langle\widetilde{\phi}(X_j)\widetilde{\psi}(Y_i),f_2\right\rangle^2\right] \lesssim \frac{1}{n}\mathbb{E}\big[\widetilde{g}(Z_1,Z_2)^2\big],
\end{aligned}$$

due to the upper bound in (31). Thus the main term to investigate is

$$\begin{aligned}
s_{\mathrm{main},n}^2 := \;&\frac{4}{n(n-1)(n-2)}\sum_{\substack{1\leq i,j,q\leq n\\ i,j,q\text{ distinct}}}\langle h(Z_i,Z_j),f_2\rangle\langle h(Z_i,Z_q),f_2\rangle \\
= \;&\frac{1}{n(n-1)(n-2)}\sum_{\substack{1\leq i,j,q\leq n\\ i,j,q\text{ distinct}}}\left\langle\big\{\widetilde{\phi}(X_i)-\widetilde{\phi}(X_j)\big\}\big\{\widetilde{\psi}(Y_i)-\widetilde{\psi}(Y_j)\big\},f_2\right\rangle \\
&\times\left\langle\big\{\widetilde{\phi}(X_i)-\widetilde{\phi}(X_q)\big\}\big\{\widetilde{\psi}(Y_i)-\widetilde{\psi}(Y_q)\big\},f_2\right\rangle.
\end{aligned}$$

We also claim that

$$s_{\mathrm{main},n}^2 = \frac{1}{n}\sum_{i=1}^{n}\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), \frac{1}{n}\sum_{t=n+1}^{2n}\widetilde{\phi}(X_t)\widetilde{\psi}(Y_t)\right\rangle^2 + o_P(n^{-1}), \tag{32}$$

which yields the desired result in Step 2. Note that $s_{\mathrm{main},n}^2$ is a U-statistic of order 3 with a varying kernel in n conditional on f_2. Therefore we are not able to directly apply the usual approximation theory of U-statistics, which focuses on a fixed kernel, in the process of obtaining approximation (32). It turns out that the analysis is non-trivial especially only with the finite second moment of $\widetilde{g}$. We postpone the detailed (long) analysis to Appendix C.5.

C.4 Proof of Step 3

By the Cramér–Wold device, the bivariate central limit theorem holds if for each $t_1, t_2 \in \mathbb{R}$,

$$\underbrace{t_1\sum_{k=1}^{\infty}\lambda_k\bigg(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}e_k(Z_i)\bigg)\bigg(\frac{1}{\sqrt{n}}\sum_{t=n+1}^{2n}e_k(Z_t)\bigg) + t_2\frac{1}{n}\sum_{i=1}^{n}\bigg\{\sum_{k=1}^{\infty}\lambda_k e_k(Z_i)\bigg(\frac{1}{\sqrt{n}}\sum_{t=n+1}^{2n}e_k(Z_t)\bigg)\bigg\}^2}_{T_n}$$

$$\xrightarrow{d} \underbrace{t_1\sum_{k=1}^{\infty}\lambda_k W_k\widetilde{W}_k + t_2\sum_{k=1}^{\infty}\lambda_k^2\widetilde{W}_k^2}_{T}.$$

In order to establish this, we make use of the truncation argument as in Chapter 5 of Serfling (2009). First of all, for some fixed K, define

$$\begin{aligned} T_{n,K} &:= t_1\sum_{k=1}^{K}\lambda_k\bigg(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}e_k(Z_i)\bigg)\bigg(\frac{1}{\sqrt{n}}\sum_{t=n+1}^{2n}e_k(Z_t)\bigg) \\ &\qquad\qquad + t_2\frac{1}{n}\sum_{i=1}^{n}\bigg\{\sum_{k=1}^{K}\lambda_k e_k(Z_i)\bigg(\frac{1}{\sqrt{n}}\sum_{t=n+1}^{2n}e_k(Z_t)\bigg)\bigg\}^2, \\ T_K &:= t_1\sum_{k=1}^{K}\lambda_k W_k\widetilde{W}_k + t_2\sum_{k=1}^{K}\lambda_k^2\widetilde{W}_k^2. \end{aligned}$$

Then by the triangle inequality

$$\begin{aligned} \mathbb{E}[|T_n - T_{n,K}|] \leq\ & |t_1|\mathbb{E}\bigg[\bigg|\sum_{k=K+1}^{\infty}\lambda_k\bigg(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}e_k(Z_i)\bigg)\bigg(\frac{1}{\sqrt{n}}\sum_{t=n+1}^{2n}e_k(Z_t)\bigg)\bigg|\bigg] \\ & |t_2|\mathbb{E}\bigg[\bigg|\frac{1}{n}\sum_{i=1}^{n}\bigg\{\sum_{k=1}^{\infty}\lambda_k e_k(Z_i)\bigg(\frac{1}{\sqrt{n}}\sum_{t=n+1}^{2n}e_k(Z_t)\bigg)\bigg\}^2 \\ &\qquad\qquad - \frac{1}{n}\sum_{i=1}^{n}\bigg\{\sum_{k=1}^{K}\lambda_k e_k(Z_i)\bigg(\frac{1}{\sqrt{n}}\sum_{t=n+1}^{2n}e_k(Z_t)\bigg)\bigg\}^2\bigg|\bigg] \\ =\ & |t_1|(\mathrm{I}) + |t_2|(\mathrm{II}). \end{aligned}$$

By the Cauchy–Schwarz inequality and using that $\{e_k\}_{k=1}^{\infty}$ are orthonormal and $\mathbb{E}[e_k(Z)] = 0$,

$$(\mathrm{I}) \leq \sqrt{\sum_{k=K+1}^{\infty} \lambda_k^2}.$$

On the other hand, letting

$$\begin{aligned}
\sum_{k=1}^{\infty} \lambda_k e_k(Z_i)\bigg(\frac{1}{\sqrt{n}}\sum_{t=n+1}^{2n} e_k(Z_t)\bigg) \ &= \ \sum_{k=1}^{K} \lambda_k e_k(Z_i)\bigg(\frac{1}{\sqrt{n}}\sum_{t=n+1}^{2n} e_k(Z_t)\bigg) \\
&\quad + \sum_{k=K+1}^{\infty} \lambda_k e_k(Z_i)\bigg(\frac{1}{\sqrt{n}}\sum_{t=n+1}^{2n} e_k(Z_t)\bigg) \\
&= \ A_i + B_i
\end{aligned}$$

and using $|(A_i + B_i)^2 - A_i^2| \leq B_i^2 + 2|A_iB_i|$, we have

$$(\mathrm{II}) \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[B_i^2 + 2|A_iB_i|\big] \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[B_i^2\big] + 2\frac{1}{n}\sum_{i=1}^{n} \sqrt{\mathbb{E}\big[A_i^2\big]}\sqrt{\mathbb{E}\big[B_i^2\big]},$$

where the last inequality uses the Cauchy–Schwarz inequality. Again by using the orthonormal property of $\{e_k\}_{k=1}^{\infty}$ and $\mathbb{E}[e_k(Z)] = 0$,

$$\mathbb{E}\big[A_i^2\big] = \sum_{k=1}^{K} \lambda_k^2 \quad \text{and} \quad \mathbb{E}\big[B_i^2\big] = \sum_{k=K+1}^{\infty} \lambda_k^2.$$

Therefore, noting that $\mathbb{E}[\widetilde{g}^2(Z_1, Z_2)] = \sum_{k=1}^{K} \lambda_k^2 < \infty$,

$$\mathbb{E}[|T_n - T_{n,K}|] \leq |t_1|\sqrt{\sum_{k=K+1}^{\infty} \lambda_k^2} + |t_2| \sum_{k=K+1}^{\infty} \lambda_k^2 + 2|t_2|\sqrt{\sum_{k=1}^{K} \lambda_k^2}\sqrt{\sum_{k=K+1}^{\infty} \lambda_k^2},$$

which goes to zero as $K \to \infty$ uniformly over n. Hence $T_{n,K}$ converges to T_n in distribution as $K \to \infty$ uniformly over n. Similarly, we obtain

$$\begin{aligned}
\mathbb{E}[|T - T_K|] \ &\leq \ |t_1|\mathbb{E}\bigg[\bigg|\sum_{k=K+1}^{\infty} \lambda_k W_k \widetilde{W}_k\bigg|\bigg] + |t_2|\mathbb{E}\bigg[\sum_{k=K+1}^{\infty} \lambda_k^2 \widetilde{W}_k^2\bigg] \\
&\leq \ |t_1|\sqrt{\sum_{k=K+1}^{\infty} \lambda_k^2} + |t_2| \sum_{k=K+1}^{\infty} \lambda_k^2,
\end{aligned}$$

which goes to zero as $K \to \infty$. In addition, for each fixed K, the multivariate central limit theorem yields

$$\begin{pmatrix} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} e_1(Z_i) \\ \vdots \\ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} e_K(Z_i) \\ \frac{1}{\sqrt{n}} \sum_{t=n+1}^{2n} e_1(Z_t) \\ \vdots \\ \frac{1}{\sqrt{n}} \sum_{t=n+1}^{2n} e_K(Z_t) \end{pmatrix} \xrightarrow{d} N_{2K}(0, I),$$

and the law of large numbers shows

$$\frac{1}{n} \sum_{i=1}^{n} e_k(Z_i) e_{k'}(Z_i) \xrightarrow{p} \begin{cases} 1 & k = k', \\ 0 & \text{otherwise.} \end{cases}$$

Using these results, Slutsky's theorem together with the continuous mapping theorem proves that $T_{n,K} \xrightarrow{d} T_K$ for each fixed K.

Having these preliminary results in place, we finish the proof of Step 3 using the argument in Serfling (2009) as follows. Let φ_n, $\varphi_{n,K}$, φ and φ_K be the characteristic functions of T_n, $T_{n,K}$, T and T_K, respectively. Let $\epsilon > 0$ be some fixed number. Then we can choose $K_1 > 0$ such that for all $K \geq K_1$ and for all n,

$$\begin{aligned} |\varphi_n(s) - \varphi_{n,K}(s)| &= |\mathbb{E}[e^{isT_n} - e^{isT_{n,K}}| \\ &\leq \mathbb{E}|e^{is(T_n - T_{n,K})} - 1| \\ &\leq |s|\mathbb{E}[|T_n - T_{n,K}|] < \frac{\epsilon}{3}. \end{aligned}$$

We can also choose $K_2 > 0$ such that for all $K \geq K_2$,

$$|\varphi(s) - \varphi_K(s)| < \frac{\epsilon}{3}.$$

Since $T_{n,K} \xrightarrow{d} T_K$ for each K, we can choose N such that for all $n \geq N$ and $K_0 = \max\{K_1, K_2\}$,

$$|\varphi_{n,K}(s) - \varphi_K| \leq \frac{\epsilon}{n}.$$

Therefore, by the triangle inequality,

$$|\varphi_n(s) - \varphi(s)| \leq |\varphi_n(s) - \varphi_{n,K_0}(s)| + |\varphi_{n,K_0}(s) - \varphi_{K_0}| + |\varphi(s) - \varphi_{K_0}(s)| \leq \epsilon.$$

Since ϵ was arbitrary, we conclude that $\lim_{n\to\infty} \varphi_n(s) = \varphi(s)$ as desired.

C.5 Details of approximation (32)

The aim of this section is to establish approximation (32). First note that $s^2_{\mathrm{main},n}$ has the following decomposition:

$$\begin{aligned}
s^2_{\mathrm{main},n} = \; & \frac{1}{n}\sum_{i=1}^{n}\Big\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_2\Big\rangle^2 - \frac{2}{n(n-1)}\sum_{1\leq i\neq j\leq n}\Big\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_2\Big\rangle\Big\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_2\Big\rangle \\
& -\frac{2}{n(n-1)}\sum_{1\leq i\neq j\leq n}\Big\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_2\Big\rangle\Big\langle \widetilde{\phi}(X_j)\widetilde{\psi}(Y_i), f_2\Big\rangle \\
& +\frac{3}{n(n-1)}\sum_{1\leq i\neq j\leq n}\Big\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_2\Big\rangle\Big\langle \widetilde{\phi}(X_j)\widetilde{\psi}(Y_j), f_2\Big\rangle \\
& -\frac{4}{n(n-1)(n-2)}\sum_{\substack{1\leq i,j,q\leq n\\ i,j,q \text{ distinct}}}\Big\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_2\Big\rangle\Big\langle \widetilde{\phi}(X_j)\widetilde{\psi}(Y_q), f_2\Big\rangle \\
& +\frac{1}{n(n-1)(n-2)}\sum_{\substack{1\leq i,j,q\leq n\\ i,j,q \text{ distinct}}}\Big\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_2\Big\rangle\Big\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_q), f_2\Big\rangle \\
& +\frac{1}{n(n-1)(n-2)}\sum_{\substack{1\leq i,j,q\leq n\\ i,j,q \text{ distinct}}}\Big\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_2\Big\rangle\Big\langle \widetilde{\phi}(X_q)\widetilde{\psi}(Y_i), f_2\Big\rangle \\
& +\frac{1}{n(n-1)(n-2)}\sum_{\substack{1\leq i,j,q\leq n\\ i,j,q \text{ distinct}}}\Big\langle \widetilde{\phi}(X_j)\widetilde{\psi}(Y_i), f_2\Big\rangle\Big\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_q), f_2\Big\rangle \\
& +\frac{1}{n(n-1)(n-2)}\sum_{\substack{1\leq i,j,q\leq n\\ i,j,q \text{ distinct}}}\Big\langle \widetilde{\phi}(X_j)\widetilde{\psi}(Y_i), f_2\Big\rangle\Big\langle \widetilde{\phi}(X_q)\widetilde{\psi}(Y_i), f_2\Big\rangle \\
:= \; & \sum_{i=1}^{9} J_{(i)}.
\end{aligned}$$

By defining $f_{2,A}$, $f_{2,B}$ and $f_{2,C}$ as

$$\begin{aligned}
f_2 \; &= \; \frac{1}{n}\sum_{t=n+1}^{2n}\widetilde{\phi}(X_t)\widetilde{\psi}(Y_t) - \frac{1}{n(n-1)}\sum_{n+1\leq t\neq u\leq 2n}\widetilde{\phi}(X_t)\widetilde{\psi}(Y_u) \\
&= \; \frac{1}{n}\sum_{t=n+1}^{2n}\widetilde{\phi}(X_t)\widetilde{\psi}(Y_t) - \frac{n}{n-1}\Bigg\{\frac{1}{n}\sum_{n+1\leq t\leq 2n}\widetilde{\phi}(X_t)\Bigg\}\Bigg\{\frac{1}{n}\sum_{n+1\leq u\leq 2n}\widetilde{\psi}(Y_u)\Bigg\} \\
& \qquad + \frac{1}{n(n-1)}\sum_{t=n+1}^{2n}\widetilde{\phi}(X_t)\widetilde{\psi}(Y_t) \\
&:= \; f_{2,A} + f_{2,B} + f_{2,C}.
\end{aligned}$$

we analyze each $J_{(i)}$ term, separately.

1. Analyzing $J_{(1)}$. Starting with $J_{(1)}$, we will show that

$$J_{(1)} = \frac{1}{n}\sum_{i=1}^{n}\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), \frac{1}{n}\sum_{t=n+1}^{2n}\widetilde{\phi}(X_t)\widetilde{\psi}(Y_t)\right\rangle^2 + o_P(n^{-1}). \tag{33}$$

First note that

$$\begin{aligned} J_{(1)} = \; & \frac{1}{n}\sum_{i=1}^{n}\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A}\right\rangle^2 + \frac{2}{n}\sum_{i=1}^{n}\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A}\right\rangle\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B}\right\rangle \\ & + \frac{1}{n}\sum_{i=1}^{n}\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B}\right\rangle^2 + \frac{2}{n}\sum_{i=1}^{n}\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B}\right\rangle\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,C}\right\rangle \\ & + \frac{1}{n}\sum_{i=1}^{n}\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,C}\right\rangle^2 + \frac{2}{n}\sum_{i=1}^{n}\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A}\right\rangle\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,C}\right\rangle. \end{aligned}$$

We can ignore the terms involving $f_{2,C}$ since they have a faster convergence rate than the same terms replacing $f_{2,C}$ with $f_{2,A}$. By noting that

$$\frac{1}{n}\sum_{i=1}^{n}\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A}\right\rangle^2 = \frac{1}{n}\sum_{i=1}^{n}\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), \frac{1}{n}\sum_{t=n+1}^{2n}\widetilde{\phi}(X_t)\widetilde{\psi}(Y_t)\right\rangle^2,$$

we shall prove

$$\underbrace{\frac{2}{n}\sum_{i=1}^{n}\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A}\right\rangle\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B}\right\rangle}_{:=J_{(1),a}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B}\right\rangle^2}_{:=J_{(1),b}} = o_P(n^{-1}),$$

and thus establish approximation (33).

Focusing on $J_{(1),b}$, notice that

$$\begin{aligned} nJ_{(1),b} \; = \; & n \times \frac{1}{n}\sum_{i=1}^{n}\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B}\right\rangle^2 \\ = \; & \frac{n^2}{(n-1)^2}\frac{1}{n^2}\sum_{i=1}^{n}\left\{\sum_{k=1}^{\infty}\lambda_{X,k}e_{X,k}(X_i)\left(\frac{1}{\sqrt{n}}\sum_{n+1\le t\le 2n}e_{X,k}(X_t)\right)\right\}^2 \\ & \left\{\sum_{k'=1}^{\infty}\lambda_{Y,k'}e_{Y,k'}(Y_i)\left(\frac{1}{\sqrt{n}}\sum_{n+1\le t\le 2n}e_{Y,k'}(Y_t)\right)\right\}^2 := \frac{n^2}{(n-1)^2}\frac{1}{n}G_{(1)}. \end{aligned}$$

In addition, the orthonormal property of eigenfunctions yields

$$\begin{aligned} & \mathbb{E}\left[\left\{\sum_{k=1}^{\infty}\lambda_{X,k}e_{X,k}(X_i)\left(\frac{1}{\sqrt{n}}\sum_{n+1\le t\le 2n}e_{X,k}(X_t)\right)\right\}^2\right] = \sum_{k=1}^{\infty}\lambda_{X,k}^2 \quad \text{and} \\ & \mathbb{E}\left[\left\{\sum_{k'=1}^{\infty}\lambda_{Y,k'}e_{Y,k'}(Y_i)\left(\frac{1}{\sqrt{n}}\sum_{n+1\le t\le 2n}e_{Y,k'}(Y_t)\right)\right\}^2\right] = \sum_{k'=1}^{\infty}\lambda_{Y,k'}^2. \end{aligned}$$

This implies that $\mathbb{E}[G_{(1)}] = \sum_{k=1}^{\infty} \lambda_k^2 < \infty$ since X and Y are independent. Therefore using Markov's inequality,

$$\frac{1}{n}\sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B} \right\rangle^2 = O_P(n^{-2}) = o_P(n^{-1}). \tag{34}$$

Next, we turn to $J_{(1),a}$. As shown before in Appendix C.4,

$$\sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle^2 \xrightarrow{d} \sum_{k=1}^{\infty} \lambda_k^2, \widetilde{W}_k^2$$

which implies

$$\frac{1}{n}\sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle^2 = O_P(n^{-1}). \tag{35}$$

Hence, by the Cauchy–Schwarz inequality,

$$\begin{aligned}
|J_{(1),a}| &= \left| \frac{2}{n}\sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B} \right\rangle \right| \\
&\leq 2\sqrt{\frac{1}{n}\sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle^2} \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B} \right\rangle^2} \\
&= 2\sqrt{O_P(n^{-1})}\sqrt{O_P(n^{-2})} = o_P(n^{-1}),
\end{aligned}$$

where we use the approximation result (34) for the second term in the upper bound. Combining results establishes the approximation (33).

2. Analyzing $J_{(2)}$**.** By the same logic as in the analysis of $J_{(1)}$, we can ignore the terms involving $f_{2,C}$ throughout the proof, and we only need to handle three terms

$$\begin{aligned}
J_{(2),a} &:= \frac{2}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_{2,A} \right\rangle, \\
J_{(2),b} &:= \frac{2}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B} \right\rangle \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_{2,B} \right\rangle, \quad \text{and} \\
J_{(2),c} &:= \frac{2}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_{2,B} \right\rangle.
\end{aligned}$$

For the first term $J_{(2),a}$,

$$\begin{aligned}
J_{(2),a} = & \frac{2}{n(n-1)} \sum_{1 \le i \neq j \le n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_{2,A} \right\rangle \\
= & \frac{2}{n(n-1)} \sum_{i=1}^{n}\sum_{j=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_{2,A} \right\rangle \\
& - \frac{2}{n(n-1)} \sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \\
= & \underbrace{\frac{2n}{n-1} \times \frac{1}{n} \sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_i)\left(\frac{1}{n}\sum_{j=1}^{n}\widetilde{\psi}(Y_j)\right), f_{2,A} \right\rangle}_{O_P(n^{-3/2})} \\
& \underbrace{- \frac{2}{n(n-1)} \sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle^2}_{O_P(n^{-2})},
\end{aligned}$$

where the second approximation result $O_P(n^{-2})$ holds by (35). On the other hand, the first approximation $O_P(n^{-3/2})$ holds since

$$\begin{aligned}
& \left| \frac{1}{n} \sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_i)\left(\frac{1}{n}\sum_{j=1}^{n}\widetilde{\psi}(Y_j)\right), f_{2,A} \right\rangle \right| \\
\le & \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle^2} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\left(\frac{1}{n}\sum_{j=1}^{n}\widetilde{\psi}(Y_j)\right), f_{2,A} \right\rangle^2} \\
= & O_P(n^{-1/2}) O_P(n^{-1}) = O_P(n^{-3/2}),
\end{aligned}$$

by the Cauchy–Schwarz inequality and Markov's inequality. In particular, it can be seen that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_{2,A} \right\rangle^2\right] = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}\left[\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_{2,A} \right\rangle^2\right] \lesssim \frac{1}{n^2}\mathbb{E}[\widetilde{g}^2(Z_1, Z_2)].$$

Therefore $J_{(2),a} = o_P(n^{-1})$.

For the second term $J_{(2),b}$,

$$\begin{aligned}
J_{(2),b} =\ & \frac{2}{n(n-1)} \sum_{1\leq i\neq j\leq n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B}\right\rangle \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_{2,B}\right\rangle \\
=\ & \underbrace{\frac{2n}{n-1} \times \frac{1}{n}\sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B}\right\rangle \left\langle \widetilde{\phi}(X_i)\bigg(\frac{1}{n}\sum_{j=1}^{n}\widetilde{\psi}(Y_j)\bigg), f_{2,B}\right\rangle}_{O_P(n^{-3/2})} \\
& \underbrace{- \frac{2}{n(n-1)} \sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B}\right\rangle^2}_{O_P(n^{-2})},
\end{aligned}$$

where the second approximation $O_P(n^{-2})$ uses (34). The first approximation follows by the Cauchy–Schwarz inequality and Markov's inequality as

$$\begin{aligned}
& \left|\frac{1}{n}\sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B}\right\rangle \left\langle \widetilde{\phi}(X_i)\bigg(\frac{1}{n}\sum_{j=1}^{n}\widetilde{\psi}(Y_j)\bigg), f_{2,B}\right\rangle\right| \\
\leq\ & \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B}\right\rangle^2} \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\bigg(\frac{1}{n}\sum_{j=1}^{n}\widetilde{\psi}(Y_j)\bigg), f_{2,B}\right\rangle^2} \\
=\ & O_P(n^{-1})O_P(n^{-1}) = O_P(n^{-2}).
\end{aligned}$$

Therefore $J_{(2),b} = o_P(n^{-1})$. For the last term $J_{(2),c}$,

$$\begin{aligned}
J_{(2),c} =\ & \frac{2}{n(n-1)} \sum_{1\leq i\neq j\leq n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A}\right\rangle \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_{2,B}\right\rangle \\
=\ & \underbrace{\frac{2n}{n-1} \times \frac{1}{n}\sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A}\right\rangle \left\langle \widetilde{\phi}(X_i)\bigg(\frac{1}{n}\sum_{j=1}^{n}\widetilde{\psi}(Y_j)\bigg), f_{2,B}\right\rangle}_{O_P(n^{-3/2})} \\
& \underbrace{- \frac{2}{n(n-1)} \sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A}\right\rangle \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B}\right\rangle}_{O_P(n^{-2})},
\end{aligned}$$

which can be established by the Cauchy–Schwarz inequality and the previous results, and thus $J_{(2),c} = o_P(n^{-1})$. In summary, it holds that $J_{(2)} = o_P(n^{-1})$.

3. Analyzing $J_{(3)}$**.** By symmetry, $J_{(3)}$ has the same convergence rate as $J_{(2)}$ and thus $J_{(3)} = o_P(n^{-1})$.

4. Analyzing $J_{(4)}$. For the fourth term $J_{(4)}$, we will show that $J_{(4)} = o_P(n^{-1})$. To this end, we only need to handle three terms

$$\begin{aligned} J_{(4),a} :=& \frac{3}{n(n-1)} \sum_{1 \le i \ne j \le n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_j)\widetilde{\psi}(Y_j), f_{2,A} \right\rangle, \\ J_{(4),b} :=& \frac{3}{n(n-1)} \sum_{1 \le i \ne j \le n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B} \right\rangle \left\langle \widetilde{\phi}(X_j)\widetilde{\psi}(Y_j), f_{2,B} \right\rangle \quad \text{and} \\ J_{(4),c} :=& \frac{3}{n(n-1)} \sum_{1 \le i \ne j \le n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_j)\widetilde{\psi}(Y_j), f_{2,B} \right\rangle. \end{aligned}$$

Starting with $J_{(4),a}$,

$$J_{(4),a} = \underbrace{\frac{3n^2}{n(n-1)} \left\langle \frac{1}{n} \sum_{i=1}^{n} \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle^2}_{O_P(n^{-2})} - \underbrace{\frac{3}{n(n-1)} \sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle^2}_{O_P(n^{-2})},$$

where the second approximation is due to (35) and the first approximation is by Markov's inequality. Indeed, we have shown in Appendix C.2 and C.4 that

$$n \left\langle \frac{1}{n} \sum_{i=1}^{n} \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \xrightarrow{d} \sum_{k=1}^{\infty} \lambda_k W_k \widetilde{W}_k.$$

This establishes $J_{(4),a} = o_P(n^{-1})$.

Next, for the second term $J_{(4),b}$,

$$J_{(4),b} = \underbrace{\frac{3n^2}{n(n-1)} \left\langle \frac{1}{n} \sum_{i=1}^{n} \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B} \right\rangle^2}_{O_P(n^{-2})} - \underbrace{\frac{3}{n(n-1)} \sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B} \right\rangle^2}_{O_P(n^{-3})},$$

where the second approximation uses (34). For the first approximation, we simply use Jensen's inequality along with (34):

$$\left\langle \frac{1}{n} \sum_{i=1}^{n} \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B} \right\rangle^2 \le \frac{1}{n} \sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B} \right\rangle^2 = O_P(n^{-2}).$$

Thus we have $J_{(4),b} = o_P(n^{-1})$.

For the last term $J_{(4),c}$,

$$\begin{aligned} J_{(4),c} =& \underbrace{\frac{3n^2}{n(n-1)} \left\langle \frac{1}{n} \sum_{i=1}^{n} \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \left\langle \frac{1}{n} \sum_{i=1}^{n} \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B} \right\rangle}_{O_P(n^{-2})} \\ & \underbrace{- \frac{3}{n(n-1)} \sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B} \right\rangle}_{O_P(n^{-2})}, \end{aligned}$$

which can be seen using the Cauchy–Schwarz inequality. Therefore it holds that $J_{(4)} = o_P(n^{-1})$.

5. Analyzing $J_{(5)}$. For the fifth term $J_{(5)}$, similarly as before, we need to study

$$J_{(5),a} := \frac{1}{n(n-1)(n-2)} \sum_{\substack{1 \le i,j,q \le n \\ i,j,q \text{ distinct}}} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_j)\widetilde{\psi}(Y_q), f_{2,A} \right\rangle,$$

$$J_{(5),b} := \frac{1}{n(n-1)(n-2)} \sum_{\substack{1 \le i,j,q \le n \\ i,j,q \text{ distinct}}} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B} \right\rangle \left\langle \widetilde{\phi}(X_j)\widetilde{\psi}(Y_q), f_{2,B} \right\rangle \quad \text{and}$$

$$J_{(5),c} := \frac{1}{n(n-1)(n-2)} \sum_{\substack{1 \le i,j,q \le n \\ i,j,q \text{ distinct}}} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_j)\widetilde{\psi}(Y_q), f_{2,B} \right\rangle.$$

Starting with $J_{(5),a}$, note that

$$\sum_{\substack{1 \le i,j,q \le n \\ i,j,q \text{ distinct}}} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_j)\widetilde{\psi}(Y_q), f_{2,A} \right\rangle = \left\langle \sum_{i=1}^{n} \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \left\langle \sum_{1 \le j \ne q \le n} \widetilde{\phi}(X_j)\widetilde{\psi}(Y_q), f_{2,A} \right\rangle$$
$$-2 \sum_{1 \le j \ne q \le n} \left\langle \widetilde{\phi}(X_j)\widetilde{\psi}(Y_j), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_j)\widetilde{\psi}(Y_q), f_{2,A} \right\rangle.$$

Thus

$$\begin{aligned} J_{(5),a} \;=\; & O(1) \times \underbrace{\left\langle \frac{1}{n} \sum_{i=1}^{n} \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle}_{O_P(n^{-1})} \underbrace{\left\langle \frac{1}{n(n-1)} \sum_{1 \le j \ne q \le n} \widetilde{\phi}(X_j)\widetilde{\psi}(Y_q), f_{2,A} \right\rangle}_{O_P(n^{-1})} \\ & + O(n^{-1}) \times \underbrace{\frac{1}{n(n-1)} \sum_{1 \le j \ne q \le n} \left\langle \widetilde{\phi}(X_j)\widetilde{\psi}(Y_j), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_j)\widetilde{\psi}(Y_q), f_{2,A} \right\rangle}_{o_P(n^{-1})}, \end{aligned}$$

where the first approximation holds since

$$n \times \left\langle \frac{1}{n} \sum_{i=1}^{n} \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \xrightarrow{d} \sum_{k=1}^{K} \lambda_k W_k \widetilde{W}_k,$$

as established in Appendix C.4. For the second approximation, we have

$$\begin{aligned} \left\langle \frac{1}{n(n-1)} \sum_{1 \le j \ne q \le n} \widetilde{\phi}(X_j)\widetilde{\psi}(Y_q), f_{2,A} \right\rangle \;=\; & O(1) \times \left\langle \left(\frac{1}{n} \sum_{i=1}^{n} \widetilde{\phi}(X_i) \right) \left(\frac{1}{n} \sum_{i=1}^{n} \widetilde{\psi}(Y_i) \right), f_{2,A} \right\rangle \\ & - O(n^{-1}) \times \underbrace{\left\langle \frac{1}{n} \sum_{i=1}^{n} \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle}_{O_P(n^{-1})}. \end{aligned}$$

Moreover the following term

$$\left\langle \left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{\phi}(X_i)\right)\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{\psi}(Y_i)\right), f_{2,A}\right\rangle$$

has the same convergence rate as

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B}\right\rangle\right| \le \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B}\right\rangle^2} = O_P(n^{-1}).$$

Thereby, the second approximation holds. The last approximation was established in the analysis of $J_{(2),a}$, and thus $J_{(5),a} = o_P(n^{-1})$.

For the second term $J_{(5),b}$, we simply use the Cauchy–Schwarz inequality and Markov's inequality, and see

$$|J_{(5),b}| \ \le \ O(1)\underbrace{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,B}\right\rangle^2}}_{O_P(n^{-1})}\underbrace{\sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_{2,B}\right\rangle^2}}_{O_P(n^{-1})} = o_P(n^{-1}).$$

Similarly, for the third term $J_{(5),c}$, we apply the Cauchy–Schwarz inequality and see

$$|J_{(5),c}| \ \le \ O(1)\underbrace{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A}\right\rangle^2}}_{O_P(n^{-1/2})}\underbrace{\sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_{2,B}\right\rangle^2}}_{O_P(n^{-1})} = o_P(n^{-1}).$$

Therefore we conclude $J_{(5)} = o_P(n^{-1})$.

6. Analyzing $J_{(6)}$. For the sixth term $J_{(6)}$, similarly as before, we need to study

$$\begin{aligned}
J_{(6),a} &:= \frac{1}{n(n-1)(n-2)}\sum_{\substack{1\le i,j,q\le n\\ i,j,q \text{ distinct}}}\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_{2,A}\right\rangle\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_q), f_{2,A}\right\rangle,\\
J_{(6),b} &:= \frac{1}{n(n-1)(n-2)}\sum_{\substack{1\le i,j,q\le n\\ i,j,q \text{ distinct}}}\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_{2,B}\right\rangle\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_q), f_{2,B}\right\rangle \quad \text{and}\\
J_{(6),c} &:= \frac{1}{n(n-1)(n-2)}\sum_{\substack{1\le i,j,q\le n\\ i,j,q \text{ distinct}}}\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_{2,A}\right\rangle\left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_q), f_{2,B}\right\rangle.
\end{aligned}$$

Starting with $J_{(6),a}$, there exist some constants $C_1, \ldots, C_4$ such that

$$\begin{aligned}
&\sum_{\substack{1 \le i,j,q \le n \\ i,j,q \text{ distinct}}} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_q), f_{2,A} \right\rangle \\
= \; & \sum_{i=1}^n \left\langle \widetilde{\phi}(X_i) \sum_{j=1}^n \widetilde{\psi}(Y_j), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_i) \sum_{q=1}^n \widetilde{\psi}(Y_q), f_{2,A} \right\rangle \\
& - C_1 \sum_{1 \le i \ne j \le n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle - C_2 \sum_{i=1}^n \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle^2 \\
& - C_3 \sum_{1 \le i \ne q \le n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_q), f_{2,A} \right\rangle - C_4 \sum_{1 \le i \ne q \le n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_q), f_{2,A} \right\rangle^2 .
\end{aligned}$$

By the Cauchy–Schwarz inequality and Markov's inequality,

$$\begin{aligned}
& \left| \frac{1}{n} \sum_{i=1}^n \left\langle \widetilde{\phi}(X_i) \frac{1}{n} \sum_{j=1}^n \widetilde{\psi}(Y_j), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_i) \bigg(\frac{1}{n} \sum_{q=1}^n \widetilde{\psi}(Y_q) \bigg), f_{2,A} \right\rangle \right|^2 \\
\le \; & \underbrace{\frac{1}{n} \sum_{i=1}^n \left\langle \widetilde{\phi}(X_i) \frac{1}{n} \sum_{j=1}^n \widetilde{\psi}(Y_j), f_{2,A} \right\rangle^2}_{O_P(n^{-2})} \underbrace{\frac{1}{n} \sum_{i=1}^n \left\langle \widetilde{\phi}(X_i) \bigg(\frac{1}{n} \sum_{q=1}^n \widetilde{\psi}(Y_q) \bigg), f_{2,A} \right\rangle^2}_{O_P(n^{-2})} .
\end{aligned}$$

Thus

$$\frac{1}{n^3} \sum_{i=1}^n \left\langle \widetilde{\phi}(X_i) \sum_{j=1}^n \widetilde{\psi}(Y_j), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_i) \sum_{q=1}^n \widetilde{\psi}(Y_q), f_{2,A} \right\rangle = O_P(n^{-2}).$$

Based on the previous results, we also have

$$\begin{aligned}
& \frac{1}{n^3} \sum_{1 \le i \ne j \le n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle = o_P(n^{-2}), \\
& \frac{1}{n^3} \sum_{i=1}^n \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle^2 = o_P(n^{-2}), \\
& \frac{1}{n^3} \sum_{1 \le i \ne q \le n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_q), f_{2,A} \right\rangle = o_P(n^{-2}), \\
& \frac{1}{n^3} \sum_{1 \le i \ne q \le n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_q), f_{2,A} \right\rangle^2 = o_P(n^{-2}),
\end{aligned}$$

which concludes $J_{(6),a} = o_P(n^{-1})$. In addition $J_{(6),b}$ and $J_{(6),c}$ are shown to be $o_P(n^{-1})$ by the Cauchy–Schwarz inequality as in the analysis of $J_{(5),b}$ and $J_{(5),c}$. This proves that $J_{(6)} = o_P(n^{-1})$.

7. Analyzing $J_{(7)}$. For the sixth term $J_{(7)}$, similarly as before, we need to study

$$J_{(7),a} := \frac{1}{n(n-1)(n-2)} \sum_{\substack{1 \le i,j,q \le n \\ i,j,q \text{ distinct}}} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_q)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle,$$

$$J_{(7),b} := \frac{1}{n(n-1)(n-2)} \sum_{\substack{1 \le i,j,q \le n \\ i,j,q \text{ distinct}}} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_{2,B} \right\rangle \left\langle \widetilde{\phi}(X_q)\widetilde{\psi}(Y_i), f_{2,B} \right\rangle \quad \text{and}$$

$$J_{(7),c} := \frac{1}{n(n-1)(n-2)} \sum_{\substack{1 \le i,j,q \le n \\ i,j,q \text{ distinct}}} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_q)\widetilde{\psi}(Y_i), f_{2,B} \right\rangle.$$

Starting with $J_{(7),a}$, there exist some constants $C_1', \ldots, C_4'$ such that

$$\begin{aligned}
&\sum_{\substack{1 \le i,j,q \le n \\ i,j,q \text{ distinct}}} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_q)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \\
&= \sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i) \sum_{j=1}^{n} \widetilde{\psi}(Y_j), f_{2,A} \right\rangle \left\langle \sum_{q=1}^{n} \widetilde{\phi}(X_q)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \\
&\quad - C_1' \sum_{1 \le i \ne j \le n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_j), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle - C_2' \sum_{i=1}^{n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle^2 \\
&\quad - C_3' \sum_{1 \le i \ne q \le n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle \left\langle \widetilde{\phi}(X_q)\widetilde{\psi}(Y_i), f_{2,A} \right\rangle - C_4' \sum_{1 \le i \ne q \le n} \left\langle \widetilde{\phi}(X_i)\widetilde{\psi}(Y_q), f_{2,A} \right\rangle^2 \Bigg].
\end{aligned}$$

From here, we can follow the same steps as in the analysis for $J_{(6)}$ and conclude that $J_{(7)} = o_P(n^{-1})$.

8. Analyzing $J_{(8)}$. By switching the role between X and Y, $J_{(8)}$ can be analyzed similarly as $J_{(7)}$ and it can be shown that $J_{(8)} = o_P(n^{-1})$.

9. Analyzing $J_{(9)}$. By switching the role between X and Y, $J_{(9)}$ can be analyzed similarly as $J_{(6)}$ and it can be shown that $J_{(9)} = o_P(n^{-1})$.

Throughout, we have shown that $\sum_{i=2}^{9} J_{(i)} = o_P(n^{-1})$ and $J_{(1)}$ satisfies (33), which concludes approximation (32).

Appendix D. Uniform Asymptotic Null Distribution (Theorem 7)

We first describe an outline of the general steps of the proof in Appendix D.1, breaking down the argument into three parts. Then, we present the details of the steps in the subsequent subsections.

Before proceeding, we recall some notation. We use $\widetilde{a}_{ij} = \widetilde{k}(X_i, \cdot)\widetilde{\ell}(Y_j, \cdot)$, $\widetilde{b}_{tu} = \widetilde{a}_{n+t,n+u}$ for $1 \le i, j, t, u \le n$, and $\widetilde{g}_{12}$ to denote $\widetilde{g}(Z_1, Z_2) = \langle \widetilde{a}_{11}, \widetilde{a}_{22} \rangle$. As before, $\langle \cdot, \cdot \rangle$ refers to the RKHS inner-product $\langle \cdot, \cdot \rangle_{k \times \ell}$ throughout this section.

D.1 Outline of the proof

To show the asymptotic normality of $\overline{\text{x}}\text{HSIC}_n$, we verify that the sufficient conditions for the Berry–Esseen theorem for studentized U-statistics are satisfied. In particular, by Jing et al. (2000, Theorem 3.1), it suffices to show that

$$C_n := \frac{\mathbb{E}[|\langle h(Z_1, Z_2), f_2\rangle|^3|\mathcal{D}_{n+1}^{2n}]}{\sqrt{n}\sigma_g^3} \xrightarrow{p} 0, \tag{36}$$

where we have used the notation $\sigma_g^2 := \mathbb{E}[\{\mathbb{E}[\langle h(Z_1, Z_2), f_2\rangle |Z_2, \mathcal{D}_{n+1}^{2n}]\}^2|\mathcal{D}_{n+1}^{2n}]$ following Jing et al. (2000).

To establish this result, we proceed in the following steps:

- Step 1: First, we observe in Theorem 26, that a sufficient condition for (36) to hold is if the following holds:

$$B_n := \frac{\mathbb{E}[\langle f_2, \widetilde{k}(X_{1,n}, \cdot)\widetilde{\ell}(Y_{1,n}, \cdot)\rangle^4|\mathcal{D}_{n+1}^{2n}]}{n\big\{\mathbb{E}[\langle f_2, \widetilde{k}(X_{1,n}, \cdot)\widetilde{\ell}(Y_{1,n}, \cdot)\rangle^2|\mathcal{D}_{n+1}^{2n}]\big\}^2} \xrightarrow{p} 0, \tag{37}$$

- Step 2: In the next step, we introduce the term $B_{1,n}$ defined as

$$B_{1,n} := \frac{n\mathbb{E}[\langle f_2, \widetilde{k}(X_{1,n}, \cdot)\widetilde{\ell}(Y_{1,n}, \cdot)\rangle^4|\mathcal{D}_{n+1}^{2n}]}{\mathbb{E}[\widetilde{g}(Z_{1,n}, Z_{2,n})^2]^2}, \tag{38}$$

  and show that $B_{1,n} \xrightarrow{p} 0$ in Theorem 27.

- Step 3: Finally, in we introduce the term $B_{2,n}$ defined as

$$B_{2,n} := \frac{\mathbb{E}[\widetilde{g}(Z_{1,n}, Z_{2,n})^2]^2}{n^2\big\{\mathbb{E}[\langle f_2, \widetilde{k}(X_{1,n}, \cdot)\widetilde{\ell}(Y_{1,n}, \cdot)\rangle^2|\mathcal{D}_{n+1}^{2n}]\big\}^2}, \tag{39}$$

  and show that $B_{2,n} = \mathcal{O}_P(1)$ in Theorem 28

Since $B_n = B_{1,n} \times B_{2,n}$, together (38) and (39) imply (37), to complete the proof. The details are in Appendix D.

D.2 Proof of Step 1

Lemma 26 *Introduce the term* $B_n := \frac{\mathbb{E}[\langle f_2, \widetilde{k}(X_{1,n}, \cdot)\widetilde{\ell}(Y_{1,n}, \cdot)\rangle^4|\mathcal{D}_{n+1}^{2n}]}{n\big\{\mathbb{E}[\langle f_2, \widetilde{k}(X_{1,n}, \cdot)\widetilde{\ell}(Y_{1,n}, \cdot)\rangle^2|\mathcal{D}_{n+1}^{2n}]\big\}^2}$. *Then, we have the following:*

$$B_n \xrightarrow{p} 0 \quad \textit{implies} \quad C_n \xrightarrow{p} 0.$$

Proof We begin by noting that due to Cauchy–Schwarz inequality, we have

$$\begin{aligned}\mathbb{E}[|\langle h(Z_1, Z_2), f_2\rangle|^3|\mathcal{D}_{n+1}^{2n}] &= \mathbb{E}[|\langle h(Z_1, Z_2), f_2\rangle| \times |\langle h(Z_1, Z_2), f_2\rangle|^2|\mathcal{D}_{n+1}^{2n}] \\ &\leq \big\{\mathbb{E}[\langle h(Z_1, Z_2), f_2\rangle^2|\mathcal{D}_{n+1}^{2n}]\big\}^{1/2}\big\{\mathbb{E}[\langle h(Z_1, Z_2), f_2\rangle^4|\mathcal{D}_{n+1}^{2n}]\big\}^{1/2}.\end{aligned}$$

Further note

$$
\begin{aligned}
\mathbb{E}[\langle h(Z_1,Z_2),f_2\rangle^2\,|\mathcal{D}_{n+1}^{2n}] &= \frac{1}{4}\mathbb{E}[\langle \widetilde{f}_2, \widetilde{a}_{11}+\widetilde{a}_{22}-\widetilde{a}_{12}-\widetilde{a}_{21}\rangle^2|\mathcal{D}_{n+1}^{2n}] \\
&\lesssim \mathbb{E}[\langle \widetilde{f}_2,\widetilde{a}_{11}\rangle^2+\langle \widetilde{f}_2,\widetilde{a}_{22}\rangle^2+\langle \widetilde{f}_2,\widetilde{a}_{12}\rangle^2+\langle \widetilde{f}_2,\widetilde{a}_{21}\rangle^2|\mathcal{D}_{n+1}^{2n}] \\
&\lesssim \mathbb{E}[\langle \widetilde{f}_2,\widetilde{a}_{11}\rangle^2|\mathcal{D}_{n+1}^{2n}], \qquad (40)
\end{aligned}
$$

where the first inequality uses Jensen's inequality, while the second inequality relies on the observation under the null:

$$
\mathbb{E}[\langle \widetilde{f}_2,\widetilde{a}_{11}\rangle^2|\mathcal{D}_{n+1}^{2n}] = \mathbb{E}[\langle \widetilde{f}_2,\widetilde{a}_{ij}\rangle^2|\mathcal{D}_{n+1}^{2n}] \quad \text{for any } i,j\in\{1,2\}.
$$

By the same logic along with Jensen's inequality,

$$
\begin{aligned}
\mathbb{E}[\langle h(Z_1,Z_2),f_2\rangle^4\,|\mathcal{D}_{n+1}^{2n}] &\lesssim \mathbb{E}[\langle f_2,\widetilde{a}_{11}\rangle^4+\langle f_2,\widetilde{a}_{22}\rangle^4+\langle f_2,\widetilde{a}_{12}\rangle^4+\langle f_2,\widetilde{a}_{21}\rangle^4\,|\mathcal{D}_{n+1}^{2n}] \\
&\lesssim \mathbb{E}[\langle \widetilde{f}_2,\widetilde{a}_{11}\rangle^4|\mathcal{D}_{n+1}^{2n}]. \qquad (41)
\end{aligned}
$$

Thus, combining (40) and (41), we get the following bound on the numerator of the term C_n:

$$
\left(\mathbb{E}[|\,\langle h(Z_1,Z_2),f_2\rangle\,|^3|\mathcal{D}_{n+1}^{2n}]\right)^2 \le \mathbb{E}[\langle \widetilde{f}_2,\widetilde{a}_{11}\rangle^2|\mathcal{D}_{n+1}^{2n}]\times\mathbb{E}[\langle \widetilde{f}_2,\widetilde{a}_{11}\rangle^4|\mathcal{D}_{n+1}^{2n}] \qquad (42)
$$

We now evaluate σ_g^2 from the denominator of C_n.

$$
\sigma_g^2 = \frac{1}{4}\mathbb{E}[\langle \widetilde{f}_2,\widetilde{a}_{11}\rangle^2|\mathcal{D}_{n+1}^{2n}]. \qquad (43)
$$

Combining the pieces, we obtain the following:

$$
\begin{aligned}
C_n = \frac{\mathbb{E}[|\,\langle h(Z_1,Z_2),f_2\rangle\,|^3|\mathcal{D}_{n+1}^{2n}]}{\sqrt{n}\sigma_g^3} &\le \frac{\left(\mathbb{E}[\langle \widetilde{f}_2,\widetilde{a}_{11}\rangle^2|\mathcal{D}_{n+1}^{2n}]\right)^{1/2}\times\left(\mathbb{E}[\langle \widetilde{f}_2,\widetilde{a}_{11}\rangle^4|\mathcal{D}_{n+1}^{2n}]\right)^{1/2}}{\sqrt{n}\sigma_g^3} \\
= \left(\frac{\mathbb{E}[\langle \widetilde{f}_2,\widetilde{a}_{11}\rangle^4|\mathcal{D}_{n+1}^{2n}]}{n\left\{\mathbb{E}[\langle \widetilde{f}_2,\widetilde{a}_{11}\rangle^2|\mathcal{D}_{n+1}^{2n}]\right\}^2}\right)^{1/2} &= \sqrt{B_n}.
\end{aligned}
$$

In the first inequality above, we used (42), while the second equality uses (43). This completes the proof. ■

D.3 Proof of Step 2

Lemma 27 *Under Assumption 1, we have* $B_{1,n} \xrightarrow{p} 0$.

Proof

It suffices to show that $\mathbb{E}[B_{1,n}] \to 0$, which in turn will imply the convergence in probability by an application of Markov's inequality. To verify this, we observe the following:

$$\begin{aligned}
\mathbb{E}[B_{1,n}] &= \frac{n}{\mathbb{E}[\widetilde{g}_{12}^2]^2}\mathcal{O}\left(\mathbb{E}\left[\left\langle \widetilde{a}_{ii}, \frac{1}{n}\sum_{t=1}^{n}\widetilde{b}_{tt} - \frac{1}{n(n-1)}\sum_{t\neq u}\widetilde{b}_{tu}\right\rangle^4\right]\right) \\
&\leq \frac{16n}{\mathbb{E}[\widetilde{g}_{12}^2]^2}\left(\mathbb{E}\left[\left\langle \widetilde{a}_{ii},\ \frac{1}{n}\sum_{t=1}^{n}\widetilde{b}_{tt}\right\rangle^4\right] + \mathbb{E}\left[\left\langle \widetilde{a}_{ii},\ \frac{1}{n(n-1)}\sum_{t\neq u}\widetilde{b}_{tu}\right\rangle^4\right]\right) && (44) \\
&= \frac{16n}{\mathbb{E}[\widetilde{g}_{12}^2]^2}\mathcal{O}\left(\frac{1}{n^4}\mathbb{E}\left[\left(\sum_{t=1}^{n}\langle \widetilde{a}_{ii}, \widetilde{b}_{tt}\rangle\right)^4\right] + \frac{1}{n^8}\mathbb{E}\left[\left(\sum_{u\neq t}\langle \widetilde{a}_{ii}, \widetilde{b}_{tu}\rangle\right)^4\right]\right) && (45) \\
&= \frac{16n}{\mathbb{E}[\widetilde{g}_{12}^2]^2}\mathcal{O}\left(\mathbb{E}[\widetilde{g}_{12}^4]\left(\frac{1}{n^3} + \frac{1}{n^6}\right) + \mathbb{E}[\widetilde{g}_{12}^2\widetilde{g}_{13}^2]\left(\frac{1}{n^2} + \frac{1}{n^4}\right)\right) && (46) \\
&= \mathcal{O}\left(\frac{\mathbb{E}[\widetilde{g}_{12}^4]n^{-1} + \mathbb{E}[\widetilde{g}_{12}^2\widetilde{g}_{13}^2]}{n\mathbb{E}[\widetilde{g}_{12}^2]^2}\right) \to 0. && (47)
\end{aligned}$$

In the above display, (44) uses the fact that $(x-y)^4 \leq 16(x^4+y^4)$. To obtain (46), we note that $\mathbb{E}[(\sum_{t=1}^{n}\langle \widetilde{a}_{ii}, \widetilde{b}_{tt}\rangle^4]$ can be expanded into n^4 terms, each of the form $\mathbb{E}[\widetilde{g}_{it_1}\widetilde{g}_{it_2}\widetilde{g}_{it_3}\widetilde{g}_{it_4}]$, for $1 \leq t_1, t_2, t_3, t_4 \leq n$. Of these n^4 terms, only the terms with two or four common $t's$ have non-zero expectation under the null. There are a total of n terms of the form $\mathbb{E}[\widetilde{g}_{it}^4]$ and $\mathcal{O}(n^2)$ terms of the form $\mathbb{E}[\widetilde{g}_{it}^2\widetilde{g}_{iu}^2]$ for $t \neq u$. Such terms appear as $\mathbb{E}[\widetilde{g}_{it}^4]/n^3$ and $\mathbb{E}[\widetilde{g}_{it}^2\widetilde{g}_{iu}^2]/n^2$ respectively in (46). Repeating the same argument for $\mathbb{E}[\sum_{t\neq u}\langle \widetilde{a}_{ii}, \widetilde{b}_{tu}\rangle^4]$ gives us the other two terms in (46).

■

D.4 Proof of Step 3

Lemma 28 *Under Assumption 1, we have $B_{2,n} = \mathcal{O}_P(1)$.*

Proof Introduce the notation $f_{21} = \frac{1}{n}\sum_{t=1}^{n}\widetilde{b}_{tt}$ and $f_{22} = \frac{1}{n(n-1)}\sum_{t=1}^{n}\sum_{u\neq t}\widetilde{b}_{tu}$, and observe that $f_2 = f_{21} - f_{22}$. Next, we define the following terms:

$$B_{3,n} = \frac{1}{\sqrt{B_{2,n}}} = \frac{n\mathbb{E}[\langle \widetilde{a}_{ii}, f_2\rangle^2 | \mathcal{D}_{n+1}^{2n}]}{\mathbb{E}[\widetilde{h}_{it}^2]}, \quad \text{and} \quad B_{4,n} = \frac{n\mathbb{E}[\langle \widetilde{a}_{ii}, f_{21}\rangle^2 | \mathcal{D}_{n+1}^{2n}]}{\mathbb{E}[\widetilde{h}_{it}^2]}$$

Next, we observe that the random variable $B_{3,n} - B_{4,n}$ converges in probability to 0. We do this by proving that the second moment of $B_{3,n} - B_{4,n}$ converges to 0 with n under Assumption 1.

$$\begin{aligned}
\mathbb{E}[(B_{3,n} - B_{4,n})^2] &= \frac{n^2}{\mathbb{E}[\widetilde{h}_{tt}^2]}\mathbb{E}\left[\left(\mathbb{E}[\langle \widetilde{a}_{ii}, f_{22}\rangle^2 | \mathcal{D}_{n+1}^{2n}] - 2\mathbb{E}[\langle \widetilde{a}_{ii}, f_{21}\rangle\langle \widetilde{a}_{ii}, f_{22}\rangle | \mathcal{D}_{n+1}^{2n}]\right)^2\right] && (48)\\
&\leq \frac{n^2}{\mathbb{E}[\widetilde{h}_{tt}^2]}\mathbb{E}\left[\left(\langle \widetilde{a}_{ii}, f_{22}\rangle^2 - 2\langle \widetilde{a}_{ii}, f_{21}\rangle\langle \widetilde{a}_{ii}, f_{22}\rangle\right)^2\right] && (49)\\
&\leq \frac{2n^2}{\mathbb{E}[\widetilde{h}_{tt}^2]}\mathbb{E}\left[\langle \widetilde{a}_{ii}, f_{22}\rangle^4 + 4\langle \widetilde{a}_{ii}, f_{21}\rangle^2\langle \widetilde{a}_{ii}, f_{22}\rangle^2\right] && (50)\\
&\leq \frac{2n^2}{\mathbb{E}[\widetilde{h}_{tt}^2]}\left(\mathbb{E}\left[\langle \widetilde{a}_{ii}, f_{22}\rangle^4\right] + 4\mathbb{E}\left[\langle \widetilde{a}_{ii}, f_{21}\rangle^4\right]^{1/2}\mathbb{E}\left[\langle \widetilde{a}_{ii}, f_{22}\rangle^4\right]^{1/2}\right). && (51)\\
&= \mathcal{O}\left(\frac{\mathbb{E}[n^{-1}\widetilde{h}_{it}^4 + \widetilde{h}_{it}^2 + \widetilde{h}_{iu}^2]}{n\mathbb{E}[\widetilde{h}_{tt}^2]}\right) \to 0. && (52)
\end{aligned}$$

In the above display,
(49) uses the (conditional) version of Jensen's inequality along with the convexity of $x \mapsto x^2$,
(50) uses the fact that $(x+y)^2 \leq 2(x^2+y^2)$,
(51) uses the Cauchy-Schwarz inequality, and
(52) follows by expanding the terms f_{21} and f_{22} in terms of $\widetilde{b}_{tu}$, and simplifying the expressions exploiting the fact that the terms containing odd powers of $\langle \widetilde{a}_{ii}, \widetilde{b}_{tu}\rangle$ are zero in expectation. The final terms in (52) is exactly the condition in Assumption 1.

Note that $1/B_{4,n}$ can be written as follows:

$$\frac{1}{B_{4,n}} = \frac{\mathbb{E}[\widetilde{h}_{it}^2]}{n\mathbb{E}\left[\langle \widetilde{a}_{ii}, \frac{1}{n}\sum_{t=1}^n \widetilde{b}_{tt}\rangle^2\right]},$$

which can be shown to be stochastically bounded under the conditions of Assumption 1, by following the exact same argument used by Kim and Ramdas (2023) in proving that the term (II) in their Eq.(53) is stochastically boundeded.

To complete the proof, we will show that the combination of the two facts proved above; that is, **(i)** $B_{3,n} - B_{4,n} \xrightarrow{p} 0$ and **(ii)** $1/B_{4,n} = \mathcal{O}_P(1)$, are sufficient to conclude that $1/B_{3,n} = \sqrt{B_{2,n}}$ is also stochastically bounded. In particular, these two results imply that for any $\epsilon > 0$, we can find a real number $1 \leq m < \infty$, and two integers $n_1, n_2 < \infty$, such that the following statements hold:

$$\mathbb{P}(1/B_{4,n} > 2m) \leq \epsilon/2, \quad \text{and} \quad \mathbb{P}(|B_{3,n} - B_{4,n}| > m) \leq \epsilon/2.$$

Hence, we have the following for any $n \geq n_\epsilon := \max\{n_1, n_2\}$:

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{B_{3,n}} > m\right) &\leq \mathbb{P}\left(\frac{1}{B_{4,n} - |B_{4,n} - B_{3,n}|} > m\right)\\
&\leq \mathbb{P}\left(\frac{1}{B_{4,n}} > 2m\right) + \mathbb{P}\left(|B_{3,n} - B_{4,n}| > m\right) \leq \epsilon.
\end{aligned}$$

Thus, we have shown that $1/B_{3,n}$ is stochastically bounded; that is, for every $\epsilon > 0$, there exists an $m < \infty$, such that for all $n \geq n_\epsilon$, we have $\mathbb{P}\left(1/B_{3,n} > m\right) \leq \epsilon$.

■

Appendix E. Power of the cross-HSIC Test

In this section, we prove the results on consistency against fixed and local alternatives (Theorem 9 and Theorem 10) of our cross-HSIC test, stated in Section 6. The proofs of both of these results can be obtained from a more abstract result, identifying sufficient condtions for the cross-HSIC test to be consistent, which we state and prove in Appendix E.1. Then, in the next two subsections, we use this general result to prove Theorem 9 and Theorem 10.

Additional Notation. Throughout this section, we will use shorthand notation for some commonly used terms. For any $1 \leq i, j, \leq n$, we use h_{ij} to represent $h(Z_i, Z_j) = \frac{1}{2}a_{ii} + a_{jj} - a_{ij} - a_{ij}$, use $\widetilde{h}_{ij}$ for denoting the centered version of h_{ij}, i.e., $\widetilde{h}_{ij} = h_{ij} - (\omega - \mu\nu)$. Recall that ω, μ and ν denote the kernel mean embeddings of P_{XY}, P_X and P_Y, and $a_{ij} = k(X_i, \cdot)\ell(Y_j, \cdot)$. Furthermore, recall the definition of the cross-HSIC statistic,

$$\mathrm{xHSIC}_n = \langle f_1, f_2 \rangle, \quad \text{where } f_1 = \frac{1}{n(n-1)} \sum_{i \neq j} h_{ij}, \quad \text{and } f_2 = \frac{1}{n(n-1)} \sum_{t \neq u} h_{tu}.$$

We will use $\widetilde{f}_1$ and $\widetilde{f}_2$ to denote the centered versions of f_1 and f_2 respectively; that is, $\widetilde{f}_1 = f_1 - (\omega - \mu\nu)$ and $\widetilde{f}_2 = f_2 - (\omega - \mu\nu)$. Finally, introduce the following term, for any $1 \leq i \leq n$,

$$A_i = \frac{1}{n-1} \sum_{j=1}^{n} h_{ij} = \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} h_{ij}.$$

As before, we use $\widetilde{A}_i$ to denote the centered version of A_i, and use $\bar{A}_n$ to denote the average of the $A_1, \ldots, A_n$.

E.1 General Conditions for Consistency

To identify sufficient conditions for the consistency of the cross-HSIC test, $\Psi = \mathbf{1}_{\overline{\mathrm{x}}\mathrm{HSIC}_n > z_{1-\alpha}}$, we first study this problem in a general scenario, in which, the distribution as well as the kernels can change with n. In particular, we consider a sequence of distributions $\{P_{XY,n} : n \geq 1\}$ and kernels $\{(k_n, \ell_n) : n \geq 1\}$, and let $\mathcal{D}_1^{2n}$ denote $2n$ i.i.d. draws from $P_{XY,n}$, for $n \geq 1$. As in our previous proofs, we drop the $k \times \ell$ from the subscript of the inner products, while referring to $\langle \cdot, \cdot \rangle_{k \times \ell}$.

Theorem 29 (General conditions for consistency) *Consider the independence testing problem with observations $\mathcal{D}_1^{2n} = \{(X_i, Y_i) : 1 \leq i \leq 2n\}$ drawn i.i.d. from the distribution $P_{XY,n}$, with marginals $P_{X,n}$ and $P_{Y,n}$. Let γ_n^2 denote $\mathrm{HSIC}(P_{XY,n}, \mathcal{K}, \mathcal{L})$, and suppose there exists a non-negative sequence $\{\delta_n : n \geq 1\}$, such that $\lim_{n\to\infty} \delta_n = 0$, satisfying:*

$$\lim_{n \to \infty} \frac{1}{n^2 \delta_n \gamma_n^4} \left(\mathbb{E}\left[\left\langle \widetilde{h}_{12}, \widetilde{h}_{34} \right\rangle^2 + \gamma_n^2 \left(\left\langle \widetilde{h}_{12}, \widetilde{h}_{12} \right\rangle + n \left\langle \widetilde{h}_{12}, \widetilde{h}_{13} \right\rangle \right) \right] \right) = 0. \tag{53}$$

Then, the test $\Psi = \mathbf{1}_{\overline{\mathrm{x}}\mathrm{HSIC}_n > z_{1-\alpha}}$ is consistent against the sequence of alternatives $\{P_{XY,n} : n \geq 1\}$.

Proof Recall that the test $\Psi = \mathbf{1}_{\overline{\text{x}}\text{HSIC}_n > z_{1-\alpha}}$ rejects the null if the statistic $\overline{\text{x}}\text{HSIC}_n$ exceeds the $(1-\alpha)$-quantile of the standard normal distribution. Hence, its type-II error can be bounded above, as follows:

$$\begin{aligned}\mathbb{P}\left(\Psi = 0\right) &= \mathbb{P}\left(\text{xHSIC}_n < \frac{z_{1-\alpha} s_n}{\sqrt{n}}\right) \\ &\leq \mathbb{P}\left(\text{xHSIC}_n < z_{1-\alpha}\sqrt{\mathbb{E}[s_n^2]/n\delta_n}\right) + \delta_n. \qquad (54)\end{aligned}$$

For the inequality in the above display, we introduce the event $E_n = \{s_n^2 < \mathbb{E}[s_n^2]/\delta_n\}$, and note that, by Markov's inequality, we have $\mathbb{P}(E_n^c) \leq \delta_n$. Recall that xHSIC_n and s_n were defined in (4) and (6) respectively.

We first note that under the conditions of Theorem 29, the expectation of s_n^2 grows at a rate smaller than $n\delta_n\gamma_n^4$.

Lemma 30 *Under the conditions of Theorem 29, we have* $\lim_{n\to\infty} \mathbb{E}[s_n^2]/(n\delta_n\gamma_n^4) = 0$.

As a consequence of this result, proved in Appendix E.1.1, we note that $\lim_{n\to\infty} z_{1-\alpha}\sqrt{\mathbb{E}[s_n^2]/n\delta_n\gamma_n^4} \to 0$, which implies that for some finite n_0, we have $z_{1-\alpha}\sqrt{\mathbb{E}[s_n^2]/n\delta_n} \leq \gamma_n^2/2$, for all $n \geq n_0$. Hence, for all $n \geq n_0$, we have the following,

$$\mathbb{P}\left(\Psi = 0\right) \leq \mathbb{P}\left(\text{xHSIC}_n < \gamma_n^2/2\right) + \delta_n = \mathbb{P}\left(\text{xHSIC}_n - \gamma_n^2 < -\gamma_n^2/2\right) + \delta_n.$$

Since $\mathbb{E}[\text{xHSIC}_n] = \gamma_n^2$, we have the following, by Chebyshev's inequality,

$$\mathbb{P}\left(\Psi = 0\right) \leq \frac{4\mathbb{V}(\text{xHSIC}_n)}{\gamma_n^4} + \delta_n.$$

By assumption, $\lim_{n\to\infty}\delta_n = 0$, and we complete the proof by showing that the first term in the right-hand-side of the above display also goes to zero.

Lemma 31 *Under the conditions of Theorem 29, we have* $\lim_{n\to\infty} \frac{\mathbb{V}(\text{xHSIC}_n)}{\gamma_n^4} = 0$.

The proof of this lemma is in Appendix E.1.2. ■

E.1.1 Proof of Theorem 30

Proof With the notation introduced at the beginning of this section, we have the following:

$$\begin{aligned}\mathbb{E}[s_n^2] &\lesssim \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \langle A_i - \bar{A}_n, f_2\rangle^2\right] \qquad (55)\\ &= \mathbb{E}\left[\langle A_i - \bar{A}_n, f_2\rangle^2\right] = \mathbb{E}\left[\frac{1}{n^2}\left\langle \sum_{j=1}^n A_i - A_j, f_2\right\rangle^2\right] \qquad (56)\\ &\lesssim \mathbb{E}\left[\langle A_1 - A_2, f_2\rangle^2\right] = \mathbb{E}\left[\langle \widetilde{A}_1 - \widetilde{A}_2, f_2\rangle^2\right] \qquad (57)\\ &\lesssim \mathbb{E}\left[\langle \widetilde{A}_1, f_2\rangle^2\right] = \mathbb{E}\left[\left\langle \widetilde{A}_1, \widetilde{f}_2 + (\omega - \mu\nu)\right\rangle^2\right]. \qquad (58)\end{aligned}$$

In the above display, the first equality in (56) uses the fact that $(A_i - \bar{A}_n)$ and $(A_j - \bar{A}_n)$ are equal in distribution. (57) uses Cauchy-Schwarz inequality on the 'cross-terms' in the expansion of $\langle \sum_{j=1}^n A_i - A_j, f_2 \rangle^2$. The first term in (58) follows by using $(x+y)^2 \lesssim x^2 + y^2$, and the fact that $\widetilde{A}_1$ and $\widetilde{A}_2$ are equal in distribution.

Upper bounding the last term in (58), we obtain

$$\begin{aligned}
\mathbb{E}[s_n^2] &\lesssim \mathbb{E}\left[\left\langle \widetilde{A}_1, \widetilde{f}_2 \right\rangle^2 + \left\langle \widetilde{A}_1, \omega - \mu\nu \right\rangle^2\right] \\
&\lesssim \mathbb{E}\left[\left\langle \widetilde{h}_{12}, \widetilde{f}_2 \right\rangle^2 + \left\langle \widetilde{A}_1, \omega - \mu\nu \right\rangle^2\right] && (59) \\
&= \texttt{term}_1 + \texttt{term}_2. && (60)
\end{aligned}$$

First, we upper bound $\texttt{term}_1$ as follows:

$$\begin{aligned}
\texttt{term}_1 &= \frac{1}{n^2(n-1)^2}\mathbb{E}\left[\left\langle \widetilde{h}_{12}, \sum_{t\neq u} \widetilde{h}_{tu} \right\rangle^2\right] \\
&\lesssim \frac{1}{n^4}\left(n^2 \mathbb{E}\left[\left\langle \widetilde{h}_{12}, \widetilde{h}_{34} \right\rangle^2\right] + n^3 \mathbb{E}\left[\left\langle \widetilde{h}_{12}, \widetilde{h}_{34} \right\rangle \left\langle \widetilde{h}_{12}, \widetilde{h}_{35} \right\rangle\right]\right) \\
&\lesssim \frac{1}{n}\mathbb{E}\left[\left\langle \widetilde{h}_{12}, \widetilde{h}_{34} \right\rangle^2\right]. && (61)
\end{aligned}$$

Next, we can get an upper bound on $\texttt{term}_2$ as follows:

$$\begin{aligned}
\texttt{term}_2 &= \left\langle \widetilde{A}_1, \omega - \mu\nu \right\rangle^2 \overset{(a)}{\leq} \|\widetilde{A}_1\|^2 \|\omega - \mu\nu\|^2 = \|\widetilde{A}_1\|^2 \gamma_n^2 \\
&\lesssim \frac{\gamma_n^2}{n^2}\left(n \left\langle \widetilde{h}_{12}, \widetilde{h}_{12} \right\rangle + n^2 \left\langle \widetilde{h}_{12}, \widetilde{h}_{13} \right\rangle\right) \\
&\lesssim \gamma_n^2 \left(\frac{\left\langle \widetilde{h}_{12}, \widetilde{h}_{12} \right\rangle}{n} + \left\langle \widetilde{h}_{12}, \widetilde{h}_{13} \right\rangle\right). && (62)
\end{aligned}$$

Combining (61) and (62), we get

$$\mathbb{E}[s_n^2] \lesssim \frac{1}{n}\left(\mathbb{E}\left[\left\langle \widetilde{h}_{12}, \widetilde{h}_{34} \right\rangle^2\right] + \gamma_n^2 \left(\left\langle \widetilde{h}_{12}, \widetilde{h}_{12} \right\rangle + n \left\langle \widetilde{h}_{12}, \widetilde{h}_{13} \right\rangle\right)\right),$$

which implies that

$$\frac{\mathbb{E}[s_n^2]}{n\delta_n\gamma_n^4} \lesssim \frac{1}{n^2\delta_n\gamma_n^4}\left(\mathbb{E}\left[\left\langle \widetilde{h}_{12}, \widetilde{h}_{34} \right\rangle^2\right] + \gamma_n^2 \left(\left\langle \widetilde{h}_{12}, \widetilde{h}_{12} \right\rangle + n \left\langle \widetilde{h}_{12}, \widetilde{h}_{13} \right\rangle\right)\right). \qquad (63)$$

By the assumptions of Theorem 29, the right-hand-side of (63) converges to 0. ■

E.1.2 Proof of Theorem 31

Recall that xHSIC_n can be rewritten as

$$\begin{aligned}\mathrm{xHSIC}_n &= \langle \widetilde{f}_1 + (\omega - \mu\nu), \widetilde{f}_2 + (\omega - \mu\nu)\rangle \\ &= \left\langle \widetilde{f}_1, \widetilde{f}_2 \right\rangle + \left\langle \widetilde{f}_1, \omega - \mu\nu \right\rangle + \left\langle \widetilde{f}_2, \omega - \mu\nu \right\rangle + \gamma_n^2. \end{aligned} \tag{64}$$

Using this, we can write the variance of the xHSIC_n statistic as

$$\begin{aligned}\mathbb{V}(\mathrm{xHSIC}_n) &= \mathbb{E}[(\mathrm{xHSIC}_n - \gamma_n^2)^2] \\ &= \mathbb{E}[\langle \widetilde{f}_1, \widetilde{f}_2\rangle^2] + \mathbb{E}[\langle \widetilde{f}_1, \omega - \mu\nu\rangle^2] + \mathbb{E}[\langle \widetilde{f}_2, \omega - \mu\nu\rangle^2] & (65) \\ &\lesssim \mathbb{E}[\|\widetilde{f}_1\|^2]\mathbb{E}[\|\widetilde{f}_2\|^2] + \gamma_n^2 \left(\mathbb{E}[\|\widetilde{f}_1\|^2] + \mathbb{E}[\|\widetilde{f}_2\|^2]\right) & (66) \\ &\lesssim \mathbb{E}[\|\widetilde{f}_1\|^2] \left(\mathbb{E}[\|\widetilde{f}_1\|^2] + \gamma_n^2\right). & (67)\end{aligned}$$

For the equality in (65), we used (64), along with the fact that all the 'cross-terms' in expansion of $\mathbb{E}[(\mathrm{xHSIC}_n - \gamma_n^2)^2]$ have zero expectation. (66) follows by applying Cauchy-Schwarz inequality to all terms of (65), while (67) uses the fact that $\widetilde{f}_1$ and $\widetilde{f}_2$ are equal in distribution.

Since $\widetilde{f}_1 = \frac{1}{n(n-1)} \sum_{i\neq j} \widetilde{h}_{ij}$, we can upper bound its norm as follows:

$$\begin{aligned}\mathbb{E}[\|\widetilde{f}_1\|^2] &\lesssim \frac{1}{n^4} \sum_{i\neq j}\sum_{r\neq s} \mathbb{E}[\langle \widetilde{h}_{ij}, \widetilde{h}_{rs}\rangle] \lesssim \frac{1}{n^4}\left(n^2 \mathbb{E}[\langle \widetilde{h}_{12}, \widetilde{h}_{12}\rangle] + n^3 \mathbb{E}[\langle \widetilde{h}_{12}, \widetilde{h}_{13}\rangle]\right) \\ &\lesssim \frac{1}{n^2}\left(\mathbb{E}[\langle \widetilde{h}_{12}, \widetilde{h}_{12}\rangle] + n\mathbb{E}[\langle \widetilde{h}_{12}, \widetilde{h}_{13}\rangle]\right). \end{aligned} \tag{68}$$

Combining (68) with (67), we get

$$\frac{\mathbb{V}(\mathrm{xHSIC}_n)}{\gamma_n^4} \lesssim \left(\frac{\mathbb{E}[\langle \widetilde{h}_{12}, \widetilde{h}_{12}\rangle + n\langle \widetilde{h}_{12}, \widetilde{h}_{13}\rangle]}{n^2\gamma_n^2}\right),$$

which converges to 0 under the conditions of Theorem 29.

E.2 Consistency against fixed alternatives (Theorem 9)

In the case of fixed alternatives, the term $\gamma_n^2 = \mathrm{HSIC}(P_{XY}, k, \ell)$ does not change with n. Hence, to prove Theorem 9, it suffices to verify (53) with γ_n set to some fixed $\gamma > 0$. We proceed in the following steps.

Verifying $\mathbb{E}[\langle \widetilde{h}_{12}, \widetilde{h}_{34}\rangle^2] < \infty$*.* Using the fact that $\widetilde{h}_{12} = h_{12} - (\omega - \mu\nu)$, we have the following:

$$\begin{aligned}\mathbb{E}\left[\langle \widetilde{h}_{12}, \widetilde{h}_{34}\rangle^2\right] &\leq \mathbb{E}\left[\|\widetilde{h}_{12}\|^2 \|\widetilde{h}_{34}\|^2\right] = \mathbb{E}\left[\|\widetilde{h}_{12}\|^2\right]^2 \lesssim \mathbb{E}\left[\left(\|h_{12}\|^2 + \gamma^2\right)\right]^2 \\ &\lesssim \mathbb{E}\left[\|h_{12}\|^2\right]^2 + \gamma^4. \end{aligned} \tag{69}$$

Hence it suffices to show that under the conditions of Theorem 9, $\mathbb{E}[\|h_{12}\|^2] < \infty$, which we do in Theorem 32 below.

Verifying $\mathbb{E}[\langle \widetilde{h}_{12}, \widetilde{h}_{12}\rangle] < \infty$. Expanding this term, we have

$$\mathbb{E}[\langle \widetilde{h}_{12}, \widetilde{h}_{12}\rangle] = \mathbb{E}\left[\|h_{12} - (\omega - \mu\nu)\|^2\right] \lesssim \mathbb{E}[\|h_{12}\|^2] + \gamma^2. \tag{70}$$

Again, this reduces to showing that $\mathbb{E}[\|h_{12}\|^2] < \infty$, which we do in Theorem 32.

Verifying $\mathbb{E}[\langle \widetilde{h}_{12}, \widetilde{h}_{13}\rangle] < \infty$. We again reduce this condition to Theorem 32 as follows:

$$\begin{aligned} \mathbb{E}\left[\langle \widetilde{h}_{12}, \widetilde{h}_{13}\rangle\right] &\leq \mathbb{E}\left[\|\widetilde{h}_{12}\|\|\widetilde{h}_{13}\|\right] = \mathbb{E}\left[\sqrt{\|\widetilde{h}_{12}\|^2\|\widetilde{h}_{13}\|^2}\right] \\ &\overset{(i)}{\lesssim} \mathbb{E}\left[\|\widetilde{h}_{12}\|^2 + \|\widetilde{h}_{13}\|^2\right] \lesssim \gamma^2 + \mathbb{E}\left[\|h_{12}\|^2\right]. \end{aligned} \tag{71}$$

In the above display, (i) follows from an application of the AM-GM inequality; $\sqrt{x^2y^2} \leq (x^2+y^2)/2$.

Lemma 32 *For (X, Y) drawn according to P_{XY}, if $\mathbb{E}[k(X,X)\ell(Y,Y)] + \mathbb{E}[k(X,X)]\mathbb{E}[\ell(Y,Y)] < \infty$, then we have $\mathbb{E}[\|h_{12}\|^2] < \infty$.*

Proof Recall that we have $2h_{12} = a_{11} + a_{22} - a_{12} - a_{21}$, where $a_{ij} = k(X_i, \cdot)\ell(Y_j, \cdot)$. Thus, on expanding $\|h_{12}\|^2$, we have

$$\begin{aligned} \|h_{12}\|^2 &\simeq \|a_{11} + a_{22} - a_{12} - a_{21}\|^2 \leq \|a_{11} - a_{12}\|^2 + \|a_{22} - a_{21}\|^2 \\ &\simeq \|a_{11} - a_{12}\|^2 \leq \|a_{11}\|^2 + \|a_{12}\|^2 = k(X_1, X_1)\ell(Y_1, Y_1) + k(X_1, X_1)\ell(Y_2, Y_2). \end{aligned}$$

This implies that

$$\begin{aligned} \mathbb{E}[\|h_{12}\|^2] &\leq \mathbb{E}[k(X_1, X_1)\ell(Y_1, Y_1) + k(X_1, X_1)\ell(Y_2, Y_2)] \\ &= \mathbb{E}[k(X,X)\ell(Y,Y)] + \mathbb{E}[k(X,X)]\mathbb{E}[\ell(Y,Y)], \end{aligned}$$

where (X, Y) are drawn according to P_{XY}. This completes the proof. ■

E.3 Type-I error and consistency against local alternatives (Theorem 10)

To prove this result, we will verify the conditions required by Theorem 7 and Theorem 29 to prove the type-I error control and consistency respectively. For verifying these conditions, we will need to use some facts about the Gaussian kernel that were derived by Li and Yuan (2019). We collect the required properties below.

Fact 33 *Suppose $Z_i = (X_i, Y_i) \sim P_X \times P_Y$ for $i = 1, 2, 3, 4$ be independent random variables, with X_i and Y_i taking values in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$ respectively. Assume that P_X and P_Y admit densities p_X and p_Y respectively, with $p_X \in \mathcal{W}^{\beta,2}(M_1)$ and $P_Y \in \mathcal{W}^{\beta,2}(M_2)$ with $M_1 \times M_2 = M$. Let $k_n(x, x') = \exp\left(-c_n\|x - x'\|_2^2\right)$ and $\ell_n(y, y') = \exp(-c_n\|y - y'\|_2^2)$ denote Gaussian kernels on $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_1}$ respectively, with $c_n = o(n^{4/d})$. Then, we have the following:*

$$\mathbb{E}_{P_X}\left[\widetilde{k}_n^2(X_1, X_2)\right] \asymp c_n^{-d_1/2}, \quad \textit{and} \quad \mathbb{E}_{P_Y}\left[\widetilde{\ell}_n^2(Y_1, Y_2)\right] \asymp c_n^{-d_2/2}, \tag{72}$$

$$\mathbb{E}_{P_X}\left[\widetilde{k}_n^4(X_1, X_2)\right] \asymp c_n^{-d_1/2}, \quad \textit{and} \quad \mathbb{E}_{P_Y}\left[\widetilde{\ell}_n^4(Y_1, Y_2)\right] \asymp c_n^{-d_2/2}, \tag{73}$$

$$\mathbb{E}_{P_X}\left[\widetilde{k}_n^2(X_1, X_2)\widetilde{k}_n^2(X_1, X_3)\right] \lesssim c_n^{-3d_1/4}, \quad \textit{and} \quad \mathbb{E}_{P_Y}\left[\widetilde{\ell}_n^2(Y_1, Y_2)\widetilde{\ell}_n^2(Y_1, Y_3)\right] \lesssim c_n^{-3d_2/4}. \tag{74}$$

In the expressions above, $\asymp$ and $\lesssim$ hide multiplicative factors depending on d_1, d_2, d and M.

The above stated conditions will be used to show the type-I error control by the cross-HSIC test in Appendix E.3.1. We now state the properties required for the proof of consistency.

Fact 34 *Suppose $Z_i = (X_i, Y_i) \sim P_{XY}$, for $i = 1, 2, 3$, and assume that P_{XY}, P_X and P_Y have smooth densities p_{XY}, p_X and p_Y respectively. Then, we have the following:*

$$\max\left(\mathbb{E}_{P_{XY}}[k_n(X,X)\ell_n(Y,Y)],\ \mathbb{E}_{P_{XY}}[k_n(X,X)]\mathbb{E}_{P_{XY}}[\ell_n(Y,Y)]\right) \lesssim M^2 c_n^{-d/2}, \quad (75)$$

$$\text{and} \quad \gamma_n^2 = \mathrm{HSIC}(P_{XY}, k_n, \ell_n) \gtrsim c_n^{-d/2}\|p_{XY} - p_X \times p_Y\|_{L^2}^2. \quad (76)$$

E.3.1 Type-I error bound

To show that the cross-HSIC test controls the type-I error at level α asymptotically, we verify that the two conditions of Assumption 1 are satisfied for the kernels considered in Theorem 10. In particular, using the fact that $\widetilde{g}(z_1, z_2) = \widetilde{k}(x_1, x_2) \times \widetilde{\ell}(y_1, y_2)$ for $z_i = (x_i, y_i)$ and $i = 1, 2$; we note that it suffices to verify the following conditions under the null:

$$\lim_{n\to\infty} \frac{1}{n^2} \frac{\mathbb{E}[\widetilde{k}(X_1,X_2)^4]}{\mathbb{E}[\widetilde{k}(X_1,X_2)^2]^2} \frac{\mathbb{E}[\widetilde{\ell}(Y_1,Y_2)^4]}{\mathbb{E}[\widetilde{\ell}(Y_1,Y_2)^2]^2} = 0, \quad (77)$$

$$\lim_{n\to\infty} \frac{1}{n} \frac{\mathbb{E}[\widetilde{k}(X_1,X_2)^2\widetilde{k}(X_1,X_3)^2]}{\mathbb{E}[\widetilde{k}(X_1,X_2)^2]^2} \frac{\mathbb{E}[\widetilde{\ell}(Y_1,Y_2)^2\widetilde{\ell}(Y_1,Y_3)^2]}{\mathbb{E}[\widetilde{\ell}(Y_1,Y_2)^2]^2} = 0, \quad (78)$$

$$\lim_{n\to\infty} \frac{\lambda_{1,n}^2}{\sum_{i=1}^{\infty}\lambda_{i,n}^2} = 0. \quad (79)$$

We proceed in the following steps:

Step 1: verification of (77)**.** Using the bounds stated in (72) and (73), we obtain

$$\frac{\mathbb{E}[\widetilde{k}(X_1,X_2)^4]}{\mathbb{E}[\widetilde{k}(X_1,X_2)^2]^2} \lesssim \frac{c_n^{-d_1/2}}{(c_n^{-d_1/2})^2}, \quad \text{and} \quad \frac{\mathbb{E}[\widetilde{\ell}(Y_1,Y_2)^4]}{\mathbb{E}[\widetilde{\ell}(Y_1,Y_2)^2]^2} \lesssim \frac{c_n^{-d_2/2}}{(c_n^{-d_2/2})^2},$$

which implies that

$$\frac{1}{n^2} \frac{\mathbb{E}[\widetilde{k}(X_1,X_2)^4]}{\mathbb{E}[\widetilde{k}(X_1,X_2)^2]^2} \frac{\mathbb{E}[\widetilde{\ell}(Y_1,Y_2)^4]}{\mathbb{E}[\widetilde{\ell}(Y_1,Y_2)^2]^2} \lesssim \frac{1}{n^2} \frac{c_n^{-d/2}}{(c_n^{-d/2})^2} = \frac{c_n^{d/2}}{n^2} \to 0,$$

as required in (77).

Step 2: verification of (78)**.** Considering the $\widetilde{k}$ dependent term of (78), we note that (73) and (74) together imply the following bound:

$$\frac{\mathbb{E}[\widetilde{k}(X_1,X_2)^2\widetilde{k}(X_1,X_3)^2]}{\mathbb{E}[\widetilde{k}(X_1,X_2)^2]^2} \lesssim \frac{c_n^{-3d_1/4}}{(c_n^{-d_1/2})^2} = c_n^{d_1/4}.$$

Similarly, the $\widetilde{\ell}$ dependent term is upper bounded by $c_n^{d_2/4}$, reducing the condition to

$$\frac{1}{n} \frac{\mathbb{E}[\widetilde{k}(X_1,X_2)^2\widetilde{k}(X_1,X_3)^2]}{\mathbb{E}[\widetilde{k}(X_1,X_2)^2]^2} \frac{\mathbb{E}[\widetilde{\ell}(Y_1,Y_2)^2\widetilde{\ell}(Y_1,Y_3)^2]}{\mathbb{E}[\widetilde{\ell}(Y_1,Y_2)^2]^2} \lesssim \frac{c_n^{d/4}}{n} = \left(\frac{c_n}{n^{4/d}}\right)^{d/4}.$$

Since $c_n = o(n^{4/d})$, the result follows.

Step 3: verification of (79)**.** We first note that the condition in (79) is equivalent to

$$\lim_{n\to\infty} \frac{\mathbb{E}\left[\mathbb{E}[\widetilde{g}(Z_1, Z_2)\widetilde{g}(Z_1, Z_3)|Z_2, Z_3]\right]}{\mathbb{E}[\widetilde{g}(Z_1, Z_2)^2]^2} \to 0.$$

Now, the denominator term satisfies:

$$\mathbb{E}[\widetilde{k}(X_1, X_2)^2] \times \mathbb{E}[\widetilde{\ell}(X_1, Y_2)^2] \asymp c_n^{-d_1/2} \times c_n^{-d_2/2} = c_n^{-d/2},$$

which implies that it suffices to show

$$c_n^d\, \mathbb{E}\left[\mathbb{E}[\widetilde{g}(Z_1, Z_2)\widetilde{g}(Z_1, Z_3)|Z_2, Z_3]\right] \to 0.$$

For Gaussian kernels with $c_n = o(n^{4/d})$, Li and Yuan (2019) showed that the above condition is true, in the course of proving their Theorem 1.

Hence, we have verified all the requirements for the limiting null distribution of $\overline{\mathrm{x}}\mathrm{HSIC}_n$ to be $N(0,1)$, which in turn, implies that the cross-HSIC test controls the type-I error at level-α asymptotically.

E.3.2 Consistency

To show the consistency of the cross-HSIC test against local alternatives, we need to show that

$$\lim_{n\to\infty} D_n \equiv D_n(P_{XY}, k, \ell) = 0,$$

$$\text{where} \quad D_n(P_{XY}, k, \ell) := \frac{1}{n^2\delta_n\gamma_n^4}\left(\mathbb{E}\left[\left\langle \widetilde{h}_{12}, \widetilde{h}_{34}\right\rangle^2 + \gamma_n^2\left(\left\langle \widetilde{h}_{12}, \widetilde{h}_{12}\right\rangle + n\left\langle \widetilde{h}_{12}, \widetilde{h}_{13}\right\rangle\right]\right)\right).$$

Following the bounds derived in the proof of Theorem 9 in Appendix E.2, we have

$$\delta_n D_n \lesssim \frac{1}{n^2\gamma_n^4}\left(\mathbb{E}[\|h_{12}\|^2]^2 + \gamma_n^4 + \gamma_n^2(1+n)(\mathbb{E}[\|h_{12}\|^2] + \gamma_n^2)\right). \tag{80}$$

In the above display, we used (69), (70) and (71) to upper bound the three terms involved in the definition of D_n. Now, from Theorem 32, we know that $\mathbb{E}[\|h_{12}\|^2] \leq \mathbb{E}[k(X,X)\ell(Y,Y)] + \mathbb{E}[k(X,X)]\mathbb{E}[\ell(Y,Y)]$. Now, for the case of Gaussian kernels, this term is further upper bounded as follows, using (75):

$$\max\left(\mathbb{E}[k(X,X)\ell(Y,Y)],\ \mathbb{E}[k(X,X)]\mathbb{E}[\ell(Y,Y)]\right) \lesssim M^2 c_n^{-d/2}. \tag{81}$$

The final component of the proof is the fact that under the conditions of Theorem 10, we also have the following bound on the true HSIC value using (76):

$$\gamma_n^2 \gtrsim c_n^{-d/2}\|p_{XY} - p_X \times p_Y\|_{L^2}^2 > c_n^{-d/2}\Delta_n^2. \tag{82}$$

Plugging (81) and (82) into (80), we get

$$\begin{aligned} D_n\delta_n &\lesssim \frac{\mathbb{E}[\|h_{12}\|^2]^2}{n^2\gamma_n^4} + \frac{\mathbb{E}[\|h_{12}\|^2]}{n\gamma_n^2} + \frac{1}{n^2} + \frac{1}{n} \lesssim \frac{M^4\nu_n^{-d}}{n^2 c_n^{-d}\Delta_n^4} + \frac{M^2 c_n^{-d/2}}{n c_n^{-d/2}\Delta_n^2} + \frac{1}{n} \\ &\lesssim \frac{1}{\left(n^{1/2}\Delta_n\right)^4} + \frac{1}{\left(n^{1/2}\Delta_n\right)^2} \lesssim \frac{1}{\left(n^{1/2}\Delta_n\right)^2}. \end{aligned} \tag{83}$$

Introduce the term $\beta' = \frac{1}{2} - \frac{2\beta}{d+4\beta} = \frac{d}{2(d+4\beta)} > 0$, and note that (83) implies

$$D_n \lesssim \frac{1}{\delta_n n^{2\beta'} \left(\Delta_n n^{2\beta/(d+4\beta)}\right)^2}.$$

By selecting $\delta_n = n^{-2\beta''}$ for any $0 < \beta'' < \beta'$, and using the assumptions that **(i)** $\lim_{n\to\infty} \Delta_n n^{2\beta/(d+4\beta)} = \infty$, and **(ii)** $\|p_{XY} - p_X \times p_Y\|_{L^2} > \Delta_n$ for all $P_{XY} \in \mathcal{P}_n^{(1)}$ with density p_{XY}; we have

$$\lim_{n\to\infty} \sup_{P_{XY} \in \mathcal{P}_n^{(1)}} D_n = 0.$$

By Theorem 29, the above condition implies the required consistency against smooth local alternatives of our cross-HSIC test.