Nyström M-Hilbert-Schmidt Independence Criterion (Supplementary Material)

Florian Kalinke[1] **Zoltán Szabó**[2]

[1]Institute for Program Structures and Data Organization, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2] Department of Statistics, London School of Economics, London, UK

A APPENDIX

Section A.1 contains two external theorems and lemmas that we use. Section A.2 is about our proofs.

A.1 EXTERNAL THEOREMS AND LEMMAS

In this section two theorems and lemmas are recalled for self-completeness, Theorem A.1 is about bounding the error of Nyström mean embeddings [Chatalic et al., 2022, Theorem 4.1], Theorem A.2 is a well-known result [Serfling, 1980, Section 5.6, Theorem A] for bounding the deviation of U-statistics. Lemma A.1 is about connection between U- and V-statistics. Lemma A.2 recalls Markov's inequality.

Theorem A.1 (Bound on mean embeddings)**.** *Let $\mathcal{X}$ be a locally compact second-countable topological space, X a random variable supported on $\mathcal{X}$ with Borel probability measure $\mathbb{P}$, and let $\mathcal{H}_k$ be a RKHS on $\mathcal{X}$ with kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and feature map ϕ_k. Assume that there exists a constant $K \in (0, \infty)$ such that $\sup_{x\in\mathcal{X}} \sqrt{k(x,x)} \leqslant K$. Let $C_k = \mathbb{E}\left[\phi_k(X) \otimes \phi_k(X)\right]$. Furthermore, assume that the data points $\hat{\mathbb{P}}_n = \{x_1, \ldots, x_n\}$ are drawn i.i.d. from the distribution $\mathbb{P}$ and that $n' \leqslant n$ subsamples $\tilde{\mathbb{P}}_{n'} = \{\tilde{x}_1, \ldots, \tilde{x}_{n'}\}$ are drawn uniformly with replacement from the dataset $\hat{\mathbb{P}}_n$. Then for any $\delta \in (0, 1)$ it holds that*

$$\left\|\mu_k\left(\mathbb{P}\right) - \mu_k\left(\tilde{\mathbb{P}}_{n'}\right)\right\|_{\mathcal{H}_k} \leqslant \frac{c_1}{\sqrt{n}} + \frac{c_2}{n'} + \frac{c_3\sqrt{\log(n'/\delta)}}{n'}\sqrt{\mathcal{N}_X\left(\frac{12K^2\log(n'/\delta)}{n'}\right)},$$

with probability at least $1 - \delta$ provided that

$$n' \geqslant \max\left(67, 12K^2\left\|C_k\right\|_{\mathrm{op}}^{-1}\right)\log\left(\frac{n'}{\delta}\right),$$

where $c_1 = 2K\sqrt{2\log(6/\delta)}$, $c_2 = 4\sqrt{3}K\log(12/\delta)$, and $c_3 = 12\sqrt{3\log(12/\delta)}K$.

Recall that a U-statistic is the average of a (symmetric) core function $h = h(x_1, \ldots, x_m)$ over the observations $X_1, \ldots, X_n \sim \mathbb{P}$ $(n \geqslant m)$ with form

$$U_n = U(X_1, \ldots, X_m) = \frac{1}{\binom{n}{m}}\sum_c h(X_{i_1}, \ldots, X_{i_m}), \tag{1}$$

where c is the set of the $\binom{n}{m}$ combinations of m distinct elements $\{i_1, \ldots, i_m\}$ from $\{1, \ldots, n\}$. U_n is an unbiased estimator of $\theta = \theta(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}[h(X_1, \ldots, X_m)]$.

Theorem A.2 (Hoeffding's inequality for U-statistics)**.** *Let $h = h(x_1, \ldots, x_m)$ be a core function for $\theta = \theta(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}\left[h(X_1, \ldots, X_m)\right]$ with $a \leqslant h(x_1, \ldots, x_m) \leqslant b$. Then, for any $u > 0$ and $n \geqslant m$,*

$$\mathbb{P}(U_n - \theta \geqslant u) \leqslant \exp\left(-\frac{2nu^2}{m(b-a)^2}\right).$$

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

Similar to (1) one can consider an alternative (slightly biased) estimator of θ, which is called V-statistic:

$$V_n = V(X_1, \ldots, X_m) = \frac{1}{n^m} \sum_{(i_1,\ldots,i_m)\in T_m(n)} h(X_{i_1}, \ldots, X_{i_m}), \tag{2}$$

where $T_m(n)$ is the m-fold Cartesian product of the set $[n]$.

There is a close relation between U- and V-statistics, as it is made explicit by the following lemma [Serfling, 1980, Lemma, Section 5.7.3].

Lemma A.1 (Connection between U- and V-statistics)**.** *Let $\mathbb{P}$ be a probability measure on a metric space $\mathcal{X}$. Let $(X_i)_{i\in[n]} \overset{i.i.d.}{\sim} \mathbb{P}$. Let m denote any element of $[n]$. Let h be a core function satisfying $\mathbb{E}\left[|h(X_1, \ldots, X_m)|^r\right] < \infty$ with some $r \in \mathbb{Z}_+$. Let U_n and V_n denote the U and V-statistic associated to h as defined in* (1) *and* (2)*, respectively. Then it holds that*

$$\mathbb{E}\left[|U_n - V_n|^r\right] = \mathcal{O}\left(n^{-r}\right).$$

Lemma A.2 (Markov inequality)**.** *For a real-valued random variable X with probability distribution $\mathbb{P}$ and $a > 0$, it holds that*

$$\mathbb{P}\left(|X| \geqslant a\right) \leqslant \frac{\mathbb{E}\left(|X|\right)}{a}.$$

A.2 PROOFS

This section is dedicated to proofs. Lemma 4.2 is derived in Section A.2.1. Proposition 4.1 is proved in Section A.2.3 relying on two lemmas shown in Section A.2.2. Lemma 4.4 is proved in Section A.2.5, with an auxiliary result in Section A.2.4.

A.2.1 Proof of Lemma 4.2

Let $\mu_k\left(\tilde{\mathbb{P}}_{n'}\right) = \sum_{i=1}^{n'} \alpha_k^i \otimes_{m=1}^M \phi_{k_m}(x_m^i)$, and let $\mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right) = \sum_{i=1}^{n'} \alpha_{k_m}^i \phi_{k_m}(x_m^i)$ for $m \in [M]$. We write

$$\begin{aligned}
\mathrm{HSIC}^2_{k,\mathrm{N}}\left(\hat{\mathbb{P}}_n\right) &= \left\|\mu_k\left(\tilde{\mathbb{P}}_{n'}\right) - \otimes_{m=1}^M \mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right)\right\|^2_{\mathcal{H}_k} \\
&= \underbrace{\left\|\mu_k\left(\tilde{\mathbb{P}}_{n'}\right)\right\|^2_{\mathcal{H}_k}}_{=:A} - 2 \cdot \underbrace{\left\langle \mu_k\left(\tilde{\mathbb{P}}_{n'}\right), \otimes_{m=1}^M \mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right)\right\rangle_{\mathcal{H}_k}}_{=:C} + \underbrace{\left\|\otimes_{m=1}^M \mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right)\right\|^2_{\mathcal{H}_k}}_{=:B},
\end{aligned}$$

and continue term-by-term. Using the definition of the tensor product, we have for term A that

$$\begin{aligned}
A &= \left\langle \mu_k\left(\tilde{\mathbb{P}}_{n'}\right), \mu_k\left(\tilde{\mathbb{P}}_{n'}\right)\right\rangle_{\mathcal{H}_k} = \sum_{i=1}^{n'}\sum_{j=1}^{n'} \alpha_k^i \alpha_k^j \left\langle \otimes_{m=1}^M \phi_{k_m}(x_m^i), \otimes_{m=1}^M \phi_{k_m}(x_m^j)\right\rangle_{\mathcal{H}_k} = \sum_{i=1}^{n'}\sum_{j=1}^{n'} \alpha_k^i \alpha_k^j \prod_{m=1}^M k_m(x_m^i, x_m^j) \\
&= \boldsymbol{\alpha}_k^\mathsf{T}\left(\circ_{m=1}^M \mathbf{K}_{k_m}\right)\boldsymbol{\alpha}_k.
\end{aligned}$$

Similarly, we obtain for term B that

$$\begin{aligned}
B &= \left\langle \otimes_{m=1}^M \mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right), \otimes_{m=1}^M \mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right)\right\rangle_{\mathcal{H}_k} \\
&= \left\langle \otimes_{m=1}^M \sum_{i^{(m)}=1}^{n'} \alpha_{k_m}^{i^{(m)}} \phi_{k_m}\left(x_m^{i^{(m)}}\right), \otimes_{m=1}^M \sum_{j^{(m)}=1}^{n'} \alpha_{k_m}^{j^{(m)}} \phi_{k_m}\left(x_m^{j^{(m)}}\right)\right\rangle_{\mathcal{H}_k} \\
&\overset{(*)}{=} \prod_{m=1}^M \sum_{i^{(m)}=1}^{n'}\sum_{j^{(m)}=1}^{n'} \alpha_{k_m}^{i^{(m)}} \alpha_{k_m}^{j^{(m)}} k_m\left(x_m^{i^{(m)}}, x_m^{j^{(m)}}\right) = \prod_{m=1}^M \boldsymbol{\alpha}_{k_m}^\mathsf{T} \mathbf{K}_{k_m} \boldsymbol{\alpha}_{k_m},
\end{aligned}$$

where in $(*)$ we used (1), the linearity of the inner product, and the reproducing property.

Last, we express term C as

$$
\begin{aligned}
C &= \left\langle \sum_{i=1}^{n'} \alpha_k^i \otimes_{m=1}^M \phi_{k_m}\left(x_m^i\right), \otimes_{m=1}^M \sum_{j^{(m)}=1}^{n'} \alpha_{k_m}^{j^{(m)}} \phi_{k_m}\left(x_m^{j^{(m)}}\right) \right\rangle_{\mathcal{H}_k} \\
&\overset{(a)}{=} \sum_{i=1}^{n'} \alpha_k^i \left\langle \otimes_{m=1}^M \phi_{k_m}\left(x_m^i\right), \otimes_{m=1}^M \sum_{j^{(m)}=1}^{n'} \alpha_{k_m}^{j^{(m)}} \phi_{k_m}\left(x_m^{j^{(m)}}\right) \right\rangle_{\mathcal{H}_k} \\
&\overset{(b)}{=} \sum_{i=1}^{n'} \alpha_k^i \prod_{m\in[M]} \left\langle \phi_{k_m}\left(x_m^i\right), \sum_{j^{(m)}=1}^{n'} \alpha_{k_m}^{j^{(m)}} \phi_{k_m}\left(x_m^{j^{(m)}}\right) \right\rangle_{\mathcal{H}_k} \\
&\overset{(c)}{=} \sum_{i=1}^{n'} \alpha_k^i \prod_{m\in[M]} \sum_{j^{(m)}=1}^{n'} \alpha_{k_m}^{j^{(m)}} \left\langle \phi_{k_m}\left(x_m^i\right), \phi_{k_m}\left(x_m^{j^{(m)}}\right) \right\rangle_{\mathcal{H}_k} \\
&\overset{(d)}{=} \sum_{i=1}^{n'} \alpha_k^i \prod_{m\in[M]} \underbrace{\sum_{j^{(m)}=1}^{n'} \alpha_{k_m}^{j^{(m)}} k_m\left(x_m^i, x_m^{j^{(m)}}\right)}_{\left(\mathbf{K}_{k_m}\right)_i \boldsymbol{\alpha}_{k_m}} = \boldsymbol{\alpha}_k^{\mathsf{T}} \left(\circ_{m=1}^M \mathbf{K}_{k_m} \boldsymbol{\alpha}_{k_m}\right),
\end{aligned}
$$

where (a) follows from the linearity of the inner product, (b) holds by (1), (c) is implied by the linearity of the inner product, (d) is valid by the reproducing property, and we refer to the i-th row of $\mathbf{K}_{k_m}$ as $\left(\mathbf{K}_{k_m}\right)_i$.

Substituting terms A, B, and C concludes the proof.

A.2.2 Two Lemmas to the Proof of Proposition 4.1

Our main result relies on two lemmas.

Lemma A.3 (Error bound for Nyström mean embedding of tensor product kernel)**.** *Let $X = (X_m)_{m=1}^M \in \mathcal{X} = \times_{m=1}^M \mathcal{X}_m$, $X \sim \mathbb{P} \in \mathcal{M}_1^+(\mathcal{X})$, and $(\mathcal{X}_m)_{m\in[M]}$ locally compact, second-countable topological spaces. Let $k_m : \mathcal{X}_m \times \mathcal{X}_m \to \mathbb{R}$ be a bounded kernel, i.e. there exists $a_{k_m} \in (0,\infty)$ such that $\sup_{x_m\in\mathcal{X}_m} \sqrt{k_m(x_m,x_m)} \leqslant a_{k_m}$ for $m \in [M]$. Let $a_k = \prod_{m=1}^M a_{k_m}$, $k = \otimes_{m=1}^M k_m$, $\mathcal{H}_k$ the RKHS associated to k, $\phi_k = \otimes_{m=1}^M \phi_{k_m}$, $C_k = \mathbb{E}\left[\phi_k(X) \otimes \phi_k(X)\right]$, $n' \leqslant n$, and $\tilde{\mathbb{P}}_{n'}$ defined according to* (14)*. Then for any $\delta \in (0,1)$ it holds that*

$$
\left\| \mu_k\left(\mathbb{P}\right) - \mu_k\left(\tilde{\mathbb{P}}_{n'}\right) \right\|_{\mathcal{H}_k} \leqslant \frac{c_{k,1}}{\sqrt{n}} + \frac{c_{k,2}}{n'} + \frac{c_{k,3}\sqrt{\log(n'/\delta)}}{n'} \sqrt{\mathcal{N}_X\left(\frac{12 a_k^2 \log(n'/\delta)}{n'}\right)},
$$

with probability at least $1-\delta$, provided that

$$
n' \geqslant \max\left(67, 12 a_k^2 \left\|C_k\right\|_{\mathrm{op}}^{-1}\right) \log\left(\frac{n'}{\delta}\right),
$$

where $c_{k,1} = 2a_k\sqrt{2\log(6/\delta)}$, $c_{k,2} = 4\sqrt{3}a_k \log(12/\delta)$, and $c_{k,3} = 12\sqrt{3\log(12/\delta)}a_k$.

Proof. With $\mathcal{X} = \times_{m\in[M]}\mathcal{X}_m$, noticing that $\mathcal{X}$ is locally compact second-countable iff. $(\mathcal{X}_m)_{m\in[M]}$ are so [Willard, 1970, Theorem 16.2(c), Theorem 18.6], $\mathcal{H}_k = \otimes_{m=1}^M \mathcal{H}_{k_m}$, $\phi_k = \otimes_{m=1}^M \phi_{k_m}$, and $\sqrt{k(x,x)} = \prod_{m=1}^M \sqrt{k_m(x_m,x_m)} \leqslant a_k$, the statement is implied by Theorem A.1. □

Proof of Lemma 4.3. To simplify notation, let $\mu_{k_m} = \mu_{k_m}\left(\mathbb{P}_m\right)$, $\tilde{\mu}_{k_m} = \mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right)$, $\mathcal{H}_k = \otimes_{m=1}^M \mathcal{H}_{k_m}$, and $d_{k_m} = \left\|\mu_{k_m} - \tilde{\mu}_{k_m}\right\|_{\mathcal{H}_{k_m}}$. The proof proceeds by induction on M: For $M = 1$ the l.h.s. = r.h.s. = $\left\|\mu_{k_1}\left(\mathbb{P}_1\right) - \mu_{k_1}\left(\tilde{\mathbb{P}}_{1,n'}\right)\right\|_{\mathcal{H}_k}$ is

satisfied, and we assume that the statement holds for $M = M - 1$, to obtain

$$\begin{aligned}
\left\|\otimes_{m=1}^{M}\mu_{k_m} - \otimes_{m=1}^{M}\tilde{\mu}_{k_m}\right\|_{\mathcal{H}_k} &= \left\| \otimes_{m=1}^{M} \mu_{k_m} - \otimes_{m=1}^{M-1}\mu_{k_m} \otimes \tilde{\mu}_{k_M} + \otimes_{m=1}^{M-1}\mu_{k_m} \otimes \tilde{\mu}_{k_M} - \otimes_{m=1}^{M}\tilde{\mu}_{k_m}\right\|_{\mathcal{H}_k} \\
&= \left\|\otimes_{m=1}^{M-1}\mu_{k_m} \otimes (\mu_{k_M} - \tilde{\mu}_{k_M}) + \left(\otimes_{m=1}^{M-1}\mu_{k_m} - \otimes_{m=1}^{M-1}\tilde{\mu}_{k_m}\right) \otimes \tilde{\mu}_{k_M}\right\|_{\mathcal{H}_k} \\
&\overset{(a)}{\leqslant} \left\|\otimes_{m=1}^{M-1}\mu_{k_m} \otimes (\mu_{k_M} - \tilde{\mu}_{k_M})\right\|_{\mathcal{H}_k} + \left\|\left(\otimes_{m=1}^{M-1}\mu_{k_m} - \otimes_{m=1}^{M-1}\tilde{\mu}_{k_m}\right) \otimes \tilde{\mu}_{k_M}\right\|_{\mathcal{H}_k} \\
&\overset{(b)}{=} \left(\prod_{m\in[M-1]} \|\mu_{k_m}\|_{\mathcal{H}_{k_m}}\right) d_{k_M} + \left\|\otimes_{m=1}^{M-1}\mu_{k_m} - \otimes_{m=1}^{M-1}\tilde{\mu}_{k_m}\right\|_{\otimes_{m=1}^{M-1}\mathcal{H}_{k_m}} \|\tilde{\mu}_{k_M}\|_{\mathcal{H}_{k_M}} \\
&\overset{(c)}{\leqslant} d_{k_M} \prod_{m\in[M-1]} a_{k_m} + \left\|\otimes_{m=1}^{M-1}\mu_{k_m} - \otimes_{m=1}^{M-1}\tilde{\mu}_{k_m}\right\|_{\otimes_{m=1}^{M-1}\mathcal{H}_{k_m}} (a_{k_M} + d_{k_M}) \\
&\overset{(d)}{\leqslant} d_{k_M} \prod_{m\in[M-1]} a_{k_m} + \left\{\prod_{m\in[M-1]} (a_{k_m} + d_{k_m}) - \prod_{m\in[M-1]} a_{k_m}\right\} (a_{k_M} + d_{k_M}) \\
&= d_{k_M} \prod_{m\in[M-1]} a_{k_m} + \prod_{m\in[M]} (a_{k_m} + d_{k_m}) - \prod_{m\in[M]} a_{k_m} - d_{k_M} \prod_{m\in[M-1]} a_{k_m} \\
&= \prod_{m\in[M]} (a_{k_m} + d_{k_m}) - \prod_{m\in[M]} a_{k_m},
\end{aligned}$$

where (a) holds by the triangle inequality, (b) is implied by (2) and the definition of d_{k_M}, (c) follows from

$$\|\mu_{k_m}\|_{\mathcal{H}_{k_m}} = \left\|\int_{\mathcal{X}_m} k_m(\cdot, x_m)\mathrm{d}\mathbb{P}_m(x_m)\right\|_{\mathcal{H}_{k_m}} \overset{(e)}{\leqslant} \int_{\mathcal{X}_m} \underbrace{\|k_m(\cdot, x_m)\|_{\mathcal{H}_{k_m}}}_{\overset{(f)}{=}\sqrt{k_m(x_m,x_m)}\overset{(g)}{\leqslant} a_{k_m}} \mathrm{d}\mathbb{P}_m(x_m) \leqslant a_{k_m}, \tag{3}$$

$$\|\tilde{\mu}_{k_M}\|_{\mathcal{H}_{k_M}} = \|\tilde{\mu}_{k_M} - \mu_{k_M} + \mu_{k_M}\|_{\mathcal{H}_{k_M}} \overset{(h)}{\leqslant} \|\tilde{\mu}_{k_M} - \mu_{k_M}\|_{\mathcal{H}_{k_M}} + \|\mu_{k_M}\|_{\mathcal{H}_{k_M}} \overset{(i)}{\leqslant} d_{k_M} + a_{k_M},$$

(d) is valid by the induction statement holding for $M - 1$, (e) is a property of Bochner integrals, (f) is implied by the reproducing property, (g) comes from the definition of a_{k_m}, the triangle inequality implies (h), (i) follows from (3) and the definition of d_{k_M}. □

A.2.3 Proof of Proposition 4.1

Let $k = \otimes_{m=1}^{M} k_m$, and let $\mathcal{H}_k = \otimes_{m=1}^{M}\mathcal{H}_{k_m}$. We note that $\mathcal{X} = \times_{m\in[M]}\mathcal{X}_m$ is locally compact second-countable as $(\mathcal{X}_m)_{m\in[M]}$ are so [Willard, 1970, Theorem 16.2(c), Theorem 18.6].

We decompose the error of the Nyström approximation as

$$\begin{aligned}
\left|\mathrm{HSIC}_k(\mathbb{P}) - \mathrm{HSIC}_{k,\mathrm{N}}\left(\hat{\mathbb{P}}_n\right)\right| &= \left|\left\|\mu_k(\mathbb{P}) - \otimes_{m=1}^{M}\mu_{k_m}(\mathbb{P}_m)\right\|_{\mathcal{H}_k} - \left\|\mu_k\left(\tilde{\mathbb{P}}_{n'}\right) - \otimes_{m=1}^{M}\mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right)\right\|_{\mathcal{H}_k}\right| \\
&\overset{(a)}{\leqslant} \left\|\mu_k(\mathbb{P}) - \otimes_{m=1}^{M}\mu_{k_m}(\mathbb{P}_m) - \mu_k\left(\tilde{\mathbb{P}}_{n'}\right) + \otimes_{m=1}^{M}\mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right)\right\|_{\mathcal{H}_k} \\
&\overset{(b)}{\leqslant} \underbrace{\left\|\mu_k(\mathbb{P}) - \mu_k\left(\tilde{\mathbb{P}}_{n'}\right)\right\|_{\mathcal{H}_k}}_{t_1} + \underbrace{\left\|\otimes_{m=1}^{M}\mu_{k_m}(\mathbb{P}_m) - \otimes_{m=1}^{M}\mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right)\right\|_{\mathcal{H}_k}}_{t_2},
\end{aligned}$$

where (a) holds by the reverse triangle inequality, and (b) follows from the triangle inequality.

First term (t_1): One can bound the error of the first term by Lemma A.3; in other words, for any $\delta \in (0, 1)$ with probability at least $(1 - \delta)$ it holds that

$$\left\|\mu_k(\mathbb{P}) - \mu_k\left(\tilde{\mathbb{P}}_{n'}\right)\right\|_{\mathcal{H}_k} \leqslant \frac{c_{k,1}}{\sqrt{n}} + \frac{c_{k,2}}{n'} + \frac{c_{k,3}\sqrt{\log(n'/\delta)}}{n'}\sqrt{\mathcal{N}_{X_m}\left(\frac{12a_{k_m}^2\log(n'/\delta)}{n'}\right)}$$

provided that $n' \geqslant \max\left(67, 12a_k^2 \left\|C_k\right\|_{\mathrm{op}}^{-1}\right) \log\left(\frac{n'}{\delta}\right)$, with the constants $c_{k,1} = 2a_k\sqrt{2\log(6/\delta)}$, $c_{k,2} = 4\sqrt{3}a_k \log(12/\delta)$, $c_{k,3} = 12\sqrt{3\log(12/\delta)}a_k$.

Second term (t_2): Applying Lemma 4.3 to the second term gives

$$\left\|\otimes_{m=1}^M \mu_{k_m}(\mathbb{P}_m) - \otimes_{m=1}^M \mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right)\right\|_{\mathcal{H}_k} \leqslant \prod_{m\in[M]}\left(a_{k_m} + \left\|\mu_{k_m}(\mathbb{P}_m) - \mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right)\right\|_{\mathcal{H}_{k_m}}\right) - \prod_{m\in[M]} a_{k_m}.$$

We now bound the error of each of the M factors by Theorem A.1, i.e., for fixed $m \in [M]$; particularly we get that for any $\delta \in (0,1)$ with probability at least $1-\delta$

$$\left\|\mu_{k_m}(\mathbb{P}_m) - \mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right)\right\|_{\mathcal{H}_{k_m}} \leqslant \frac{c_{k_m,1}}{\sqrt{n}} + \frac{c_{k_m,2}}{n'} + \frac{c_{k_m,3}\sqrt{\log(n'/\delta)}}{n'}\sqrt{\mathcal{N}_{X_m}\left(\frac{12a_{k_m}^2\log(n'/\delta)}{n'}\right)}, \text{ hence}$$

$$a_{k_m} + \left\|\mu_{k_m}(\mathbb{P}_m) - \mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right)\right\|_{\mathcal{H}_{k_m}} \leqslant a_{k_m} + \frac{c_{k_m,1}}{\sqrt{n}} + \frac{c_{k_m,2}}{n'} + \frac{c_{k_m,3}\sqrt{\log(n'/\delta)}}{n'}\sqrt{\mathcal{N}_{X_m}\left(\frac{12a_{k_m}^2\log(n'/\delta)}{n'}\right)},$$

and by union bound that their product is for any $\delta \in (0, \frac{1}{M})$ with probability at least $1 - M\delta$

$$\prod_{m\in[M]}\left[a_{k_m} + \left\|\mu_{k_m}(\mathbb{P}_m) - \mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right)\right\|_{\mathcal{H}_{k_m}}\right] \leqslant$$

$$\leqslant \prod_{m\in[M]}\left[a_{k_m} + \frac{c_{k_m,1}}{\sqrt{n}} + \frac{c_{k_m,2}}{n'} + \frac{c_{k_m,3}\sqrt{\log(n'/\delta)}}{n'}\sqrt{\mathcal{N}_{X_m}\left(\frac{12a_{k_m}^2\log(n'/\delta)}{n'}\right)}\right],$$

$$\prod_{m\in[M]}\left[a_{k_m} + \left\|\mu_{k_m}(\mathbb{P}_m) - \mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right)\right\|_{\mathcal{H}_{k_m}}\right] - \prod_{m\in[M]} a_{k_m} \leqslant$$

$$\leqslant \prod_{m\in[M]}\left[a_{k_m} + \frac{c_{k_m,1}}{\sqrt{n}} + \frac{c_{k_m,2}}{n'} + \frac{c_{k_m,3}\sqrt{\log(n'/\delta)}}{n'}\sqrt{\mathcal{N}_{X_m}\left(\frac{12a_{k_m}^2\log(n'/\delta)}{n'}\right)}\right] - \prod_{m\in[M]} a_{k_m},$$

provided that $n' \geqslant \max\left(67, 12a_{k_m}^2\left\|C_{k_m}\right\|_{\mathrm{op}}^{-1}\right)\log\left(\frac{n'}{\delta}\right)$ for all $m \in [M]$, with $C_{k_m} = \mathbb{E}\left[\phi_{k_m}(X_m) \otimes \phi_{k_m}(X_m)\right]$ and constants $c_{k_m,1} = 2a_{k_m}\sqrt{2\log(6/\delta)}$, $c_{k_m,2} = 4\sqrt{3}a_{k_m}\log(12/\delta)$, $c_{k_m,3} = 12\sqrt{3\log(12/\delta)}a_{k_m}$, with $m \in [M]$.

Combining the $M+1$ terms by union bound yields the stated result.

A.2.4 Lemma to the Proof of Lemma 4.4

Lemma A.4 (Deviation bound for U-statistics based HSIC estimator)**.** *It holds that*

$$\left|\mathrm{HSIC}_{k,u}^2\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}_k^2\left(\mathbb{P}\right)\right| = \mathcal{O}_P\left(\frac{1}{\sqrt{n}}\right),$$

where $\mathrm{HSIC}_{k,u}^2$ *is the U-statistic based estimator of* HSIC_k^2.

Proof. We show that (3) can be expressed as a sum of U-statistics and then bound the terms individually. First, square (3) to obtain

$$\mathrm{HSIC}_k^2(\mathbb{P}) = \underbrace{\mathbb{E}_{(x_1,\ldots,x_M),(x_1',\ldots,x_M')\sim\mathbb{P}}\left[\prod_{m\in[M]} k_m\left(x_m, x_m'\right)\right]}_{A} + \underbrace{\mathbb{E}_{x_1,x_1'\sim\mathbb{P}_1,\ldots,x_M,x_M'\sim\mathbb{P}_M}\left[\prod_{m\in[M]} k_m\left(x_m, x_m'\right)\right]}_{B}$$

$$\underbrace{- 2\,\mathbb{E}_{(x_1,\ldots,x_M)\sim\mathbb{P},x_1'\sim\mathbb{P}_1,\ldots,x_M'\sim\mathbb{P}_M}\left[\prod_{m\in[M]} k_m(x_m, x_m')\right]}_{C},$$

where A, B, and C can be estimated by U-statistics A'_n, B'_n, and C'_n, respectively. Let $\mathrm{HSIC}^2_{k,u}\left(\hat{\mathbb{P}}_n\right) = A'_n + B'_n - 2C'_n$, and split t as $\alpha t + \beta t + (1-\alpha-\beta)t$, with $\alpha, \beta > 0$ and $\alpha + \beta < 1$. One obtains

$$P\left(\left|\mathrm{HSIC}^2_k(\mathbb{P}) - \mathrm{HSIC}^2_{k,u}\left(\hat{\mathbb{P}}_n\right)\right| \geqslant t\right) \leqslant P\left(\left|A - A'_n\right| \geqslant \alpha t\right) + P\left(\left|B - B'_n\right| \geqslant \beta t\right) + P\left(2\left|C - C'_n\right| \geqslant (1-\alpha-\beta)t\right).$$

Doubling and rewriting Theorem A.2, we have that for U-statistics and any $\delta \in (0,1)$

$$\mathbb{P}\left(|U_n - \theta| \geqslant \sqrt{\frac{m(b-a)^2 \ln(\frac{2}{\delta})}{2n}}\right) \leqslant \delta.$$

Now, choosing the (θ, U_n, u) triplet to be $(A, A'_n, \alpha t)$, $(B, B'_n, \beta t)$, $\left(C, C'_n, \frac{(1-\alpha-\beta)t}{2}\right)$, respectively, setting $m = 2M$, and observing that $a \leqslant k(x,y) \leqslant b$ as k is bounded, we obtain that $|A'_n - A|\sqrt{n}$, $|B'_n - B|\sqrt{n}$, and $|C'_n - C|\sqrt{n}$ are bounded in probability and so is their sum. □

A.2.5 Proof of Lemma 4.4

We consider the decomposition

$$\left|\mathrm{HSIC}^2_k\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}^2_k\left(\mathbb{P}\right)\right| \leqslant \underbrace{\left|\mathrm{HSIC}^2_k\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}^2_{k,u}\left(\hat{\mathbb{P}}_n\right)\right|}_{t_1} + \underbrace{\left|\mathrm{HSIC}^2_{k,u}\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}^2_k\left(\mathbb{P}\right)\right|}_{t_2}, \tag{4}$$

by using the triangle inequality, where $\mathrm{HSIC}_{k,u}$ is the U-statistic based HSIC estimator.

Second term (t_2): Lemma A.4 establishes that $t_2 = \mathcal{O}_{\mathrm{P}}\left(\frac{1}{\sqrt{n}}\right)$.

First term (t_1): To bound t_1, first, by Markov's inequality (Lemma A.2) observe that

$$\mathbb{P}\left(\left|\mathrm{HSIC}^2_k\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}^2_{k,u}\left(\hat{\mathbb{P}}_n\right)\right| \geqslant a\right) \leqslant \underbrace{\frac{\mathbb{E}\left(\left|\mathrm{HSIC}^2_k\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}^2_{k,u}\left(\hat{\mathbb{P}}_n\right)\right|\right)}{a}}_{=:\epsilon},$$

$$\mathbb{P}\left(\left|\mathrm{HSIC}^2_k\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}^2_{k,u}\left(\hat{\mathbb{P}}_n\right)\right| \geqslant \frac{\mathbb{E}\left(\left|\mathrm{HSIC}^2_k\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}^2_{k,u}\left(\hat{\mathbb{P}}_n\right)\right|\right)}{\epsilon}\right) \leqslant \epsilon,$$

$$\mathbb{P}\left(\left|\mathrm{HSIC}^2_k\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}^2_{k,u}\left(\hat{\mathbb{P}}_n\right)\right| < \frac{\mathbb{E}\left(\left|\mathrm{HSIC}^2_k\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}^2_{k,u}\left(\hat{\mathbb{P}}_n\right)\right|\right)}{\epsilon}\right) \geqslant 1 - \epsilon,$$

$$\mathbb{P}\left(\left|\mathrm{HSIC}^2_k\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}^2_{k,u}\left(\hat{\mathbb{P}}_n\right)\right| < \frac{C}{n\epsilon}\right) \overset{(*)}{\geqslant} 1 - \epsilon,$$

$$\mathbb{P}\left(\left|\mathrm{HSIC}^2_k\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}^2_{k,u}\left(\hat{\mathbb{P}}_n\right)\right| \geqslant \frac{C}{n\epsilon}\right) \leqslant \epsilon, \tag{5}$$

for constant $C > 0$ and n large enough, where $(*)$ follows from Lemma A.1 (with $r = 1$). (5) implies that

$$\left|\mathrm{HSIC}^2_k\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}^2_{k,u}\left(\hat{\mathbb{P}}_n\right)\right| = \mathcal{O}_{\mathrm{P}}\left(\frac{1}{n}\right).$$

Combining the terms ($t_1 + t_2$): Combining the obtained results for the two terms, one gets that

$$\begin{aligned}\left|\mathrm{HSIC}^2_k\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}^2_k\left(\mathbb{P}\right)\right| &\overset{(4)}{\leqslant} \left|\mathrm{HSIC}^2_k\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}^2_{k,u}\left(\hat{\mathbb{P}}_n\right)\right| + \left|\mathrm{HSIC}^2_{k,u}\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}^2_k\left(\mathbb{P}\right)\right| \\ &= \mathcal{O}_{\mathrm{P}}\left(\frac{1}{n}\right) + \mathcal{O}_{\mathrm{P}}\left(\frac{1}{\sqrt{n}}\right) = \mathcal{O}_{\mathrm{P}}\left(\frac{1}{\sqrt{n}}\right).\end{aligned} \tag{6}$$

Hence

$$\mathcal{O}_{\mathrm{P}}\left(\frac{1}{\sqrt{n}}\right) \overset{(6)}{\geqslant} \left|\mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}_k^2\left(\mathbb{P}\right)\right| = \left|\mathrm{HSIC}_k\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}_k\left(\mathbb{P}\right)\right| \left|\underbrace{\mathrm{HSIC}_k\left(\hat{\mathbb{P}}_n\right)}_{\overset{(5)}{\geqslant} 0} + \underbrace{\mathrm{HSIC}_k\left(\mathbb{P}\right)}_{\geqslant 0}\right|$$

$$\geqslant \left|\mathrm{HSIC}_k\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}_k\left(\mathbb{P}\right)\right| \mathrm{HSIC}_k\left(\mathbb{P}\right),$$

which by dividing with the constant $\mathrm{HSIC}_k\left(\mathbb{P}\right) > 0$ implies the statement.

References

Antoine Chatalic, Nicolas Schreuder, Alessandro Rudi, and Lorenzo Rosasco. Nyström kernel mean embeddings. In *International Conference on Machine Learning (ICML)*, pages 3006–3024, 2022.

Robert J. Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 1980.

Stephen Willard. *General Topology*. Addison-Wesley, 1970.