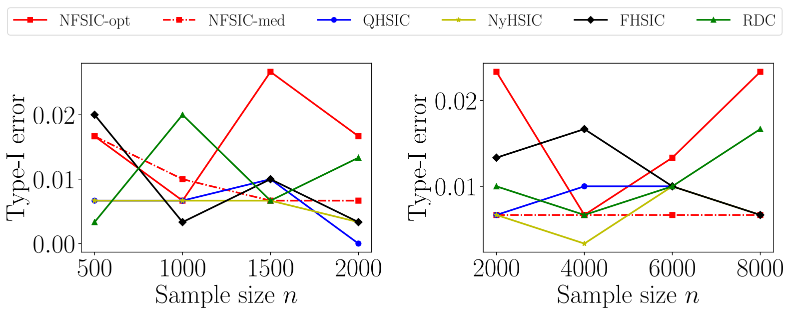
Supplementary Material

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.

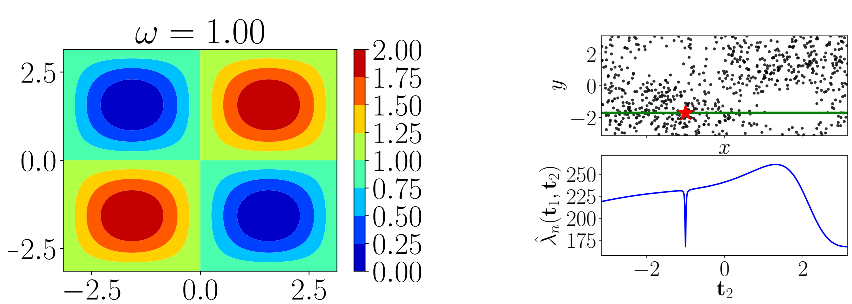


Figure 6: Plot of optimization objective values as location $\mathbf{t}_2$ moves along the green line. The objective sharply drops when the two locations are in the same neighborhood.

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.

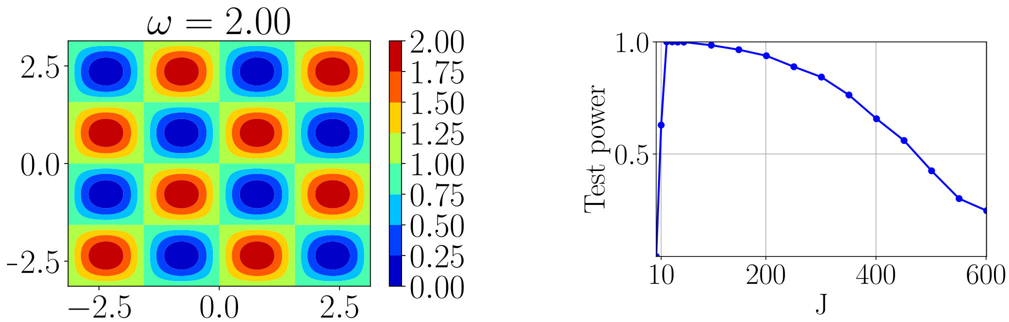


Figure 7: The Sinusoid problem and the plot of test power vs. the number of test locations.

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 Proposition 3

Recall Proposition 3,

Proposition (A product of Gaussian kernels is characteristic and analytic)**.** *Let* $k(\mathbf{x}, \mathbf{x}') = \exp\left(-(\mathbf{x}-\mathbf{x}')^\top \mathbf{A}(\mathbf{x}-\mathbf{x}')\right)$ *and* $l(\mathbf{y}, \mathbf{y}') = \exp\left(-(\mathbf{y}-\mathbf{y}')^\top \mathbf{B}(\mathbf{y}-\mathbf{y}')\right)$ *be Gaussian kernels on* $\mathbb{R}^{d_x}\times\mathbb{R}^{d_x}$ *and* $\mathbb{R}^{d_y}\times\mathbb{R}^{d_y}$ *respectively, for positive definite matrices* $\mathbf{A}$ *and* $\mathbf{B}$*. Then,* $g((\mathbf{x},\mathbf{y}),(\mathbf{x}',\mathbf{y}')) = k(\mathbf{x},\mathbf{x}')l(\mathbf{y},\mathbf{y}')$ *is characteristic and analytic on* $(\mathbb{R}^{d_x}\times\mathbb{R}^{d_y})\times(\mathbb{R}^{d_x}\times\mathbb{R}^{d_y})$*.*

Proof. Let $\mathbf{z} := (\mathbf{x}^\top, \mathbf{y}^\top)^\top$ and $\mathbf{z}' := (\mathbf{x}'^\top, \mathbf{y}'^\top)^\top$ be vectors in $\mathbb{R}^{d_x+d_y}$. We prove by reducing the product kernel to one Gaussian kernel with $g(\mathbf{z},\mathbf{z}') = \exp\left(-(\mathbf{z}-\mathbf{z}')^\top \mathbf{C}(\mathbf{z}-\mathbf{z}')\right)$ where $\mathbf{C} := \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix}$. Write $g(\mathbf{z},\mathbf{z}') = \Psi(\mathbf{z}-\mathbf{z}')$ where $\Psi(\mathbf{t}) := \exp\left(-\mathbf{t}^\top\mathbf{C}\mathbf{t}\right)$. Since $\mathbf{C}$ is positive definite, we see that the finite measure ζ corresponding to Ψ as defined in Lemma 12 has support everywhere in $\mathbb{R}^{d_x+d_y}$. Thus, Sriperumbudur et al. (2010, Theorem 9) implies that g is characteristic.

To see that g is analytic, we observe that for each $\mathbf{z}' \in \mathbb{R}^{d_x+d_y}$, $\mathbf{z} \mapsto -(\mathbf{z}-\mathbf{z}')^\top\mathbf{C}(\mathbf{z}-\mathbf{z}')$ is a multivariate polynomial in $\mathbf{z}$, which is known to be analytic. Using the fact that $t \mapsto \exp(t)$ is analytic on $\mathbb{R}$, and that a composition of analytic functions is analytic, we see that $\mathbf{z} \mapsto \exp\left(-(\mathbf{z}-\mathbf{z}')^\top\mathbf{C}(\mathbf{z}-\mathbf{z}')\right)$ is analytic on $\mathbb{R}^{d_x+d_y}$ for each $\mathbf{z}'$. □

E. Proof of Theorem 5

Recall Theorem 5,

Theorem 5 (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 4. 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, $\hat{\lambda}_n \stackrel{d}{\to} \chi^2(J)$ as $n \to \infty$.*
2. *Under H_1, for any $r \in \mathbb{R}$, $\lim_{n\to\infty} \mathbb{P}\left(\hat{\lambda}_n \geq r\right) = 1$ 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 $\left(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1} \stackrel{p}{\to} \mathbf{\Sigma}^{-1}$ 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 $\left[\sqrt{n}\hat{\mathbf{u}}, \left(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1}\right] \stackrel{d}{\to} \left[\mathbf{a}, \mathbf{\Sigma}^{-1}\right]$ where $\mathbf{u} = 0$ almost surely by Proposition 2, and $\sqrt{n}\hat{\mathbf{u}} \stackrel{d}{\to} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ by Proposition 4. Since $f(\mathbf{x}, \mathbf{S}) := \mathbf{x}^\top \mathbf{S}\mathbf{x}$ is continuous, $f\left(\sqrt{n}\hat{\mathbf{u}}, \left(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1}\right) \stackrel{d}{\to} f(\mathbf{a}, \mathbf{\Sigma}^{-1})$. Equivalently, $n\hat{\mathbf{u}}^\top \left(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}} \stackrel{d}{\to} \mathbf{a}^\top \mathbf{\Sigma}^{-1}\mathbf{a} \sim \chi^2(J)$ 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. □

F. Proof of Theorem 7

Recall Theorem 7,

Theorem 7 (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, $\xi_2 := 72c_2^2 JB^2$, $B := B_k B_l$, $\xi_3 := 8c_1 B^2 J$, $c_3 := 4B^2 J\tilde{c}^2$, $\xi_4 := 2^8 B^4 J^2 c_1^2$, $c_1 := 4B^2 J\sqrt{J}\tilde{c}$, 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 $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F$. 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 $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F$, we decompose it into a sum of smaller components, and bound each term with a product variant of the Hoeffding's inequality (Lemma 9). $L(\lambda_n)$ is obtained by combining all the bounds with the union bound.

F.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 $B_2(r) := \{\mathbf{x} \mid \|\mathbf{x}\|_2 \le r\}$ be a closed ball with radius r centered at the origin. Similarly, define $B_F(r) := \{\mathbf{A} \mid \|\mathbf{A}\|_F \le r\}$ 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*.

F.2. Proof

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

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

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

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

$$\leq \frac{\sqrt{J}}{\gamma_n}\|\hat{\mathbf{u}}\|_2^2\|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}\|_F\|\boldsymbol{\Sigma}^{-1}\|_F + \left(\|\hat{\mathbf{u}}\|_2+\|\mathbf{u}\|_2\right)\|\hat{\mathbf{u}}-\mathbf{u}\|_2\|\boldsymbol{\Sigma}^{-1}\|_F, \tag{5}$$

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

For $(\bigstar)_2$, we have

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

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

Combining (5) and (6), we have

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

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.

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

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

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

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

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

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

F.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 $\mathbf{u} = (u(\mathbf{t}_1),\ldots,u(\mathbf{t}_J))^\top = (u_1,\ldots,u_J)^\top$.

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

where at (a) we used $\|\mathbf{a}\|_1 \le \sqrt{J}\|\mathbf{a}\|_2$ 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

$$
|\hat{\lambda}_n - \lambda_n| \le \frac{c_1}{\gamma_n} n\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_F + c_2 n\sqrt{J}|\hat{u}(\mathbf{t}^*) - u(\mathbf{t}^*)| + c_3 n\gamma_n. \qquad (14)
$$

F.2.2. Bounding $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F$

The plan is to write $\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{S}} - \hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top}, \boldsymbol{\Sigma} = \mathbf{S} - \mathbf{u}\mathbf{u}^\top$, so that $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F \le \|\hat{\mathbf{S}} - \mathbf{S}\|_F + \|\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \mathbf{u}\mathbf{u}^\top\|_F$ 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 $\Sigma_{ij} = \eta(\mathbf{t}_i,\mathbf{t}_j), \eta(\mathbf{t},\mathbf{t}') = \mathbb{E}_{\mathbf{xy}}[(\tilde{k}(\mathbf{x},\mathbf{v})\tilde{l}(\mathbf{y},\mathbf{w}) - u(\mathbf{v},\mathbf{w}))(\tilde{k}(\mathbf{x},\mathbf{v}')\tilde{l}(\mathbf{y},\mathbf{w}') - u(\mathbf{v}',\mathbf{w}'))]$ 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 6) is $\hat{\Sigma}_{ij} = \hat{\eta}(\mathbf{t}_i,\mathbf{t}_j)$ where

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

$\overline{k}(\mathbf{x},\mathbf{v}) := k(\mathbf{x},\mathbf{v}) - \frac{1}{n}\sum_{i=1}^n k(\mathbf{x}_i,\mathbf{v})$, 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

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

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

F.2.3. Bounding $\|\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \mathbf{u}\mathbf{u}^\top\|_F$

$$
\|\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \mathbf{u}\mathbf{u}^\top\|_F = \|\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \hat{\mathbf{u}}^b\mathbf{u}^\top + \hat{\mathbf{u}}^b\mathbf{u}^\top - \mathbf{u}\mathbf{u}^\top\|_F
$$

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

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

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

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

$$
\begin{aligned}
\|\hat{\mathbf{u}}^b-\mathbf{u}\|_2 &\overset{(a)}{\leq} \sqrt{J}|\hat{u}^b(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})| \\
&= \sqrt{J}\left|\hat{\mu}_{xy}(\tilde{\mathbf{t}}) - \hat{\mu}_x(\tilde{\mathbf{v}})\hat{\mu}_y(\tilde{\mathbf{w}}) - u(\tilde{\mathbf{t}})\right| \\
&= \sqrt{J}\left|\hat{\mu}_{xy}(\tilde{\mathbf{t}}) - \widehat{\mu_x\mu_y}(\tilde{\mathbf{t}}) + \widehat{\mu_x\mu_y}(\tilde{\mathbf{t}}) - \hat{\mu}_x(\tilde{\mathbf{v}})\hat{\mu}_y(\tilde{\mathbf{w}}) - u(\tilde{\mathbf{t}})\right| \\
&\leq \sqrt{J}\left|\hat{\mu}_{xy}(\tilde{\mathbf{t}}) - \widehat{\mu_x\mu_y}(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})\right| + \sqrt{J}\left|\widehat{\mu_x\mu_y}(\tilde{\mathbf{t}}) - \hat{\mu}_x(\tilde{\mathbf{v}})\hat{\mu}_y(\tilde{\mathbf{w}})\right| \\
&= \sqrt{J}\left|\hat{u}(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})\right| + \sqrt{J}\left|\widehat{\mu_x\mu_y}(\tilde{\mathbf{t}}) - \hat{\mu}_x(\tilde{\mathbf{v}})\hat{\mu}_y(\tilde{\mathbf{w}})\right|,
\end{aligned}
\tag{18}
$$

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

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

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

$$
|\hat{\lambda}_n - \lambda_n| \leq \frac{c_1 n}{\gamma_n}\|\hat{\mathbf{S}} - \mathbf{S}\|_F + \frac{4BJc_1n}{\gamma_n}\left|\hat{u}(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})\right| + \frac{c_1 n}{\gamma_n}\frac{8B^2J}{n-1} + c_2 n\sqrt{J}|\hat{u}(\mathbf{t}^*) - u(\mathbf{t}^*)| + c_3 n\gamma_n. \tag{20}
$$

F.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 $(\mathbf{t}^{(1)},\mathbf{t}^{(2)}) = \arg\max_{(\mathbf{s},\mathbf{t})\in V_J\times V_J}|\hat{S}(\mathbf{s},\mathbf{t}) - S(\mathbf{s},\mathbf{t})|$.

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

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

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

F.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 $\widetilde{\cdot}$ 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

$$S(\mathbf{t},\mathbf{t}') = \mathbb{E}_{\mathbf{xy}}\big[+aba'b' - aba'\tilde{b}' - ab\tilde{a}'b' + ab\tilde{a}'\tilde{b}'$$

$$
\begin{aligned}
&- a\tilde{b}a'b' + a\tilde{b}a'\tilde{b}' + a\tilde{b}\tilde{a}'b' - a\tilde{b}\tilde{a}'\tilde{b}' \\
&- \tilde{a}ba'b' + \tilde{a}ba'\tilde{b}' + \tilde{a}b\tilde{a}'b' - \tilde{a}b\tilde{a}'\tilde{b}' \\
&+ \tilde{a}\tilde{b}a'b' - \tilde{a}\tilde{b}a'\tilde{b}' - \tilde{a}\tilde{b}\tilde{a}'\tilde{b}' + \tilde{a}\tilde{b}\tilde{a}'\tilde{b}'] \\
= &+\widetilde{aba'b'} - \widetilde{aba'}\tilde{b}' - \widetilde{abb'}\tilde{a}' + \widetilde{ab}\tilde{a}'\tilde{b}' \\
&- \widetilde{aa'b'}\tilde{b} + \widetilde{aa'}\tilde{b}\tilde{b}' + \widetilde{ab'}\tilde{a}'\tilde{b} - \tilde{a}\tilde{b}\tilde{a}'\tilde{b}' \\
&- \widetilde{a'bb'}\tilde{a} + \widetilde{a'b}\tilde{a}\tilde{b}' + \tilde{a}\tilde{a}'\widetilde{bb'} - \tilde{a}\tilde{b}\tilde{a}'\tilde{b}' \\
&+ \widetilde{a'b'}\tilde{a}\tilde{b} - \tilde{a}\tilde{b}\tilde{a}'\tilde{b}' - \tilde{a}\tilde{b}\tilde{a}'\tilde{b}' + \tilde{a}\tilde{b}\tilde{a}'\tilde{b}' \\
= &+\widetilde{aba'b'} - \widetilde{aba'}\tilde{b}' - \widetilde{abb'}\tilde{a}' + \widetilde{ab}\tilde{a}'\tilde{b}' \\
&- \widetilde{aa'b'}\tilde{b} + \widetilde{aa'}\tilde{b}\tilde{b}' + \widetilde{ab'}\tilde{a}'\tilde{b} + \widetilde{a'b'}\tilde{a}\tilde{b} \\
&- \widetilde{a'bb'}\tilde{a} + \widetilde{a'b}\tilde{a}\tilde{b}' + \tilde{a}\tilde{a}'\widetilde{bb'} - 3\tilde{a}\tilde{b}\tilde{a}'\tilde{b}'.
\end{aligned}
$$

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

$$
\begin{aligned}
\left|\hat{S}(\mathbf{t},\mathbf{t}') - S(\mathbf{t},\mathbf{t}')\right| \le &\left|\overline{aba'b'} - \widetilde{aba'b'}\right| + \left|\overline{aba'}\,\overline{b}' - \widetilde{aba'}\tilde{b}'\right| + \left|\overline{abb'}\overline{a}' - \widetilde{abb'}\tilde{a}'\right| + \left|\overline{ab}\overline{a}'\overline{b}' - \widetilde{ab}\tilde{a}'\tilde{b}'\right| \\
&\left|\overline{aa'b'}\,\overline{b} - \widetilde{aa'b'}\tilde{b}\right| + \left|\overline{aa'}\,\overline{b}\,\overline{b}' - \widetilde{aa'}\tilde{b}\tilde{b}'\right| + \left|\overline{ab'}\overline{a}'\overline{b} - \widetilde{ab'}\tilde{a}'\tilde{b}\right| + \left|\overline{a'b'}\overline{a}\overline{b} - \widetilde{a'b'}\tilde{a}\tilde{b}\right| \\
&\left|\overline{a'bb'}\overline{a} - \widetilde{a'bb'}\tilde{a}\right| + \left|\overline{a'b}\overline{a}\overline{b}' - \widetilde{a'b}\tilde{a}\tilde{b}'\right| + \left|\overline{a}\,\overline{a}'\overline{bb'} - \tilde{a}\tilde{a}'\widetilde{bb'}\right| + 3\left|\overline{a}\overline{b}\overline{a}'\overline{b}' - \tilde{a}\tilde{b}\tilde{a}'\tilde{b}'\right|.
\end{aligned}
$$

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 9. 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 14), 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 9 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 9 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 9 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

$$(3^{rd}\text{ term})\quad \mathbb{P}\left(\left|\overline{abb'}\overline{a}' - \widetilde{abb'}\tilde{a}'\right| \le t\right) \ge 1 - 4\exp\left(-\frac{nt^2}{8(BB_l, B_k)_+^4}\right),$$

$$(5^{th}\text{ term})\quad \mathbb{P}\left(\left|\overline{aa'b'}\,\overline{b}-\widetilde{aa'b'}\tilde{b}\right|\le t\right)\ge 1-4\exp\left(-\frac{nt^2}{8(BB_k,B_l)_+^4}\right),$$

$$(6^{th}\text{ term})\quad \mathbb{P}\left(\left|\overline{aa'}\,\overline{b}\,\overline{b}'-\widetilde{aa'}\tilde{b}\tilde{b}'\right|\le t\right)\ge 1-6\exp\left(-\frac{nt^2}{18(B_k^2,B_l)_+^6}\right),$$

$$(7^{th}\text{ term})\quad \mathbb{P}\left(\left|\overline{ab'}\overline{a}'\overline{b}-\widetilde{ab'}\tilde{a}'\tilde{b}\right|\le t\right)\ge 1-6\exp\left(-\frac{nt^2}{18(B,B_k,B_l)_+^6}\right),$$

$$(8^{th}\text{ term})\quad \mathbb{P}\left(\left|\overline{a'b'}\overline{a}\overline{b}-\widetilde{a'b'}\tilde{a}\tilde{b}\right|\le t\right)\ge 1-6\exp\left(-\frac{nt^2}{18(B,B_k,B_l)_+^6}\right),$$

$$(9^{th}\text{ term})\quad \mathbb{P}\left(\left|\overline{a'bb'}\overline{a}-\widetilde{a'bb'}\tilde{a}\right|\le t\right)\ge 1-4\exp\left(-\frac{nt^2}{8(BB_l,B_k)_+^4}\right),$$

$$(10^{th}\text{ term})\quad \mathbb{P}\left(\left|\overline{a'b}\overline{a}\overline{b}'-\widetilde{a'b}\tilde{a}\tilde{b}'\right|\le t\right)\ge 1-6\exp\left(-\frac{nt^2}{18(B,B_k,B_l)_+^6}\right),$$

$$(11^{th}\text{ term})\quad \mathbb{P}\left(\left|\overline{a}\,\overline{a}'\overline{bb'}-\tilde{a}\tilde{a}'\widetilde{bb'}\right|\le t\right)\ge 1-6\exp\left(-\frac{nt^2}{18(B_k,B_l^2)_+^6}\right).$$

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

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

F.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 13 (with $m = 2$), we have

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

Bounding $\frac{c_1 n}{\gamma_n}\frac{8B^2 J}{n-1} + c_3 n\gamma_n + \frac{4BJc_1 n}{\gamma_n}\left|\hat{u}(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})\right|$. By Lemma 13 (with $m = 2$), it follows that

$$\begin{aligned}
&\mathbb{P}\left(\frac{c_1 n}{\gamma_n}\frac{8B^2 J}{n-1} + c_3 n\gamma_n + \frac{4BJc_1 n}{\gamma_n}\left|\hat{u}(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})\right| \le t\right) \\
&\ge 1 - 2\exp\left(-\frac{\lfloor 0.5n \rfloor \gamma_n^2\left[t - \frac{c_1 n}{\gamma_n}\frac{8B^2 J}{n-1} - c_3 n\gamma_n\right]^2}{2^7 B^4 J^2 c_1^2 n^2}\right) \\
&= 1 - 2\exp\left(-\frac{\lfloor 0.5n \rfloor\left[t\gamma_n(n-1) - 8c_1 B^2 nJ - c_3 n(n-1)\gamma_n^2\right]^2}{2^7 B^4 J^2 c_1^2 n^2 (n-1)^2}\right) \\
&\overset{(a)}{\ge} 1 - 2\exp\left(-\frac{\left[t\gamma_n(n-1) - 8c_1 B^2 nJ - c_3 n(n-1)\gamma_n^2\right]^2}{2^8 B^4 J^2 c_1^2 n^2 (n-1)}\right),
\end{aligned} \tag{25}$$

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

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

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

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

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

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

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

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 $\mathbf{u}^\top\boldsymbol{\Sigma}^{-1}\mathbf{u}$ in $\lambda_n = n\mathbf{u}^\top\boldsymbol{\Sigma}^{-1}\mathbf{u}$ will increase $L(\lambda_n)$.

G. Helper Lemmas

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

Lemma 8 (Product to sum)**.** *Assume that $|a_i| \le B$, $|b_i| \le B$ for $i = 1, \ldots, E$. Then $\left|\prod_{i=1}^E a_i - \prod_{i=1}^E b_i\right| \le B^{E-1}\sum_{j=1}^E |a_j - b_j|$.*

Proof.

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

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

Lemma 9 (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 $|f_i(\mathbf{x})| \leq B < \infty$ 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,*

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

Proof. By Lemma 8, we have

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

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

H. External Lemmas

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

Lemma 10 (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 11 (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*

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

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 12 (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 13 (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 $a \leq b < \infty$. 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 $n \geq m$,*

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

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

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 14 (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* $\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$*. Then,*

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

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.

[sup14] 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.

[sup20] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 2009.