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 $O(n)$ and $O(n \log n)$ tests.

1. Introduction

We consider the design of adaptive, nonparametric statistical tests of dependence: that is, tests of whether a joint distribution $P_{X,Y}$ 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 $\tau$, 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.

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 distance covariance and Kendall’s $\tau$, are able to detect monotonic relations between univariate variables; for the 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 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 $\chi^2$ variables; for the block-averaged statistic, the null distribution is asymptotically normal; for Nyström, 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 $\chi^2$ 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 $O(n^2)$ 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,
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 basis functions, chosen for instance to be step functions or low order B-splines. The cost of this approach is $O(n)$. 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 $O(n \log n)$. 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 cost of this approach is $O(n)$. 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 $O(n \log n)$. 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).

In these experiments, we outperform competing linear and $O(n \log n)$ 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 $(a \otimes b) = a (b, c)$. Assume that $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ and $l : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ 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 $\mathcal{H}_k \rightarrow \mathcal{H}_k$ Hilbert-Schmidt operators. Then, HSIC between $X$ and $Y$ is defined as

$$
\text{HSIC}(X,Y) = \| \mu_{xy} - \mu_x \otimes \mu_y \|_{HS}^2
$$

where

$$
\begin{align*}
\mu_{xy} &= \mathbb{E}_{(x,y) \sim P_{xy}} [k(x, x') l(y, y')] \\
\mu_x &= \mathbb{E}_{x \sim P_x} \mathbb{E}_{y \sim P_y} [k(x, x') l(y, y')] \\
\mu_y &= \mathbb{E}_{x \sim P_x} \mathbb{E}_{y \sim P_y} [k(x, x') l(y, y')]
\end{align*}
$$

and $x'$ is an independent copy of $x$. The mean embedding of $P_{xy}$ belongs to the space of Hilbert-Schmidt operators from $\mathcal{H}_l$ to $\mathcal{H}_k$, $\mu_{xy} := \int_{\mathcal{X} \times \mathcal{Y}} k(x, \cdot) \otimes l(y, \cdot) \ dP_{xy}(x,y) \in \text{HS}(\mathcal{H}_l, \mathcal{H}_k)$, and the marginal mean embeddings are $\mu_x := \int_{\mathcal{X}} k(x, \cdot) \ dP_X(x) \in \mathcal{H}_k$ and $\mu_y := \int_{\mathcal{Y}} l(y, \cdot) \ dP_Y(y) \in \mathcal{H}_l$ (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 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 $Z_n = \{(x_i, y_i)\}_{i=1}^n \sim P_{xy}$, an empirical estimator of HSIC can be computed in $O(n^2)$ 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 $\mu_{x,y}(x,y) := \mu_x(x)\mu_y(y)$. The idea is to see $\mu_{x,y}(v,w) = \mathbb{E}_{x,y}[k(x,v)l(y,w)]$, $\mu_x(v) = \mathbb{E}_{x}[k(x,v)]$ and $\mu_y(w) = \mathbb{E}_{y}[l(y,w)]$ as smooth functions, and consider a new distance between $\mu_{x,y}$ and $\mu_{x,y}$ 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 $\mu_{x,y}$ and $\mu_{x,y}$ evaluated at $J$ random test locations $V_J := \{(v_i,w_i)\}_{i=1}^J \subset \mathcal{X} \times \mathcal{Y}$.

$\text{FSIC}^2(X,Y) := \frac{1}{J} \sum_{i=1}^J u^2(v_i,w_i) = \frac{1}{J} \|u\|_2^2$, where

$$u(v,w) := \mu_{xy}(v,w) - \mu_x(v)\mu_y(w) = \mathbb{E}_{xy}[k(x,v)l(y,w)] - \mathbb{E}_x[k(x,v)\mathbb{E}_y[l(y,w)]],$$

$$= \text{cov}_{xy}[k(x,v),l(y,w)].$$

$u := (u(v_1,w_1), \ldots, u(v_J,w_J))^T$, and $\{(v_i,w_i)\}_{i=1}^J$ are realizations from an absolutely continuous distribution (wrt the Lebesgue measure).

Our first result in Proposition 2 states that $\text{FSIC}(X,Y)$ 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 A (A0 kernels).** Let $X$ 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 $\sup_{x,x' \in \mathcal{X}} k(x,x') \leq B$), analytic (Chwialkowski et al., 2015) and vanishes at infinity: for all $v \in \mathcal{X}$, $f(x) := k(x,v)$ is bounded, real analytic on $\mathcal{X}$, and for all $\epsilon > 0$ the set $\{x \mid |f(x)| \geq \epsilon\}$ is compact.

**Assumption A.** The kernels $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $l: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ 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 $k$ and $l$ such that $k(x,x') = k(x-x',x)$ for all $x, x' \in \mathcal{X}$, and $l(y,y') = l(y-y',y')$ for all $y, y' \in \mathcal{Y}$.

**Proposition 2 (FSIC is a dependence measure).** Assume that assumption A holds, and that the test locations $V_J = \{(v_i,w_i)\}_{i=1}^J$ are drawn from an absolutely continuous distribution $\eta$. Then, $\eta$-almost surely, it holds that $\text{FSIC}(X,Y) = \frac{1}{\sqrt{2J}} \|u\|_2 = 0$ 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 $g((x,y),(v,w)) := k(x,v)l(y,w)$. Since $k$ and $l$ are $C_0$-kernels, characteristic and translation invariant, Gretton (2015, Theorem 2) implies that $u = 0$ if and only if $P_{xy} = P_xP_y$. Since for all $v \in \mathcal{X}, w \in \mathcal{Y}$, $x \mapsto k(x,v)$ and $y \mapsto l(y,w)$ are real analytic, it follows from Krantz & Parks (2002, Proposition 2.2.10) that $(x,y) \mapsto k(x,v)l(y,w)$ 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 $u \neq 0$, then the set of roots $R_u := \{(v,w) \mid u(v,w) = 0\}$ has Lebesgue measure zero (Mityagin, 2015). Hence, it is sufficient to draw $(v,w)$ from an absolutely continuous distribution to have $(v,w) \notin R_u \eta$-almost surely, and consequently if $X$ and $Y$ are dependent, then $\text{FSIC}(X,Y) > 0$, $\eta$-almost surely.

Examples of kernels $k$ and $l$ which satisfy Assumption A are the Gaussian kernels. FSIC uses $\mu_{xy}$ as a proxy for $P_{xy}$, and $\mu_{x,y}$ as a proxy for $P_xP_y$. Proposition 2 states that, to detect the dependence between $X$ and $Y$, it is sufficient to evaluate the difference of the population joint embedding $\mu_{xy}$ and the embedding of the product of the marginal distributions $\mu_{x,y}$ at a finite number of locations (defined by $V_J$). The intuitive explanation of this property is as follows. If $P_{xy} = P_xP_y$, then $u(v,w) = 0$ everywhere, and $\text{FSIC}(X,Y) = 0$ for any $V_J$. If $P_{xy} \neq P_xP_y$, 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 $R_u := \{(v,w) \mid u(v,w) = 0\}$ has Lebesgue measure zero (Mityagin, 2015). Thus, it is sufficient to draw $(v,w)$ from an absolutely continuous distribution to have $(v,w) \notin R_u \eta$-almost surely, and hence $\text{FSIC}(X,Y) > 0$. 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 $(v,w) \in R_u$, 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 $g((x,y),(v,w)) := k(x,v)l(y,w)$ which would have been much more restrictive.

**Plug-in Estimator** Assume that we observe a joint sample $Z_n := \{(x_i,y_i)\}_{i=1}^n$ i.i.d. $P_{xy}$. Unbiased estimators of $\mu_{xy}(v,w)$ and $\mu_{x,y}(v,w)$ are $\hat{\mu}_{xy}(v,w) := \frac{1}{n} \sum_{i=1}^n k(x_i,v)l(y_i,w)$ and $\hat{\mu}_{x,y}(v,w) := \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n k(x_i,v)l(y_j,w)$, respectively. A straightforward empirical estimator of $\text{FSIC}^2$ is then given by

$$\text{FSIC}^2(Z_n) = \frac{1}{J} \sum_{i=1}^J \hat{u}(v_i,w_i)^2,$$

$$\hat{u}(v,w) := \hat{\mu}_{xy}(v,w) - \hat{\mu}_{x,y}(v,w)$$

$$= \frac{2}{n(n-1)} \sum_{i,j} h_{(v,w)}((x_i,y_i),(x_j,y_j))$$

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

The idea is to see $\text{FSIC}^2(X,Y)$ as a proxy for $\text{HSIC}(X,Y)$, and we refer to Gretton et al. (2005) for a discussion of the properties of this estimator.
whose asymptotic null distribution takes a more convenient
form linear in
coordinates
Thus, by Lehmann (1999, Theorem 6.1.6) and Kowalski
Lemma
\[ \Sigma = \frac{1}{n} \sum \{ \hat{u}(v_i, w_i) \} . \]

Since FSIC satisfies \( \text{FSIC}(X, Y) = 0 \iff X \perp Y \), in principle its empirical estimator can be used as a test statistic for an independence test proposing a null hypothesis \( H_0 : "X \text{ and } Y \text{ are independent}" 

An Adaptive Test of Independence with Analytic Kernel Embeddings

Proposition 3 (Asymptotic distribution of \( \hat{u} \)). Define \( \hat{u} := (\hat{u}(t_1), \ldots, \hat{u}(t_T))^\top \), \( \hat{k}(x, v) := k(x, v) - \mathbb{E}_x k(x', v) \), and \( \hat{l}(y, w) := l(y, w) - \mathbb{E}_w l(y', w) \). Let \( \Sigma = \{ \Sigma_{ij} \} \in \mathbb{R}^{J \times J} \) be the positive semi-definite matrix with entries \( \Sigma_{ij} = \text{cov}_x (\hat{u}(t_i), \hat{u}(t_j)) = \mathbb{E}_x \{ \hat{k}(x, v_i) \hat{l}(y, w_j) \} - u(t_i)u(t_j) \). Then, under both \( H_0 \) and \( H_1 \), for any fixed test locations \( \{ t_1, \ldots, t_J \} \) for which \( \Sigma \) is full rank, and \( 0 < \mathbb{V}_x \{ h_{t_j}(u) \} < \infty \) for \( j = 1, \ldots, J \), it holds that \( \sqrt{n} (\hat{u} - u) \stackrel{d}{\to} \mathcal{N}(0, \Sigma) \). 

Proof. For a fixed \( \{ t_1, \ldots, t_J \} \), \( \hat{u} \) is a one-sample second-order multivariate \( U \)-statistic with a \( U \)-statistic kernel \( h_t \). Thus, by Lehmann (1999, Theorem 6.1.6) and Kowalski & Tu (2008, Section 5.1, Theorem 1), it follows directly that \( \sqrt{n} (\hat{u} - u) \stackrel{d}{\to} \mathcal{N}(0, \Sigma) \) where we note that \( \mathbb{E}_{xy} \{ \hat{k}(x, v) \hat{l}(y, w) \} = u(v, w) \).

Recall from Proposition 2 that \( u = 0 \) holds almost surely under \( H_0 \). The asymptotic normality described in Proposition 3 implies that \( n \text{FSIC}^2 = \frac{1}{n} \hat{u}^\top \hat{u} \) converges in distribution to a sum of \( J \) dependent weighted \( \chi^2 \) random variables. The dependence comes from the fact that the coordinates \( \hat{u}_1, \ldots, \hat{u}_J \) of \( \hat{u} \) 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 \( T_n \) for a given significance value \( \alpha \).

2.2. Normalized FSIC and Adaptive Test

For the purpose of an independence test, we will consider a normalized variant of FSIC2, which we call \( \text{FSIC}^2 \), 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 \( \text{FSIC}^2 \) is consistent. These results are given in Theorem 4.

Theorem 4 (Independence test based on \( \text{FSIC}^2 \) is consistent). Let \( \Sigma \) be a consistent estimate of \( \Sigma \) based on the joint sample \( Z_n \), where \( \Sigma \) is defined in Proposition 3. Assume that \( V_J = \{ (v_i, w_i) \} \} \sim \eta \) where \( \eta \) is absolutely continuous wrt the Lebesgue measure. The \( \text{FSIC}^2 \) statistic is defined as \( \hat{\lambda}_n := n \hat{u}^\top \left( \hat{\Sigma} + \gamma_n I \right)^{-1} \hat{u} \) where \( \gamma_n \geq 0 \) is a regularization parameter. Assume that

1. Assumption A holds.
2. \( \Sigma \) is invertible \( \eta \)-almost surely.
3. \( \lim_{n \to \infty} \gamma_n = 0 \).

Then, for any \( k, l \) and \( V_J \) satisfying the assumptions,

1. Under \( H_0 \), \( \hat{\lambda}_n \overset{d}{\to} \chi^2(J) \) as \( n \to \infty \).
2. Under \( H_1 \), for any \( r \in \mathbb{R} \), \( \lim_{n \to \infty} \mathbb{P} (\hat{\lambda}_n \geq r) = 1 \) \( \eta \)-almost surely. That is, the independence test based on \( \text{FSIC}^2 \) is consistent.

Proof (sketch). Under \( H_0 \), \( n \hat{u}^\top \left( \hat{\Sigma} + \gamma_n I \right)^{-1} \hat{u} \) asymptotically follows \( \chi^2(J) \) because \( \sqrt{n} \hat{u} \) is asymptotically normally distributed (see Proposition 3). Claim 2 builds on the result in Proposition 2 stating that \( \hat{u} \neq 0 \) under \( H_1 \); it follows using the convergence of \( \hat{u} \) to \( u \). 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 \( T_n \) is given by the \( (1 - \alpha) \)-quantile of \( \chi^2(J) \) and is independent of \( n \). The assumption on the consistency of \( \Sigma \) is required to obtain the asymptotic chi-squared distribution. The regularization parameter \( \gamma_n \) is to ensure that \( \left( \hat{\Sigma} + \gamma_n I \right)^{-1} \) can be stably computed. In practice, \( \gamma_n \) 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 \( \text{FSIC}^2 \) estimator is linear in both the input dimension and sample size, and that it can be expressed in terms of the \( K = \{ K_{ij} \} = \{ k(v_i, x_j) \} \in \mathbb{R}^{J \times J} \), \( L = \{ L_{ij} \} = \{ l(w_i, y_j) \} \in \mathbb{R}^{J \times J} \) matrices. In contrast to typical kernel methods, a large Gram matrix of size \( n \times n \) is not needed to compute \( \text{FSIC}^2 \).

Proposition 5 (An empirical estimator of \( \text{FSIC}^2 \)). Let \( 1_n := (1, \ldots, 1)^\top \in \mathbb{R}^n \). Denote by \( \circ \) the element-wise matrix product. Then,
1. \( \hat{u} = (K \circ L)_{n^{-1}} - (K_{1n} \circ (L1_n))_{n(n-1)} \).

2. A consistent estimator for \( \Sigma \) is \( \hat{\Sigma} = \frac{\Gamma \Gamma^\top}{n} \) where
   \[
   \Gamma := (K - n^{-1} K_{1n} 1_n^\top) \circ (L - n^{-1} L_{1n} 1_n^\top) - \hat{u} \hat{u}^\top, \]
   \[
   \hat{u} = n^{-1} (K \circ L) 1_n - n^{-2} (K_{1n} \circ 1_n^n) .
   \]

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

**Proof (sketch).** Claim 1 for \( \hat{u} \) is straightforward. The expression for \( \hat{\Sigma} \) in claim 2 follows directly from the asymptotic covariance expression in Proposition 3. The consistency of \( \hat{\Sigma} \) can be obtained by noting that the finite sample bound for \( P(\| \hat{\Sigma} - \Sigma \|_F > t) \) 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 \( X \times Y \) of \( p_{xy} \) and \( p_{x,y} \) 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 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 NFSIC\(^2\)(\( X, Y \)) := \( \lambda_n := nu^\top \Sigma^{-1} u \). Let \( K \) be a kernel class for \( k \), \( L \) be a kernel class for \( l \), and \( V \) 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
   \[
   \sup_{k \in K} \sup_{x, x' \in X} |k(x, x')| \leq B_k \quad \text{and} \quad \sup_{l \in L} \sup_{y, y' \in Y} |l(y, y')| \leq B_l .
   \]

2. \( \hat{\gamma} := \sup_{k \in K} \sup_{x, x' \in X} \sup_{l \in L} \sup_{y, y' \in Y} \| \Sigma^{-1} l \|_F < \infty .
   \]

Then, for any \( k \in K \), \( l \in L \), \( V \) \( J \) is \( \mathcal{F} \), and \( \lambda_n \geq r \), the test power satisfies \( P \left( \lambda_n \geq r \right) \geq L(\lambda_n) \) where
   \[
   L(\lambda_n) := 1 - 62 e^{-\xi_1 (\lambda_n - r)^2/2} - 2 e^{-|0.5 n | (\lambda_n - r)^2 / [\xi_2 n^2]} - 2 e^{-\left( |\lambda_n - r| \gamma_n (n-1)/3 - e^{-\xi_3 n - \xi_4 n^2} x_n (n-1)^2 / [\xi_5 n^2 (n-1)^2] \right)} .
   \]

\( [\cdot] \) is the floor function, \( \xi_1 := \frac{1}{2(\xi_4 + \xi_5)} \), \( B^* \) is a constant depending on only \( B_k \) and \( B_l \), \( \xi_2 := 72 \xi_5 B_k^2 \), \( B := B_k B_l \), \( \xi_3 := 8c_1 B^2 J \), \( c_1 := 4B^2 J^2 c_2 \), \( c_2 := 4B^2 J^2 c_2 \), \( c_1 := 4B^2 J^2 \mathcal{J} \), and \( c_2 := 4B^2 J^2 \mathcal{J} \). Moreover, for sufficiently large fixed \( n \), \( L(\lambda_n) \) is increasing in \( \lambda_n \).

We provide the proof in Appendix E. To put Theorem 6 into perspective, assume that \( K = \{(x, v) \mapsto \exp \left(-\frac{|x-\bar{v}|^2}{2\sigma_1^2} \right) \mid (\sigma_1^2, \sigma_2^2) \} := K_{\sigma_2} \) for some \( 0 < \sigma_1^2 < \sigma_2^2 < \infty \) and \( L = \{(y, w) \mapsto \exp \left(-\frac{|y-\bar{w}|^2}{2\sigma_2^2} \right) \mid (\sigma_1^2, \sigma_2^2) \} := L_{\sigma_2} \) for some \( 0 < \sigma_1^2 < \sigma_2^2 < \infty \) are Gaussian kernel classes. Then, in Theorem 6, \( B = B_k = B_l = 1 \), and \( B^* = 2 \). The assumption \( \hat{c} < \infty \) is a technical condition to guarantee that the test power lower bound is finite for all \( \theta \) defined by the feasible sets \( K, L, \mathcal{F} \). Let \( \mathcal{V}_{\theta, r} := \{ V_j \mid \| v_i - v_j \|_2^2 \leq r \) and \( \mathcal{V}_{\theta, r} \) for all \( \theta \). If we set \( K = K_{\sigma_2}, L = L_{\sigma_2}, \mathcal{V} = \mathcal{V}_{\theta, r} \) and \( \lambda_n \) is increasing in \( \lambda_n \). One can therefore think of \( \lambda_n \) (a function of \( \theta \)) as representing how easily the test rejects \( H_0 \) given a problem \( P_{xy} \). The higher the \( \lambda_n \), 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 \( \theta * \) by maximizing the lower bound on the test power i.e., set \( \theta * = \arg \max_{\theta} L(\lambda_n) \). Assume that \( n \) is sufficiently large so that \( \lambda_n \mapsto L(\lambda_n) \) is an increasing function. Then, \( \arg \max_{\theta} L(\lambda_n) \) is a constant \( \lambda_n \). That this procedure is also valid under \( H_0 \) can be seen as follows. Under \( H_0 \), \( \theta * = \arg \max_{\theta} L(\lambda_n) \) will be arbitrary. Since Theorem 6 guarantees that \( \lambda_n \mapsto \chi_2(J) \) as \( n \to \infty \) for any \( \theta \), the asymptotic null distribution does not change by using \( \theta * \). In practice, \( \lambda_n \) 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 \( \lambda_n \) (an estimate of \( \lambda_n \)) to optimize for \( \theta * \), and the test set is used for the actual independence test with the optimized \( \theta * \). The splitting is to guarantee the independence of \( \theta * \) and the test sample to avoid overfitting.
An Adaptive Test of Independence with Analytic Kernel Embeddings

To better understand the behaviour of NFSIC, we visualize $\hat{\mu}_{xy}(v, w), \hat{\mu}_{x\mu_y}(v, w)$ and $\hat{\Sigma}(v, w)$ as a function of one test location $(v, w)$ on a simple toy problem. In this problem, $Y = -X + Z$ where $Z \sim N(0, 0.3^2)$ is an independent noise variable. As we consider only one location $(J = 1)$, $\hat{\Sigma}(v, w)$ is a scalar. The statistic can be written as $\hat{\lambda}_n = n \frac{(\hat{\mu}_{xy}(v, w) - \hat{\mu}_{x\mu_y}(v, w))^2}{\hat{\Sigma}(v, w)}$. 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 $v \in \mathbb{R}$ and $w \in \mathbb{R}$, respectively.

Intuitively, $\hat{u}(v, w) = \hat{\mu}_{xy}(v, w) - \hat{\mu}_{x\mu_y}(v, w)$ captures the difference of the joint distribution and the product of the marginals as a function of $(v, w)$. Squaring $\hat{u}(v, w)$ 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 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 $O(n)$, 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 $O((d_x + d_y)n \log n)$.

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 $\sigma_x$ and $\sigma_y$ are set according to the widely used median heuristic i.e., $\sigma_x = \text{median}\left\{\left\|x_i - x_j\right\|_2 \mid 1 \leq i < j \leq n\right\}$, and $\sigma_y$ is set in the same way using $\{y_i\}_{i=1}^n$. 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 HSIC 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 $\sigma_x, \sigma_y$, and $J$ locations of size $(d_x + d_y)J$. We treat all the parameters as a long vector in $\mathbb{R}^{2n + (d_x + d_y)J}$ 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 $\gamma_n$ 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.$^3$

$^3$We use a permutation test for RDC, following the authors’ implementation (https://github.com/lopezpaz/randomized_dependence_coefficient, referred commit: b0ac6c0).

$^2$Our claim on linear runtime (with respect to $n$) is for the gradient ascent procedure to find a local optimum for $\theta$. We do not
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 contrast, 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., $X \sim \mathcal{N}(0, I_d)$ and $Y \sim \mathcal{N}(0, I_d)$ where $I_d$ 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 $X, Y = (-\pi, \pi)$ and $\omega$ is the frequency of the sinusoid. As the frequency $\omega$ increases, the drawn sample becomes more similar to a sample drawn from Uniform($(-\pi, \pi)^2$). That is, the higher $\omega$, 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 $\omega$ 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, $Y = |Z| \prod_{i=1}^{d} \text{sgn}(X_i)$, where $X \sim \mathcal{N}(0, I_d)$, $\text{sgn}(\cdot)$ is the sign function, and $Z \sim \mathcal{N}(0, 1)$ serves as a source of noise. The full interaction of $X = (X_1, \ldots, X_d)$ 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 $\alpha$ 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 $X \times Y$. 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 $(v, w)$. 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.
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 contrast, 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 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.

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


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.

![Figure 5: Probability of rejecting $H_0$ as $n$ increases. $\alpha = 0.01$.](image)

B. Redundant Test Locations

Here, we provide a simple illustration to show that two locations $t_1 = (v_1, w_1)$ and $t_2 = (v_2, w_2)$ 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, $t_1$ is fixed at the red star, while $t_2$ is varied along the horizontal line. The objective value $\hat{\lambda}_n$ as a function of $t_2$ is shown in the bottom figure. It can be seen that $\hat{\lambda}_n$ decreases sharply when $t_2$ is in the neighborhood of $t_1$. 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 $t_2$ moves along the green line. The objective sharply drops when the two locations are in the same neighborhood.](image)

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 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.
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 $\Sigma$ harder to estimate accurately, it also increases the computation as the complexity on $J$ is $O(J^3)$.

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 NFSIC$^2$ is consistent). Let $\hat{\Sigma}$ be a consistent estimate of $\Sigma$ based on the joint sample $Z_n$, where $\Sigma$ is defined in Proposition 3. Assume that $V_i = \{(v_i, w_i)\}_{i=1}^l \sim \eta$ where $\eta$ is absolutely continuous wrt the Lebesgue measure. The NFSIC$^2$ statistic is defined as $\hat{\lambda}_n := n\hat{\mu}^\top \left(\hat{\Sigma} + \gamma_n I\right)^{-1} \hat{\mu}$ where $\gamma_n \geq 0$ is a regularization parameter. Assume that

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

Then, for any $k, l$ and $V_i$ satisfying the assumptions,

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

**Proof.** Assume that $H_0$ holds. The consistency of $\hat{\Sigma}$ and the continuous mapping theorem imply that $\left(\hat{\Sigma} + \gamma_n I\right)^{-1} \overset{p}{\to} \Sigma^{-1}$ which is a constant. Let $a$ be a random vector in $\mathbb{R}^J$ following $\mathcal{N}(0, \Sigma)$. By van der Vaart (2000, Theorem 2.7 (v)), it follows that $\left[\sqrt{n} \hat{\mu}, \left(\hat{\Sigma} + \gamma_n I\right)^{-1}\right] \overset{d}{\to} \left[a, \Sigma^{-1}\right]$ where $\hat{\mu} = 0$ almost surely by Proposition 2, and $\sqrt{n} \hat{\mu} \overset{d}{\to} \mathcal{N}(0, \Sigma)$ by Proposition 3. Since $f(x, S) := x^\top Sx$ is continuous, $f\left(\sqrt{n} \hat{\mu}, \left(\hat{\Sigma} + \gamma_n I\right)^{-1}\right) \overset{d}{\to} f(a, \Sigma^{-1})$. Equivalently, $n\hat{\mu}^\top \left(\hat{\Sigma} + \gamma_n I\right)^{-1} \hat{\mu} \overset{d}{\to} a^\top \Sigma^{-1} a \sim \chi^2(J)$ 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, \( u \neq 0 \) almost surely by Proposition 2. Since \( k \) and \( l \) are bounded, it follows that
\[
|\hat{h}_n(z, z')| \leq 2d^2 B_1 \text{ for any } z, z' \text{ (see (8))},
\]
and we have that \( \hat{u} \overset{d}{=} u \) by Serfling (2009, Section 5.4, Theorem A). Thus, \( \hat{u}^\top (\Sigma + \gamma_n k)^{-1} \hat{u} - \frac{r}{n} \overset{d}{=} u^\top \Sigma^{-1} u \) by the continuous mapping theorem, and the consistency of \( \hat{\Sigma} \). Consequently,
\[
\lim_{n \to \infty} \mathbb{P} \left( \hat{\lambda}_n \geq r \right) = 1 - \lim_{n \to \infty} \mathbb{P} \left( \hat{u}^\top (\Sigma + \gamma_n I)^{-1} \hat{u} - \frac{r}{n} < 0 \right) = 1 - \mathbb{P} \left( u^\top \Sigma^{-1} u < 0 \right) = 1,
\]
where at (a) we use the Portmanteau theorem (van der Vaart, 2000, Lemma 2.2 (i)) guaranteeing that \( x_n \to x \) if and only if \( \mathbb{P}(x_n < t) \to \mathbb{P}(x < t) \) for all continuity points of \( t \to \mathbb{P}(x < t) \). Step (b) is justified by noting that the covariance matrix \( \Sigma \) is positive definite so that \( u^\top \Sigma^{-1} u > 0 \), and \( t \to \mathbb{P}(u^\top \Sigma^{-1} u < t) \) (a step function) is continuous at 0.

\[\square\]

### E. Proof of Theorem 6

Recall Theorem 6.

**Theorem 6** (A lower bound on the test power). Let NFSIC\(^2\)(\(X, Y\)) := \( \lambda_n := n u^\top \Sigma^{-1} u \). Let \( K \) be a kernel class for \( k \), \( L \) be a kernel class for \( l \), and \( V \) 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 \( \sup_{k \in K} \sup_{x, x' \in \mathcal{X}} |k(x, x')| \leq B_k \) and \( \sup_{l \in L} \sup_{y, y' \in \mathcal{Y}} |l(y, y')| \leq B_l \).
2. \( c := \sup_{k \in K} \sup_{l \in L} \sup_{v \in V} \|\Sigma^{-1}\|_F < \infty \).

Then, for any \( k \in K, l \in L, V_j \in V \), and \( \lambda_n \geq r \), the test power satisfies \( \mathbb{P} \left( \hat{\lambda}_n \geq r \right) \geq L(\lambda_n) \) where
\[
L(\lambda_n) = 1 - 62e^{-\xi_1^2(\lambda_n - r)^2/n} - 2e^{-0.5n[(\lambda_n - r)^2]/[\xi_1^2n^2]} - 2e^{-[(\lambda_n - r)\gamma n(n-1)/3 - \xi_2^2n^2c_3n(n-1)]^2/[\xi_3^2n^2(n-1)]},
\]
\( \lfloor \cdot \rfloor \) is the floor function, \( \xi_1 := \frac{1}{\sqrt{2c_3^2JB^3}}, B^* \) is a constant depending on only \( B_k \) and \( B_l \), \( \xi_2 := 72c_3^2JB^2, B := B_k B_l, \)
\( \xi_3 := 8c_3 B^2 J, c_3 := 4B^2 J_2^2, \xi_4 := 28B^2 J_2^2, c_1 := 4B^2 J \sqrt{Jc}, \) and \( c_2 := 4B \sqrt{Jc} \). Moreover, for sufficiently large fixed \( n \), \( L(\lambda_n) \) is increasing in \( \lambda_n \).

**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 \( \| \hat{u} - u \|_2 \) and \( \| \Sigma - \Sigma \|_F \). The difference \( \| \hat{u} - u \|_2 \) can be bounded by applying the bound for U-statistics given in Serfling (2009, Theorem A, p. 201). For \( \| \Sigma - \Sigma \|_F \), 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 \( (A, B)_F := \text{tr}(A^\top B) \) denote the Frobenius inner product, and \( \|A\|_F := \sqrt{\text{tr}(A^\top A)} \) be the Frobenius norm. Write \( z := (x, y) \) to denote a pair of points from \( \mathcal{X} \times \mathcal{Y} \). We write \( t := (v, w) \) to denote a pair of test locations from \( \mathcal{X} \times \mathcal{Y} \). For brevity, an expectation over \( (x, y) \) (i.e., \( E_{(x,y) \sim P_{xy}} \)) will be written as \( E_x \) or \( E_{xy} \). Define \( \hat{k}(x, v) := \hat{k}(x, v) - E_x k(x', v), \) and \( \hat{l}(y, w) := \hat{l}(y, w) - E_y l(y', w) \). Let \( B_j(r) := \{ x \mid \|x\|_2 \leq r \} \) be a closed ball with radius \( r \) centered at the origin. Similarly, define \( B_F(r) := \{ A \mid \|A\|_F \leq r \} \) to be a closed ball with radius \( r \) of \( J \times J \) matrices under the Frobenius norm. Denote the max operation by \( \max(x_1, \ldots, x_m) := \max(x_1, \ldots, x_m) \).

For a product of marginal mean embeddings \( \mu_x(v) \mu_y(w) \), we write \( \hat{\mu}_x \hat{\mu}_y(v, w) := \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i} k(x_i, v)l(y_j, w) \) to denote the unbiased plug-in estimator, and write \( \hat{\mu}_x(v) \mu_y(w) := \frac{1}{n} \sum_{i=1}^n k(x_i, v)l(y_j, w) \) which is a biased estimator. Define \( \hat{u}^b(v, w) := \hat{\mu}_x(v, w) - \mu_x(v) \mu_y(w) \) so that \( \hat{u}^b := (\hat{u}^b(t_1), \ldots, \hat{u}^b(t_J)) \) where the superscript \( b \) stands for “biased”. To avoid 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 \( P(\hat{\lambda}_n - \lambda_n \geq t) \), which will then be reparametrized to get a bound for the target quantity \( P(\hat{\lambda}_n \geq r) \). We closely follow the proof in Jitkrittum et al. (2016, Section C.1) up to (12), then we diverge. We start by considering \( |\hat{\lambda}_n - \lambda_n|/n \).

\[
|\hat{\lambda}_n - \lambda_n|/n = \left| \hat{u}^\top (\hat{\Sigma} + \gamma_n I)^{-1} \hat{u} - u^\top \Sigma^{-1} u \right|
= \left| \hat{u}^\top \left( \hat{\Sigma} + \gamma_n I \right)^{-1} \hat{u} - u^\top \left( \Sigma + \gamma_n I \right)^{-1} u + u^\top \left( \Sigma + \gamma_n I \right)^{-1} u - u^\top \Sigma^{-1} u \right|
\leq \left| \hat{u}^\top \left( \hat{\Sigma} + \gamma_n I \right)^{-1} \hat{u} - u^\top \left( \Sigma + \gamma_n I \right)^{-1} u \right| + \left| u^\top \left( \Sigma + \gamma_n I \right)^{-1} u - u^\top \Sigma^{-1} u \right|
:= (\star)_1 + (\star)_2.
\]

We next bound \((\star)_1\) and \((\star)_2\) separately.

\[
(\star)_1 = \left| \hat{u}^\top \left( \hat{\Sigma} + \gamma_n I \right)^{-1} \hat{u} - u^\top \left( \Sigma + \gamma_n I \right)^{-1} u \right|
= \left| \hat{u}^\top \left( \hat{\Sigma} + \gamma_n I \right)^{-1} \hat{u} - u^\top \left( \Sigma + \gamma_n I \right)^{-1} \hat{u} + \hat{u}^\top (\Sigma + \gamma_n I)^{-1} \hat{u} - u^\top (\Sigma + \gamma_n I)^{-1} u \right|
\leq \left| \hat{u}^\top \left( \hat{\Sigma} + \gamma_n I \right)^{-1} \hat{u} - u^\top \left( \Sigma + \gamma_n I \right)^{-1} \hat{u} \right| + \left| \hat{u}^\top (\Sigma + \gamma_n I)^{-1} \hat{u} - u^\top (\Sigma + \gamma_n I)^{-1} u \right|
= \left| \hat{u}^\top \left( \hat{\Sigma} + \gamma_n I \right)^{-1} \hat{u} - (\Sigma + \gamma_n I)^{-1} \hat{u} \right| + \left| \hat{u}^\top (\Sigma + \gamma_n I)^{-1} \hat{u} - \left( \Sigma + \gamma_n I \right)^{-1} u \right|
\leq \left\| \hat{u}^\top \right\|_F \left\| \left( \hat{\Sigma} + \gamma_n I \right)^{-1} - (\Sigma + \gamma_n I)^{-1} \right\|_F \left\| \hat{u} \right\|_F + \left\| \hat{u}^\top - \hat{u}^\top \right\|_F \left\| (\Sigma + \gamma_n I)^{-1} \right\|_F
\leq (a) \left\| \hat{u}^\top \right\|_F \left\| (\hat{\Sigma} + \gamma_n I)^{-1} \right\|_F \left\| \hat{\Sigma} - \Sigma \right\|_F \left\| \Sigma^{-1} \right\|_F + \left\| \hat{u}^\top - \hat{u}^\top \right\|_F \left\| (\hat{\Sigma} + \gamma_n I)^{-1} \right\|_F \left\| \Sigma^{-1} \right\|_F
\leq (b) \frac{\sqrt{J}}{\gamma_n} \left\| \hat{u} \right\|_2 \left\| \Sigma - \hat{\Sigma} \right\|_F \left\| \Sigma^{-1} \right\|_F + \left( \left\| \hat{u} - u \right\|_F + \left\| \hat{u} - u \right\|_F \right) \left\| \Sigma^{-1} \right\|_F
\leq \frac{\sqrt{J}}{\gamma_n} \left\| \hat{u} \right\|_2 \left\| \Sigma - \hat{\Sigma} \right\|_F \left\| \Sigma^{-1} \right\|_F + \left( \left\| \hat{u} \right\|_2 + \left\| \hat{u} \right\|_2 \right) \left\| \hat{u} - u \right\|_2 \left\| \Sigma^{-1} \right\|_F,
\]

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

For \((\star)_2\), we have

\[
(\star)_2 = \left| u^\top \left( \Sigma + \gamma_n I \right)^{-1} u - u^\top \Sigma^{-1} u \right|
= \left| \langle uu^\top, (\Sigma + \gamma_n I)^{-1} - \Sigma^{-1} \rangle_F \right|
\leq \left\| uu^\top \right\|_F \left\| (\Sigma + \gamma_n I)^{-1} - \Sigma^{-1} \right\|_F \left\| uu^\top \right\|_F
\leq \gamma_n \left\| uu^\top \right\|_F \left\| (\Sigma + \gamma_n I)^{-1} \right\|_F \left\| \Sigma^{-1} \right\|_F
\leq (a) \gamma_n \left\| uu^\top \right\|_2 \left\| (\Sigma + \gamma_n I)^{-1} \right\|_F \left\| \Sigma^{-1} \right\|_F
\leq (a) \gamma_n \left\| uu^\top \right\|_2 \left\| \Sigma^{-1} \right\|_2^2,
\]

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

Combining (5) and (6), we have

\[
\left| \hat{u}^\top \left( \hat{\Sigma} + \gamma_n I \right)^{-1} \hat{u} - u^\top \Sigma^{-1} u \right|
\]
We will separately upper bound \( \| \hat{u} \|_2^2 \Sigma - \hat{\Sigma} \|_F \Sigma^{-1} \|_F + (\| \hat{u} \|_2 + \| u \|_2) \| \hat{u} - u \|_2 \Sigma^{-1} \|_F + \gamma_n \| u \|_2^2 \Sigma^{-1} \|_F^2. \) (7)

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

\[
|h_t((x, y), (x', y'))| = \left| \frac{1}{2} (k(x, v) - k(x', v))(l(y, w) - l(y', w)) \right| \\
\leq \frac{1}{2} \left( |k(x, v)| + |k(x', v)| \left| l(y, w) + l(y', w) \right| \right) \\
\leq 2B_kB_l := 2B,
\]

where we define \( B := B_kB_l \). It follows that

\[
\| \hat{u} \|_2^2 = \sum_{m=1}^J \left( \frac{2}{n(n-1)} \sum_{i < j} h_{t_m}(z_i, z_j) \right)^2 \leq \sum_{m=1}^J \| 2B_kB_l \|^2 = 4B^2J,
\]

\[
\| u \|_2^2 = \sum_{m=1}^J \| E_n E_n h_{t_m}(z, z') \|^2 \leq 4B^2J.
\]

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

\[
\left| \hat{u}^\top \left( \Sigma + \gamma_nI \right)^{-1} \hat{u} - u^\top \Sigma^{-1} u \right| \\
\leq \sqrt{J} 4B^2J\bar{c} \| \Sigma - \hat{\Sigma} \|_F + 4B\sqrt{J}\bar{c} \| \hat{u} - u \|_2 + 4B^2J\bar{c}^2\gamma_n \\
=: \frac{c_1}{\gamma_n} \| \Sigma - \hat{\Sigma} \|_F + c_2\| \hat{u} - u \|_2 + c_3\gamma_n,
\]

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

\[
|\hat{\lambda}_n - \lambda_n| \leq \frac{c_1}{\gamma_n} n \| \Sigma - \hat{\Sigma} \|_F + c_2n\| \hat{u} - u \|_2 + c_3n\gamma_n.
\]

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

E.2.1. Bounding \( \| \hat{u} - u \|_2 \)

Let \( t^* = \arg \max_{t \in \{t_1, \ldots, t_J\}} |\hat{u}(t) - u(t)| \). Recall that \( u = (u(t_1), \ldots, u(t_J))^\top = (u_1, \ldots, u_J)^\top \).

\[
\| \hat{u} - u \|_2 = \sup_{b \in B_2(1)} \| b, \hat{u} - u \|_2 \leq \sup_{b \in B_2(1)} \sum_{j=1}^J |b_j| |\hat{u}(t_j) - u(t_j)| \\
\leq |\hat{u}(t^*) - u(t^*)| \sup_{b \in B_2(1)} \sum_{j=1}^J |b_j| \\
\leq \sqrt{J}|\hat{u}(t^*) - u(t^*)| \sup_{b \in B_2(1)} \| b \|_2 \\
= \sqrt{J}|\hat{u}(t^*) - u(t^*)|,
\]

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

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

(14)
E.2.2. **Bounding $\|\hat{\Sigma} - \Sigma\|_F$**

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

Recall that $\Sigma_{ij} = \eta(t_i, t_j)$, $\eta(t, t') = \mathbb{E}_{xy} (\langle \tilde{k}(x, v) \tilde{l}(y, w) - u(v, w) \rangle \langle \hat{k}(x, v') \hat{l}(y, w') - u(v', w') \rangle)$ where $\tilde{k}(x, v) = k(x, v) - \mathbb{E}_x k(x', v)$, and $\hat{l}(y, w) = l(y, w) - \mathbb{E}_y l(y', w)$. Its empirical estimator (see Proposition 5) is $\hat{\Sigma}_{ij} = \hat{\eta}(t_i, t_j)$ where

$$\hat{\eta}(t, t') = \frac{1}{n} \sum_{i=1}^n (\langle \tilde{k}(x_i, v) \tilde{l}(y_i, w) - \hat{u}^b(v, w) \rangle \langle \hat{k}(x_i, v') \hat{l}(y_i, w') - \hat{u}^b(v', w') \rangle) = \frac{1}{n} \sum_{i=1}^n \hat{k}(x_i, v) \hat{l}(y_i, w) \tilde{k}(x_i, v') \tilde{l}(y_i, w') - \hat{u}^b(v, w) \hat{u}^b(v', w'),$$

$\hat{k}(x, v) := k(x, v) - \frac{1}{n} \sum_{i=1}^n k(x_i, v)$, and $\tilde{l}(y, w) := l(y, w) - \frac{1}{n} \sum_{i=1}^n l(y_i, w)$. We note that $\frac{1}{n} \sum_{i=1}^n \hat{k}(x_i, v) \tilde{l}(y_i, w) = \hat{u}^b(v, w)$. We define $S \in \mathbb{R}^{d \times J}$ such that $\hat{S}_{ij} := \frac{1}{n} \sum_{m=1}^n \hat{k}(x_m, v_i) \tilde{l}(y_m, w_j) \tilde{k}(x_m, v_j) \tilde{l}(y_j, w_j)$, and define similarly its population counterpart $S$ such that $S_{ij} := \mathbb{E}_{xy} \langle \tilde{k}(x, v) \tilde{l}(y, w) k(x, v') \tilde{l}(y, w') \rangle$. We have

$$\hat{\Sigma} = \hat{S} - \hat{u}^b \hat{u}^T,$$

$$\Sigma = S - uu^T,$$

$$\|\hat{\Sigma} - \Sigma\|_F = \|\hat{S} - S - (\hat{u}^b \hat{u}^T - uu^T)\|_F \leq \|\hat{S} - S\|_F + \|\hat{u}^b \hat{u}^T - uu^T\|_F. \quad (15)$$

(16)

With (16), (14) becomes

$$|\hat{\lambda}_n - \lambda_n| \leq \frac{c_1 n}{\gamma_n} \|\hat{S} - S\|_F + \frac{c_1 n}{\gamma_n} \|\hat{u}^b \hat{u}^T - uu^T\|_F + c_2 n \sqrt{J} |\hat{u}(t^*) - u(t^*)| + c_3 n \gamma_n. \quad (17)$$

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

E.2.3. **Bounding $\|\hat{u}^b \hat{u}^T - uu^T\|_F$**

$$\|\hat{u}^b \hat{u}^T - uu^T\|_F = \|\hat{u}^b \hat{u}^T - \hat{u}^b u^T + \hat{u}^b u^T - uu^T\|_F$$

$$\leq \|\hat{u}^b (\hat{u} - u)^T\|_F + \|u^T\|_F + \|u^T\|_F + \|u^T\|_F$$

$$= \|\hat{u}^b\|_2 \|\hat{u} - u\|_2 + \|u\|_2 \|u\|_2$$

$$\leq 4B \sqrt{J} |\hat{u} - u|_2,$$

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

$$\|\hat{u}^b\|_2^2 = \sum_{m=1}^J \left[ \|\hat{\mu}_{x^b}(v_m, w_m) - \hat{\mu}_x(u_m) \|_{\mathcal{M}} \right]^2 = \sum_{m=1}^J \left[ \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n h_{t_m}(z_i, z_j) \right] \leq \sum_{m=1}^J \|2B_k B_l\|^2 = 4B^2 J.$$

Let $\langle \tilde{v}, \tilde{w} \rangle := \tilde{t} = \arg \max t \in \{t_1, \ldots, t_J\} |\hat{u}^b(t) - u(t)|$. We bound $\|\hat{u}^b - u\|_2$ by

$$\|\hat{u}^b - u\|_2 \leq \|\hat{u}^b - u|_{\mathcal{M}} \leq \sqrt{J} |\hat{\mu}_{x^b} - \hat{\mu}_x|$$

$$= \sqrt{J} |\hat{\mu}_{x^b} - \hat{\mu}_x(v) \hat{\mu}_y - u(t)|$$

$$= \sqrt{J} |\hat{\mu}_{x^b} - \hat{\mu}_x(v) | + \hat{\mu}_y(v) \hat{\mu}_y(v) - u(t)|$$

$$\leq \sqrt{J} |\hat{\mu}_x(v) - \hat{\mu}_x(t) - u(t)| + \sqrt{J} |\hat{\mu}_x(t) - \hat{\mu}_x(v) \hat{\mu}_y(v) - u(t)|$$
\[
\begin{align*}
&= \sqrt{J} \left| \hat{u}(\hat{t}) - u(\hat{t}) \right| + \sqrt{J} \left| \hat{\mu}_x \hat{\mu}_y (\hat{t}) - \hat{\mu}_x (\hat{\nu}) \hat{\mu}_y (\hat{w}) \right|, \\
\text{where at (a) we used the same reasoning as in (13). The bias} \left| \hat{\mu}_x \hat{\mu}_y (\hat{t}) - \hat{\mu}_x (\hat{\nu}) \hat{\mu}_y (\hat{w}) \right| \text{in the second term can be bounded as} \\
&\left| \hat{\mu}_x \hat{\mu}_y (\hat{t}) - \hat{\mu}_x (\hat{\nu}) \hat{\mu}_y (\hat{w}) \right| \\
&= \left| \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} k(x_i, \hat{\nu})l(y_j, \hat{w}) - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, \hat{\nu})l(y_j, \hat{w}) \right| \\
&= \left| \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, \hat{\nu})l(y_j, \hat{w}) - \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, \hat{\nu})l(y_i, \hat{w}) \right| \\
&= \left( 1 - \frac{n}{n-1} \right) \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, \hat{\nu})l(y_j, \hat{w}) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, \hat{\nu})l(y_i, \hat{w}) \\
&\leq \left( 1 - \frac{n}{n-1} \right) \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, \hat{\nu})l(y_j, \hat{w}) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, \hat{\nu})l(y_i, \hat{w}) \\
&\leq \frac{B}{n-1} + \frac{B}{n-1} = \frac{2B}{n-1}. 
\end{align*}
\]
Having an upper bound for $|\hat{S}(t, t') - S(t, t')|$ will allow us to bound (22). To keep the notations uncluttered, we will define the following shorthands.

<table>
<thead>
<tr>
<th>Expression</th>
<th>Shorthand</th>
<th>Expression</th>
<th>Shorthand</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k(x, v)$</td>
<td>$a$</td>
<td>$l(y, w)$</td>
<td>$b$</td>
</tr>
<tr>
<td>$k(x', v')$</td>
<td>$a'$</td>
<td>$l(y, w')$</td>
<td>$b'$</td>
</tr>
<tr>
<td>$k(x_i, v)$</td>
<td>$a_i$</td>
<td>$l(y_i, w)$</td>
<td>$b_i$</td>
</tr>
<tr>
<td>$k(x, v')$</td>
<td>$a_1$</td>
<td>$l(y_i, w')$</td>
<td>$b'_i$</td>
</tr>
<tr>
<td>$\mathbb{E}_{x \sim p} k(x, v)$</td>
<td>$\bar{a}$</td>
<td>$\mathbb{E}_{y \sim p} l(y, w)$</td>
<td>$\bar{b}$</td>
</tr>
<tr>
<td>$\mathbb{E}_{x \sim p} k(x, v')$</td>
<td>$\bar{a}'$</td>
<td>$\mathbb{E}_{y \sim p} l(y, w')$</td>
<td>$\bar{b}'$</td>
</tr>
<tr>
<td>$\frac{1}{n} \sum_{i=1}^{n} k(x_i, v)$</td>
<td>$\bar{a}$</td>
<td>$\frac{1}{n} \sum_{i=1}^{n} l(y_i, w)$</td>
<td>$\bar{b}$</td>
</tr>
<tr>
<td>$\frac{1}{n} \sum_{i=1}^{n} k(x_i, v')$</td>
<td>$\bar{a}'$</td>
<td>$\frac{1}{n} \sum_{i=1}^{n} l(y_i, w')$</td>
<td>$\bar{b}'$</td>
</tr>
</tbody>
</table>

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

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

$$\hat{S}(t, t') = \frac{1}{n} \sum_{i=1}^{n} (a_i - \bar{a})(b_i - \bar{b})(a'_i - \bar{a'})(b'_i - \bar{b}')$$

$$S(t, t') = \mathbb{E}_{xy} \left[ (a - \bar{a})(b - \bar{b})(a' - \bar{a'})(b' - \bar{b}') \right]$$

By expanding $S(t, t')$, we have

$$S(t, t') = \mathbb{E}_{xy} \left[ + a \bar{a} \bar{b}' - a \bar{a} \bar{b}' - a \bar{a} \bar{b}' + a \bar{a} \bar{b}' \
- a \bar{a} \bar{b}' + a \bar{a} \bar{b}' + a \bar{a} \bar{b}' - a \bar{a} \bar{b}' \
- \bar{a} \bar{a} \bar{b}' + \bar{a} \bar{a} \bar{b}' + \bar{a} \bar{a} \bar{b}' - \bar{a} \bar{a} \bar{b}' \
+ a \bar{a} \bar{b}' - a \bar{a} \bar{b}' - a \bar{a} \bar{b}' + a \bar{a} \bar{b}' ] $$

$$= +a \bar{a} \bar{b}' - a \bar{a} \bar{b}' - a \bar{a} \bar{b}' + a \bar{a} \bar{b}' \
- a \bar{a} \bar{b}' + a \bar{a} \bar{b}' + a \bar{a} \bar{b}' - a \bar{a} \bar{b}' \
- a \bar{a} \bar{b}' + a \bar{a} \bar{b}' + a \bar{a} \bar{b}' - a \bar{a} \bar{b}' \
+ a \bar{a} \bar{b}' - a \bar{a} \bar{b}' - a \bar{a} \bar{b}' + a \bar{a} \bar{b}' $$

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

$$|\hat{S}(t, t') - S(t, t')| \leq |a \bar{a} \bar{b}' - a \bar{a} \bar{b}'| + |a \bar{a} \bar{b}' - a \bar{a} \bar{b}'| + |a \bar{a} \bar{b}' - a \bar{a} \bar{b}'| + |a \bar{a} \bar{b}' - a \bar{a} \bar{b}'|$$

$$= |a \bar{a} \bar{b}' - a \bar{a} \bar{b}'| + |a \bar{a} \bar{b}' - a \bar{a} \bar{b}'| + |a \bar{a} \bar{b}' - a \bar{a} \bar{b}'| + |a \bar{a} \bar{b}' - a \bar{a} \bar{b}'|$$
\[ |\tilde{a}b\tilde{a} - \tilde{a}b\tilde{a}| + |\tilde{a}\tilde{b}\tilde{a} - \tilde{a}\tilde{b}\tilde{a}| + |\tilde{a}b\tilde{a} - \tilde{a}b\tilde{a}| + 3 |\tilde{a}b\tilde{a} - \tilde{a}b\tilde{a}|. \]

The first term \( |\tilde{a}b\tilde{a} - \tilde{a}b\tilde{a}| \) 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 \( |\tilde{a}b\tilde{a} - \tilde{a}b\tilde{a}| \) (1st term).** Since \(-B^2 \leq ab\tilde{a} \leq B^2\), by the Hoeffding’s inequality (Lemma 13), we have

\[ P \left( |\tilde{a}b\tilde{a} - \tilde{a}b\tilde{a}| \leq t \right) \geq 1 - 2 \exp \left( -\frac{nt^2}{2B^2} \right). \]

**Bounding \( |\tilde{a}b\tilde{a} - \tilde{a}b\tilde{a}| \) (2nd term).** Let \(f_1(x, y) = ab\tilde{a} = k(x, v)l(y, w)k(x, v')\) and \(f_2(y) = b' = l(y, w')\). We note that \(|f_1(x, y)| \leq (BB_k, B_l)_+ \) and \(|f_2(y)| \leq (BB_k, B_l)_+\). Thus, by Lemma 8 with \(E = 2\), we have

\[ P \left( |\tilde{a}b\tilde{a} - \tilde{a}b\tilde{a}| \leq t \right) \geq 1 - 6 \exp \left( -\frac{nt^2}{18B, B_l} \right). \]

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

\[ P \left( |\tilde{a}b\tilde{a} - \tilde{a}b\tilde{a}| \leq t \right) \geq 1 - 8 \exp \left( -\frac{nt^2}{32 \cdot 3^2(B, B_k)_+} \right). \]

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

- **(3rd term)** \( P \left( |\tilde{a}b\tilde{a} - \tilde{a}b\tilde{a}| \leq t \right) \geq 1 - 4 \exp \left( -\frac{nt^2}{8(B, B_k)_+} \right), \)
- **(5th term)** \( P \left( |\tilde{a}b\tilde{a} - \tilde{a}b\tilde{a}| \leq t \right) \geq 1 - 4 \exp \left( -\frac{nt^2}{8(B, B_k)_+} \right), \)
- **(6th term)** \( P \left( |\tilde{a}b\tilde{a} - \tilde{a}b\tilde{a}| \leq t \right) \geq 1 - 6 \exp \left( -\frac{nt^2}{18(B, B_k)_+} \right), \)
- **(7th term)** \( P \left( |\tilde{a}b\tilde{a} - \tilde{a}b\tilde{a}| \leq t \right) \geq 1 - 6 \exp \left( -\frac{nt^2}{18(B, B_k)_+} \right), \)
- **(8th term)** \( P \left( |\tilde{a}b\tilde{a} - \tilde{a}b\tilde{a}| \leq t \right) \geq 1 - 6 \exp \left( -\frac{nt^2}{18(B, B_k)_+} \right), \)
- **(9th term)** \( P \left( |\tilde{a}b\tilde{a} - \tilde{a}b\tilde{a}| \leq t \right) \geq 1 - 4 \exp \left( -\frac{nt^2}{8(B, B_k)_+} \right), \)
- **(10th term)** \( P \left( |\tilde{a}b\tilde{a} - \tilde{a}b\tilde{a}| \leq t \right) \geq 1 - 6 \exp \left( -\frac{nt^2}{18(B, B_k)_+} \right), \)
- **(11th term)** \( P \left( |\tilde{a}b\tilde{a} - \tilde{a}b\tilde{a}| \leq t \right) \geq 1 - 6 \exp \left( -\frac{nt^2}{18(B, B_k)_+} \right). \)
By the union bound, we have
\[ P\left(\left|\hat{S}(t, t') - S(t, t')\right| \leq 2t\right) \]
\[ \geq 1 - \left[ 2 \exp\left( -\frac{nt^2}{2B^2} \right) + 4 \exp\left( -\frac{nt^2}{8(BB_k, B_k)_{t}^+} \right) + 4 \exp\left( -\frac{nt^2}{18(B, B_k)_{t}^+} \right) + 6 \exp\left( -\frac{nt^2}{18(B, B_k)_{t}^+} \right) \right] \nonumber \]
\[ \geq 1 - \left[ 2 \exp\left( -\frac{nt^2}{2B^2} \right) + 8 \exp\left( -\frac{nt^2}{8(BB_k, B_k)_{t}^+} \right) + 8 \exp\left( -\frac{nt^2}{8(BB_k, B_k)_{t}^+} \right) + 24 \exp\left( -\frac{nt^2}{18(B, B_k)_{t}^+} \right) + 6 \exp\left( -\frac{nt^2}{18(B, B_k)_{t}^+} \right) + 8 \exp\left( -\frac{nt^2}{18(B, B_k)_{t}^+} \right) \right] \nonumber \]
\[ \geq 1 - \left[ 2 \exp\left( -\frac{12^2nt^2}{B^2} \right) + 8 \exp\left( -\frac{12^2nt^2}{B^2} \right) + 8 \exp\left( -\frac{12^2nt^2}{B^2} \right) + 24 \exp\left( -\frac{12^2nt^2}{B^2} \right) + 6 \exp\left( -\frac{12^2nt^2}{B^2} \right) + 8 \exp\left( -\frac{12^2nt^2}{B^2} \right) \right] \]
\[ = 1 - 62 \exp\left( -\frac{12^2nt^2}{B^2} \right), \]
where
\[ B^* := \frac{1}{12^2} \max(2B^4, 8(BB_k, B_k)_{t}^+, 8(BB_k, B_k)_{t}^+, 18(B, B_k)_{t}^+, 18(B, B_k)_{t}^+, 18(B, B_k)_{t}^+, 18(B, B_k)_{t}^+, 18(B, B_k)_{t}^+, 32 \cdot 3^2(B, B_k)_{t}^9). \]

By reparameterization, it follows that
\[ P\left(\left|\frac{c_1Jn}{\gamma_n} \hat{S}(t, t') - S(t, t')\right| \leq t\right) \geq 1 - 62 \exp\left( -\frac{\gamma_n^2t^2}{c_1^2J^2nB^*} \right). \] (23)

E.2.6. UNION BOUND FOR \( |\hat{\lambda}_n - \lambda_n| \) AND FINAL LOWER BOUND

Recall from (22) that
\[ |\hat{\lambda}_n - \lambda_n| \leq \frac{c_1Jn}{\gamma_n} \left| \hat{S}(t^{(1)}, t^{(2)}) - S(t^{(1)}, t^{(2)}) \right| + \frac{4BJc_1n}{\gamma_n} |\hat{u}(t) - u(t)| \]
\[ + \frac{c_1n 8B^2J}{\gamma_n n - 1} + c_2n \sqrt{J} |\hat{u}(t^*) - u(t^*)| + c_3n\gamma_n. \]

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
\[ P\left(\left|\frac{c_2n \sqrt{J} |\hat{u}(t^*) - u(t^*)| \leq t\right| \leq t\right) \geq 1 - 2 \exp\left( -\frac{|0.5n|t^2}{8c_2^2n^2JB^2} \right). \] (24)

Bounding \( \frac{c_1n 8B^2J}{\gamma_n n - 1} + c_3n\gamma_n + \frac{4BJc_1n}{\gamma_n} |\hat{u}(t) - u(t)|, \) By Lemma 12 (with \( m = 2 \), it follows that
\[ P\left(\left|\frac{c_1n 8B^2J}{\gamma_n n - 1} + c_3n\gamma_n + \frac{4BJc_1n}{\gamma_n} |\hat{u}(t) - u(t)| \leq t\right| \leq t\right) \geq 1 - 2 \exp\left( -\frac{|0.5n|\gamma_n^2 (t - \frac{c_1n 8B^2J}{\gamma_n n - 1} - c_3n\gamma_n)^2}{2^2B^4J^2c_1^4n^2} \right) \]
where at (a) we used $\lceil 0.5n \rceil \geq (n-1)/2$. Combining (23), (24), and (25) with the union bound (set $T = 3t$), we can bound (22) with

$$
\mathbb{P}\left( |\hat{\lambda}_n - \lambda_n| \leq T \right) \geq 1 - 2\exp\left( -\frac{\gamma^2 T^2}{3^2 c_1^2 J^2 n B^*} \right) - 2\exp\left( -\frac{[0.5n]T^2}{72c_2^2 n^2 J B^2} \right) - 2\exp\left( -\frac{[\gamma(n-1)/3 - 8c_1B^2nJ - c_3\gamma^2 n(n-1)]^2}{2^8 B^2 J^2 c_1^2 n^2 (n-1)} \right).
$$

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

$$
\mathbb{P}\left( \hat{\lambda}_n \geq r \right) \geq 1 - 2\exp\left( -\frac{\gamma^2 (\lambda_n - r)^2}{3^2 c_1^2 J^2 n B^*} \right) - 2\exp\left( -\frac{[0.5n](\lambda_n - r)^2}{72c_2^2 n^2 J B^2} \right) - 2\exp\left( -\frac{[(\lambda_n - r)\gamma(n-1)/3 - 8c_1B^2nJ - c_3\gamma^2 n(n-1)]^2}{2^8 B^2 J^2 c_1^2 n^2 (n-1)} \right) := L(\lambda_n).
$$

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

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

$$
1 - 2\exp\left( -C_1(\lambda_n - T_\alpha)^2 \right) - 2\exp\left( -C_2(\lambda_n - T_\alpha)^2 \right) - 2\exp\left( -\frac{[(\lambda_n - T_\alpha)C_3 - C_4]^2}{C^5} \right),
$$

where $C_1, \ldots, C_5$ are positive constants. For fixed large enough $n$ such that $\lambda_n > T_\alpha$, and fixed significance level $\alpha$, increasing $\lambda_n$ will increase $L(\lambda_n)$. Specifically, since $n$ is fixed, increasing $u^\top \Sigma^{-1} u$ in $\lambda_n = nu^\top \Sigma^{-1} u$ 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 $|\prod_{i=1}^E a_i - \prod_{i=1}^E b_i| \leq B^{E-1} \sum_{j=1}^E |a_j - b_j|$.

**Proof.**

$$
\left| \prod_{i=1}^E a_i - \prod_{j=1}^E b_j \right| \leq \left| \prod_{i=1}^E a_i - \prod_{i=1}^{E-1} a_i b_E \right| + \left| \prod_{i=1}^{E-1} a_i b_E - \prod_{i=1}^{E-2} a_i b_{E-1} b_E \right| + \ldots + \left| a_1 \prod_{j=2}^E b_j - \prod_{j=1}^E b_j \right|
\leq |a_E - b_E| \left| \prod_{i=1}^{E-1} a_i \right| + |a_{E-1} - b_{E-1}| \left| \prod_{i=1}^{E-2} a_i \right| b_E + \ldots + |a_1 - b_1| \left| \prod_{j=2}^E b_j \right|
\leq |a_E - b_E| B^{E-1} + |a_{E-1} - b_{E-1}| B^{E-1} + \ldots + |a_1 - b_1| B^{E-1}
= B^{E-1} \sum_{j=1}^E |a_j - b_j|
$$

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

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

Proof. By Lemma 7, we have

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

By applying the Hoeffding’s inequality to each term in the sum, we have

\[
P \left( \left| \mathbb{E}_{X_i^{(i)} \sim \hat{P}_i} f_i(x^{(i)}) - \mathbb{E}_{X_i^{(i)} \sim P_i} f_i(x^{(i)}) \right| \leq t \right) \geq 1 - 2 \exp \left( - \frac{2n_i t^2}{4B^2} \right)
\]

The result is obtained with a union bound.

\[\square\]

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 \( \Lambda \) be an injective mapping from the space of probability measures into a space of analytic functions on \( \mathbb{R}^d \). Define

\[
d_{\nu_{\Lambda}}(P, Q) = \sum_{j=1}^{J} \| [\Lambda P](v_j) - [\Lambda Q](v_j) \|^2,
\]

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

Lemma 11 (Bochner’s theorem (Rudin, 2011)). A continuous function \( \Psi : \mathbb{R}^d \rightarrow \mathbb{R} \) is positive definite if and only if it is the Fourier transform of a finite nonnegative Borel measure \( \zeta \) on \( \mathbb{R}^d \), that is, \( \Psi(x) = \int_{\mathbb{R}^d} e^{-i \langle x, \omega \rangle} d\zeta(\omega), x \in \mathbb{R}^d \).

Lemma 12 (A bound for U-statistics (Serfling, 2009, Theorem A, p. 201)). Let \( h(x_1, \ldots, x_m) \) be a U-statistic kernel for an \( m \)-order U-statistic such that \( h(\{x_i\}, \ldots, \{x_m\}) \in [a, b] \) where \( a \leq b < \infty \). Let \( U_n = \binom{n}{m}^{-1} \sum_{i_1 < \cdots < i_m} h(x_{i_1}, \ldots, x_{i_m}) \) 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 \( n \geq m \),

\[
P(U_n - \mathbb{E}h(\{x_i\}, \ldots, \{x_m\}) \geq t) \leq \exp \left( -2|n/m| t^2 / (b-a)^2 \right),
\]

\[
P(|U_n - \mathbb{E}h(\{x_i\}, \ldots, \{x_m\})| \geq t) \leq 2 \exp \left( -2|n/m| t^2 / (b-a)^2 \right),
\]

where \( \lfloor x \rfloor \) denotes the greatest integer which is smaller than or equal to \( x \). Hoeffding’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 \( a \leq X_i \leq b \) almost surely. Define \( \bar{X} := \frac{1}{n} \sum_{i=1}^{n} X_i \). Then,

\[
P \left( \left| \bar{X} - \mathbb{E}[\bar{X}] \right| \leq \alpha \right) \geq 1 - 2 \exp \left( - \frac{2n\alpha^2}{(b-a)^2} \right).
\]

References

