An Adaptive Test of Independence with Analytic Kernel Embeddings

Wittawat Jitkrittum [1] **Zoltán Szabó** [2] **Arthur Gretton** [1]

Abstract

A new computationally efficient dependence measure, and an adaptive statistical test of independence, are proposed. The dependence measure is the difference between analytic embeddings of the joint distribution and the product of the marginals, evaluated at a finite set of locations (features). These features are chosen so as to maximize a lower bound on the test power, resulting in a test that is data-efficient, and that runs in linear time (with respect to the sample size n). The optimized features can be interpreted as evidence to reject the null hypothesis, indicating regions in the joint domain where the joint distribution and the product of the marginals differ most. Consistency of the independence test is established, for an appropriate choice of features. In real-world benchmarks, independence tests using the optimized features perform comparably to the state-of-the-art quadratic-time HSIC test, and outperform competing $\mathcal{O}(n)$ and $\mathcal{O}(n \log n)$ tests.

1. Introduction

We consider the design of adaptive, nonparametric statistical tests of dependence: that is, tests of whether a joint distribution P_{xy} factorizes into the product of marginals $P_x P_y$ with the null hypothesis that $H_0 : X$ and Y are independent. While classical tests of dependence, such as Pearson's correlation and Kendall's τ, are able to detect monotonic relations between univariate variables, more modern tests can address complex interactions, for instance changes in variance of X with the value of Y. Key to many recent tests is to examine covariance or correlation between data features. These interactions become significantly harder to detect, and the features are more difficult to design, when the data reside in high dimensions.

Zoltán Szabó's ORCID ID: 0000-0001-6183-7603. Arthur Gretton's ORCID ID: 0000-0003-3169-7624. [1]Gatsby Unit, University College London, UK. [2]CMAP, École Polytechnique, France. Correspondence to: Wittawat Jitkrittum <wittawatj@gmail.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

A basic nonlinear dependence measure is the Hilbert-Schmidt Independence Criterion (HSIC), which is the Hilbert-Schmidt norm of the covariance operator between feature mappings of the random variables (Gretton et al., 2005; 2008). Each random variable X and Y is mapped to a respective reproducing kernel Hilbert space $\mathcal{H}_k$ and $\mathcal{H}_l$. For sufficiently rich mappings, the covariance operator norm is zero if and only if the variables are independent. A second basic nonlinear dependence measure is the smoothed difference between the characteristic function of the joint distribution, and that of the product of marginals. When a particular smoothing function is used, the statistic corresponds to the covariance between distances of X and Y variable pairs (Feuerverger, 1993; Székely et al., 2007; Székely & Rizzo, 2009), yielding a simple test statistic based on pairwise distances. It has been shown by Sejdinovic et al. (2013) that the distance covariance (and its generalization to semi-metrics) is an instance of HSIC for an appropriate choice of kernels. A disadvantage of these feature covariance statistics, however, is that they require quadratic time to compute (besides in the special case of the distance covariance with univariate real-valued variables, where Huo & Székely (2016) achieve an $\mathcal{O}(n \log n)$ cost). Moreover, the feature covariance statistics have intractable null distributions, and either a permutation approach or the solution of an expensive eigenvalue problem (e.g. Zhang et al., 2011) is required for consistent estimation of the quantiles. Several approaches were proposed by Zhang et al. (2017) to obtain faster tests along the lines of HSIC. These include computing HSIC on finite-dimensional feature mappings chosen as random Fourier features (RFFs) (Rahimi & Recht, 2008), a block-averaged statistic, and a Nyström approximation to the statistic. Key to each of these approaches is a more efficient computation of the statistic and its threshold under the null distribution: for RFFs, the null distribution is a finite weighted sum of χ^2 variables; for the block-averaged statistic, the null distribution is asymptotically normal; for Nyström, either a permutation approach is employed, or the spectrum of the Nyström approximation to the kernel matrix is used in approximating the null distribution. Each of these methods costs significantly less than the $\mathcal{O}(n^2)$ cost of the full HSIC (the cost is linear in n, but also depends quadratically on the number of features retained). A potential disadvantage of the Nyström and Fourier approaches is that the features are not optimized to maximize test power, but are chosen randomly. The block statistic performs worse than both, due to the large variance of the statistic under the null (which can be mitigated by observing more data).

In addition to feature covariances, correlation measures have also been developed in infinite dimensional feature spaces: in particular, Bach & Jordan (2002); Fukumizu et al. (2008) proposed statistics on the correlation operator in a reproducing kernel Hilbert space. While convergence has been established for certain of these statistics, their computational cost is high at $\mathcal{O}(n^3)$, and test thresholds have relied on permutation. A number of much faster approaches to testing based on feature correlations have been proposed, however. For instance, Dauxois & Nkiet (1998) compute statistics of the correlation between finite sets of basis functions, chosen for instance to be step functions or low order B-splines. The cost of this approach is $\mathcal{O}(n)$. This idea was extended by Lopez-Paz et al. (2013), who computed the canonical correlation between finite sets of basis functions chosen as random Fourier features; in addition, they performed a copula transform on the inputs, with a total cost of $\mathcal{O}(n \log n)$. Finally, space partitioning approaches have also been proposed, based on statistics such as the KL divergence, however these apply only to univariate variables (Heller et al., 2016), or to multivariate variables of low dimension (Gretton & Györfi, 2010) (that said, these tests have other advantages of theoretical interest, notably distribution-independent test thresholds).

The approach we take is most closely related to HSIC on a finite set of features. Our simplest test statistic, the Finite Set Independence Criterion (FSIC), is an average of covariances of analytic functions (i.e., features) defined on each of X and Y. A normalized version of the statistic (NFSIC) yields a distribution-independent asymptotic test threshold. We show that our test is consistent, despite a finite number of analytic features being used, via a generalization of arguments in Chwialkowski et al. (2015). As in recent work on two-sample testing by Jitkrittum et al. (2016), our test is *adaptive* in the sense that we choose our features on a held-out validation set to optimize a lower bound on the test power. The design of features for independence testing turns out to be quite different to the case of two-sample testing, however: the task is to find correlated feature *pairs* on the respective marginal domains, rather than attempting to find a single, high-dimensional feature representation on the *tensor product* of the marginals, as we would need to do if we were comparing distributions P_{xy} and Q_{xy}. While the use of coupled feature pairs on the marginals entails a smaller feature space dimension, it introduces significant complications in the proof of the lower bound, compared with the two-sample case. We demonstrate the performance of our tests on several challenging artificial and real-world datasets, including detection of dependence between music and its year of appearance, and between videos and captions. In these experiments, we outperform competing linear and $\mathcal{O}(n \log n)$ time tests.

2. Independence Criteria and Statistical Tests

We introduce two test statistics: first, the Finite Set Independence Criterion (FSIC), which builds on the principle that dependence can be measured in terms of the covariance between data features. Next, we propose a normalized version of this statistic (NFSIC), with a simpler asymptotic distribution when $P_{xy} = P_x P_y$. We show how to select features for the latter statistic to maximize a lower bound on the power of its corresponding statistical test.

2.1. The Finite Set Independence Criterion

We begin by recalling the Hilbert-Schmidt Independence Criterion (HSIC) as proposed in Gretton et al. (2005), since our unnormalized statistic is built along similar lines. Consider two random variables $X \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $Y \in \mathcal{Y} \subseteq \mathbb{R}^{d_y}$. Denote by P_{xy} the joint distribution between X and Y; P_x and P_y are the marginal distributions of X and Y. Let $\otimes$ denote the tensor product, such that $(a \otimes b)\, c = a \langle b, c\rangle$. Assume that $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ are positive definite kernels associated with reproducing kernel Hilbert spaces (RKHS) $\mathcal{H}_k$ and $\mathcal{H}_l$, respectively. Let $\|\cdot\|_{HS}$ be the norm on the space of $\mathcal{H}_l \to \mathcal{H}_k$ Hilbert-Schmidt operators. Then, HSIC between X and Y is defined as

$$\begin{aligned}\mathrm{HSIC}(X, Y) &= \left\|\mu_{xy} - \mu_x \otimes \mu_y\right\|^2_{\mathrm{HS}} \\ &= \mathbb{E}_{(\mathbf{x},\mathbf{y}),(\mathbf{x}',\mathbf{y}')}\left[k(\mathbf{x},\mathbf{x}')l(\mathbf{y},\mathbf{y}')\right] \\ &\quad + \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{x}'}[k(\mathbf{x},\mathbf{x}')]\mathbb{E}_{\mathbf{y}}\mathbb{E}_{\mathbf{y}'}[l(\mathbf{y},\mathbf{y}')] \\ &\quad - 2\mathbb{E}_{(\mathbf{x},\mathbf{y})}\left[\mathbb{E}_{\mathbf{x}'}[k(\mathbf{x},\mathbf{x}')]\mathbb{E}_{\mathbf{y}'}[l(\mathbf{y},\mathbf{y}')]\right], \end{aligned} \tag{1}$$

where $\mathbb{E}_{\mathbf{x}} := \mathbb{E}_{\mathbf{x}\sim P_x}$, $\mathbb{E}_{\mathbf{y}} := \mathbb{E}_{\mathbf{y}\sim P_y}$, $\mathbb{E}_{\mathbf{xy}} := \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P_{xy}}$, and $\mathbf{x}'$ is an independent copy of $\mathbf{x}$. The mean embedding of P_{xy} belongs to the space of Hilbert-Schmidt operators from $\mathcal{H}_l$ to $\mathcal{H}_k$, $\mu_{xy} := \int_{\mathcal{X}\times\mathcal{Y}} k(\mathbf{x},\cdot) \otimes l(\mathbf{y},\cdot)\,\mathrm{d}P_{xy}(\mathbf{x},\mathbf{y}) \in \mathrm{HS}(\mathcal{H}_l, \mathcal{H}_k)$, and the marginal mean embeddings are $\mu_x := \int_{\mathcal{X}} k(\mathbf{x},\cdot)\,\mathrm{d}P_x(\mathbf{x}) \in \mathcal{H}_k$ and $\mu_y := \int_{\mathcal{Y}} l(\mathbf{y},\cdot)\,\mathrm{d}P_y(\mathbf{y}) \in \mathcal{H}_l$ (Smola et al., 2007). Gretton et al. (2005, Theorem 4) show that if the kernels k and l are universal (Steinwart & Christmann, 2008) on compact domains $\mathcal{X}$ and $\mathcal{Y}$, then $\mathrm{HSIC}(X, Y) = 0$ if and only if X and Y are independent. Alternatively, Gretton (2015) shows that it is sufficient for each of k and l to be characteristic to their respective domains (meaning that distribution embeddings are injective in each marginal domain: see Sriperumbudur et al. (2010)). Given a joint sample $\mathsf{Z}_n = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \sim P_{xy}$, an empirical estimator of HSIC can be computed in $\mathcal{O}(n^2)$ time by replacing the population expectations in (1) with their corresponding empirical expectations based on Z_n.

We now propose our new linear-time dependence measure, the Finite Set Independence Criterion (FSIC). Let

$\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ be open sets. Let $\mu_x\mu_y(\mathbf{x}, \mathbf{y}) := \mu_x(\mathbf{x})\mu_y(\mathbf{y})$ The idea is to see $\mu_{xy}(\mathbf{v}, \mathbf{w}) = \mathbb{E}_{\mathbf{xy}}[k(\mathbf{x}, \mathbf{v})l(\mathbf{y}, \mathbf{w})]$, $\mu_x(\mathbf{v}) = \mathbb{E}_{\mathbf{x}}[k(\mathbf{x}, \mathbf{v})]$ and $\mu_y(\mathbf{w}) = \mathbb{E}_{\mathbf{y}}[l(\mathbf{y}, \mathbf{w})]$ as smooth functions, and consider a new distance between μ_{xy} and $\mu_x\mu_y$ instead of a Hilbert-Schmidt distance as in HSIC (Gretton et al., 2005). The new measure is given by the average of squared differences between μ_{xy} and $\mu_x\mu_y$, evaluated at J random test locations $V_J := \{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J \subset \mathcal{X} \times \mathcal{Y}$.

$$\mathrm{FSIC}^2(X, Y) := \frac{1}{J}\sum_{i=1}^{J} u^2(\mathbf{v}_i, \mathbf{w}_i) = \frac{1}{J}\|\mathbf{u}\|_2^2, \text{ where}$$

$$\begin{aligned} u(\mathbf{v}, \mathbf{w}) &:= \mu_{xy}(\mathbf{v}, \mathbf{w}) - \mu_x(\mathbf{v})\mu_y(\mathbf{w}) \\ &= \mathbb{E}_{\mathbf{xy}}[k(\mathbf{x}, \mathbf{v})l(\mathbf{y}, \mathbf{w})] - \mathbb{E}_{\mathbf{x}}[k(\mathbf{x}, \mathbf{v})]\mathbb{E}_{\mathbf{y}}[l(\mathbf{y}, \mathbf{w})], \quad (2) \\ &= \mathrm{cov}_{\mathbf{xy}}[k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w})], \end{aligned}$$

$\mathbf{u} := (u(\mathbf{v}_1, \mathbf{w}_1), \ldots, u(\mathbf{v}_J, \mathbf{w}_J))^\top$, and $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$ are realizations from an absolutely continuous distribution (wrt the Lebesgue measure).

Our first result in Proposition 2 states that $\mathrm{FSIC}(X, Y)$ almost surely defines a dependence measure for the random variables X and Y, provided that the kernels k and l satisfy some conditions summarized in Assumption A.

Definition 1 (A_0 kernels)**.** Let $\mathcal{X}$ be an open set in $\mathbb{R}^d$. A positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be A_0 if it is bounded (i.e., there exists $B \in \mathbb{R}$ such that $\sup_{\mathbf{x},\mathbf{x}' \in \mathcal{X}} k(\mathbf{x}, \mathbf{x}') \le B$), analytic (Chwialkowski et al., 2015) and vanishes at infinity: for all $\mathbf{v} \in \mathcal{X}$, $f(\mathbf{x}) := k(\mathbf{x}, \mathbf{v})$ is bounded, real analytic on $\mathcal{X}$, and for all $\epsilon > 0$ the set $\{\mathbf{x} \mid |f(\mathbf{x})| \ge \epsilon\}$ is compact.[1]

Assumption A. *The kernels $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $l\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ are A_0 (assumed to be bounded by B_k and B_l respectively), characteristic (Sriperumbudur et al., 2010, Definition 6), and translation invariant i.e., there exist $\check{k}$ and $\check{l}$ such that $k(\mathbf{x}, \mathbf{x}') = \check{k}(\mathbf{x} - \mathbf{x}')$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, and $l(\mathbf{y}, \mathbf{y}') = \check{l}(\mathbf{y} - \mathbf{y}')$ for all $\mathbf{y}, \mathbf{y}' \in \mathcal{Y}$.*

Proposition 2 (FSIC is a dependence measure)**.** *Assume that assumption A holds, and that the test locations $V_J = \{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$ are drawn from an absolutely continuous distribution η. Then, η-almost surely, it holds that $\mathrm{FSIC}(X, Y) = \frac{1}{\sqrt{J}}\|\mathbf{u}\|_2 = 0$ if and only if X and Y are independent.*

Proof. We will prove the forward direction. The backward direction is obvious. Let $u := \mu_{xy} - \mu_x \otimes \mu_y$, a member of the RKHS $\mathcal{H}_k \times \mathcal{H}_l$ associated with the product kernel $g((\mathbf{x}, \mathbf{y}), (\mathbf{v}, \mathbf{w})) := k(\mathbf{x}, \mathbf{v})l(\mathbf{y}, \mathbf{w})$. Since k and l are C_0-kernels, characteristic and translation invariant, Gretton (2015, Theorem 2) implies that $u = 0$ if and only if $P_{xy} = P_x P_y$. Since for all $\mathbf{v} \in \mathcal{X}, \mathbf{w} \in \mathcal{Y}$, $\mathbf{x} \mapsto k(\mathbf{x}, \mathbf{v})$ and $\mathbf{y} \mapsto l(\mathbf{y}, \mathbf{w})$ are real analytic, it follows from Krantz & Parks (2002, Proposition 2.2.10) that $(\mathbf{x}, \mathbf{y}) \mapsto k(\mathbf{x}, \mathbf{v})l(\mathbf{y}, \mathbf{w})$ is analytic on $\mathcal{X} \times \mathcal{Y}$. Since g is analytic, and bounded, Lemma 9 ensures that u is a real analytic function. It is known that if $u \neq 0$, the set of roots $R_u := \{(\mathbf{v}, \mathbf{w}) \mid u(\mathbf{v}, \mathbf{w}) = 0\}$ has Lebesgue measure zero (Mityagin, 2015). Hence, it is sufficient to draw $(\mathbf{v}, \mathbf{w})$ from an absolutely continuous distribution to have $(\mathbf{v}, \mathbf{w}) \notin R_u$ η-almost surely, and consequently if X and Y are dependent, then $\mathrm{FSIC}(X, Y) > 0$, η-almost surely. □

Examples of kernels k and l which satisfy Assumption A are the Gaussian kernels. FSIC uses μ_{xy} as a proxy for P_{xy}, and $\mu_x\mu_y$ as a proxy for $P_x P_y$. Proposition 2 states that, to detect the dependence between X and Y, it is sufficient to evaluate the difference of the population joint embedding μ_{xy} and the embedding of the product of the marginal distributions $\mu_x\mu_y$ at a finite number of locations (defined by V_J). The intuitive explanation of this property is as follows. If $P_{xy} = P_x P_y$, then $u(\mathbf{v}, \mathbf{w}) = 0$ everywhere, and $\mathrm{FSIC}(X, Y) = 0$ for any V_J. If $P_{xy} \neq P_x P_y$, then u will not be a zero function. Using the same argument as in Chwialkowski et al. (2015), since k and l are analytic, u is also analytic, and the set of roots $R_u := \{(\mathbf{v}, \mathbf{w}) \mid u(\mathbf{v}, \mathbf{w}) = 0\}$ has Lebesgue measure zero (Mityagin, 2015). Thus, it is sufficient to draw $(\mathbf{v}, \mathbf{w})$ from an absolutely continuous distribution to have $(\mathbf{v}, \mathbf{w}) \notin R_u$ η-almost surely, and hence $\mathrm{FSIC}(X, Y) > 0$. We note that a characteristic kernel which is not analytic may produce u such that R_u has a positive Lebesgue measure. In this case, there is a positive probability that $(\mathbf{v}, \mathbf{w}) \in R_u$, resulting in a potential failure to detect the dependence. The required Assumption A only imposes conditions separately on each of the marginal kernels k and l. In particular, there is no condition on the product kernel $g((\mathbf{x}, \mathbf{y}), (\mathbf{v}, \mathbf{w})) := k(\mathbf{x}, \mathbf{v})l(\mathbf{y}, \mathbf{w})$ which would have been much more restrictive.

Plug-in Estimator Assume that we observe a joint sample $\mathsf{Z}_n := \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P_{xy}$. Unbiased estimators of $\mu_{xy}(\mathbf{v}, \mathbf{w})$ and $\mu_x\mu_y(\mathbf{v}, \mathbf{w})$ are $\hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) := \frac{1}{n}\sum_{i=1}^n k(\mathbf{x}_i, \mathbf{v})l(\mathbf{y}_i, \mathbf{w})$ and $\widehat{\mu_x\mu_y}(\mathbf{v}, \mathbf{w}) := \frac{1}{n(n-1)}\sum_{i=1}^n \sum_{j \neq i} k(\mathbf{x}_i, \mathbf{v})l(\mathbf{y}_j, \mathbf{w})$, respectively. A straightforward empirical estimator of FSIC^2 is then given by

$$\widehat{\mathrm{FSIC}^2}(\mathsf{Z}_n) = \frac{1}{J}\sum_{i=1}^{J} \hat{u}(\mathbf{v}_i, \mathbf{w}_i)^2,$$

$$\begin{aligned} \hat{u}(\mathbf{v}, \mathbf{w}) &:= \hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) - \widehat{\mu_x\mu_y}(\mathbf{v}, \mathbf{w}) \quad (3) \\ &= \frac{2}{n(n-1)}\sum_{i<j} h_{(\mathbf{v},\mathbf{w})}((\mathbf{x}_i, \mathbf{y}_i), (\mathbf{x}_j, \mathbf{y}_j)), \quad (4) \end{aligned}$$

where $h_{(\mathbf{v},\mathbf{w})}((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) := \frac{1}{2}(k(\mathbf{x}, \mathbf{v}) -$

[1] A related class is the set of C_0 kernels (Sriperumbudur et al., 2010, Section 2). A kernel k is said to be C_0 if it is bounded with $k(\cdot, \mathbf{v}) \in C_0(\mathcal{X})$ for all $\mathbf{v} \in \mathcal{X}$ where $C_0(\mathcal{X})$ contains the set of functions that vanish at infinity. Note that if k is A_0, then it is C_0.

$k(\mathbf{x}', \mathbf{v}))(l(\mathbf{y}, \mathbf{w}) - l(\mathbf{y}', \mathbf{w}))$. For conciseness, we define $\hat{\mathbf{u}} := (\hat{u}_1, \ldots, \hat{u}_J)^\top \in \mathbb{R}^J$ where $\hat{u}_i := \hat{u}(\mathbf{v}_i, \mathbf{w}_i)$ so that $\widehat{\mathrm{FSIC}^2}(\mathsf{Z}_n) = \frac{1}{J}\hat{\mathbf{u}}^\top \hat{\mathbf{u}}$.

$\widehat{\mathrm{FSIC}^2}$ can be efficiently computed in $\mathcal{O}((d_x + d_y)Jn)$ time which is linear in n [see (3) which does not have nested double sums], assuming that the runtime complexity of evaluating $k(\mathbf{x}, \mathbf{v})$ is $\mathcal{O}(d_x)$ and that of $l(\mathbf{y}, \mathbf{w})$ is $\mathcal{O}(d_y)$.

Since FSIC satisfies $\mathrm{FSIC}(X, Y) = 0 \iff X \perp Y$, in principle its empirical estimator can be used as a test statistic for an independence test proposing a null hypothesis H_0 : "X and Y are independent" against an alternative H_1 : "X and Y are dependent." The null distribution (i.e., distribution of the test statistic assuming that H_0 is true) is challenging to obtain, however, and depends on the unknown P_{xy}. This prompts us to consider a normalized version of FSIC whose asymptotic null distribution takes a more convenient form. We first derive the asymptotic distribution of $\hat{\mathbf{u}}$ in Proposition 3, which we use to derive the normalized test statistic in Theorem 4. As a shorthand, we write $\mathbf{z} := (\mathbf{x}, \mathbf{y})$, $\mathbf{t} := (\mathbf{v}, \mathbf{w})$, $\mathrm{cov}_\mathbf{z}$ is covariance,$\mathbb{V}_\mathbf{z}$ stands for variance.

Proposition 3 (Asymptotic distribution of $\hat{\mathbf{u}}$). *Define $\mathbf{u} := (u(\mathbf{t}_1), \ldots, u(\mathbf{t}_J))^\top$, $\tilde{k}(\mathbf{x}, \mathbf{v}) := k(\mathbf{x}, \mathbf{v}) - \mathbb{E}_{\mathbf{x}'} k(\mathbf{x}', \mathbf{v})$, and $\tilde{l}(\mathbf{y}, \mathbf{w}) := l(\mathbf{y}, \mathbf{w}) - \mathbb{E}_{\mathbf{y}'} l(\mathbf{y}', \mathbf{w})$. Let $\mathbf{\Sigma} = [\Sigma_{ij}] \in \mathbb{R}^{J \times J}$ be the positive semi-definite matrix with entries $\Sigma_{ij} = \mathrm{cov}_\mathbf{z}(\hat{u}(\mathbf{t}_i), \hat{u}(\mathbf{t}_j)) = \mathbb{E}_{\mathbf{xy}}[\tilde{k}(\mathbf{x}, \mathbf{v}_i)\tilde{l}(\mathbf{y}, \mathbf{w}_i)\tilde{k}(\mathbf{x}, \mathbf{v}_j)\tilde{l}(\mathbf{y}, \mathbf{w}_j)] - u(\mathbf{t}_i)u(\mathbf{t}_j)$. Then, under both H_0 and H_1, for any fixed test locations $\{\mathbf{t}_1, \ldots, \mathbf{t}_J\}$ for which $\mathbf{\Sigma}$ is full rank, and $0 < \mathbb{V}_\mathbf{z}[h_{\mathbf{t}_j}(\mathbf{z})] < \infty$ for $j = 1, \ldots, J$, it holds that $\sqrt{n}(\hat{\mathbf{u}} - \mathbf{u}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$.*

Proof. For a fixed $\{\mathbf{t}_1, \ldots, \mathbf{t}_J\}$, $\hat{\mathbf{u}}$ is a one-sample second-order multivariate U-statistic with a U-statistic kernel $h_\mathbf{t}$. Thus, by Lehmann (1999, Theorem 6.1.6) and Kowalski & Tu (2008, Section 5.1, Theorem 1), it follows directly that $\sqrt{n}(\hat{\mathbf{u}} - \mathbf{u}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ where we note that $\mathbb{E}_{\mathbf{xy}}[\tilde{k}(\mathbf{x}, \mathbf{v})\tilde{l}(\mathbf{y}, \mathbf{w})] = u(\mathbf{v}, \mathbf{w})$. □

Recall from Proposition 2 that $\mathbf{u} = \mathbf{0}$ holds almost surely under H_0. The asymptotic normality described in Proposition 3 implies that $n\widehat{\mathrm{FSIC}^2} = \frac{n}{J}\hat{\mathbf{u}}^\top \hat{\mathbf{u}}$ converges in distribution to a sum of J dependent weighted χ^2 random variables. The dependence comes from the fact that the coordinates $\hat{u}_1 \ldots, \hat{u}_J$ of $\hat{\mathbf{u}}$ all depend on the sample Z_n. This null distribution is not analytically tractable, and requires a large number of simulations to compute the rejection threshold T_α for a given significance value α.

2.2. Normalized FSIC and Adaptive Test

For the purpose of an independence test, we will consider a normalized variant of $\widehat{\mathrm{FSIC}^2}$, which we call $\widehat{\mathrm{NFSIC}^2}$, whose tractable asymptotic null distribution is $\chi^2(J)$, the chi-squared distribution with J degrees of freedom. We then show that the independence test defined by $\widehat{\mathrm{NFSIC}^2}$ is consistent. These results are given in Theorem 4.

Theorem 4 (Independence test based on $\widehat{\mathrm{NFSIC}^2}$ is consistent). *Let $\hat{\mathbf{\Sigma}}$ be a consistent estimate of $\mathbf{\Sigma}$ based on the joint sample Z_n, where $\mathbf{\Sigma}$ is defined in Proposition 3. Assume that $V_J = \{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J \sim \eta$ where η is absolutely continuous wrt the Lebesgue measure. The $\widehat{\mathrm{NFSIC}^2}$ statistic is defined as $\hat{\lambda}_n := n\hat{\mathbf{u}}^\top \left(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}}$ where $\gamma_n \geq 0$ is a regularization parameter. Assume that*

1. *Assumption A holds.*
2. *$\mathbf{\Sigma}$ is invertible η-almost surely.*
3. *$\lim_{n\to\infty} \gamma_n = 0$.*

Then, for any k, l and V_J satisfying the assumptions,

1. *Under H_0, $\hat{\lambda}_n \xrightarrow{d} \chi^2(J)$ as $n \to \infty$.*
2. *Under H_1, for any $r \in \mathbb{R}$, $\lim_{n\to\infty} \mathbb{P}\left(\hat{\lambda}_n \geq r\right) = 1$ η-almost surely. That is, the independence test based on $\widehat{\mathrm{NFSIC}^2}$ is consistent.*

Proof (sketch). Under H_0, $n\hat{\mathbf{u}}^\top(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I})^{-1}\hat{\mathbf{u}}$ asymptotically follows $\chi^2(J)$ because $\sqrt{n}\hat{\mathbf{u}}$ is asymptotically normally distributed (see Proposition 3). Claim 2 builds on the result in Proposition 2 stating that $\mathbf{u} \neq \mathbf{0}$ under H_1; it follows using the convergence of $\hat{\mathbf{u}}$ to $\mathbf{u}$. The full proof can be found in Appendix D. □

Theorem 4 states that if H_1 holds, the statistic can be arbitrarily large as n increases, allowing H_0 to be rejected for any fixed threshold. Asymptotically the test threshold T_α is given by the $(1 - \alpha)$-quantile of $\chi^2(J)$ and is independent of n. The assumption on the consistency of $\hat{\mathbf{\Sigma}}$ is required to obtain the asymptotic chi-squared distribution. The regularization parameter γ_n is to ensure that $(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I})^{-1}$ can be stably computed. In practice, γ_n requires no tuning, and can be set to be a very small constant. We emphasize that J need not increase with n for test consistency.

The next proposition states that the computational complexity of the $\widehat{\mathrm{NFSIC}^2}$ estimator is linear in both the input dimension and sample size, and that it can be expressed in terms of the $\mathbf{K} = [K_{ij}] = [k(\mathbf{v}_i, \mathbf{x}_j)] \in \mathbb{R}^{J \times n}$, $\mathbf{L} = [L_{ij}] = [l(\mathbf{w}_i, \mathbf{y}_j)] \in \mathbb{R}^{J \times n}$ matrices. In contrast to typical kernel methods, a large Gram matrix of size $n \times n$ is not needed to compute $\widehat{\mathrm{NFSIC}^2}$.

Proposition 5 (An empirical estimator of $\widehat{\mathrm{NFSIC}^2}$). *Let $\mathbf{1}_n := (1, \ldots, 1)^\top \in \mathbb{R}^n$. Denote by $\circ$ the element-wise matrix product. Then,*

1. $\hat{\mathbf{u}} = \frac{(\mathbf{K}\circ\mathbf{L})\mathbf{1}_n}{n-1} - \frac{(\mathbf{K}\mathbf{1}_n)\circ(\mathbf{L}\mathbf{1}_n)}{n(n-1)}$.
2. *A consistent estimator for* $\mathbf{\Sigma}$ *is* $\hat{\mathbf{\Sigma}} = \frac{\mathbf{\Gamma}\mathbf{\Gamma}^\top}{n}$ *where*

$$\mathbf{\Gamma} := (\mathbf{K} - n^{-1}\mathbf{K}\mathbf{1}_n\mathbf{1}_n^\top) \circ (\mathbf{L} - n^{-1}\mathbf{L}\mathbf{1}_n\mathbf{1}_n^\top) - \hat{\mathbf{u}}^b\mathbf{1}_n^\top,$$
$$\hat{\mathbf{u}}^b = n^{-1}(\mathbf{K}\circ\mathbf{L})\mathbf{1}_n - n^{-2}(\mathbf{K}\mathbf{1}_n)\circ(\mathbf{L}\mathbf{1}_n).$$

Assume that the complexity of the kernel evaluation is linear in the input dimension. Then the test statistic $\hat{\lambda}_n = n\hat{\mathbf{u}}^\top\left(\hat{\mathbf{\Sigma}} + \gamma_n\mathbf{I}\right)^{-1}\hat{\mathbf{u}}$ *can be computed in* $\mathcal{O}(J^3 + J^2 n + (d_x + d_y)Jn)$ *time.*

Proof (sketch). Claim 1 for $\hat{\mathbf{u}}$ is straightforward. The expression for $\hat{\mathbf{\Sigma}}$ in claim 2 follows directly from the asymptotic covariance expression in Proposition 3. The consistency of $\hat{\mathbf{\Sigma}}$ can be obtained by noting that the finite sample bound for $\mathbb{P}(\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F > t)$ decreases as n increases. This is implicitly shown in Appendix E.2.2 and its following sections. □

Although the dependency of the estimator on J is cubic, we empirically observe that only a small value of J is required (see Section 3). The number of test locations J relates to the number of regions in $\mathcal{X} \times \mathcal{Y}$ of p_{xy} and $p_x p_y$ that differ (see Figure 1).

Theorem 4 asserts the consistency of the test for any test locations V_J drawn from an absolutely continuous distribution. In practice, V_J can be further optimized to increase the test power for a fixed sample size. Our final theoretical result gives a lower bound on the test power of $\widehat{\mathrm{NFSIC}}^2$ i.e., the probability of correctly rejecting H_0. We will use this lower bound as the objective function to determine V_J and the kernel parameters. Let $\|\cdot\|_F$ be the Frobenius norm.

Theorem 6 (A lower bound on the test power)**.** *Let* $\mathrm{NFSIC}^2(X,Y) := \lambda_n := n\mathbf{u}^\top\mathbf{\Sigma}^{-1}\mathbf{u}$. *Let* $\mathcal{K}$ *be a kernel class for* k, $\mathcal{L}$ *be a kernel class for* l, *and* $\mathcal{V}$ *be a collection with each element being a set of* J *locations. Assume that*

1. *There exist finite* B_k *and* B_l *such that* $\sup_{k\in\mathcal{K}}\sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}}|k(\mathbf{x},\mathbf{x}')| \le B_k$ *and* $\sup_{l\in\mathcal{L}}\sup_{\mathbf{y},\mathbf{y}'\in\mathcal{Y}}|l(\mathbf{y},\mathbf{y}')| \le B_l$.
2. $\tilde{c} := \sup_{k\in\mathcal{K}}\sup_{l\in\mathcal{L}}\sup_{V_J\in\mathcal{V}}\|\mathbf{\Sigma}^{-1}\|_F < \infty$.

Then, for any $k \in \mathcal{K}, l \in \mathcal{L}, V_J \in \mathcal{V}$, *and* $\lambda_n \ge r$, *the test power satisfies* $\mathbb{P}\left(\hat{\lambda}_n \ge r\right) \ge L(\lambda_n)$ *where*

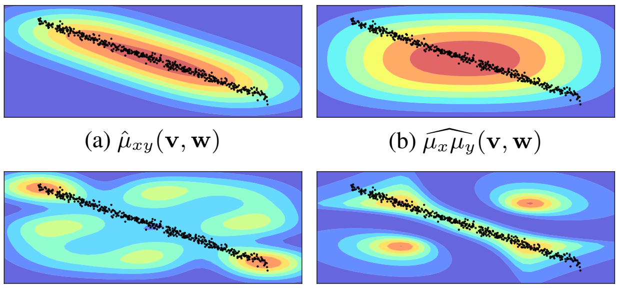
$$L(\lambda_n) = 1 - 62e^{-\xi_1\gamma_n^2(\lambda_n - r)^2/n} - 2e^{-\lfloor 0.5n\rfloor(\lambda_n-r)^2/\left[\xi_2 n^2\right]} - 2e^{-\left[(\lambda_n-r)\gamma_n(n-1)/3-\xi_3 n - c_3\gamma_n^2 n(n-1)\right]^2/\left[\xi_4 n^2(n-1)\right]},$$

$\lfloor\cdot\rfloor$ *is the floor function,* $\xi_1 := \frac{1}{3^2c_1^2J^2B^*}$, B^* *is a constant depending on only* B_k *and* B_l, $\xi_2 := 72c_2^2JB^2$, $B := B_kB_l$, $\xi_3 := 8c_1B^2J$, $c_3 := 4B^2J\tilde{c}^2$, $\xi_4 := 2^8B^4J^2c_1^2$, $c_1 := 4B^2J\sqrt{J}\tilde{c}$, *and* $c_2 := 4B\sqrt{J}\tilde{c}$. *Moreover, for sufficiently large fixed* n, $L(\lambda_n)$ *is increasing in* λ_n.

We provide the proof in Appendix E. To put Theorem 6 into perspective, assume that $\mathcal{K} = \left\{(\mathbf{x},\mathbf{v}) \mapsto \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma_x^2}\right) \mid \sigma_x^2 \in [\sigma_{x,l}^2, \sigma_{x,u}^2]\right\} =: \mathcal{K}_g$ for some $0 < \sigma_{x,l}^2 < \sigma_{x,u}^2 < \infty$ and $\mathcal{L} = \left\{(\mathbf{y},\mathbf{w}) \mapsto \exp\left(-\frac{\|\mathbf{y}-\mathbf{w}\|^2}{2\sigma_y^2}\right) \mid \sigma_y^2 \in [\sigma_{y,l}^2, \sigma_{y,u}^2]\right\} =: \mathcal{L}_g$ for some $0 < \sigma_{y,l}^2 < \sigma_{y,u}^2 < \infty$ are Gaussian kernel classes. Then, in Theorem 6, $B = B_k = B_l = 1$, and $B^* = 2$. The assumption $\tilde{c} < \infty$ is a technical condition to guarantee that the test power lower bound is finite for all θ defined by the feasible sets $\mathcal{K}, \mathcal{L}$, and $\mathcal{V}$. Let $\mathcal{V}_{\epsilon,r} := \left\{V_J \mid \|\mathbf{v}_i\|^2, \|\mathbf{w}_i\|^2 \le r \text{ and } \|\mathbf{v}_i - \mathbf{v}_j\|_2^2 + \|\mathbf{w}_i - \mathbf{w}_j\|_2^2 \ge \epsilon, \text{ for all } i \ne j\right\}$. If we set $\mathcal{K} = \mathcal{K}_g, \mathcal{L} = \mathcal{L}_g$, and $\mathcal{V} = \mathcal{V}_{\epsilon,r}$ for some $\epsilon, r > 0$, then $\tilde{c} < \infty$ as $\mathcal{K}_g, \mathcal{L}_g$, and $\mathcal{V}_{\epsilon,r}$ are compact. In practice, these conditions do not necessarily create restrictions as they almost always hold implicitly. We show in Appendix C that the objective function used to choose V_J will discourage any two locations to be in the same neighborhood.

Parameter Tuning Let θ be the collection of all tuning parameters of the test. If $k \in \mathcal{K}_g$ and $l \in \mathcal{L}_g$ (i.e., Gaussian kernels), then $\theta = \{\sigma_x^2, \sigma_y^2, V_J\}$. The test power lower bound $L(\lambda_n)$ in Theorem 6 is a function of $\lambda_n = n\mathbf{u}^\top\mathbf{\Sigma}^{-1}\mathbf{u}$ which is the population counterpart of the test statistic $\hat{\lambda}_n$. As in FSIC, it can be shown that $\lambda_n = 0$ if and only if X are Y are independent (from Proposition 2). According to Theorem 6, for a sufficiently large n, the test power lower bound is increasing in λ_n. One can therefore think of λ_n (a function of θ) as representing how easily the test rejects H_0 given a problem P_{xy}. The higher the λ_n, the greater the lower bound on the test power, and thus the more likely it is that the test will reject H_0 when it is false.

In light of this reasoning, we propose to set θ by maximizing the lower bound on the test power i.e., set θ to $\theta^* = \arg\max_\theta L(\lambda_n)$. Assume that n is sufficiently large so that $\lambda_n \mapsto L(\lambda_n)$ is an increasing function. Then, $\arg\max_\theta L(\lambda_n) = \arg\max_\theta \lambda_n$. That this procedure is also valid under H_0 can be seen as follows. Under H_0, $\theta^* = \arg\max_\theta 0$ will be arbitrary. Since Theorem 6 guarantees that $\hat{\lambda}_n \xrightarrow{d} \chi^2(J)$ as $n \to \infty$ for any θ, the asymptotic null distribution does not change by using θ^*. In practice, λ_n is a population quantity which is unknown. We propose dividing the sample Z_n into two disjoint sets: training and test sets. The training set is used to compute $\hat{\lambda}_n$ (an estimate of λ_n) to optimize for θ^*, and the test set is used for the actual independence test with the optimized θ^*. The splitting is to guarantee the independence of θ^* and the test sample to avoid overfitting.



(a) $\hat{\mu}_{xy}(\mathbf{v}, \mathbf{w})$ (b) $\widehat{\mu_x \mu_y}(\mathbf{v}, \mathbf{w})$

(c) $\widehat{\mathbf{\Sigma}}(\mathbf{v}, \mathbf{w})$ (d) Statistic $\hat{\lambda}_n(\mathbf{v}, \mathbf{w})$

Figure 1: Illustration of $\widehat{\mathrm{NFSIC}}^2$.

To better understand the behaviour of $\widehat{\mathrm{NFSIC}}^2$, we visualize $\hat{\mu}_{xy}(\mathbf{v}, \mathbf{w})$, $\widehat{\mu_x \mu_y}(\mathbf{v}, \mathbf{w})$ and $\hat{\mathbf{\Sigma}}(\mathbf{v}, \mathbf{w})$ as a function of one test location $(\mathbf{v}, \mathbf{w})$ on a simple toy problem. In this problem, $Y = -X + Z$ where $Z \sim \mathcal{N}(0, 0.3^2)$ is an independent noise variable. As we consider only one location $(J = 1)$, $\hat{\mathbf{\Sigma}}(\mathbf{v}, \mathbf{w})$ is a scalar. The statistic can be written as $\hat{\lambda}_n = n \frac{\left(\hat{\mu}_{xy}(\mathbf{v},\mathbf{w}) - \widehat{\mu_x \mu_y}(\mathbf{v},\mathbf{w})\right)^2}{\hat{\mathbf{\Sigma}}(\mathbf{v},\mathbf{w})}$. These components are shown in Figure 1, where we use Gaussian kernels for both X and Y, and the horizontal and vertical axes correspond to $\mathbf{v} \in \mathbb{R}$ and $\mathbf{w} \in \mathbb{R}$, respectively.

Intuitively, $\hat{u}(\mathbf{v}, \mathbf{w}) = \hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) - \widehat{\mu_x \mu_y}(\mathbf{v}, \mathbf{w})$ captures the difference of the joint distribution and the product of the marginals as a function of $(\mathbf{v}, \mathbf{w})$. Squaring $\hat{u}(\mathbf{v}, \mathbf{w})$ and dividing it by the variance shown in Figure 1c gives the statistic (also the parameter tuning objective) shown in Figure 1d. The latter figure illustrates that the parameter tuning objective function can be non-convex: non-convexity arises since there are multiple ways to detect the difference between the joint distribution and the product of the marginals. In this case, the lower left and upper right regions equally indicate the largest difference. A convex objective would not be able to capture this phenomenon.

3. Experiments

In this section, we empirically study the performance of the proposed method on both toy (Section 3.1) and real problems (Section 3.2). We are interested in challenging problems requiring a large number of samples, where a quadratic-time test might be computationally infeasible. Our goal is not to outperform a quadratic-time test with a linear-time test uniformly over *all* testing problems. We will find, however, that our test does outperform the quadratic-time test in some cases. Code is available at `https://github.com/wittawatj/fsic-test`.

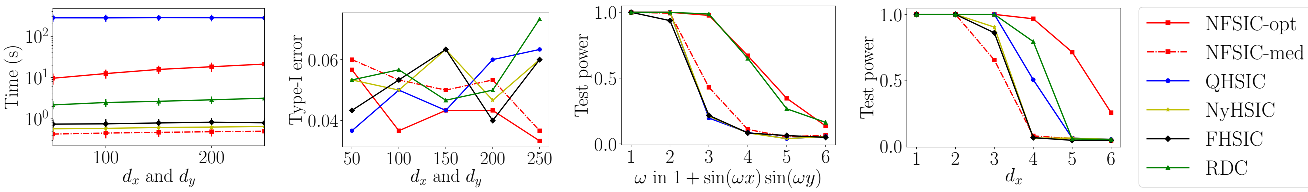
We compare the proposed NFSIC with optimization (NFSIC-opt) to five multivariate nonparametric tests. The $\widehat{\mathrm{NFSIC}}^2$ test without optimization (NFSIC-med) acts as a baseline, allowing the effect of parameter optimization to be clearly seen. For pedagogical reason, we consider the original HSIC test of Gretton et al. (2005) denoted by QHSIC, which is a quadratic-time test. Nyström HSIC (NyHSIC) uses a Nyström approximation to the kernel matrices of X and Y when computing the HSIC statistic. FHSIC is another variant of HSIC in which a random Fourier feature approximation (Rahimi & Recht, 2008) to the kernel is used. NyHSIC and FHSIC are studied in Zhang et al. (2017) and can be computed in $\mathcal{O}(n)$, with quadratic dependency on the number of inducing points in NyHSIC, and quadratic dependency on the number of random features in FHSIC. Finally, the Randomized Dependence Coefficient (RDC) proposed in Lopez-Paz et al. (2013) is also considered. The RDC can be seen as the primal form (with random Fourier features) of the kernel canonical correlation analysis of Bach & Jordan (2002) on copula-transformed data. We consider RDC as a linear-time test even though preprocessing by an empirical copula transform costs $\mathcal{O}((d_x + d_y) n \log n)$.

We use Gaussian kernel classes $\mathcal{K}_g$ and $\mathcal{L}_g$ for both X and Y in all the methods. Except NFSIC-opt, all other tests use full sample to conduct the independence test, where the Gaussian widths σ_x and σ_y are set according to the widely used median heuristic i.e., $\sigma_x = \mathrm{median}\left(\{\|\mathbf{x}_i - \mathbf{x}_j\|_2 \mid 1 \le i < j \le n\}\right)$, and σ_y is set in the same way using $\{\mathbf{y}_i\}_{i=1}^n$. The J locations for NFSIC-med are randomly drawn from the standard multivariate normal distribution in each trial. For a sample of size n, NFSIC-opt uses half the sample for parameter tuning, and the other disjoint half for the test. We permute the sample 300 times in RDC[2] and HSIC to simulate from the null distribution and compute the test threshold. The null distributions for FHSIC and NyHSIC are given by a finite sum of weighted $\chi^2(1)$ random variables given in Eq. 8 of Zhang et al. (2017). Unless stated otherwise, we set the test threshold of the two NFSIC tests to be the $(1-\alpha)$-quantile of $\chi^2(J)$. To provide a fair comparison, we set $J = 10$, use 10 inducing points in NyHSIC, and 10 random Fourier features in FHSIC and RDC.

Optimization of NFSIC-opt The parameters of NFSIC-opt are σ_x, σ_y, and J locations of size $(d_x + d_y)J$. We treat all the parameters as a long vector in $\mathbb{R}^{2+(d_x+d_y)J}$ and use gradient ascent to optimize $\hat{\lambda}_{n/2}$. We observe that initializing V_J by randomly picking J points from the training sample yields good performance. The regularization parameter γ_n in NFSIC is fixed to a small value, and is not optimized. It is worth emphasizing that the complexity of the optimization procedure is still linear-time.[3]

[2] We use a permutation test for RDC, following the authors' implementation (`https://github.com/lopezpaz/randomized_dependence_coefficient`, referred commit: b0ac6c0).

[3] Our claim on linear runtime (with respect to n) is for the gradient ascent procedure to find a local optimum for θ. We do not



(a) SG ($\alpha = 0.05$) (b) SG ($\alpha = 0.05$) (c) Sin (d) GSign

Figure 2: (a): Runtime. (b): Probability of rejecting H_0 as problem parameters vary. Fix $n = 4000$.

Since FSIC, NyHFSIC and RDC rely on a finite-dimensional kernel approximation, these tests are consistent only if both the number of features increases with n. By constrast, the proposed NFSIC requires only n to go to infinity to achieve consistency i.e., J can be fixed. We refer the reader to Appendix C for a brief investigation of the test power vs. increasing J. The test power does not necessarily monotonically increase with J.

3.1. Toy Problems

We consider three toy problems.

1. Same Gaussian (SG). The two variables are independently drawn from the standard multivariate normal distribution i.e., $X \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_x})$ and $Y \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_y})$ where $\mathbf{I}_d$ is the $d \times d$ identity matrix. This problem represents a case in which H_0 holds.

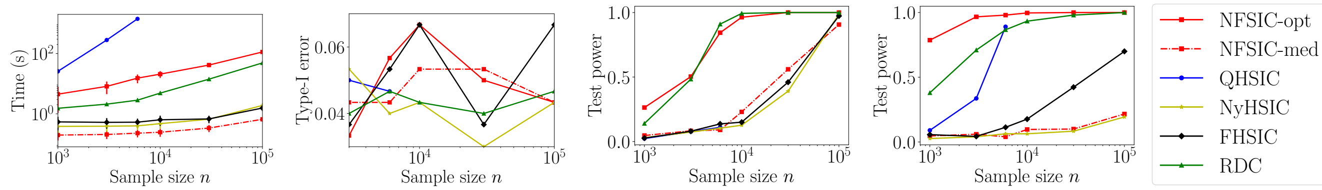
2. Sinusoid (Sin). Let p_{xy} be the probability density of P_{xy}. In the Sinusoid problem, the dependency of X and Y is characterized by $(X, Y) \sim p_{xy}(x, y) \propto 1 + \sin(\omega x)\sin(\omega y)$, where the domains of $\mathcal{X}, \mathcal{Y} = (-\pi, \pi)$ and ω is the frequency of the sinusoid. As the frequency ω increases, the drawn sample becomes more similar to a sample drawn from $\mathrm{Uniform}((-\pi, \pi)^2)$. That is, the higher ω, the harder to detect the dependency between X and Y. This problem was studied in Sejdinovic et al. (2013). Plots of the density for a few values of ω are shown in Figures 6 and 7 in the appendix. The main characteristic of interest in this problem is the local change in the density function.

3. Gaussian Sign (GSign). In this problem, $Y = |Z| \prod_{i=1}^{d_x} \mathrm{sgn}(X_i)$, where $X \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_x})$, $\mathrm{sgn}(\cdot)$ is the sign function, and $Z \sim \mathcal{N}(0, 1)$ serves as a source of noise. The full interaction of $X = (X_1, \ldots, X_{d_x})$ is what makes the problem challenging. That is, Y is dependent on X, yet it is independent of any proper subset of $\{X_1, \ldots, X_d\}$. Thus, simultaneous consideration of all the coordinates of X is required to successfully detect the dependency.

We fix $n = 4000$ and vary the problem parameters. Each problem is repeated for 300 trials, and the sample is redrawn each time. The significance level α is set to 0.05. The results are shown in Figure 2. It can be seen that in the SG problem (Figure 2b) where H_0 holds, all the tests achieve roughly correct type-I errors at $\alpha = 0.05$. In particular, we point out that NFSIC-opt's rejection rate is well controlled as the sample used for testing and the sample used for parameter tuning are independent. The rejection rate would have been much higher had we done the optimization and testing on the same sample (i.e., overfitting). In the Sin problem, NFSIC-opt achieves high test power for all considered $\omega = 1, \ldots, 6$, highlighting its strength in detecting local changes in the joint density. The performance of NFSIC-med is significantly lower than that of NFSIC-opt. This phenomenon clearly emphasizes the importance of the optimization to place the locations at the relevant regions in $\mathcal{X} \times \mathcal{Y}$. RDC has a remarkably high performance in both Sin and GSign (Figure 2c, 2d) despite no parameter tuning. The ability to simultaneously consider interacting features of NFSIC-opt is indicated by its superior test power in GSign, especially at the challenging settings of $d_x = 5, 6$.

NFSIC vs. QHSIC. We observe that NFSIC-opt outperforms the quadratic-time QHSIC in these two problems. QHSIC is defined as the RKHS norm of the witness function u (see (2)). Intuitively, one can think of the RKHS norm as taking into account all the locations $(\mathbf{v}, \mathbf{w})$. By contrast, the proposed NFSIC evaluates the witness function at J locations. If the differences in p_{xy} and $p_x p_y$ are local (e.g., Sin problem), or there are interacting features (e.g., GSign problem), then only small regions in the space of (X, Y) are relevant in detecting the difference of p_{xy} and $p_x p_y$. In these cases, pinpointing exact test locations by the optimization of NFSIC performs well. On the other hand, taking into account all possible test locations as done implicitly in QHSIC also integrates over regions where the difference between p_{xy} and $p_x p_y$ is small, resulting in a weaker indication of dependence. Whether QHSIC is better than NFSIC depends heavily on the problem, and there is no one best answer. If the difference between p_{xy} and $p_x p_y$ is large only in localized regions, then the proposed linear time statistic has an advantage. If the difference is spatially diffuse, then QHSIC has an advantage. No existing work has proposed a procedure to optimally tune kernel parameters for QHSIC; by contrast, NFSIC has a clearly defined objective for parameter tuning.

claim a linear runtime to find a global optimum.



(a) SG. $d_x = d_y = 250$. (b) SG. $d_x = d_y = 250$. (c) Sin. $\omega = 4$. (d) GSign. $d_x = 4$.

Figure 3: (a) Runtime. (b): Probability of rejecting H_0 as n increases in the toy problems.

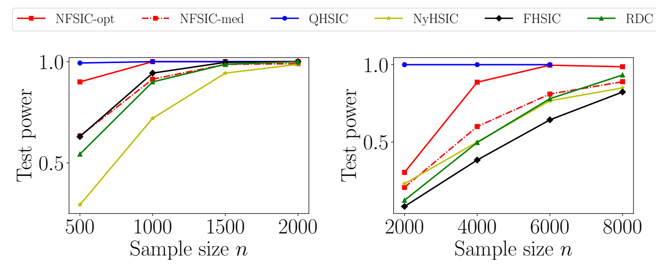
To investigate the sample efficiency of all the tests, we fix $d_x = d_y = 250$ in SG, $\omega = 4$ in Sin, $d_x = 4$ in GSign, and increase n. Figure 3 shows the results. The quadratic dependency on n in QHSIC makes it infeasible both in terms of memory and runtime to consider n larger than 6000 (Figure 3a). By constrast, although not the most time-efficient, NFSIC-opt has the highest sample-efficiency for GSign, and for Sin in the low-sample regime, significantly outperforming QHSIC. Despite the small additional overhead from the optimization, we are yet able to conduct an accurate test with $n = 10^5, d_x = d_y = 250$ in less than 100 seconds. We observe in Figure 3b that the two NFSIC variants have correct type-I errors across all sample sizes. We recall from Theorem 4 that the NFSIC test with random test locations will asymptotically reject H_0 if it is false. A demonstration of this property is given in Figure 3c, where the test power of NFSIC-med eventually reaches 1 with n higher than 10^5.

3.2. Real Problems

We now examine the performance of our proposed test on real problems.

Million Song Data (MSD) We consider a subset of the Million Song Data[4] (Bertin-Mahieux et al., 2011), in which each song (X) out of 515,345 is represented by 90 features, of which 12 features are timbre average (over all segments) of the song, and 78 features are timbre covariance. Most of the songs are western commercial tracks from 1922 to 2011. The goal is to detect the dependency between each song and its year of release (Y). We set $\alpha = 0.01$, and repeat for 300 trials where the full sample is randomly subsampled to n points in each trial. Other settings are the same as in the toy problems. To make sure that the type-I error is correct, we use the permutation approach in the NFSIC tests to compute the threshold. Figure 4b shows the test powers as n increases from 500 to 2000. To simulate the case where H_0 holds in the problem, we permute the sample to break the dependency of X and Y. The results are shown in Figure 5 in the appendix.

Evidently, NFSIC-opt has the highest test power among all



(a) MSD problem. (b) Videos & Captions problem.

Figure 4: Probability of rejecting H_0 as n increases in the two real problems. $\alpha = 0.01$.

the linear-time tests for all the sample sizes. Its test power is second to only QHSIC. We recall that NFSIC-opt uses half of the sample for parameter tuning. Thus, at $n = 500$, the actual sample for testing is 250, which is relatively small. The fact that there is a vast power gain from 0.4 (NFSIC-med) to 0.8 (NFSIC-opt) at $n = 500$ suggests that the optimization procedure can perform well even at a lower sample sizes.

Videos and Captions Our last problem is based on the VideoStory46K[5] dataset (Habibian et al., 2014). The dataset contains 45,826 Youtube videos (X) of an average length of roughly one minute, and their corresponding text captions (Y) uploaded by the users. Each video is represented as a $d_x = 2000$ dimensional Fisher vector encoding of motion boundary histograms (MBH) descriptors of Wang & Schmid (2013). Each caption is represented as a bag of words with each feature being the frequency of one word. After filtering only words which occur in at least six video captions, we obtain $d_y = 1878$ words. We examine the test powers as n increases from 2000 to 8000. The results are given in Figure 4. The problem is sufficiently challenging that all linear-time tests achieve a low power at $n = 2000$. QHSIC performs exceptionally well on this problem, achieving a maximum power throughout. NFSIC-opt has the highest sample efficiency among the linear-time tests, showing that the optimization procedure is also practical in a high dimensional setting.

[4]Million Song Data subset: https://archive.ics.uci.edu/ml/datasets/YearPredictionMSD.

[5]VideoStory46K dataset: https://ivi.fnwi.uva.nl/isis/mediamill/datasets/videostory.php.

Acknowledgement

We thank the Gatsby Charitable Foundation for the financial support. The major part of this work was carried out while Zoltán Szabó was a research associate at the Gatsby Computational Neuroscience Unit, University College London.

References

Anderson, Theodore W. *An Introduction to Multivariate Statistical Analysis*. Wiley, 2003.

Bach, Francis R. and Jordan, Michael I. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

Bertin-Mahieux, Thierry, Ellis, Daniel P.W., Whitman, Brian, and Lamere, Paul. The million song dataset. In *International Conference on Music Information Retrieval (ISMIR)*, 2011.

Chwialkowski, Kacper P., Ramdas, Aaditya, Sejdinovic, Dino, and Gretton, Arthur. Fast Two-Sample Testing with Analytic Representations of Probability Measures. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1981–1989. 2015.

Dauxois, Jacques and Nkiet, Guy Martial. Nonlinear canonical analysis and independence tests. *The Annals of Statistics*, 26(4):1254–1278, 1998.

Feuerverger, Andrey. A consistent test for bivariate dependence. *International Statistical Review*, 61(3):419–433, 1993.

Fukumizu, Kenji, Gretton, Arthur, Sun, Xiaohai, and Schölkopf, Bernhard. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 489–496, 2008.

Gretton, Arthur. A simpler condition for consistency of a kernel independence test. Technical report, 2015. URL http://arxiv.org/abs/1501.06103.

Gretton, Arthur and Györfi, László. Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11:1391–1423, 2010.

Gretton, Arthur, Bousquet, Olivier, Smola, Alex, and Schölkopf, Bernhard. Measuring Statistical Dependence with Hilbert-Schmidt Norms. In *Algorithmic Learning Theory (ALT)*, pp. 63–77. 2005.

Gretton, Arthur, Fukumizu, Kenji, Teo, Choon H., Song, Le, Schölkopf, Bernhard, and Smola, Alex J. A Kernel Statistical Test of Independence. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 585–592. 2008.

Habibian, Amirhossein, Mensink, Thomas, and Snoek, Cees GM. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *ACM International Conference on Multimedia*, pp. 17–26, 2014.

Heller, Ruth, Heller, Yair, Kaufman, Shachar, Brill, Barak, and Gorfine, Malka. Consistent distribution-free k-sample and independence tests for univariate random variables. *Journal of Machine Learning Research*, 17 (29):1–54, 2016.

Huo, Xiaoming and Székely, Gábor J. Fast computing for distance covariance. *Technometrics*, 58(4):435–447, 2016.

Jitkrittum, Wittawat, Szabó, Zoltán, Chwialkowski, Kacper, and Gretton, Arthur. Interpretable Distribution Features with Maximum Testing Power. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 181–189. 2016.

Kowalski, Jeanne and Tu, Xin M. *Modern Applied U-Statistics*. John Wiley & Sons, 2008.

Krantz, Steven G. and Parks, Harold R. *A Primer of Real Analytic Functions*. Springer Science & Business Media, 2002.

Lehmann, Eric L. *Elements of Large-Sample Theory*. Springer Science & Business Media, 1999.

Lopez-Paz, David, Hennig, Philipp, and Schölkopf, Bernhard. The Randomized Dependence Coefficient. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1–9. 2013.

Mityagin, Boris. The Zero Set of a Real Analytic Function. December 2015. arXiv: 1512.07276.

Rahimi, Ali and Recht, Benjamin. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1177–1184. 2008.

Sejdinovic, Dino, Sriperumbudur, Bharath, Gretton, Arthur, and Fukumizu, Kenji. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

Serfling, Robert J. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 2009.

Smola, Alex, Gretton, Arthur, Song, Le, and Schölkopf, Bernhard. A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory (ALT)*, pp. 13–31, 2007.

Sriperumbudur, Bharath K., Gretton, Arthur, Fukumizu, Kenji, Schölkopf, Bernhard, and Lanckriet, Gert R. G. Hilbert Space Embeddings and Metrics on Probability Measures. *Journal of Machine Learning Research*, 11: 1517–1561, 2010.

Steinwart, Ingo and Christmann, Andreas. *Support vector machines*. Springer Science & Business Media, 2008.

Székely, Gábor J. and Rizzo, Maria L. Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236–1265, 2009.

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

van der Vaart, Aad. *Asymptotic Statistics*. Cambridge University Press, 2000.

Wang, Heng and Schmid, Cordelia. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 3551–3558, 2013.

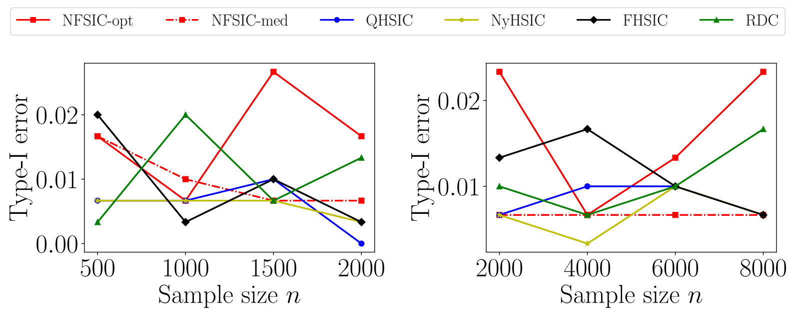
Zhang, Kun, Peters, Jonas, Janzing, Dominik, and Schölkopf, Bernhard. Kernel-based conditional independence test and application in causal discovery. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 804–813, 2011.

Zhang, Qinyi, Filippi, Sarah, Gretton, Arthur, and Sejdinovic, Dino. Large-Scale Kernel Methods for Independence Testing. *Statistics and Computing*, pp. 1–18, 2017.

Supplementary Material

A. Type-I Errors

In this section, we show that all the tests have correct type-I errors (i.e., the probability of reject H_0 when it is true) in real problems. We permute the joint sample so that the dependency is broken to simulate cases in which H_0 holds. The results are shown in Figure 5.



(a) MSD problem (permuted sample). (b) Videos & Captions problem (permuted sample).

Figure 5: Probability of rejecting H_0 as n increases. $\alpha = 0.01$.

B. Redundant Test Locations

Here, we provide a simple illustration to show that two locations $\mathbf{t}_1 = (\mathbf{v}_1, \mathbf{w}_1)$ and $\mathbf{t}_2 = (\mathbf{v}_2, \mathbf{w}_2)$ which are too close to each other will reduce the optimization objective. We consider the Sinusoid problem described in Section 3.1 with $\omega = 1$, and use $J = 2$ test locations. In Figure 6, $\mathbf{t}_1$ is fixed at the red star, while $\mathbf{t}_2$ is varied along the horizontal line. The objective value $\hat{\lambda}_n$ as a function of $\mathbf{t}_2$ is shown in the bottom figure. It can be seen that $\hat{\lambda}_n$ decreases sharply when $\mathbf{t}_2$ is in the neighborhood of $\mathbf{t}_1$. This property implies that two locations which are too close will not maximize the objective function (i.e., the second feature contains no additional information when it matches the first). For $J > 2$, the objective sharply decreases if any two locations are in the same neighborhood.

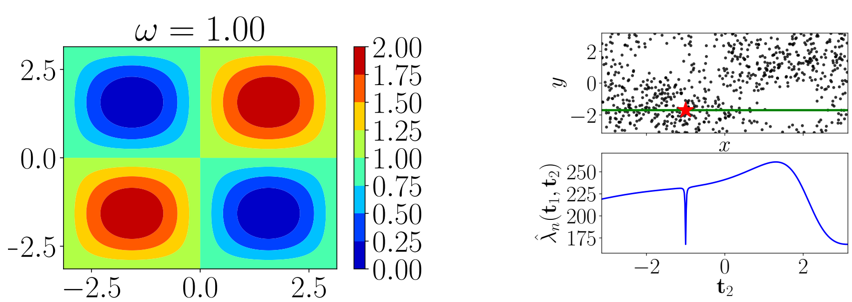


Figure 6: Plot of optimization objective values as location $\mathbf{t}_2$ moves along the green line. The objective sharply drops when the two locations are in the same neighborhood.

C. Test Power vs. J

It might seem intuitive that as the number of locations J increases, the test power should also increase. Here, we empirically show that this statement is *not* always true. Consider the Sinusoid toy example described in Section 3.1 with $\omega = 2$ (also see the left figure of Figure 7). By construction, X and Y are dependent in this problem. We run NFSIC test with a sample size of $n = 800$, varying J from 1 to 600. For each value of J, the test is repeated for 500 times. In each trial, the sample is redrawn and the J test locations are drawn from $\mathrm{Uniform}((-\pi, \pi)^2)$. There is no optimization of the test locations. We use Gaussian kernels for both X and Y, and use the median heuristic to set the Gaussian widths to 1.8. Figure 7 shows the test power as J increases.

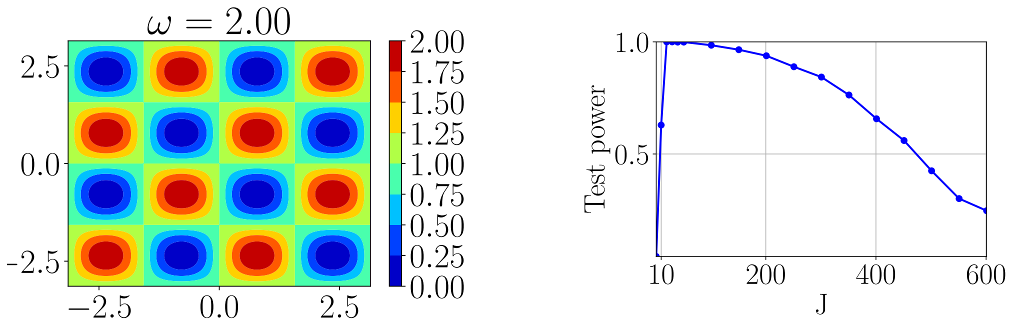


Figure 7: The Sinusoid problem and the plot of test power vs. the number of test locations.

We observe that the test power does not monotonically increase as J increases. When $J = 1$, the difference of p_{xy} and $p_x p_y$ cannot be adequately captured, resulting in a low power. The power increases rapidly to roughly 0.6 at $J = 10$, and stays at 1 until about $J = 100$. Then, the power starts to drop sharply when J is higher than 400 in this problem.

Unlike random Fourier features, the number of test locations in NFSIC is not the number of Monte Carlo particles used to approximate an expectation. There is a tradeoff: if the test locations are in key regions (i.e., regions in which there is a big difference between p_{xy} and $p_x p_y$), then they increase power; yet the statistic gains in variance (thus reducing test power) as J increases. As can be seen in Figure 7, there are eight key regions (in blue) that can reveal the difference of p_{xy} and $p_x p_y$. Using an unnecessarily high J not only makes the covariance matrix $\hat{\mathbf{\Sigma}}$ harder to estimate accurately, it also increases the computation as the complexity on J is $\mathcal{O}(J^3)$.

We note that NFSIC is not intended to be used with a large J. In practice, it should be set to be large enough so as to capture the key regions as stated. As a practical guide, with optimization of the test locations, a good starting point is $J = 5$ or 10.

D. Proof of Theorem 4

Recall Theorem 4,

Theorem 4 (Independence test based on $\widehat{\mathrm{NFSIC}}^2$ is consistent)**.** *Let $\hat{\mathbf{\Sigma}}$ be a consistent estimate of $\mathbf{\Sigma}$ based on the joint sample Z_n, where $\mathbf{\Sigma}$ is defined in Proposition 3. Assume that $V_J = \{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J \sim \eta$ where η is absolutely continuous wrt the Lebesgue measure. The $\widehat{\mathrm{NFSIC}}^2$ statistic is defined as $\hat{\lambda}_n := n\hat{\mathbf{u}}^\top \left(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}}$ where $\gamma_n \geq 0$ is a regularization parameter. Assume that*

1. *Assumption A holds.*
2. *$\mathbf{\Sigma}$ is invertible η-almost surely.*
3. *$\lim_{n\to\infty} \gamma_n = 0$.*

Then, for any k, l and V_J satisfying the assumptions,

1. *Under H_0, $\hat{\lambda}_n \xrightarrow{d} \chi^2(J)$ as $n \to \infty$.*
2. *Under H_1, for any $r \in \mathbb{R}$, $\lim_{n\to\infty} \mathbb{P}\left(\hat{\lambda}_n \geq r\right) = 1$ η-almost surely. That is, the independence test based on $\widehat{\mathrm{NFSIC}}^2$ is consistent.*

Proof. Assume that H_0 holds. The consistency of $\hat{\mathbf{\Sigma}}$ and the continuous mapping theorem imply that $\left(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1} \xrightarrow{p} \mathbf{\Sigma}^{-1}$ which is a constant. Let $\mathbf{a}$ be a random vector in $\mathbb{R}^J$ following $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. By van der Vaart (2000, Theorem 2.7 (v)), it follows that $\left[\sqrt{n}\hat{\mathbf{u}}, \left(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1}\right] \xrightarrow{d} \left[\mathbf{a}, \mathbf{\Sigma}^{-1}\right]$ where $\mathbf{u} = 0$ almost surely by Proposition 2, and $\sqrt{n}\hat{\mathbf{u}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ by Proposition 3. Since $f(\mathbf{x}, \mathbf{S}) := \mathbf{x}^\top \mathbf{S}\mathbf{x}$ is continuous, $f\left(\sqrt{n}\hat{\mathbf{u}}, \left(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1}\right) \xrightarrow{d} f(\mathbf{a}, \mathbf{\Sigma}^{-1})$. Equivalently, $n\hat{\mathbf{u}}^\top \left(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}} \xrightarrow{d} \mathbf{a}^\top \mathbf{\Sigma}^{-1} \mathbf{a} \sim \chi^2(J)$ by Anderson (2003, Theorem 3.3.3). This proves the first claim.

The proof of the second claim has a very similar structure to the proof of Proposition 2 of Chwialkowski et al. (2015). Assume that H_1 holds. Then, $\mathbf{u} \neq \mathbf{0}$ almost surely by Proposition 2. Since k and l are bounded, it follows that $|h_{\mathbf{t}}(\mathbf{z}, \mathbf{z}')| \leq 2B_k B_l$ for any $\mathbf{z}, \mathbf{z}'$ (see (8)), and we have that $\hat{\mathbf{u}} \stackrel{a.s.}{\to} \mathbf{u}$ by Serfling (2009, Section 5.4, Theorem A). Thus, $\hat{\mathbf{u}}^\top \left(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}} - \frac{r}{n} \stackrel{d}{\to} \mathbf{u}^\top \mathbf{\Sigma}^{-1} \mathbf{u}$ by the continuous mapping theorem, and the consistency of $\hat{\mathbf{\Sigma}}$. Consequently,

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}\left(\hat{\lambda}_n \geq r\right) \\
&= 1 - \lim_{n\to\infty} \mathbb{P}\left(\hat{\mathbf{u}}^\top \left(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}} - \frac{r}{n} < 0\right) \\
&\stackrel{(a)}{=} 1 - \mathbb{P}\left(\mathbf{u}^\top \mathbf{\Sigma}^{-1} \mathbf{u} < 0\right) \stackrel{(b)}{=} 1,
\end{aligned}$$

where at (a) we use the Portmanteau theorem (van der Vaart, 2000, Lemma 2.2 (i)) guaranteeing that $x_n \stackrel{d}{\to} x$ if and only if $\mathbb{P}(x_n < t) \to \mathbb{P}(x < t)$ for all continuity points of $t \mapsto \mathbb{P}(x < t)$. Step (b) is justified by noting that the covariance matrix $\mathbf{\Sigma}$ is positive definite so that $\mathbf{u}^\top \mathbf{\Sigma}^{-1} \mathbf{u} > 0$, and $t \mapsto \mathbb{P}(\mathbf{u}^\top \mathbf{\Sigma}^{-1} \mathbf{u} < t)$ (a step function) is continuous at 0. □

E. Proof of Theorem 6

Recall Theorem 6,

Theorem 6 (A lower bound on the test power). *Let* $\mathrm{NFSIC}^2(X, Y) := \lambda_n := n\mathbf{u}^\top \mathbf{\Sigma}^{-1} \mathbf{u}$. *Let* $\mathcal{K}$ *be a kernel class for* k, $\mathcal{L}$ *be a kernel class for* l, *and* $\mathcal{V}$ *be a collection with each element being a set of* J *locations. Assume that*

1. *There exist finite* B_k *and* B_l *such that* $\sup_{k\in\mathcal{K}} \sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} |k(\mathbf{x}, \mathbf{x}')| \leq B_k$ *and* $\sup_{l\in\mathcal{L}} \sup_{\mathbf{y},\mathbf{y}'\in\mathcal{Y}} |l(\mathbf{y}, \mathbf{y}')| \leq B_l$.
2. $\tilde{c} := \sup_{k\in\mathcal{K}} \sup_{l\in\mathcal{L}} \sup_{V_J\in\mathcal{V}} \|\mathbf{\Sigma}^{-1}\|_F < \infty$.

Then, for any $k \in \mathcal{K}, l \in \mathcal{L}, V_J \in \mathcal{V}$, *and* $\lambda_n \geq r$, *the test power satisfies* $\mathbb{P}\left(\hat{\lambda}_n \geq r\right) \geq L(\lambda_n)$ *where*

$$\begin{aligned}
L(\lambda_n) = 1 &- 62e^{-\xi_1 \gamma_n^2 (\lambda_n - r)^2/n} - 2e^{-\lfloor 0.5n \rfloor (\lambda_n - r)^2 / \left[\xi_2 n^2\right]} \\
&- 2e^{-\left[(\lambda_n - r)\gamma_n(n-1)/3 - \xi_3 n - c_3 \gamma_n^2 n(n-1)\right]^2 / \left[\xi_4 n^2 (n-1)\right]},
\end{aligned}$$

$\lfloor\cdot\rfloor$ *is the floor function,* $\xi_1 := \frac{1}{3^2 c_1^2 J^2 B^*}$, B^* *is a constant depending on only* B_k *and* B_l, $\xi_2 := 72c_2^2 J B^2$, $B := B_k B_l$, $\xi_3 := 8c_1 B^2 J$, $c_3 := 4B^2 J \tilde{c}^2$, $\xi_4 := 2^8 B^4 J^2 c_1^2$, $c_1 := 4B^2 J\sqrt{J}\tilde{c}$, *and* $c_2 := 4B\sqrt{J}\tilde{c}$. *Moreover, for sufficiently large fixed* n, $L(\lambda_n)$ *is increasing in* λ_n.

Overview of the proof We first derive a probabilistic bound for $|\hat{\lambda}_n - \lambda_n|/n$. The bound is in turn upper bounded by an expression involving $\|\hat{\mathbf{u}} - \mathbf{u}\|_2$ and $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F$. The difference $\|\hat{\mathbf{u}} - \mathbf{u}\|_2$ can be bounded by applying the bound for U-statistics given in Serfling (2009, Theorem A, p. 201). For $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F$, we decompose it into a sum of smaller components, and bound each term with a product variant of the Hoeffding's inequality (Lemma 8). $L(\lambda_n)$ is obtained by combining all the bounds with the union bound.

E.1. Notations

Let $\langle \mathbf{A}, \mathbf{B}\rangle_F := \mathrm{tr}(\mathbf{A}^\top \mathbf{B})$ denote the Frobenius inner product, and $\|\mathbf{A}\|_F := \sqrt{\mathrm{tr}(\mathbf{A}^\top \mathbf{A})}$ be the Frobenius norm. Write $\mathbf{z} := (\mathbf{x}, \mathbf{y})$ to denote a pair of points from $\mathcal{X} \times \mathcal{Y}$. We write $\mathbf{t} := (\mathbf{v}, \mathbf{w})$ to denote a pair of test locations from $\mathcal{X} \times \mathcal{Y}$. For brevity, an expectation over $(\mathbf{x}, \mathbf{y})$ (i.e., $\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P_{xy}}$) will be written as $\mathbb{E}_{\mathbf{z}}$ or $\mathbb{E}_{\mathbf{xy}}$. Define $\tilde{k}(\mathbf{x}, \mathbf{v}) := k(\mathbf{x}, \mathbf{v}) - \mathbb{E}_{\mathbf{x}'} k(\mathbf{x}', \mathbf{v})$, and $\tilde{l}(\mathbf{y}, \mathbf{w}) := l(\mathbf{y}, \mathbf{w}) - \mathbb{E}_{\mathbf{y}'} l(\mathbf{y}', \mathbf{w})$. Let $B_2(r) := \{\mathbf{x} \mid \|\mathbf{x}\|_2 \leq r\}$ be a closed ball with radius r centered at the origin. Similarly, define $B_F(r) := \{\mathbf{A} \mid \|\mathbf{A}\|_F \leq r\}$ to be a closed ball with radius r of $J \times J$ matrices under the Frobenius norm. Denote the max operation by $(x_1, \ldots, x_m)_+ = \max(x_1, \ldots, x_m)$.

For a product of marginal mean embeddings $\mu_x(\mathbf{v})\mu_y(\mathbf{w})$, we write $\widehat{\mu_x \mu_y}(\mathbf{v}, \mathbf{w}) := \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j\neq i} k(\mathbf{x}_i, \mathbf{v}) l(\mathbf{y}_j, \mathbf{w})$ to denote the unbiased plug-in estimator, and write $\hat{\mu}_x(\mathbf{v})\hat{\mu}_y(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^n k(\mathbf{x}_i, \mathbf{v}) \frac{1}{n}\sum_{j=1}^n l(\mathbf{y}_j, \mathbf{w})$ which is a biased estimator. Define $\hat{u}^b(\mathbf{v}, \mathbf{w}) := \hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) - \hat{\mu}_x(\mathbf{v})\hat{\mu}_y(\mathbf{w})$ so that $\hat{\mathbf{u}}^b := \left(\hat{u}^b(\mathbf{t}_1), \ldots, \hat{u}^b(\mathbf{t}_J)\right)^\top$ where the superscript b stands for "biased". To avoid confusing with a positive definite kernel, we will refer to a U-statistic kernel as a *core*.

E.2. Proof

We will first derive a bound for $\mathbb{P}(|\hat{\lambda}_n - \lambda_n| \geq t)$, which will then be reparametrized to get a bound for the target quantity $\mathbb{P}(\hat{\lambda}_n \geq r)$. We closely follow the proof in Jitkrittum et al. (2016, Section C.1) up to (12), then we diverge. We start by considering $|\hat{\lambda}_n - \lambda_n|/n$.

$$
\begin{aligned}
|\hat{\lambda}_n - \lambda_n|/n &= \left|\hat{\mathbf{u}}^\top (\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I})^{-1}\hat{\mathbf{u}} - \mathbf{u}^\top \mathbf{\Sigma}^{-1}\mathbf{u}\right| \\
&= \left|\hat{\mathbf{u}}^\top \left(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1}\hat{\mathbf{u}} - \mathbf{u}^\top \left(\mathbf{\Sigma} + \gamma_n \mathbf{I}\right)^{-1}\mathbf{u} + \mathbf{u}^\top \left(\mathbf{\Sigma} + \gamma_n \mathbf{I}\right)^{-1}\mathbf{u} - \mathbf{u}^\top \mathbf{\Sigma}^{-1}\mathbf{u}\right| \\
&\leq \left|\hat{\mathbf{u}}^\top \left(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1}\hat{\mathbf{u}} - \mathbf{u}^\top \left(\mathbf{\Sigma} + \gamma_n \mathbf{I}\right)^{-1}\mathbf{u}\right| + \left|\mathbf{u}^\top \left(\mathbf{\Sigma} + \gamma_n \mathbf{I}\right)^{-1}\mathbf{u} - \mathbf{u}^\top \mathbf{\Sigma}^{-1}\mathbf{u}\right| \\
&:= (\bigstar)_1 + (\bigstar)_2 .
\end{aligned}
$$

We next bound $(\bigstar_1)$ and $(\bigstar_2)$ separately.

$$
\begin{aligned}
(\bigstar)_1 &= \left|\hat{\mathbf{u}}^\top \left(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1}\hat{\mathbf{u}} - \mathbf{u}^\top \left(\mathbf{\Sigma} + \gamma_n \mathbf{I}\right)^{-1}\mathbf{u}\right| \\
&= \left|\hat{\mathbf{u}}^\top \left(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1}\hat{\mathbf{u}} - \hat{\mathbf{u}}^\top \left(\mathbf{\Sigma} + \gamma_n \mathbf{I}\right)^{-1}\hat{\mathbf{u}} + \hat{\mathbf{u}}^\top \left(\mathbf{\Sigma} + \gamma_n \mathbf{I}\right)^{-1}\hat{\mathbf{u}} - \mathbf{u}^\top \left(\mathbf{\Sigma} + \gamma_n \mathbf{I}\right)^{-1}\mathbf{u}\right| \\
&\leq \left|\hat{\mathbf{u}}^\top \left(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1}\hat{\mathbf{u}} - \hat{\mathbf{u}}^\top \left(\mathbf{\Sigma} + \gamma_n \mathbf{I}\right)^{-1}\hat{\mathbf{u}}\right| + \left|\hat{\mathbf{u}}^\top \left(\mathbf{\Sigma} + \gamma_n \mathbf{I}\right)^{-1}\hat{\mathbf{u}} - \mathbf{u}^\top \left(\mathbf{\Sigma} + \gamma_n \mathbf{I}\right)^{-1}\mathbf{u}\right| \\
&= \left|\left\langle \hat{\mathbf{u}}\hat{\mathbf{u}}^\top, \left(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1} - \left(\mathbf{\Sigma} + \gamma_n \mathbf{I}\right)^{-1}\right\rangle_F\right| + \left|\left\langle \hat{\mathbf{u}}\hat{\mathbf{u}}^\top - \mathbf{u}\mathbf{u}^\top, \left(\mathbf{\Sigma} + \gamma_n \mathbf{I}\right)^{-1}\right\rangle_F\right| \\
&\leq \|\hat{\mathbf{u}}\hat{\mathbf{u}}^\top\|_F \|(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I})^{-1} - (\mathbf{\Sigma} + \gamma_n \mathbf{I})^{-1}\|_F + \|\hat{\mathbf{u}}\hat{\mathbf{u}}^\top - \mathbf{u}\mathbf{u}^\top\|_F \|(\mathbf{\Sigma} + \gamma_n \mathbf{I})^{-1}\|_F \\
&= \|\hat{\mathbf{u}}\hat{\mathbf{u}}^\top\|_F \|(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I})^{-1}[(\mathbf{\Sigma} + \gamma_n \mathbf{I}) - (\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I})](\mathbf{\Sigma} + \gamma_n \mathbf{I})^{-1}\|_F + \|\hat{\mathbf{u}}\hat{\mathbf{u}}^\top - \hat{\mathbf{u}}\mathbf{u}^\top + \hat{\mathbf{u}}\mathbf{u}^\top - \mathbf{u}\mathbf{u}^\top\|_F \|(\mathbf{\Sigma} + \gamma_n \mathbf{I})^{-1}\|_F \\
&\overset{(a)}{\leq} \|\hat{\mathbf{u}}\hat{\mathbf{u}}^\top\|_F \|(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I})^{-1}\|_F \|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_F \|\mathbf{\Sigma}^{-1}\|_F + \|\hat{\mathbf{u}}\hat{\mathbf{u}}^\top - \hat{\mathbf{u}}\mathbf{u}^\top + \hat{\mathbf{u}}\mathbf{u}^\top - \mathbf{u}\mathbf{u}^\top\|_F \|\mathbf{\Sigma}^{-1}\|_F \\
&\overset{(b)}{\leq} \frac{\sqrt{J}}{\gamma_n}\|\hat{\mathbf{u}}\|_2^2 \|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_F \|\mathbf{\Sigma}^{-1}\|_F + \left(\|\hat{\mathbf{u}}(\hat{\mathbf{u}} - \mathbf{u})^\top\|_F + \|(\hat{\mathbf{u}} - \mathbf{u})\mathbf{u}^\top\|_F\right) \|\mathbf{\Sigma}^{-1}\|_F \\
&\leq \frac{\sqrt{J}}{\gamma_n}\|\hat{\mathbf{u}}\|_2^2 \|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_F \|\mathbf{\Sigma}^{-1}\|_F + \left(\|\hat{\mathbf{u}}\|_2 + \|\mathbf{u}\|_2\right) \|\hat{\mathbf{u}} - \mathbf{u}\|_2 \|\mathbf{\Sigma}^{-1}\|_F,
\end{aligned} \tag{5}
$$

where at (a) we used $\|(\mathbf{\Sigma} + \gamma_n \mathbf{I})^{-1}\|_F \leq \|\mathbf{\Sigma}^{-1}\|_F$, at (b) we used $\|(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I})^{-1}\|_F \leq \sqrt{J}\|(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I})^{-1}\|_2 \leq \sqrt{J}/\gamma_n$.

For $(\bigstar)_2$, we have

$$
\begin{aligned}
(\bigstar)_2 &= \left|\mathbf{u}^\top \left(\mathbf{\Sigma} + \gamma_n \mathbf{I}\right)^{-1}\mathbf{u} - \mathbf{u}^\top \mathbf{\Sigma}^{-1}\mathbf{u}\right| \\
&= \left|\left\langle \mathbf{u}\mathbf{u}^\top, (\mathbf{\Sigma} + \gamma_n \mathbf{I})^{-1} - \mathbf{\Sigma}^{-1}\right\rangle_F\right| \\
&\leq \|\mathbf{u}\mathbf{u}^\top\|_F \|(\mathbf{\Sigma} + \gamma_n \mathbf{I})^{-1} - \mathbf{\Sigma}^{-1}\|_F \\
&= \|\mathbf{u}\|_2^2 \|(\mathbf{\Sigma} + \gamma_n \mathbf{I})^{-1}\left[\mathbf{\Sigma} - (\mathbf{\Sigma} + \gamma_n \mathbf{I})\right]\mathbf{\Sigma}^{-1}\|_F \\
&\leq \gamma_n \|\mathbf{u}\|_2^2 \|(\mathbf{\Sigma} + \gamma_n \mathbf{I})^{-1}\|_F \|\mathbf{\Sigma}^{-1}\|_F \\
&\overset{(a)}{\leq} \gamma_n \|\mathbf{u}\|_2^2 \|\mathbf{\Sigma}^{-1}\|_F^2,
\end{aligned} \tag{6}
$$

where at (a) we used $\|(\mathbf{\Sigma} + \gamma_n \mathbf{I})^{-1}\|_F \leq \|\mathbf{\Sigma}^{-1}\|_F$.

Combining (5) and (6), we have

$$
\left|\hat{\mathbf{u}}^\top (\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I})^{-1}\hat{\mathbf{u}} - \mathbf{u}^\top \mathbf{\Sigma}^{-1}\mathbf{u}\right|
$$

$$\leq \frac{\sqrt{J}}{\gamma_n}\|\hat{\mathbf{u}}\|^2\|\mathbf{\Sigma}-\hat{\mathbf{\Sigma}}\|_F\|\mathbf{\Sigma}^{-1}\|_F + \left(\|\hat{\mathbf{u}}\|_2+\|\mathbf{u}\|_2\right)\|\hat{\mathbf{u}}-\mathbf{u}\|_2\|\mathbf{\Sigma}^{-1}\|_F + \gamma_n\|\mathbf{u}\|_2^2\|\mathbf{\Sigma}^{-1}\|_F^2. \tag{7}$$

Bounding $\|\hat{\mathbf{u}}\|_2^2$ and $\|\mathbf{u}\|_2^2$ Here, we show that by the boundedness of the kernels k and l, it follows that $\|\hat{\mathbf{u}}\|_2^2$ is bounded. Recall that $\sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}}|k(\mathbf{x},\mathbf{x}')|\leq B_k$, $\sup_{\mathbf{y},\mathbf{y}'}|l(\mathbf{y},\mathbf{y}')|\leq B_l$, our notation $\mathbf{t}=(\mathbf{v},\mathbf{w})$ for the test locations, and $\mathbf{z}_i:=(\mathbf{x}_i,\mathbf{y}_i)$. We first show that the U-statistic core h is bounded.

$$\begin{aligned}
|h_{\mathbf{t}}((\mathbf{x},\mathbf{y}),(\mathbf{x}',\mathbf{y}'))| &= \left|\frac{1}{2}(k(\mathbf{x},\mathbf{v})-k(\mathbf{x}',\mathbf{v}))(l(\mathbf{y},\mathbf{w})-l(\mathbf{y}',\mathbf{w}))\right| \\
&\leq \frac{1}{2}\left(|k(\mathbf{x},\mathbf{v})|+|k(\mathbf{x}',\mathbf{v})|\right)\left(|l(\mathbf{y},\mathbf{w})|+|l(\mathbf{y}',\mathbf{w})|\right) \\
&\leq 2B_kB_l := 2B, \qquad (8)
\end{aligned}$$

where we define $B:=B_kB_l$. It follows that

$$\|\hat{\mathbf{u}}\|_2^2 = \sum_{m=1}^{J}\left[\frac{2}{n(n-1)}\sum_{i<j}h_{\mathbf{t}_m}(\mathbf{z}_i,\mathbf{z}_j)\right]^2 \leq \sum_{m=1}^{J}\left[2B_kB_l\right]^2 = 4B^2J, \tag{9}$$

$$\|\mathbf{u}\|_2^2 = \sum_{m=1}^{J}\left[\mathbb{E}_{\mathbf{z}}\mathbb{E}_{\mathbf{z}'}h_{\mathbf{t}_m}(\mathbf{z},\mathbf{z}')\right]^2 \leq 4B^2J. \tag{10}$$

Using the upper bounds on $\|\hat{\mathbf{u}}\|_2^2$, $\|\mathbf{u}\|_2^2$,(7) and the definition of $\tilde{c}$, we have

$$\begin{aligned}
&\left|\hat{\mathbf{u}}^\top(\hat{\mathbf{\Sigma}}+\gamma_n\mathbf{I})^{-1}\hat{\mathbf{u}}-\mathbf{u}^\top\mathbf{\Sigma}^{-1}\mathbf{u}\right| \\
&\leq \frac{\sqrt{J}}{\gamma_n}4B^2J\tilde{c}\|\mathbf{\Sigma}-\hat{\mathbf{\Sigma}}\|_F + 4B\sqrt{J}\tilde{c}\|\hat{\mathbf{u}}-\mathbf{u}\|_2 + 4B^2J\tilde{c}^2\gamma_n \\
&=: \frac{c_1}{\gamma_n}\|\mathbf{\Sigma}-\hat{\mathbf{\Sigma}}\|_F + c_2\|\hat{\mathbf{u}}-\mathbf{u}\|_2 + c_3\gamma_n, \qquad (11)
\end{aligned}$$

where we define $c_1:=4B^2J\sqrt{J}\tilde{c}$, $c_2:=4B\sqrt{J}\tilde{c}$, and $c_3:=4B^2J\tilde{c}^2$. This upper bound implies that

$$|\hat{\lambda}_n-\lambda_n| \leq \frac{c_1}{\gamma_n}n\|\mathbf{\Sigma}-\hat{\mathbf{\Sigma}}\|_F + c_2n\|\hat{\mathbf{u}}-\mathbf{u}\|_2 + c_3n\gamma_n. \tag{12}$$

We will separately upper bound $\|\mathbf{\Sigma}-\hat{\mathbf{\Sigma}}\|_F$ and $\|\hat{\mathbf{u}}-\mathbf{u}\|_2$, and combine them with a union bound.

E.2.1. Bounding $\|\hat{\mathbf{u}}-\mathbf{u}\|_2$

Let $\mathbf{t}^*=\arg\max_{\mathbf{t}\in\{\mathbf{t}_1,\ldots,\mathbf{t}_J\}}|\hat{u}(\mathbf{t})-u(\mathbf{t})|$. Recall that $\mathbf{u}=(u(\mathbf{t}_1),\ldots,u(\mathbf{t}_J))^\top=(u_1,\ldots,u_J)^\top$.

$$\begin{aligned}
\|\hat{\mathbf{u}}-\mathbf{u}\|_2 &= \sup_{\mathbf{b}\in B_2(1)}\langle\mathbf{b},\hat{\mathbf{u}}-\mathbf{u}\rangle_2 \leq \sup_{\mathbf{b}\in B_2(1)}\sum_{j=1}^{J}|b_j||\hat{u}(\mathbf{t}_j)-u(\mathbf{t}_j)| \\
&\leq |\hat{u}(\mathbf{t}^*)-u(\mathbf{t}^*)|\sup_{\mathbf{b}\in B_2(1)}\sum_{j=1}^{J}|b_j| \\
&\overset{(a)}{\leq} \sqrt{J}|\hat{u}(\mathbf{t}^*)-u(\mathbf{t}^*)|\sup_{\mathbf{b}\in B_2(1)}\|\mathbf{b}\|_2 \\
&= \sqrt{J}|\hat{u}(\mathbf{t}^*)-u(\mathbf{t}^*)|, \qquad (13)
\end{aligned}$$

where at (a) we used $\|\mathbf{a}\|_1\leq\sqrt{J}\|\mathbf{a}\|_2$ for any $\mathbf{a}\in\mathbb{R}^J$. From (13), it can be seen that bounding $\|\hat{\mathbf{u}}-\mathbf{u}\|_2$ amounts to bounding the difference of a U-statistic $\hat{u}(\mathbf{t}^*)$ (see (4)) to its expectation $u(\mathbf{t}^*)$. Combining (13) and (12), we have

$$|\hat{\lambda}_n-\lambda_n| \leq \frac{c_1}{\gamma_n}n\|\mathbf{\Sigma}-\hat{\mathbf{\Sigma}}\|_F + c_2n\sqrt{J}|\hat{u}(\mathbf{t}^*)-u(\mathbf{t}^*)| + c_3n\gamma_n. \tag{14}$$

E.2.2. BOUNDING $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F$

The plan is to write $\hat{\mathbf{\Sigma}} = \hat{\mathbf{S}} - \hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top}, \mathbf{\Sigma} = \mathbf{S} - \mathbf{u}\mathbf{u}^\top$, so that $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F \le \|\hat{\mathbf{S}} - \mathbf{S}\|_F + \|\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \mathbf{u}\mathbf{u}^\top\|_F$ and bound separately $\|\hat{\mathbf{S}} - \mathbf{S}\|_F$ and $\|\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \mathbf{u}\mathbf{u}^\top\|_F$.

Recall that $\Sigma_{ij} = \eta(\mathbf{t}_i, \mathbf{t}_j), \eta(\mathbf{t}, \mathbf{t}') = \mathbb{E}_{\mathbf{xy}}[\left(\tilde{k}(\mathbf{x}, \mathbf{v})\tilde{l}(\mathbf{y}, \mathbf{w}) - u(\mathbf{v}, \mathbf{w})\right)\left(\tilde{k}(\mathbf{x}, \mathbf{v}')\tilde{l}(\mathbf{y}, \mathbf{w}') - u(\mathbf{v}', \mathbf{w}')\right)]$ where $\tilde{k}(\mathbf{x}, \mathbf{v}) = k(\mathbf{x}, \mathbf{v}) - \mathbb{E}_{\mathbf{x}'}k(\mathbf{x}', \mathbf{v})$, and $\tilde{l}(\mathbf{y}, \mathbf{w}) = l(\mathbf{y}, \mathbf{w}) - \mathbb{E}_{\mathbf{y}'}l(\mathbf{y}', \mathbf{w})$. Its empirical estimator (see Proposition 5) is $\hat{\Sigma}_{ij} = \hat{\eta}(\mathbf{t}_i, \mathbf{t}_j)$ where

$$
\begin{aligned}
\hat{\eta}(\mathbf{t}, \mathbf{t}') &= \frac{1}{n}\sum_{i=1}^{n}[\left(\overline{k}(\mathbf{x}_i, \mathbf{v})\overline{l}(\mathbf{y}_i, \mathbf{w}) - \hat{u}^b(\mathbf{v}, \mathbf{w})\right)\left(\overline{k}(\mathbf{x}_i, \mathbf{v}')\overline{l}(\mathbf{y}_i, \mathbf{w}') - \hat{u}^b(\mathbf{v}', \mathbf{w}')\right)] \\
&= \frac{1}{n}\sum_{i=1}^{n}\overline{k}(\mathbf{x}_i, \mathbf{v})\overline{l}(\mathbf{y}_i, \mathbf{w})\overline{k}(\mathbf{x}_i, \mathbf{v}')\overline{l}(\mathbf{y}_i, \mathbf{w}') - \hat{u}^b(\mathbf{v}, \mathbf{w})\hat{u}^b(\mathbf{v}', \mathbf{w}'),
\end{aligned}
$$

$\overline{k}(\mathbf{x}, \mathbf{v}) := k(\mathbf{x}, \mathbf{v}) - \frac{1}{n}\sum_{i=1}^n k(\mathbf{x}_i, \mathbf{v})$, and $\overline{l}(\mathbf{y}, \mathbf{w}) := l(\mathbf{y}, \mathbf{w}) - \frac{1}{n}\sum_{i=1}^n l(\mathbf{y}_i, \mathbf{w})$. We note that $\frac{1}{n}\sum_{i=1}^n \overline{k}(\mathbf{x}_i, \mathbf{v})\overline{l}(\mathbf{y}_i, \mathbf{w}) = \hat{u}^b(\mathbf{v}, \mathbf{w})$. We define $\hat{\mathbf{S}} \in \mathbb{R}^{J\times J}$ such that $\hat{S}_{ij} := \frac{1}{n}\sum_{m=1}^n \overline{k}(\mathbf{x}_m, \mathbf{v}_i)\overline{l}(\mathbf{y}_m, \mathbf{w}_i)\overline{k}(\mathbf{x}_m, \mathbf{v}_j)\overline{l}(\mathbf{y}_i, \mathbf{w}_j)$, and define similarly its population counterpart $\mathbf{S}$ such that $S_{ij} := \mathbb{E}_{\mathbf{xy}}[\tilde{k}(\mathbf{x}, \mathbf{v})\tilde{l}(\mathbf{y}, \mathbf{w})\tilde{k}(\mathbf{x}, \mathbf{v}')\tilde{l}(\mathbf{y}, \mathbf{w}')]$. We have

$$
\begin{aligned}
\hat{\mathbf{\Sigma}} &= \hat{\mathbf{S}} - \hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top}, \\
\mathbf{\Sigma} &= \mathbf{S} - \mathbf{u}\mathbf{u}^\top, \\
\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F &= \|\hat{\mathbf{S}} - \mathbf{S} - (\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \mathbf{u}\mathbf{u}^\top)\|_F \qquad (15) \\
&\le \|\hat{\mathbf{S}} - \mathbf{S}\|_F + \|\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \mathbf{u}\mathbf{u}^\top\|_F. \qquad (16)
\end{aligned}
$$

With (16), (14) becomes

$$
|\hat{\lambda}_n - \lambda_n| \le \frac{c_1 n}{\gamma_n}\|\hat{\mathbf{S}} - \mathbf{S}\|_F + \frac{c_1 n}{\gamma_n}\|\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \mathbf{u}\mathbf{u}^\top\|_F + c_2 n\sqrt{J}|\hat{u}(\mathbf{t}^*) - u(\mathbf{t}^*)| + c_3 n\gamma_n. \qquad (17)
$$

We will further separately bound $\|\hat{\mathbf{S}} - \mathbf{S}\|_F$ and $\|\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \mathbf{u}\mathbf{u}^\top\|_F$.

E.2.3. BOUNDING $\|\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \mathbf{u}\mathbf{u}^\top\|_F$

$$
\begin{aligned}
\|\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \mathbf{u}\mathbf{u}^\top\|_F &= \|\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \hat{\mathbf{u}}^b\mathbf{u}^\top + \hat{\mathbf{u}}^b\mathbf{u}^\top - \mathbf{u}\mathbf{u}^\top\|_F \\
&\le \|\hat{\mathbf{u}}^b(\hat{\mathbf{u}}^b - \mathbf{u})^\top\|_F + \|(\hat{\mathbf{u}}^b - \mathbf{u})\mathbf{u}^\top\|_F \\
&= \|\hat{\mathbf{u}}^b\|_2\|\hat{\mathbf{u}}^b - \mathbf{u}\|_2 + \|\hat{\mathbf{u}}^b - \mathbf{u}\|_2\|\mathbf{u}\|_2 \\
&\le 4B\sqrt{J}\|\hat{\mathbf{u}}^b - \mathbf{u}\|_2,
\end{aligned}
$$

where we used (10) and the fact that $\|\hat{\mathbf{u}}^b\|_2 \le 2B\sqrt{J}$ which can be shown similarly to (9) as

$$
\|\hat{\mathbf{u}}^b\|_2^2 = \sum_{m=1}^{J}\left[\hat{\mu}_{xy}(\mathbf{v}_m, \mathbf{w}_m) - \hat{\mu}_x(\mathbf{v}_m)\hat{\mu}_y(\mathbf{w}_m)\right]^2 = \sum_{m=1}^{J}\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}h_{\mathbf{t}_m}(\mathbf{z}_i, \mathbf{z}_j)\right]^2 \le \sum_{m=1}^{J}\left[2B_k B_l\right]^2 = 4B^2 J.
$$

Let $(\tilde{\mathbf{v}}, \tilde{\mathbf{w}}) := \tilde{\mathbf{t}} = \arg\max_{\mathbf{t}\in\{\mathbf{t}_1,\ldots,\mathbf{t}_J\}}|\hat{u}^b(\mathbf{t}) - u(\mathbf{t})|$. We bound $\|\hat{\mathbf{u}}^b - \mathbf{u}\|_2$ by

$$
\begin{aligned}
\|\hat{\mathbf{u}}^b - \mathbf{u}\|_2 &\overset{(a)}{\le} \sqrt{J}|\hat{u}^b(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})| \\
&= \sqrt{J}\left|\hat{\mu}_{xy}(\tilde{\mathbf{t}}) - \hat{\mu}_x(\tilde{\mathbf{v}})\hat{\mu}_y(\tilde{\mathbf{w}}) - u(\tilde{\mathbf{t}})\right| \\
&= \sqrt{J}\left|\hat{\mu}_{xy}(\tilde{\mathbf{t}}) - \widehat{\mu_x\mu_y}(\tilde{\mathbf{t}}) + \widehat{\mu_x\mu_y}(\tilde{\mathbf{t}}) - \hat{\mu}_x(\tilde{\mathbf{v}})\hat{\mu}_y(\tilde{\mathbf{w}}) - u(\tilde{\mathbf{t}})\right| \\
&\le \sqrt{J}\left|\hat{\mu}_{xy}(\tilde{\mathbf{t}}) - \widehat{\mu_x\mu_y}(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})\right| + \sqrt{J}\left|\widehat{\mu_x\mu_y}(\tilde{\mathbf{t}}) - \hat{\mu}_x(\tilde{\mathbf{v}})\hat{\mu}_y(\tilde{\mathbf{w}})\right|
\end{aligned}
$$

$$= \sqrt{J}\left|\hat{u}(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})\right| + \sqrt{J}\left|\widehat{\mu_x\mu_y}(\tilde{\mathbf{t}}) - \hat{\mu}_x(\tilde{\mathbf{v}})\hat{\mu}_y(\tilde{\mathbf{w}})\right|, \tag{18}$$

where at (a) we used the same reasoning as in (13). The bias $\left|\widehat{\mu_x\mu_y}(\tilde{\mathbf{t}}) - \hat{\mu}_x(\tilde{\mathbf{v}})\hat{\mu}_y(\tilde{\mathbf{w}})\right|$ in the second term can be bounded as

$$\begin{aligned}
&\left|\widehat{\mu_x\mu_y}(\tilde{\mathbf{t}}) - \hat{\mu}_x(\tilde{\mathbf{v}})\hat{\mu}_y(\tilde{\mathbf{w}})\right| \\
&= \left|\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} k(\mathbf{x}_i, \tilde{\mathbf{v}})l(\mathbf{y}_j, \tilde{\mathbf{w}}) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(\mathbf{x}_i, \tilde{\mathbf{v}})l(\mathbf{y}_j, \tilde{\mathbf{w}})\right| \\
&= \left|\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n} k(\mathbf{x}_i, \tilde{\mathbf{v}})l(\mathbf{y}_j, \tilde{\mathbf{w}}) - \frac{1}{n(n-1)}\sum_{i=1}^{n} k(\mathbf{x}_i, \tilde{\mathbf{v}})l(\mathbf{y}_i, \tilde{\mathbf{w}}) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(\mathbf{x}_i, \tilde{\mathbf{v}})l(\mathbf{y}_j, \tilde{\mathbf{w}})\right| \\
&= \left|\left(1 - \frac{n}{n-1}\right)\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(\mathbf{x}_i, \tilde{\mathbf{v}})l(\mathbf{y}_j, \tilde{\mathbf{w}}) + \frac{1}{n(n-1)}\sum_{i=1}^{n} k(\mathbf{x}_i, \tilde{\mathbf{v}})l(\mathbf{y}_i, \tilde{\mathbf{w}})\right| \\
&\leq \left|\left(1 - \frac{n}{n-1}\right)\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(\mathbf{x}_i, \tilde{\mathbf{v}})l(\mathbf{y}_j, \tilde{\mathbf{w}})\right| + \left|\frac{1}{n(n-1)}\sum_{i=1}^{n} k(\mathbf{x}_i, \tilde{\mathbf{v}})l(\mathbf{y}_i, \tilde{\mathbf{w}})\right| \\
&\leq \frac{B}{n-1} + \frac{B}{n-1} = \frac{2B}{n-1}.
\end{aligned}$$

Combining this upper bound with (18), we have

$$\|\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \mathbf{u}\mathbf{u}^\top\|_F \leq 4BJ\left|\hat{u}(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})\right| + \frac{8B^2J}{n-1}. \tag{19}$$

With (19), (17) becomes

$$|\hat{\lambda}_n - \lambda_n| \leq \frac{c_1 n}{\gamma_n}\|\hat{\mathbf{S}} - \mathbf{S}\|_F + \frac{4BJc_1 n}{\gamma_n}\left|\hat{u}(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})\right| + \frac{c_1 n}{\gamma_n}\frac{8B^2J}{n-1} + c_2 n\sqrt{J}|\hat{u}(\mathbf{t}^*) - u(\mathbf{t}^*)| + c_3 n\gamma_n. \tag{20}$$

E.2.4. BOUNDING $\|\hat{\mathbf{S}} - \mathbf{S}\|_F$

Recall that $V_J = \{\mathbf{t}_1, \ldots, \mathbf{t}_J\}$, $\hat{S}_{ij} = \hat{S}(\mathbf{t}_i, \mathbf{t}_j) = \frac{1}{n}\sum_{m=1}^{n} \overline{k}(\mathbf{x}_m, \mathbf{v}_i)\overline{l}(\mathbf{y}_m, \mathbf{w}_i)\overline{k}(\mathbf{x}_m, \mathbf{v}_j)\overline{l}(\mathbf{y}_m, \mathbf{w}_j)$, and $S_{ij} = S(\mathbf{t}_i, \mathbf{t}_j) = \mathbb{E}_{\mathbf{xy}}[\tilde{k}(\mathbf{x}, \mathbf{v}_i)\tilde{l}(\mathbf{y}, \mathbf{w}_i)\tilde{k}(\mathbf{x}, \mathbf{v}_j)\tilde{l}(\mathbf{y}, \mathbf{w}_j)]$. Let $(\mathbf{t}^{(1)}, \mathbf{t}^{(2)}) = \arg\max_{(\mathbf{s},\mathbf{t})\in V_J\times V_J} |\hat{S}(\mathbf{s}, \mathbf{t}) - S(\mathbf{s}, \mathbf{t})|$.

$$\begin{aligned}
\|\hat{\mathbf{S}} - \mathbf{S}\|_F &= \sup_{\mathbf{B}\in B_F(1)} \left\langle \mathbf{B}, \hat{\mathbf{S}} - \mathbf{S}\right\rangle_F \\
&\leq \sup_{\mathbf{B}\in B_F(1)} \sum_{i=1}^{J}\sum_{j=1}^{J} |B_{ij}||\hat{S}_{ij} - S_{ij}| \\
&\leq \left|\hat{S}(\mathbf{t}^{(1)}, \mathbf{t}^{(2)}) - S(\mathbf{t}^{(1)}, \mathbf{t}^{(2)})\right| \sup_{\mathbf{B}\in B_F(1)} \sum_{i=1}^{J}\sum_{j=1}^{J} |B_{ij}| \\
&\overset{(a)}{\leq} J\left|\hat{S}(\mathbf{t}^{(1)}, \mathbf{t}^{(2)}) - S(\mathbf{t}^{(1)}, \mathbf{t}^{(2)})\right| \sup_{\mathbf{B}\in B_F(1)} \|\mathbf{B}\|_F \\
&= J\left|\hat{S}(\mathbf{t}^{(1)}, \mathbf{t}^{(2)}) - S(\mathbf{t}^{(1)}, \mathbf{t}^{(2)})\right|,
\end{aligned} \tag{21}$$

where at (a) we used $\sum_{i=1}^{J}\sum_{j=1}^{J} |A_{ij}| \leq J\|\mathbf{A}\|_F$ for any matrix $\mathbf{A} \in \mathbb{R}^{J\times J}$. We arrive at

$$\begin{aligned}
|\hat{\lambda}_n - \lambda_n| \leq{}& \frac{c_1 J n}{\gamma_n}\left|\hat{S}(\mathbf{t}^{(1)}, \mathbf{t}^{(2)}) - S(\mathbf{t}^{(1)}, \mathbf{t}^{(2)})\right| + \frac{4BJc_1 n}{\gamma_n}\left|\hat{u}(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})\right| \\
&+ \frac{c_1 n}{\gamma_n}\frac{8B^2J}{n-1} + c_2 n\sqrt{J}|\hat{u}(\mathbf{t}^*) - u(\mathbf{t}^*)| + c_3 n\gamma_n.
\end{aligned} \tag{22}$$

E.2.5. Bounding $\left|\hat{S}(\mathbf{t},\mathbf{t}') - S(\mathbf{t},\mathbf{t}')\right|$

Having an upper bound for $\left|\hat{S}(\mathbf{t},\mathbf{t}') - S(\mathbf{t},\mathbf{t}')\right|$ will allow us to bound (22). To keep the notations uncluttered, we will define the following shorthands.

Expression	Shorthand
$k(\mathbf{x},\mathbf{v})$	a
$k(\mathbf{x},\mathbf{v}')$	a'
$k(\mathbf{x}_i,\mathbf{v})$	a_i
$k(\mathbf{x}_i,\mathbf{v}')$	a'_i
$\mathbb{E}_{\mathbf{x}\sim P_x} k(\mathbf{x},\mathbf{v})$	$\tilde{a}$
$\mathbb{E}_{\mathbf{x}\sim P_x} k(\mathbf{x},\mathbf{v}')$	$\tilde{a}'$
$\frac{1}{n}\sum_{i=1}^n k(\mathbf{x}_i,\mathbf{v})$	$\overline{a}$
$\frac{1}{n}\sum_{i=1}^n k(\mathbf{x}_i,\mathbf{v}')$	$\overline{a}'$

Expression	Shorthand
$l(\mathbf{y},\mathbf{w})$	b
$l(\mathbf{y},\mathbf{w}')$	b'
$l(\mathbf{y}_i,\mathbf{w})$	b_i
$l(\mathbf{y}_i,\mathbf{w}')$	b'_i
$\mathbb{E}_{\mathbf{y}\sim P_y} l(\mathbf{y},\mathbf{w})$	$\tilde{b}$
$\mathbb{E}_{\mathbf{y}\sim P_y} l(\mathbf{y},\mathbf{w}')$	$\tilde{b}'$
$\frac{1}{n}\sum_{i=1}^n l(\mathbf{y}_i,\mathbf{w})$	$\overline{b}$
$\frac{1}{n}\sum_{i=1}^n l(\mathbf{y}_i,\mathbf{w}')$	$\overline{b}'$

We will also use $\overline{\cdot}$ to denote a empirical expectation over $\mathbf{x}$, or $\mathbf{y}$, or $(\mathbf{x},\mathbf{y})$. The argument under $\overline{\cdot}$ will determine the variable over which we take the expectation. For instance, $\overline{aa'} = \frac{1}{n}\sum_{i=1}^n k(\mathbf{x}_i,\mathbf{v})k(\mathbf{x}_i,\mathbf{v}')$ and $\overline{aba'} = \frac{1}{n}\sum_{i=1}^n k(\mathbf{x}_i,\mathbf{v})l(\mathbf{y}_i,\mathbf{w})k(\mathbf{x}_i,\mathbf{v}')$, and so on. We define in the same way for the population expectation using $\tilde{\cdot}$ i.e., $\widetilde{aa'} = \mathbb{E}_{\mathbf{x}}\left[k(\mathbf{x},\mathbf{v})k(\mathbf{x},\mathbf{v}')\right]$ and $\widetilde{aba'} = \mathbb{E}_{\mathbf{xy}}\left[k(\mathbf{x},\mathbf{v})l(\mathbf{y},\mathbf{w})k(\mathbf{x},\mathbf{v}')\right]$.

With these shorthands, we can rewrite $\hat{S}(\mathbf{t},\mathbf{t}')$ and $S(\mathbf{t},\mathbf{t}')$ as

$$\begin{aligned}
\hat{S}(\mathbf{t},\mathbf{t}') &= \frac{1}{n}\sum_{i=1}^n (a_i - \overline{a})(b_i - \overline{b})(a'_i - \overline{a}')(b'_i - \overline{b}'), \\
S(\mathbf{t},\mathbf{t}') &= \mathbb{E}_{\mathbf{xy}}\left[(a-\tilde{a})(b-\tilde{b})(a'-\tilde{a}')(b'-\tilde{b}')\right].
\end{aligned}$$

By expanding $S(\mathbf{t},\mathbf{t}')$, we have

$$\begin{aligned}
S(\mathbf{t},\mathbf{t}') = \mathbb{E}_{\mathbf{xy}}\big[&+ aba'b' - aba'\tilde{b}' - ab\tilde{a}'b' + ab\tilde{a}'\tilde{b}' \\
&- a\tilde{b}a'b' + a\tilde{b}a'\tilde{b}' + a\tilde{b}\tilde{a}'b' - a\tilde{b}\tilde{a}'\tilde{b}' \\
&- \tilde{a}ba'b' + \tilde{a}ba'\tilde{b}' + \tilde{a}b\tilde{a}'b' - \tilde{a}b\tilde{a}'\tilde{b}' \\
&+ \tilde{a}\tilde{b}a'b' - \tilde{a}\tilde{b}a'\tilde{b}' - \tilde{a}\tilde{b}\tilde{a}'\tilde{b}' + \tilde{a}\tilde{b}\tilde{a}'\tilde{b}'\big] \\
= &+\widetilde{aba'b'} - \widetilde{aba'}\tilde{b}' - \widetilde{abb'}\tilde{a}' + \widetilde{ab}\tilde{a}'\tilde{b}' \\
&- \widetilde{aa'b'}\tilde{b} + \widetilde{aa'}\tilde{b}\tilde{b}' + \widetilde{ab'}\tilde{a}'\tilde{b} - {\color{blue}\tilde{a}\tilde{b}\tilde{a}'\tilde{b}'} \\
&- \widetilde{a'bb'}\tilde{a} + \widetilde{a'b}\tilde{a}\tilde{b}' + \tilde{a}\tilde{a}'\widetilde{bb'} - {\color{purple}\tilde{a}\tilde{b}\tilde{a}'\tilde{b}'} \\
&+ \widetilde{a'b'}\tilde{a}\tilde{b} - {\color{purple}\tilde{a}\tilde{b}\tilde{a}'\tilde{b}'} - {\color{purple}\tilde{a}\tilde{b}\tilde{a}'\tilde{b}'} + {\color{blue}\tilde{a}\tilde{b}\tilde{a}'\tilde{b}'} \\
= &+\widetilde{aba'b'} - \widetilde{aba'}\tilde{b}' - \widetilde{abb'}\tilde{a}' + \widetilde{ab}\tilde{a}'\tilde{b}' \\
&- \widetilde{aa'b'}\tilde{b} + \widetilde{aa'}\tilde{b}\tilde{b}' + \widetilde{ab'}\tilde{a}'\tilde{b} + \widetilde{a'b'}\tilde{a}\tilde{b} \\
&- \widetilde{a'bb'}\tilde{a} + \widetilde{a'b}\tilde{a}\tilde{b}' + \tilde{a}\tilde{a}'\widetilde{bb'} - 3\tilde{a}\tilde{b}\tilde{a}'\tilde{b}'.
\end{aligned}$$

The expansion of $\hat{S}(\mathbf{t},\mathbf{t}')$ can be done in the same way. By the triangle inequality, we have

$$\begin{aligned}
\left|\hat{S}(\mathbf{t},\mathbf{t}') - S(\mathbf{t},\mathbf{t}')\right| \le{} & \left|\overline{aba'b'} - \widetilde{aba'b'}\right| + \left|\overline{aba'}\,\overline{b}' - \widetilde{aba'}\tilde{b}'\right| + \left|\overline{abb'}\overline{a}' - \widetilde{abb'}\tilde{a}'\right| + \left|\overline{ab}\overline{a}'\overline{b}' - \widetilde{ab}\tilde{a}'\tilde{b}'\right| \\
& \left|\overline{aa'b'}\,\overline{b} - \widetilde{aa'b'}\tilde{b}\right| + \left|\overline{aa'}\,\overline{b}\,\overline{b}' - \widetilde{aa'}\tilde{b}\tilde{b}'\right| + \left|\overline{ab'}\overline{a}'\overline{b} - \widetilde{ab'}\tilde{a}'\tilde{b}\right| + \left|\overline{a'b'}\overline{a}\overline{b} - \widetilde{a'b'}\tilde{a}\tilde{b}\right|
\end{aligned}$$

$$\left|\overline{a'bb'}\overline{a} - \widetilde{a'bb'}\tilde{a}\right| + \left|\overline{a'b}\overline{a}\overline{b}' - \widetilde{a'b}\tilde{a}\tilde{b}'\right| + \left|\overline{a}\,\overline{a}'\overline{bb'} - \tilde{a}\tilde{a}'\widetilde{bb'}\right| + 3\left|\overline{a}\overline{b}\overline{a}'\overline{b}' - \tilde{a}\tilde{b}\tilde{a}'\tilde{b}'\right|.$$

The first term $\left|\overline{aba'b'} - \widetilde{aba'b'}\right|$ can be bounded by applying the Hoeffding's inequality. Other terms can be bounded by applying Lemma 8. Recall that we write $(x_1, \ldots, x_m)_+$ for $\max(x_1, \ldots, x_m)$.

Bounding $\left|\overline{aba'b'} - \widetilde{aba'b'}\right|$ (1^{st} term). Since $-B^2 \le aba'b' \le B^2$, by the Hoeffding's inequality (Lemma 13), we have

$$\mathbb{P}\left(\left|\overline{aba'b'} - \widetilde{aba'b'}\right| \le t\right) \ge 1 - 2\exp\left(-\frac{nt^2}{2B^4}\right).$$

Bounding $\left|\overline{aba'}\,\overline{b}' - \widetilde{aba'}\tilde{b}'\right|$ (2^{nd} term). Let $f_1(\mathbf{x}, \mathbf{y}) = aba' = k(\mathbf{x}, \mathbf{v})l(\mathbf{y}, \mathbf{w})k(\mathbf{x}, \mathbf{v}')$ and $f_2(\mathbf{y}) = b' = l(\mathbf{y}, \mathbf{w}')$. We note that $|f_1(\mathbf{x}, \mathbf{y})| \le (BB_k, B_l)_+$ and $|f_2(\mathbf{y})| \le (BB_k, B_l)_+$. Thus, by Lemma 8 with $E = 2$, we have

$$\mathbb{P}\left(\left|\overline{aba'}\,\overline{b}' - \widetilde{aba'}\tilde{b}'\right| \le t\right) \ge 1 - 4\exp\left(-\frac{nt^2}{8(BB_k, B_l)_+^4}\right).$$

Bounding $\left|\overline{ab}\overline{a}'\overline{b}' - \widetilde{ab}\tilde{a}'\tilde{b}'\right|$ (4^{th} term). Let $f_1(\mathbf{x}, \mathbf{y}) = ab = k(\mathbf{x}, \mathbf{v})l(\mathbf{y}, \mathbf{w})$, $f_2(\mathbf{x}) = a' = k(\mathbf{x}, \mathbf{v}')$ and $f_3(\mathbf{y}) = b' = l(\mathbf{y}, \mathbf{w}')$. We can see that $|f_1(\mathbf{x}, \mathbf{y})|, |f_2(\mathbf{x})|, |f_3(\mathbf{y})| \le (B, B_k, B_l)_+$. Thus, by Lemma 8 with $E = 3$, we have

$$\mathbb{P}\left(\left|\overline{ab}\overline{a}'\overline{b}' - \widetilde{ab}\tilde{a}'\tilde{b}'\right| \le t\right) \ge 1 - 6\exp\left(-\frac{nt^2}{18(B, B_k, B_l)_+^6}\right).$$

Bounding $\left|\overline{a}\overline{b}\overline{a}'\overline{b}' - \tilde{a}\tilde{b}\tilde{a}'\tilde{b}'\right|$ (last term). Let $f_1(\mathbf{x}) = a = k(\mathbf{x}, \mathbf{v})$, $f_2(\mathbf{y}) = b = l(\mathbf{y}, \mathbf{w})$, $f_3(\mathbf{x}) = a' = k(\mathbf{x}, \mathbf{v}')$ and $f_4(\mathbf{y}) = b' = l(\mathbf{y}, \mathbf{w}')$. It can be seen that $|f_1(\mathbf{x})|, |f_2(\mathbf{y})|, |f_3(\mathbf{x})|, |f_4(\mathbf{y})| \le (B_k, B_l)_+$. Thus, by Lemma 8 with $E = 4$, we have

$$\mathbb{P}\left(3\left|\overline{a}\overline{b}\overline{a}'\overline{b}' - \tilde{a}\tilde{b}\tilde{a}'\tilde{b}'\right| \le t\right) \ge 1 - 8\exp\left(-\frac{nt^2}{32 \cdot 3^2 (B_k, B_l)_+^8}\right).$$

Bounds for other terms can be derived in a similar way to yield

$$\begin{aligned}
&(3^{rd}\text{ term}) \quad \mathbb{P}\left(\left|\overline{abb'}\overline{a}' - \widetilde{abb'}\tilde{a}'\right| \le t\right) \ge 1 - 4\exp\left(-\frac{nt^2}{8(BB_l, B_k)_+^4}\right), \\
&(5^{th}\text{ term}) \quad \mathbb{P}\left(\left|\overline{aa'b'}\,\overline{b} - \widetilde{aa'b'}\tilde{b}\right| \le t\right) \ge 1 - 4\exp\left(-\frac{nt^2}{8(BB_k, B_l)_+^4}\right), \\
&(6^{th}\text{ term}) \quad \mathbb{P}\left(\left|\overline{aa'}\,\overline{b}\,\overline{b}' - \widetilde{aa'}\tilde{b}\tilde{b}'\right| \le t\right) \ge 1 - 6\exp\left(-\frac{nt^2}{18(B_k^2, B_l)_+^6}\right), \\
&(7^{th}\text{ term}) \quad \mathbb{P}\left(\left|\overline{ab'}\overline{a}'\overline{b} - \widetilde{ab'}\tilde{a}'\tilde{b}\right| \le t\right) \ge 1 - 6\exp\left(-\frac{nt^2}{18(B, B_k, B_l)_+^6}\right), \\
&(8^{th}\text{ term}) \quad \mathbb{P}\left(\left|\overline{a'b'}\overline{a}\overline{b} - \widetilde{a'b'}\tilde{a}\tilde{b}\right| \le t\right) \ge 1 - 6\exp\left(-\frac{nt^2}{18(B, B_k, B_l)_+^6}\right), \\
&(9^{th}\text{ term}) \quad \mathbb{P}\left(\left|\overline{a'bb'}\overline{a} - \widetilde{a'bb'}\tilde{a}\right| \le t\right) \ge 1 - 4\exp\left(-\frac{nt^2}{8(BB_l, B_k)_+^4}\right), \\
&(10^{th}\text{ term}) \quad \mathbb{P}\left(\left|\overline{a'b}\overline{a}\overline{b}' - \widetilde{a'b}\tilde{a}\tilde{b}'\right| \le t\right) \ge 1 - 6\exp\left(-\frac{nt^2}{18(B, B_k, B_l)_+^6}\right), \\
&(11^{th}\text{ term}) \quad \mathbb{P}\left(\left|\overline{a}\,\overline{a}'\overline{bb'} - \tilde{a}\tilde{a}'\widetilde{bb'}\right| \le t\right) \ge 1 - 6\exp\left(-\frac{nt^2}{18(B_k, B_l^2)_+^6}\right).
\end{aligned}$$

By the union bound, we have

$$
\begin{aligned}
&\mathbb{P}\left(\left|\hat{S}(\mathbf{t},\mathbf{t}') - S(\mathbf{t},\mathbf{t}')\right| \le 12t\right) \\
&\ge 1 - \Bigg[2\exp\left(-\frac{nt^2}{2B^4}\right) + 4\exp\left(-\frac{nt^2}{8(BB_k,B_l)_+^4}\right) + 4\exp\left(-\frac{nt^2}{8(BB_l,B_k)_+^4}\right) + 6\exp\left(-\frac{nt^2}{18(B,B_k,B_l)_+^6}\right) \\
&\qquad 4\exp\left(-\frac{nt^2}{8(BB_k,B_l)_+^4}\right) + 6\exp\left(-\frac{nt^2}{18(B_k^2,B_l)_+^6}\right) + 6\exp\left(-\frac{nt^2}{18(B,B_k,B_l)_+^6}\right) + 6\exp\left(-\frac{nt^2}{18(B,B_k,B_l)_+^6}\right) \\
&\qquad 4\exp\left(-\frac{nt^2}{8(BB_l,B_k)_+^4}\right) + 6\exp\left(-\frac{nt^2}{18(B,B_k,B_l)_+^6}\right) + 6\exp\left(-\frac{nt^2}{18(B_k,B_l^2)_+^6}\right) + 8\exp\left(-\frac{nt^2}{32\cdot 3^2(B_k,B_l)_+^8}\right)\Bigg] \\
&= 1 - \Bigg[2\exp\left(-\frac{nt^2}{2B^4}\right) + 8\exp\left(-\frac{nt^2}{8(BB_k,B_l)_+^4}\right) + 8\exp\left(-\frac{nt^2}{8(BB_l,B_k)_+^4}\right) + 24\exp\left(-\frac{nt^2}{18(B,B_k,B_l)_+^6}\right) \\
&\qquad + 6\exp\left(-\frac{nt^2}{18(B_k^2,B_l)_+^6}\right) + 6\exp\left(-\frac{nt^2}{18(B_k,B_l^2)_+^6}\right) + 8\exp\left(-\frac{nt^2}{32\cdot 3^2(B_k,B_l)_+^8}\right)\Bigg] \\
&\ge 1 - \Bigg[2\exp\left(-\frac{12^2nt^2}{B^*}\right) + 8\exp\left(-\frac{12^2nt^2}{B^*}\right) + 8\exp\left(-\frac{12^2nt^2}{B^*}\right) + 24\exp\left(-\frac{12^2nt^2}{B^*}\right) \\
&\qquad + 6\exp\left(-\frac{12^2nt^2}{B^*}\right) + 6\exp\left(-\frac{12^2nt^2}{B^*}\right) + 8\exp\left(-\frac{12^2nt^2}{B^*}\right)\Bigg] \\
&= 1 - 62\exp\left(-\frac{12^2nt^2}{B^*}\right),
\end{aligned}
$$

where

$$
B^* := \frac{1}{12^2}\max(2B^4, 8(BB_k,B_l)_+^4, 8(BB_l,B_k)_+^4, 18(B,B_k,B_l)_+^6, 18(B_k^2,B_l)_+^6, 18(B_k,B_l^2)_+^6, 32\cdot 3^2(B_k,B_l)_+^8).
$$

By reparameterization, it follows that

$$
\mathbb{P}\left(\frac{c_1Jn}{\gamma_n}\left|\hat{S}(\mathbf{t},\mathbf{t}') - S(\mathbf{t},\mathbf{t}')\right| \le t\right) \ge 1 - 62\exp\left(-\frac{\gamma_n^2t^2}{c_1^2J^2nB^*}\right). \tag{23}
$$

E.2.6. Union Bound for $\left|\hat{\lambda}_n - \lambda_n\right|$ and Final Lower Bound

Recall from (22) that

$$
\begin{aligned}
|\hat{\lambda}_n - \lambda_n| \le \frac{c_1Jn}{\gamma_n}\left|\hat{S}(\mathbf{t}^{(1)},\mathbf{t}^{(2)}) - S(\mathbf{t}^{(1)},\mathbf{t}^{(2)})\right| + \frac{4BJc_1n}{\gamma_n}\left|\hat{u}(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})\right| \\
+ \frac{c_1n}{\gamma_n}\frac{8B^2J}{n-1} + c_2n\sqrt{J}|\hat{u}(\mathbf{t}^*) - u(\mathbf{t}^*)| + c_3n\gamma_n.
\end{aligned}
$$

We will bound terms in (22) separately and combine all the bounds with the union bound. As shown in (8), the U-statistic core h is bounded between $-2B$ and $2B$. Thus, by Lemma 12 (with $m = 2$), we have

$$
\mathbb{P}\left(c_2n\sqrt{J}|\hat{u}(\mathbf{t}^*) - u(\mathbf{t}^*)| \le t\right) \ge 1 - 2\exp\left(-\frac{\lfloor 0.5n\rfloor t^2}{8c_2^2n^2JB^2}\right). \tag{24}
$$

Bounding $\frac{c_1n}{\gamma_n}\frac{8B^2J}{n-1} + c_3n\gamma_n + \frac{4BJc_1n}{\gamma_n}\left|\hat{u}(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})\right|$. By Lemma 12 (with $m = 2$), it follows that

$$
\begin{aligned}
&\mathbb{P}\left(\frac{c_1n}{\gamma_n}\frac{8B^2J}{n-1} + c_3n\gamma_n + \frac{4BJc_1n}{\gamma_n}\left|\hat{u}(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})\right| \le t\right) \\
&\ge 1 - 2\exp\left(-\frac{\lfloor 0.5n\rfloor\gamma_n^2\left[t - \frac{c_1n}{\gamma_n}\frac{8B^2J}{n-1} - c_3n\gamma_n\right]^2}{2^7B^4J^2c_1^2n^2}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= 1 - 2\exp\left(-\frac{\lfloor 0.5n \rfloor \left[t\gamma_n(n-1) - 8c_1B^2nJ - c_3n(n-1)\gamma_n^2\right]^2}{2^7B^4J^2c_1^2n^2(n-1)^2}\right) \\
&\overset{(a)}{\geq} 1 - 2\exp\left(-\frac{\left[t\gamma_n(n-1) - 8c_1B^2nJ - c_3n(n-1)\gamma_n^2\right]^2}{2^8B^4J^2c_1^2n^2(n-1)}\right),
\end{aligned} \tag{25}
$$

where at (a) we used $\lfloor 0.5n \rfloor \geq (n-1)/2$. Combining (23), (24), and (25) with the union bound (set $T = 3t$), we can bound (22) with

$$
\begin{aligned}
\mathbb{P}\left(\left|\hat{\lambda}_n - \lambda_n\right| \leq T\right) \geq 1 - 62\exp\left(-\frac{\gamma_n^2T^2}{3^2c_1^2J^2nB^*}\right) - 2\exp\left(-\frac{\lfloor 0.5n \rfloor T^2}{72c_2^2n^2JB^2}\right) \\
- 2\exp\left(-\frac{\left[T\gamma_n(n-1)/3 - 8c_1B^2nJ - c_3\gamma_n^2n(n-1)\right]^2}{2^8B^4J^2c_1^2n^2(n-1)}\right).
\end{aligned}
$$

Since $\left|\hat{\lambda}_n - \lambda_n\right| \leq T$ implies $\hat{\lambda}_n \geq \lambda_n - T$, a reparametrization with $r = \lambda_n - T$ gives

$$
\begin{aligned}
\mathbb{P}\left(\hat{\lambda}_n \geq r\right) \geq 1 - 62\exp\left(-\frac{\gamma_n^2(\lambda_n - r)^2}{3^2c_1^2J^2nB^*}\right) - 2\exp\left(-\frac{\lfloor 0.5n \rfloor(\lambda_n - r)^2}{72c_2^2n^2JB^2}\right) \\
- 2\exp\left(-\frac{\left[(\lambda_n - r)\gamma_n(n-1)/3 - 8c_1B^2nJ - c_3\gamma_n^2n(n-1)\right]^2}{2^8B^4J^2c_1^2n^2(n-1)}\right) \\
:= L(\lambda_n).
\end{aligned}
$$

Grouping constants into $\xi_1, \ldots \xi_5$ gives the result.

The lower bound $L(\lambda_n)$ takes the form

$$
1 - 62\exp\left(-C_1(\lambda_n - T_\alpha)^2\right) - 2\exp\left(-C_2(\lambda_n - T_\alpha)^2\right) - 2\exp\left(-\frac{[(\lambda_n - T_\alpha)C_3 - C_4]^2}{C_5}\right),
$$

where $C_1, \ldots, C_5$ are positive constants. For fixed large enough n such that $\lambda_n > T_\alpha$, and fixed significance level α, increasing λ_n will increase $L(\lambda_n)$. Specifically, since n is fixed, increasing $\mathbf{u}^\top\mathbf{\Sigma}^{-1}\mathbf{u}$ in $\lambda_n = n\mathbf{u}^\top\mathbf{\Sigma}^{-1}\mathbf{u}$ will increase $L(\lambda_n)$.

F. Helper Lemmas

This section contains lemmas used to prove the main results in this work.

Lemma 7 (Product to sum)**.** *Assume that* $|a_i| \leq B$, $|b_i| \leq B$ *for* $i = 1, \ldots, E$. *Then* $\left|\prod_{i=1}^E a_i - \prod_{i=1}^E b_i\right| \leq B^{E-1}\sum_{j=1}^E |a_j - b_j|$.

Proof.

$$
\begin{aligned}
\left|\prod_{i=1}^E a_i - \prod_{j=1}^E b_j\right| &\leq \left|\prod_{i=1}^E a_i - \prod_{i=1}^{E-1} a_ib_E\right| + \left|\prod_{i=1}^{E-1} a_ib_E - \prod_{i=1}^{E-2} a_ib_{E-1}b_E\right| + \ldots + \left|a_1\prod_{j=2}^E b_j - \prod_{j=1}^E b_j\right| \\
&\leq |a_E - b_E|\left|\prod_{i=1}^{E-1} a_i\right| + |a_{E-1} - b_{E-1}|\left|\left(\prod_{i=1}^{E-2} a_i\right)b_E\right| + \ldots + |a_1 - b_1|\left|\prod_{j=2}^E b_j\right| \\
&\leq |a_E - b_E|B^{E-1} + |a_{E-1} - b_{E-1}|B^{E-1} + \ldots + |a_1 - b_1|B^{E-1} \\
&= B^{E-1}\sum_{j=1}^E |a_j - b_j|
\end{aligned}
$$

applying triangle inequality, and the boundedness of a_i and b_i-s. □

Lemma 8 (Product variant of the Hoeffding's inequality)**.** *For $i = 1, \ldots, E$, let $\{\mathbf{x}_j^{(i)}\}_{j=1}^{n_i} \subset \mathcal{X}_i$ be an i.i.d. sample from a distribution P_i, and $f_i : \mathcal{X}_i \mapsto \mathbb{R}$ be a measurable function. Note that it is possible that $P_1 = P_2 = \cdots = P_E$ and $\{\mathbf{x}_j^{(1)}\}_{j=1}^{n_1} = \cdots = \{\mathbf{x}_j^{(E)}\}_{j=1}^{n_E}$. Assume that $|f_i(\mathbf{x})| \le B < \infty$ for all $\mathbf{x} \in \mathcal{X}_i$ and $i = 1, \ldots, E$. Write $\hat{P}_i$ to denote an empirical distribution based on the sample $\{\mathbf{x}_j^{(i)}\}_{j=1}^{n_i}$. Then,*

$$\mathbb{P}\left(\left|\left[\prod_{i=1}^{E} \mathbb{E}_{\mathbf{x}^{(i)} \sim \hat{P}_i} f_i(\mathbf{x}^{(i)})\right] - \left[\prod_{i=1}^{E} \mathbb{E}_{\mathbf{x}^{(i)} \sim P_i} f_i(\mathbf{x}^{(i)})\right]\right| \le T\right) \ge 1 - 2\sum_{i=1}^{E} \exp\left(-\frac{n_i T^2}{2E^2 B^{2E}}\right).$$

Proof. By Lemma 7, we have

$$\left|\left[\prod_{i=1}^{E} \mathbb{E}_{\mathbf{x}^{(i)} \sim \hat{P}_i} f_i(\mathbf{x}^{(i)})\right] - \left[\prod_{i=1}^{E} \mathbb{E}_{\mathbf{x}^{(i)} \sim P_i} f_i(\mathbf{x}^{(i)})\right]\right| \le B^{E-1} \sum_{i=1}^{E} \left|\mathbb{E}_{\mathbf{x}^{(i)} \sim \hat{P}_i} f_i(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x}^{(i)} \sim P_i} f_i(\mathbf{x}^{(i)})\right|.$$

By applying the Hoeffding's inequality to each term in the sum, we have $\mathbb{P}\left(\left|\mathbb{E}_{\mathbf{x}^{(i)} \sim \hat{P}_i} f_i(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x}^{(i)} \sim P_i} f_i(\mathbf{x}^{(i)})\right| \le t\right) \ge 1 - 2\exp\left(-\frac{2n_i t^2}{4B^2}\right)$. The result is obtained with a union bound. □

G. External Lemmas

In this section, we provide known results referred to in this work.

Lemma 9 (Chwialkowski et al. (2015, Lemma 1))**.** *If k is a bounded, analytic kernel (in the sense given in Definition 1) on $\mathbb{R}^d \times \mathbb{R}^d$, then all functions in the RKHS defined by k are analytic.*

Lemma 10 (Chwialkowski et al. (2015, Lemma 3))**.** *Let Λ be an injective mapping from the space of probability measures into a space of analytic functions on $\mathbb{R}^d$. Define*

$$d_{V_J}^2(P, Q) = \sum_{j=1}^{J} \left|[\Lambda P](\mathbf{v}_j) - [\Lambda Q](\mathbf{v}_j)\right|^2,$$

where $V_J = \{\mathbf{v}_i\}_{i=1}^{J}$ are vector-valued i.i.d. random variables from a distribution which is absolutely continuous with respect to the Lebesgue measure. Then, $d_{V_J}(P, Q)$ is almost surely (w.r.t. V_J) a metric.

Lemma 11 (Bochner's theorem (Rudin, 2011))**.** *A continuous function $\Psi : \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if it is the Fourier transform of a finite nonnegative Borel measure ζ on $\mathbb{R}^d$, that is, $\Psi(\mathbf{x}) = \int_{\mathbb{R}^d} e^{-i\mathbf{x}^\top \boldsymbol{\omega}}\, \mathrm{d}\zeta(\boldsymbol{\omega})$, $\mathbf{x} \in \mathbb{R}^d$.*

Lemma 12 (A bound for U-statistics (Serfling, 2009, Theorem A, p. 201))**.** *Let $h(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ be a U-statistic kernel for an m-order U-statistic such that $h(\mathbf{x}_1, \ldots, \mathbf{x}_m) \in [a, b]$ where $a \le b < \infty$. Let $U_n = \binom{n}{m}^{-1} \sum_{i_1 < \cdots < i_m} h(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_m})$ be a U-statistic computed with a sample of size n, where the summation is over the $\binom{n}{m}$ combinations of m distinct elements $\{i_1, \ldots, i_m\}$ from $\{1, \ldots, n\}$. Then, for $t > 0$ and $n \ge m$,*

$$\mathbb{P}(U_n - \mathbb{E}h(\mathbf{x}_1, \ldots, \mathbf{x}_m) \ge t) \le \exp\left(-2\lfloor n/m \rfloor t^2/(b-a)^2\right),$$
$$\mathbb{P}(|U_n - \mathbb{E}h(\mathbf{x}_1, \ldots, \mathbf{x}_m)| \ge t) \le 2\exp\left(-2\lfloor n/m \rfloor t^2/(b-a)^2\right),$$

where $\lfloor x \rfloor$ denotes the greatest integer which is smaller than or equal to x. Hoeffind's inequality is a special case when $m = 1$.

Lemma 13 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables such that $a \le X_i \le b$ almost surely. Define $\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$. Then,*

$$\mathbb{P}\left(\left|\overline{X} - \mathbb{E}[\overline{X}]\right| \le \alpha\right) \ge 1 - 2\exp\left(-\frac{2n\alpha^2}{(b-a)^2}\right).$$

References

[sup4] K. P. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast Two-Sample Testing with Analytic Representations of Probability Measures. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1981–1989. 2015.

[sup15] W. Jitkrittum, Z. Szabó, K. Chwialkowski, and A. Gretton. Interpretable Distribution Features with Maximum Testing Power. 2016. URL `http://arxiv.org/abs/1605.06796`.

[sup3] W. Rudin. *Fourier analysis on groups*. John Wiley & Sons, 2011.

[sup23] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 2009.