A Practical Introduction to Kernel Discrepancies: MMD, HSIC & KSD

Antonin Schrab
Centre for Artificial Intelligence
Gatsby Computational Neuroscience Unit
University College London

Abstract

This article provides a practical introduction to kernel discrepancies, focusing on the Maximum Mean Discrepancy (MMD), the Hilbert–Schmidt Independence Criterion (HSIC), and the Kernel Stein Discrepancy (KSD). Various estimators for these discrepancies are presented, including the commonly-used V-statistics and U-statistics, as well as several forms of the more computationally-efficient incomplete U-statistics. The importance of the choice of kernel bandwidth is stressed, showing how it affects the behaviour of the discrepancy estimation. Adaptive estimators are introduced, which combine multiple estimators with various kernels, addressing the problem of kernel selection.

This paper corresponds to the introduction of my PhD thesis (Schrab, 2025a, Chapter 2) and is presented as a standalone article to introduce the reader to kernel discrepancies estimators. First, in Section 1, we define kernels, Reproducing Kernel Hilbert Spaces, mean embeddings and cross-covariance operators, and present kernel properties such as characteristicity, universality and translation invariance. Then, in Section 2, we introduce the Maximum Mean Discprecancy, the Hilbert–Schmidt Independence Criterion, and the Kernel Stein Discrepancy, as well as their estimators, and we discuss the importance of the choice of kernel for such measures. We then introduce a collection of statistics in Section 3, including the commonly-used complete statistics, as well as their incomplete counterparts which trade accuracy for computational efficiency. Finally, in Section 4, we construct adaptive estimators combining multiple statistics with various kernels, which is one method to address the problem of kernel selection.

1 Kernels

In this section, we introduce kernels and RKHSs, define characteristicity and universality of kernels, present translation-invariant kernels with bandwidth parameters, and provide some classical kernel examples.

We refer the readers to Sejdinovic and Gretton (2012); Gretton (2013) for some kernel/RKHS introductions, to Aronszajn (1950); Steinwart and Christmann (2008); Berlinet and Thomas-Agnan (2011) for details about the RKHS constructions, to Fukumizu et al. (2004); Sriperumbudur et al. (2008, 2010a,b, 2011); Carmeli et al. (2010) for the characteristicity and universality of kernels, and to Muandet et al. (2017) for an in-depth review of kernel mean embedding methods.

Kernel & RKHS. We present three definitions of a *'kernel'* and then discuss their relations. First, recall that a vector space $\mathcal{H}$ is called a Hilbert space if it is equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and is complete (*i.e.* every Cauchy sequence converges with respect to the metric induced by the inner product). In the following three definitions, let $\mathcal{X}$ to be a non-empty set.

1. A function $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *kernel* if there exist a Hilbert space $\mathcal{H}$ and a function $\phi \colon \mathcal{X} \to \mathcal{H}$ (called *feature map*) such that
$$k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$$
for all $x, y \in \mathcal{X}$.

2. A symmetric function $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *positive definite kernel* if
$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$$
for all $x_1, \ldots, x_n \in \mathcal{X}$ and all $c_1, \ldots, c_n \in \mathbb{R}$, for any $n \in \mathbb{N}$.

3. For a Hilbert space $\mathcal{H}$ of real-valued functions on $\mathcal{X}$, a function $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *reproducing kernel* of $\mathcal{H}$ if
   - $k(x, \cdot) \in \mathcal{H}$ for all $x \in \mathcal{X}$,
   - $\langle f, k(x, \cdot) \rangle_{\mathcal{H}} = f(x)$ for all $f \in \mathcal{H}$ and all $x \in \mathcal{X}$.

   The resulting space $\mathcal{H}$ is then called a *Reproducing Kernel Hilbert Space* (RKHS).

These three definitions are equivalent in the following sense.

- [$1 \implies 2$] Every kernel is a positive definite kernel as $\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(x_i, x_j) = \| \sum_{i=1}^{n} c_i \phi(x_i) \|_{\mathcal{H}} \geq 0$.
- [$2 \implies 1$] Every positive definite kernel is guaranteed to be an inner product between features in an Hilbert space (Steinwart and Christmann, 2008, Theorem 4.16).
- [$2 \implies 3$] Given a positive definite kernel k on $\mathcal{X} \times \mathcal{X}$, there exists a unique Hilbert space $\mathcal{H}_k$ of real-valued functions on $\mathcal{X}$ for which k is the reproducing kernel, $\mathcal{H}_k$ is the (unique) RKHS associated to k (Aronszajn, 1950, Moore-Aronsajn Theorem).
- [$3 \implies 1$] A reproducing kernel is a kernel with feature map $\phi(x) = k(x, \cdot)$ as $\langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}} = k(x, y)$.

There exists an equivalent definition of an RKHS, which perhaps suprisingly, does not involve the notion of a kernel.

4. A Hilbert space $\mathcal{H}$ of real-valued functions on $\mathcal{X}$ is an RKHS if the evaluation functional is continuous, that is, $|f(x)| \leq C_x \|f\|_{\mathcal{H}}$ for some $C_x > 0$, for all $f \in \mathcal{H}$ and all $x \in \mathcal{X}$.

These two RKHS definitions are indeed equivalent.

- [3 $\implies$ 4] Using the reproducing property followed by Cauchy–Schwartz inequality, we obtain $|f(x)| \leq \sqrt{k(x,x)}\|f\|_{\mathcal{H}}$ for all $f \in \mathcal{H}$ and all $x \in \mathcal{X}$.
- [4 $\implies$ 3] By Riesz Representation Theorem (Rudin, 1987, Theorem 4.12), for all $x \in \mathcal{X}$ there exists $g_x \in \mathcal{H}$ such that $f(x) = \langle f, g_x \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$. The function $k(x,y) \coloneqq g_x(y)$ satisfies the two properties of definition 3 and hence is a repoducing kernel.

We note that, from the reproducing property (definition 3 above), it also follows that (Theorem 1, Zhou, 2008 and Lemma C.9, Barp et al., 2022)

$$\tfrac{\partial}{\partial x} f(x) = \left\langle f, \tfrac{\partial}{\partial x} k(x, \cdot) \right\rangle_{\mathcal{H}} \tag{1}$$

for all $f \in \mathcal{H}$ and all $x \in \mathcal{X}$, under appropriate regularity conditions.

Kernel mean embedding. Let P be a probability distribution on $\mathcal{X}$. Riesz Representation Theorem (Rudin, 1987, Theorem 4.12) guarantees the existence of a unique element $\mu_P \in \mathcal{H}$ satisfying

$$\langle f, \mu_P \rangle_{\mathcal{H}} = \mathbb{E}_P[f(X)] = \mathbb{E}_P[\langle f, k(X, \cdot) \rangle_{\mathcal{H}}]$$

for all $f \in \mathcal{H}$, where $\mathbb{E}_P[f(X)]$ denotes the expectation of f with respect to P. This element μ_P is called the *kernel mean embedding* and can be written as

$$\mu_P = \mathbb{E}_P[k(X, \cdot)]$$

so that $\langle f, \mathbb{E}_P[k(X,\cdot)] \rangle_{\mathcal{H}} = \mathbb{E}_P[\langle f, k(X,\cdot)\rangle_{\mathcal{H}}]$ is justified. Abusing notation, we write $\frac{\partial}{\partial X} f(X)$ for $\frac{\partial}{\partial x} f(x)\big|_{x=X}$, we then have

$$\mathbb{E}_P\left[\tfrac{\partial}{\partial X} f(X)\right] = \mathbb{E}_P\left[\langle f, \tfrac{\partial}{\partial X} k(X, \cdot)\rangle_{\mathcal{H}}\right] = \left\langle f, \mathbb{E}_P\left[\tfrac{\partial}{\partial X} k(X, \cdot)\right]\right\rangle_{\mathcal{H}}$$

where the existence of $\mathbb{E}_P\left[\frac{\partial}{\partial X} k(X, \cdot)\right]$ is again guaranteed by Riesz Representation Theorem.

Cross-covariance operator. Given two Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ with associated inner products $\langle \cdot, \cdot \rangle_{\mathcal{H}_1}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}_2}$, respectively, the tensor product $\otimes$ and Hilbert–Schmidt inner product $\langle \cdot, \cdot \rangle_{\text{HS}}$ are defined as $\langle f \otimes g, \tilde{f} \otimes \tilde{g} \rangle_{\text{HS}} = \langle f, \tilde{f} \rangle_{\mathcal{H}_1} \langle g, \tilde{g} \rangle_{\mathcal{H}_2}$ for all $f, \tilde{f} \in \mathcal{H}_1$ and all $g, \tilde{g} \in \mathcal{H}_2$. Let P_{XY} be a joint probability distribution on $\mathcal{X} \times \mathcal{Y}$, and consider kernels $k^{\mathcal{X}}$ and $k^{\mathcal{Y}}$ on $\mathcal{X} \times \mathcal{X}$ and $\mathcal{Y} \times \mathcal{Y}$, respectively. The cross-covariance operator is defined as the linear operator $\mathcal{C}_{P_{XY}} : \mathcal{H}_{k^{\mathcal{Y}}} \to \mathcal{H}_{k^{\mathcal{X}}}$ satisfying

$$\begin{aligned}
\langle f, \mathcal{C}_{P_{XY}} g \rangle_{\mathcal{H}_{k^{\mathcal{X}}}} &= \mathbb{E}_{P_{XY}}\left[\left(f(X) - \mathbb{E}_{P_X}[f(X')]\right) \left(g(Y) - \mathbb{E}_{P_Y}[g(Y')]\right) \right] \\
&= \mathbb{E}_{P_{XY}}\left[\langle f, k^{\mathcal{X}}(X, \cdot) - \mu_{P_X} \rangle_{\mathcal{H}_{k^{\mathcal{X}}}} \langle g, k^{\mathcal{Y}}(Y, \cdot) - \mu_{P_Y} \rangle_{\mathcal{H}_{k^{\mathcal{Y}}}} \right] \\
&= \mathbb{E}_{P_{XY}}\left[\left\langle f \otimes g, \left(k^{\mathcal{X}}(X, \cdot) - \mu_{P_X}\right) \otimes \left(k^{\mathcal{Y}}(Y, \cdot) - \mu_{P_Y}\right) \right\rangle_{\text{HS}} \right] \\
&= \mathbb{E}_{P_{XY}}\left[\left\langle f, \left(\left(k^{\mathcal{X}}(X, \cdot) - \mu_{P_X}\right) \otimes \left(k^{\mathcal{Y}}(Y, \cdot) - \mu_{P_Y}\right) \right) g \right\rangle_{\mathcal{H}_{k^{\mathcal{X}}}} \right]
\end{aligned}$$

for all $f \in \mathcal{H}_k$ and all $g \in \mathcal{H}_{k^{\mathcal{Y}}}$, where $\langle \cdot, \cdot \rangle_{\text{HS}}$ denotes the Hilbert–Schmidt inner product. The existence and uniqueness of $\mathcal{C}_{P_{XY}}$ is guaranteed by Riesz Representation Theorem. The notation

$$\mathcal{C}_{P_{XY}} = \mathbb{E}_{P_{XY}}\left[\left(k^{\mathcal{X}}(X, \cdot) - \mu_{P_X}\right) \otimes \left(k^{\mathcal{Y}}(Y, \cdot) - \mu_{P_Y}\right)\right]$$

is justified in the sense that

$$\left\langle f, \mathbb{E}_{P_{XY}}\left[\left(k^{\mathcal{X}}(X, \cdot) - \mu_{P_X}\right) \otimes \left(k^{\mathcal{Y}}(Y, \cdot) - \mu_{P_Y}\right)\right] g \right\rangle_{\mathcal{H}_{k^{\mathcal{X}}}}$$

$$= \mathbb{E}_{P_{XY}}\left[\left\langle f, \Big(\big(k^{\mathcal{X}}(X,\cdot) - \mu_{P_X}\big) \otimes \big(k^{\mathcal{Y}}(Y,\cdot) - \mu_{P_Y}\big)\Big) g \right\rangle_{\mathcal{H}_{k^{\mathcal{X}}}}\right]$$

which gives

$$\begin{aligned} &\left\langle f \otimes g, \mathbb{E}_{P_{XY}}\left[\big(k^{\mathcal{X}}(X,\cdot) - \mu_{P_X}\big) \otimes \big(k^{\mathcal{Y}}(Y,\cdot) - \mu_{P_Y}\big)\right]\right\rangle_{\mathrm{HS}} \\ &= \mathbb{E}_{P_{XY}}\left[\left\langle f \otimes g, \big(k^{\mathcal{X}}(X,\cdot) - \mu_{P_X}\big) \otimes \big(k^{\mathcal{Y}}(Y,\cdot) - \mu_{P_Y}\big)\right\rangle_{\mathrm{HS}}\right]. \end{aligned} \tag{2}$$

Characteristic kernel. A kernel k is *characteristic* (Fukumizu et al., 2004; Sriperumbudur et al., 2008, 2010b, 2011) if $\mu_P = \mu_Q$ implies $P = Q$, where $\mu_P = \mathbb{E}_P[k(X,\cdot)]$. In other words, the kernel mean embedding captures all the information about the distribution, in the sense that, if two kernel mean embeddings are the same (*i.e.* $\mu_P = \mu_Q$), then the distributions must be the same (*i.e.* $P = Q$).

Universal kernel. A kernel $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is C_0-*universal* (Carmeli et al., 2010; Sriperumbudur et al., 2010a) if its associated RKHS $\mathcal{H}_k$ is dense in $C_0(\mathcal{X}, \mathbb{R})$ (*i.e.* the space of continuous functions from $\mathcal{X}$ to $\mathbb{R}$ vanishing at infinity). There exist various notions of universality, we refer the reader to Sriperumbudur et al. (2011) for details.

Translation invariance, radial kernels & bandwidths. A kernel $k\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is *translation-invariant* if

$$k(x,y) = \psi(x-y)$$

for some (positive definite) function $\psi\colon \mathbb{R}^d \to \mathbb{R}$, often required to satisfy $\psi(0) = 1$. Then, for any *bandwidth* $\lambda > 0$, the scaled function

$$k_\lambda(x,y) = \psi\left(\frac{x-y}{\lambda}\right)$$

is also a kernel as it is equal to $k\big(\frac{x}{\lambda}, \frac{y}{\lambda}\big)$. We say that $k\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a *radial kernel*[1] if

$$k(x,y) = \Psi(\|x-y\|_r)$$

for some $r \geq 1$ and some function $\Psi\colon \mathbb{R} \to \mathbb{R}$ with $\Psi(0) = 1$, giving

$$k_\lambda(x,y) = \Psi\left(\frac{\|x-y\|_r}{\lambda}\right).$$

Note that $k_\lambda(x,x) = 1$ for all $x \in \mathbb{R}^d$ and all $\lambda > 0$. For $x \neq y$ both in $\mathbb{R}^d$, we have

$$k_\lambda(x,y) \to 0 \text{ as } \lambda \to 0 \qquad \text{and} \qquad k_\lambda(x,y) \to 1 \text{ as } \lambda \to \infty. \tag{3}$$

Kernel examples. We now present some commonly used kernels which are characteristic and C_0-universal: the *Gausssian kernel*

$$k_\lambda(x,y) = \exp\left(-\frac{\|x-y\|_2^2}{\lambda^2}\right),$$

the *Laplace kernel*

$$k_\lambda(x,y) = \exp\left(-\frac{\|x-y\|_1}{\lambda}\right),$$

[1] In the literature, a radial kernel is sometimes defined only for the special case $r = 2$.

the *inverse multiquadric IMQ kernel*

$$k_\lambda(x,y) = \frac{1}{\sqrt{\lambda^2 + \|x-y\|_2^2}} \propto \frac{1}{\sqrt{1 + \frac{\|x-y\|_2^2}{\lambda^2}}},$$

and the *Matérn kernels* with $\nu = 0.5, 1.5, 2.5, 3.5, 4.5$ and L^r-distances for $r \geq 1$

$$k_\lambda(x,y) = \exp\left(-\frac{\|x-y\|_r}{\lambda}\right),$$

$$k_\lambda(x,y) = \left(1 + \sqrt{3}\frac{\|x-y\|_r}{\lambda}\right)\exp\left(-\sqrt{3}\frac{\|x-y\|_r}{\lambda}\right),$$

$$k_\lambda(x,y) = \left(1 + \sqrt{5}\frac{\|x-y\|_r}{\lambda} + \frac{5}{3}\frac{\|x-y\|_r^2}{\lambda^2}\right)\exp\left(-\sqrt{5}\frac{\|x-y\|_r}{\lambda}\right),$$

$$k_\lambda(x,y) = \left(1 + \sqrt{7}\frac{\|x-y\|_r}{\lambda} + \frac{2\cdot 7}{5}\frac{\|x-y\|_r^2}{\lambda^2} + \frac{7\sqrt{7}}{3\cdot 5}\frac{\|x-y\|_r^3}{\lambda^3}\right)\exp\left(-\sqrt{7}\frac{\|x-y\|_r}{\lambda}\right),$$

$$k_\lambda(x,y) = \left(1 + 3\frac{\|x-y\|_r}{\lambda} + \frac{3\cdot 6^2}{28}\frac{\|x-y\|_r^2}{\lambda^2} + \frac{6^3}{84}\frac{\|x-y\|_r^3}{\lambda^3} + \frac{6^4}{1680}\frac{\|x-y\|_r^4}{\lambda^4}\right)\exp\left(-3\frac{\|x-y\|_r}{\lambda}\right).$$

2 Kernel discrepancies

We introduce the Maximum Mean Discrepancy (MMD) in Section 2.1, the Hilbert–Schmidt Independence Criterion (HSIC) in Section 2.2, and the Kernel Stein Discrepancy (KSD) in Section 2.3.

2.1 MMD: Maximum Mean Discrepancy

MMD measure. As a measure between two probability distributions P and Q, we consider the kernel-based *Maximum Mean Discrepancy* (MMD—Gretton et al., 2007, 2012). For a given RKHS $\mathcal{H}_k$ with reproducing kernel k, the MMD is defined as the integral probability metric (Müller, 1997)

$$\mathrm{MMD}_k(P,Q) \coloneqq \sup_{f\in\mathcal{H}_k\,:\,\|f\|_{\mathcal{H}_k}\leq 1} \mathbb{E}_{X\sim P}[f(X)] - \mathbb{E}_{Y\sim Q}[f(Y)]. \tag{4}$$

We often simply write MMD_k for $\mathrm{MMD}_k(P,Q)$ when the distributions are clear from the context, and similarly for other discrepancies. Using the reproducibility property, we obtain

$$\mathrm{MMD}_k = \sup_{f\in\mathcal{H}_k\,:\,\|f\|_{\mathcal{H}_k}\leq 1} \langle f, \mu_P - \mu_Q\rangle_{\mathcal{H}_k} = \|\mu_P - \mu_Q\|_{\mathcal{H}_k} \tag{5}$$

and

$$\mathrm{MMD}_k^2 = \|\mu_P - \mu_Q\|_{\mathcal{H}_k}^2 = \mathbb{E}_{P,P}[k(X,X')] - 2\mathbb{E}_{P,Q}[k(X,Y)] + \mathbb{E}_{Q,Q}[k(Y,Y')] \tag{6}$$

by the properties of kernel mean embeddings, where $X, X' \sim P$ and $Y, Y' \sim Q$ are independent copies. We observe that the MMD is the $\mathcal{H}_k$-norm of the difference between the mean embeddings. We note that the MMD can be leveraged to construct divergences for more general two-sample problems (Chau et al., 2025).

MMD V-statistic. We now introduce some estimators of the MMD given some independent samples $X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} P$ and $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} Q$. We let $\widehat{P}$ and $\widehat{Q}$ denote the empirical distributions (uniform

distributions on the datapoints). The plug-in estimator (Gretton et al., 2012, Equations 2 and 5) for $\mathrm{MMD}_k^2(P, Q)$ is $\mathrm{MMD}_k^2(\widehat{P}, \widehat{Q})$ which from Equation 6 is equal to

$$V_{\mathrm{MMD}_k^2} \coloneqq \frac{1}{m^2} \sum_{1 \le i,i' \le m} k(X_i, X_{i'}) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(X_i, Y_j) + \frac{1}{n^2} \sum_{1 \le j,j' \le n} k(Y_j, Y_{j'}) \tag{7}$$

which can be expressed as a two-sample (both of second order) V-statistic (Lee, 1990)

$$V_{\mathrm{MMD}_k^2} = \frac{1}{m^2 n^2} \sum_{1 \le i,i' \le m} \sum_{1 \le j,j' \le n} h_k^{\mathrm{MMD}}(X_i, X_{i'}; Y_j, Y_{j'}) \tag{8}$$

with core function[2]

$$h_k^{\mathrm{MMD}}(x, x'; y, y') \coloneqq k(x, x') - k(x', y) - k(x, y') + k(y, y') \tag{9}$$

for $x, x', y, y' \in \mathbb{R}^d$. Writing the estimator $V_{\mathrm{MMD}_k^2}$ as a two-sample V-statistic can be theoretically appealing but we stress that it can in fact be computed in quadratic time using Equation 7. The V-statistic incorporates the terms $\{k(X_i, Y_i) : i = 1, \dots, n\}$ and, hence, is biased. We also point out that the V-statistic is always non-negative, and taking its square root yields an estimator of the (non-squared) MMD. The global sensitivity of this MMD statistic is studied in Kim and Schrab (2023, Lemma 5), which provides robustness guarantees (Schrab and Kim, 2025).

Writing $Z_i = X_i$ for $i = 1, \dots, m$, $Z_{m+j} = Y_j$ for $j = 1, \dots, n$, and considering the $(m+n) \times (m+n)$ kernel matrix $K^{ZZ} = \big(k(Z_i, Z_j)\big)_{1 \le i,j \le m+n}$, the MMD V-statistic can be computed as

$$V_{\mathrm{MMD}_k^2} = w^\top K^{ZZ} w \tag{10}$$

where w is a vector of size $m+n$ with $w_i = 1/m$ for $i = 1, \dots, m$ and $w_{m+j} = -1/n$ for $j = 1, \dots, n$.

When the sample sizes are equal $m = n$, the estimator reduces to a one-sample second-order V-statistic

$$V_{\mathrm{MMD}_k^2} = \frac{1}{n^2} \sum_{1 \le i,i' \le n} h_k^{\mathrm{MMD}}(X_i, X_{i'}; Y_i, Y_{i'}). \tag{11}$$

MMD U-statistic. An unbiased estimator of the squared MMD (Gretton et al., 2012, Lemma 6) naturally arises from Equation 6 as

$$U_{\mathrm{MMD}_k^2} \coloneqq \frac{1}{m(m-1)} \sum_{1 \le i \ne i' \le m} k(X_i, X_{i'}) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(X_i, Y_j) + \frac{1}{n(n-1)} \sum_{1 \le j \ne j' \le n} k(Y_j, Y_{j'}), \tag{12}$$

this is actually the minimum variance unbiased MMD estimator (Serfling, 1980, Section 5.1.4). It can be expressed as a two-sample (both of second order) U-statistic (Hoeffding, 1948)

$$U_{\mathrm{MMD}_k^2} = \frac{1}{m(m-1)n(n-1)} \sum_{1 \le i \ne i' \le m} \sum_{1 \le j \ne j' \le n} h_k^{\mathrm{MMD}}(X_i, X_{i'}; Y_j, Y_{j'}). \tag{13}$$

The MMD U-statistic expression of Equation 12 cannot be expressed as a single vector-matrix-vector product, instead it needs to be computed as

$$U_{\mathrm{MMD}_k^2} = \frac{1}{m(m-1)} \mathbf{1}^\top \bar{K}^{XX} \mathbf{1} - \frac{2}{mn} \mathbf{1}^\top K^{XY} \mathbf{1} + \frac{1}{n(n-1)} \mathbf{1}^\top \bar{K}^{YY} \mathbf{1} \tag{14}$$

[2] This is more commonly referred to as a 'kernel' in the litterature, we use the term 'core' to avoid confusion with the positive-definite kernel k.

where $K^{XX} = \big(k(X_i, X_j)\big)_{1\leq i,j\leq m}$, $K^{XY} = \big(k(X_i, Y_j)\big)_{1\leq i\leq m, 1\leq j\leq n}$, $K^{YY} = \big(k(Y_i, Y_j)\big)_{1\leq i,j\leq n}$, where $\bar{K}^{XX}$ and $\bar{K}^{YY}$ denote the matrices K^{XX} and K^{YY} with diagonal entries set to zero, and where $\mathbf{1}$ is a vector of ones of size either m or n depending on the context. Efficient implementations of the MMD U-statistic are further discussed in Schrab et al. (2022b).

When $m = n$, not incorporating the terms $\{k(X_i, Y_i) : i = 1, \ldots, n\}$ (note that the order of the samples matters) in the statistic computation results in a simpler one-sample second-order U-statistic (Gretton et al., 2012, Equation 4)

$$\widetilde{U}_{\mathrm{MMD}_k^2} = \frac{1}{n(n-1)} \sum_{1\leq i\neq i'\leq n} h_k^{\mathrm{MMD}}(X_i, X_{i'}; Y_i, Y_{i'}) \tag{15}$$

which can be computed as

$$\widetilde{U}_{\mathrm{MMD}_k^2} = \frac{1}{n(n-1)} v^\top \bar{K}^{ZZ} v \tag{16}$$

where $\bar{K}^{ZZ}$ is the matrix K^{ZZ} as defined in Equation 10 but with diagonal entries set to zero, and $v \coloneqq (\mathbf{1}_n, -\mathbf{1}_n)$ is a vector of signed ones. As a result of the U-statistic being unbiased, it is not always non-negative and hence cannot be used to estimate the MMD by taking its square root (as for the V-statistic).

MMD kernel choice importance. When using the translation-invariant kernel k_λ, Equation 3 implies[3]

$$\begin{aligned}
&\mathrm{MMD}_\lambda^2 \to 0 \text{ when } \lambda \to 0 \text{ or } \lambda \to \infty, \\
&V_{\mathrm{MMD}_\lambda^2} \to \frac{1}{m} + \frac{1}{n} \text{ when } \lambda \to 0, \quad \text{and} \quad V_{\mathrm{MMD}_\lambda^2} \to 0 \text{ when } \lambda \to \infty, \\
&U_{\mathrm{MMD}_\lambda^2} \to 0 \text{ and } \widetilde{U}_{\mathrm{MMD}_\lambda^2} \to 0 \quad \text{when} \quad \lambda \to 0 \text{ or } \lambda \to \infty.
\end{aligned} \tag{17}$$

We emphasize that this holds regardless of the relation between the distributions P and Q, so even when $P \neq Q$ if the bandwidth λ is not well-calibrated (in the sense that it is either too small or too large) then the estimated MMD can be very close to zero which would fail to capture the difference in distributions (even for characteristic kernel k_λ). This observation really highlights the importance of the choice of kernel bandwidth (and more generally of kernel) in the MMD computation.

2.2 HSIC: Hilbert–Schmidt Independence Criterion

HSIC measure. For a joint probability density P_{XY} on $\mathcal{X} \times \mathcal{Y}$ with marginals P_X on $\mathcal{X}$ and P_Y on $\mathcal{Y}$, we quantify the dependence with the *Hilbert–Schmidt Independence Criterion* (HSIC) introduced by Gretton et al. (2005), which is defined as the Hilbert–Schmidt norm of the cross-covariance operator, that is[4]

$$\begin{aligned}
&\mathrm{HSIC}^2_{k^\mathcal{X}, k^\mathcal{Y}}(P_{XY}) \\
&\coloneqq \|\mathcal{C}_{P_{XY}}\|^2_{\mathrm{HS}} \\
&= \big\langle \mathcal{C}_{P_{XY}}, \mathcal{C}_{P_{XY}} \big\rangle_{\mathrm{HS}} \\
&= \Big\langle \mathbb{E}_{P_{XY}}\big[\big(k^\mathcal{X}(X,\cdot) - \mu_{P_X}\big) \otimes \big(k^\mathcal{Y}(Y,\cdot) - \mu_{P_Y}\big)\big], \mathbb{E}_{P_{XY}}\big[\big(k^\mathcal{X}(X',\cdot) - \mu_{P_X}\big) \otimes \big(k^\mathcal{Y}(Y',\cdot) - \mu_{P_Y}\big)\big] \Big\rangle_{\mathrm{HS}} \\
&= \mathbb{E}_{P_{XY}, P_{XY}}\Big[\big\langle \big(k^\mathcal{X}(X,\cdot) - \mu_{P_X}\big) \otimes \big(k^\mathcal{Y}(Y,\cdot) - \mu_{P_Y}\big), \big(k^\mathcal{X}(X',\cdot) - \mu_{P_X}\big) \otimes \big(k^\mathcal{Y}(Y',\cdot) - \mu_{P_Y}\big) \big\rangle_{\mathrm{HS}}\Big] \\
&= \mathbb{E}_{P_{XY}, P_{XY}}\Big[\big\langle \big(k^\mathcal{X}(X,\cdot) - \mu_{P_X}\big), \big(k^\mathcal{X}(X',\cdot) - \mu_{P_X}\big) \big\rangle_{\mathcal{H}_{k^\mathcal{X}}} \big\langle \big(k^\mathcal{Y}(Y,\cdot) - \mu_{P_Y}\big), \big(k^\mathcal{Y}(Y',\cdot) - \mu_{P_Y}\big) \big\rangle_{\mathcal{H}_{k^\mathcal{Y}}}\Big] \\
&= \mathbb{E}_{P_{XY}, P_{XY}}\Big[\Big(k^\mathcal{X}(X,X') - \mathbb{E}_X\big[k^\mathcal{X}(X,X')\big] - \mathbb{E}_{X'}\big[k^\mathcal{X}(X,X')\big] + \mathbb{E}_{X,X'}\big[k^\mathcal{X}(X,X')\big]\Big)
\end{aligned}$$

[3] We use the convention that $(\cdot)_\lambda$ denotes $(\cdot)_{k_\lambda}$ throughout the thesis.

[4] The HSIC is most commonly defined without the square, however, we choose to define it as such for consistence with the MMD and KSD discrepancies.

$$
\begin{aligned}
&\Big(k^{\mathcal{Y}}(Y,Y') - \mathbb{E}_Y\big[k^{\mathcal{Y}}(Y,Y')\big] - \mathbb{E}_{Y'}\big[k^{\mathcal{Y}}(Y,Y')\big] + \mathbb{E}_{Y,Y'}\big[k^{\mathcal{Y}}(Y,Y')\big]\Big)\Big] \\
&= \mathbb{E}_{P_{XY},P_{XY}}\Big[k^{\mathcal{X}}(X,X')\Big(k^{\mathcal{Y}}(Y,Y') - \mathbb{E}_Y\big[k^{\mathcal{Y}}(Y,Y')\big] - \mathbb{E}_{Y'}\big[k^{\mathcal{Y}}(Y,Y')\big] + \mathbb{E}_{Y,Y'}\big[k^{\mathcal{Y}}(Y,Y')\big]\Big)\Big] \\
&= \mathbb{E}_{P_{XY},P_{XY}}\Big[k^{\mathcal{X}}(X,X')k^{\mathcal{Y}}(Y,Y')\Big] - 2\,\mathbb{E}_{P_{XY}}\Big[\mathbb{E}_{P_X}[k^{\mathcal{X}}(X,X')]\mathbb{E}_{P_Y}[k^{\mathcal{Y}}(Y,Y')]\Big] \\
&\qquad + \mathbb{E}_{P_X,P_X}\Big[k^{\mathcal{X}}(X,X')\Big]\mathbb{E}_{P_Y,P_Y}\Big[k^{\mathcal{Y}}(Y,Y')\Big]
\end{aligned} \tag{18}
$$

for kernels $k^{\mathcal{X}}$ and $k^{\mathcal{Y}}$ on $\mathcal{X}\times\mathcal{X}$ and $\mathcal{Y}\times\mathcal{Y}$, respectively, where we have used the property of the cross-covariance operator shown in Equation 2. We also mention the related conditional HSIC quantities of Zhang et al. (2011) and Pogodin et al. (2022, 2024).

HSIC V-statistic. We now present some HSIC estimators given i.i.d. paired samples $\big((X_i,Y_i)\big)_{i=1}^N$ drawn from P_{XY}. For convenience, we use the notation $Z_i=(X_i,Y_i)$ for $i=1,\dots,N$. We also denote by $\widehat{P}_{XY}$ the empirical distribution of the joint. For notational purposes, we let $K^{\mathcal{X}}_{ij}$ and $K^{\mathcal{Y}}_{ij}$ denote $k^{\mathcal{X}}(X_i,X_j)$ and $k^{\mathcal{Y}}(Y_i,Y_j)$, respectively, for all $1\le i,j\le N$. The plug-in estimator (Gretton et al., 2008, Equation 4) of $\mathrm{HSIC}^2_{k^{\mathcal{X}}\times k^{\mathcal{Y}}}(P_{XY})$ is $\mathrm{HSIC}^2_{k^{\mathcal{X}}\times k^{\mathcal{Y}}}(\widehat{P}_{XY})$ which is equal to

$$
\begin{aligned}
V_{\mathrm{HSIC}^2_{k^{\mathcal{X}},k^{\mathcal{Y}}}} &\coloneqq \frac{1}{N^2}\sum_{1\le i,j\le N} K^{\mathcal{X}}_{ij}K^{\mathcal{Y}}_{ij} - \frac{2}{N}\sum_{i=1}^N\Big(\frac{1}{N}\sum_{j=1}^N K^{\mathcal{X}}_{ij}\Big)\Big(\frac{1}{N}\sum_{r=1}^N K^{\mathcal{Y}}_{ir}\Big) + \Big(\frac{1}{N^2}\sum_{1\le i,j\le N} K^{\mathcal{X}}_{ij}\Big)\Big(\frac{1}{N^2}\sum_{1\le r,s\le N} K^{\mathcal{Y}}_{rs}\Big) \\
&= \frac{1}{N^2}\sum_{1\le i,j\le N} K^{\mathcal{X}}_{ij}K^{\mathcal{Y}}_{ij} - \frac{2}{N^3}\sum_{1\le i,j,r\le N} K^{\mathcal{X}}_{ij}K^{\mathcal{Y}}_{ir} + \frac{1}{N^4}\sum_{1\le i,j,r,s\le N} K^{\mathcal{X}}_{ij}K^{\mathcal{Y}}_{rs} \\
&= \frac{1}{N^4}\sum_{1\le i,j,r,s\le N} K^{\mathcal{X}}_{ij}\Big(K^{\mathcal{Y}}_{ij} - K^{\mathcal{Y}}_{is} - K^{\mathcal{Y}}_{rj} + K^{\mathcal{Y}}_{rs}\Big) \qquad (19) \\
&= \frac{1}{N^4}\sum_{1\le i,j,r,s\le N} \frac{1}{4}\Big(K^{\mathcal{X}}_{ij} - K^{\mathcal{X}}_{is} - K^{\mathcal{X}}_{rj} + K^{\mathcal{X}}_{rs}\Big)\Big(K^{\mathcal{Y}}_{ij} - K^{\mathcal{Y}}_{is} - K^{\mathcal{Y}}_{rj} + K^{\mathcal{Y}}_{rs}\Big). \qquad (20)
\end{aligned}
$$

So, this HSIC estimator can be expressed as a one-sample fourth-order V-statistic

$$
V_{\mathrm{HSIC}^2_{k^{\mathcal{X}},k^{\mathcal{Y}}}} = \frac{1}{N^4}\sum_{1\le i,j,r,s\le N} h^{\mathrm{HSIC}}_{k^{\mathcal{X}},k^{\mathcal{Y}}}(Z_i,Z_j,Z_r,Z_s) \tag{21}
$$

where the core HSIC function can either be defined as

$$
h^{\mathrm{HSIC}}_{k^{\mathcal{X}},k^{\mathcal{Y}}}(Z_i,Z_j,Z_r,Z_s) = K^{\mathcal{X}}_{ij}\Big(K^{\mathcal{Y}}_{ij} - K^{\mathcal{Y}}_{is} - K^{\mathcal{Y}}_{rj} + K^{\mathcal{Y}}_{rs}\Big) = k^{\mathcal{X}}(X_i,X_j)h^{\mathrm{MMD}}_{k^{\mathcal{Y}}}(Y_i,Y_j;Y_r,Y_s) \tag{22}
$$

or

$$
\begin{aligned}
h^{\mathrm{HSIC}}_{k^{\mathcal{X}},k^{\mathcal{Y}}}(Z_i,Z_j,Z_r,Z_s) &= \frac{1}{4}\Big(K^{\mathcal{X}}_{ij} - K^{\mathcal{X}}_{is} - K^{\mathcal{X}}_{rj} + K^{\mathcal{X}}_{rs}\Big)\Big(K^{\mathcal{Y}}_{ij} - K^{\mathcal{Y}}_{is} - K^{\mathcal{Y}}_{rj} + K^{\mathcal{Y}}_{rs}\Big) \\
&= \frac{1}{4}h^{\mathrm{MMD}}_{k^{\mathcal{X}}}(X_i,X_j;X_r,X_s)h^{\mathrm{MMD}}_{k^{\mathcal{Y}}}(Y_i,Y_j;Y_r,Y_s),
\end{aligned} \tag{23}
$$

where the second expression has the benefit of being symmetric in the samples but this comes at the expense of computing four times the same quantities. In this thesis, we will use the second expression for the core HSIC function. Again, this estimator is non-negative and biased (as it includes the terms with same indices), its square root can be taken to estimate the (non-squared) HSIC directly, and its global sensitivity is studied in Kim and Schrab (2023, Lemma 6). We stress that the HISC estimator $V_{\mathrm{HSIC}^2_{k^{\mathcal{X}},k^{\mathcal{Y}}}}$ can be computed in

quadratic time and admits the following closed-form expression (Gretton et al., 2008, Equation 4)

$$V_{\mathrm{HSIC}^2_{k^{\mathcal{X}},k^{\mathcal{Y}}}} = \frac{1}{N^2}\mathrm{tr}\big(K^{\mathcal{X}}HK^{\mathcal{Y}}H\big) \tag{24}$$

where $K^{\mathcal{X}} = \big(K^{\mathcal{X}}_{ij}\big)_{1\leq i,j\leq N}$, $K^{\mathcal{Y}} = \big(K^{\mathcal{Y}}_{ij}\big)_{1\leq i,j\leq N}$ and $H = I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ is the centering matrix with I the identity matrix and $\mathbf{1}$ a vector of ones, and where tr denotes the trace of a matrix. Finally, we present a one-sample second-order V-statistic for the HSIC, which does not consider all possible terms (unlike the one presented above) but which is useful to construct efficient estimators with faster computation time, as discussed in Section 3,

$$\widetilde{V}_{\mathrm{HSIC}^2_{k^{\mathcal{X}},k^{\mathcal{Y}}}} = \frac{1}{N^2}\sum_{1\leq i,j\leq N} h^{\mathrm{HSIC}}_{k^{\mathcal{X}},k^{\mathcal{Y}}}(Z_i, Z_j, Z_{i+N/2}, Z_{j+N/2}) \tag{25}$$

where the indices are taken modulo N which is here assumed to be even. While shifting the indices by other quantities than $N/2$ is possible, this choice turns out to be particularly useful due to the property that adding this shift twice to an index simply leaves the index unchanged (useful in the setting of Schrab et al., 2022b).

HSIC U-statistic. A natural unbiased HSIC estimator (Gretton et al., 2008; Song et al., 2012) is the minimum variance one-sample fourth-order U-statistic

$$\begin{aligned} U_{\mathrm{HSIC}^2_{k^{\mathcal{X}},k^{\mathcal{Y}}}} &\coloneqq \frac{1}{\left|\mathbf{i}^N_2\right|}\sum_{(i,j)\in\mathbf{i}^N_2} K^{\mathcal{X}}_{ij}K^{\mathcal{Y}}_{ij} - \frac{2}{\left|\mathbf{i}^N_3\right|}\sum_{(i,j,r)\in\mathbf{i}^N_3} K^{\mathcal{X}}_{ij}K^{\mathcal{Y}}_{ir} + \frac{1}{\left|\mathbf{i}^N_4\right|}\sum_{(i,j,r,s)\in\mathbf{i}^N_4} K^{\mathcal{X}}_{ij}K^{\mathcal{Y}}_{rs} \\ &= \frac{1}{\left|\mathbf{i}^N_4\right|}\sum_{(i,j,r,s)\in\mathbf{i}^N_4} h^{\mathrm{HSIC}}_{k^{\mathcal{X}},k^{\mathcal{Y}}}(Z_i, Z_j, Z_r, Z_s). \end{aligned} \tag{26}$$

Here, $\mathbf{i}^N_r$ denotes the set of all r-tuples drawn without replacement from $\{1,\dots,N\}$ so that $\left|\mathbf{i}^N_r\right| = N\cdots(N-r+1)$, for example $\mathbf{i}^N_2 = \{(i,j) : 1\leq i\neq j\leq N\}$ and $\left|\mathbf{i}^N_2\right| = N(N-1)$.

We stress the fact that this HSIC U-statistic can actually be computed in quadratic time as shown by Song et al. (2012, Equation 5) who provide the following closed-form expression

$$U_{\mathrm{HSIC}^2_{k^{\mathcal{X}},k^{\mathcal{Y}}}} = \frac{1}{N(N-3)}\left(\mathrm{tr}\big(\bar{K}^{\mathcal{X}}\bar{K}^{\mathcal{Y}}\big) + \frac{\mathbf{1}^\top\bar{K}^{\mathcal{X}}\mathbf{1}\mathbf{1}^\top\bar{K}^{\mathcal{Y}}\mathbf{1}}{(N-1)(N-2)} - \frac{2}{N-2}\mathbf{1}^\top\bar{K}^{\mathcal{X}}\bar{K}^{\mathcal{Y}}\mathbf{1}\right) \tag{27}$$

where $\bar{K}^{\mathcal{X}}$ and $\bar{K}^{\mathcal{Y}}$ are the kernel matrices $K^{\mathcal{X}}$ and $K^{\mathcal{Y}}$ with diagonal entries set to 0.

HSIC kernel choice importance. For translation-invariant kernel $k^{\mathcal{X}}_\lambda$ and $k^{\mathcal{Y}}_\mu$ with bandwidths λ and μ, Equation 3 implies[5]

$$\begin{aligned} &\mathrm{HSIC}^2_{\lambda,\mu} \to 0 \text{ when } \lambda\to 0 \text{ or } \mu\to 0 \text{ or } \lambda\to\infty \text{ or } \mu\to\infty, \\ &V_{\mathrm{HSIC}^2_{\lambda,\mu}} \to \frac{1}{N} - \frac{1}{N^2} \text{ when } \lambda\to 0 \text{ and } \mu\to 0, \quad \text{and} \quad V_{\mathrm{HSIC}^2_{\lambda,\mu}} \to 0 \text{ when } \lambda\to\infty \text{ or } \mu\to\infty, \\ &U_{\mathrm{HSIC}^2_{\lambda,\mu}} \to 0 \text{ when } \lambda\to 0 \text{ or } \mu\to 0 \text{ or } \lambda\to\infty \text{ or } \mu\to\infty. \end{aligned} \tag{28}$$

Again, we stress that this holds regardless of the potential dependence in the joint distribution. This means that even if strong dependence exists, it will be failed to be captured by the (estimated) HSIC if either of the kernel bandwidths are not well-calibrated (in the sense that they are either too small or too large). These remarks emphasize the crucial role of kernel selection when using HSIC in practical applications.

[5] We use the convention that $(\cdot)_{\lambda,\mu}$ denotes $(\cdot)_{k^{\mathcal{X}}_\lambda\times k^{\mathcal{Y}}_\mu}$.

HSIC as an MMD. First, we note that the HSIC is an MMD between the joint and the product of the marginals using a product kernel defined as $(k^{\mathcal{X}} \times k^{\mathcal{Y}})\big((x,y),(x',y')\big) \coloneqq k^{\mathcal{X}}(x,x')k^{\mathcal{Y}}(y,y')$ for any $(x,y),(x',y') \in \mathcal{X} \times \mathcal{Y}$, that is

$$\begin{aligned}
&\mathrm{MMD}^2_{k^{\mathcal{X}}\times k^{\mathcal{Y}}}(P_{XY}, P_X \otimes P_Y) \\
&= \mathbb{E}_{P_{XY},P_{XY}}[k^{\mathcal{X}}(X,X')k^{\mathcal{Y}}(Y,Y')] - 2\mathbb{E}_{P_{XY},P_XP_Y}[k^{\mathcal{X}}(X,X')k^{\mathcal{Y}}(Y,Y')] + \mathbb{E}_{P_XP_Y,P_XP_Y}[k^{\mathcal{X}}(X,X')k^{\mathcal{Y}}(Y,Y')] \\
&= \mathbb{E}_{P_{XY},P_{XY}}\Big[k^{\mathcal{X}}(X,X')k^{\mathcal{Y}}(Y,Y')\Big] - 2\,\mathbb{E}_{P_{XY}}\Big[\mathbb{E}_{P_X}[k^{\mathcal{X}}(X,X')]\mathbb{E}_{P_Y}[k^{\mathcal{Y}}(Y,Y')]\Big] \\
&\qquad\qquad + \mathbb{E}_{P_X,P_X}\Big[k^{\mathcal{X}}(X,X')\Big]\mathbb{E}_{P_Y,P_Y}\Big[k^{\mathcal{Y}}(Y,Y')\Big] \\
&= \mathrm{HSIC}^2_{k^{\mathcal{X}},k^{\mathcal{Y}}}(P_{XY}).
\end{aligned} \tag{29}$$

MMD as an HSIC. Now, consider the two-sample problem again with distributions P, Q, where for clarity we use variables $A, A' \overset{\text{i.i.d.}}{\sim} P$ and $B, B' \sim Q$. Construct a joint distribution P_{XY} with marginal $P_X = w_P P + w_Q Q$ a mixture of P and Q with positive weights $w_P + w_Q = 1$, and set $Y = 1$ if X is drawn from P, or $Y = -1$ if X is drawn from Q. For the labels, use the indicator kernel $k^{\mathcal{Y}}(y,y') = \mathbf{1}(y = y')$. For the data, we simply set the kernel to be the one used for two-sample testing, that is $k^{\mathcal{Y}}(x,x') = k(x,x')$. Then, we observe that

$$\mathrm{HSIC}^2_{k^{\mathcal{X}},k^{\mathcal{Y}}}(P_{XY}) = (\mathrm{I}) + (\mathrm{II}) + (\mathrm{III})$$

where

$$(\mathrm{I}) = \mathbb{E}_{P_{XY},P_{XY}}\Big[k^{\mathcal{X}}(X,X')k^{\mathcal{Y}}(Y,Y')\Big] = w_P^2\,\mathbb{E}_{P,P}\Big[k(A,A')\Big] + w_Q^2\,\mathbb{E}_{Q,Q}\Big[k(B,B')\Big],$$

and

$$\begin{aligned}
(\mathrm{II}) &= -2\,\mathbb{E}_{P_{XY}}\Big[\mathbb{E}_{P_X}[k^{\mathcal{X}}(X,X')]\mathbb{E}_{P_Y}[k^{\mathcal{Y}}(Y,Y')]\Big] \\
&= -2w_P^2\left(w_P\,\mathbb{E}_{P,P}\Big[k(A,A')\Big] + w_Q\,\mathbb{E}_{P,Q}\Big[k(A,B)\Big]\right) - 2w_Q^2\left(w_Q\,\mathbb{E}_{Q,Q}\Big[k(B,B')\Big] + w_P\,\mathbb{E}_{P,Q}\Big[k(A,B)\Big]\right) \\
&= -2\left(w_P^3\,\mathbb{E}_{P,P}\Big[k(A,A')\Big] + w_Q^3\,\mathbb{E}_{Q,Q}\Big[k(B,B')\Big] + (w_P^2 w_Q + w_P w_Q^2)\,\mathbb{E}_{P,Q}\Big[k(A,B)\Big]\right),
\end{aligned}$$

and

$$\begin{aligned}
(\mathrm{III}) &= \mathbb{E}_{P_X,P_X}\Big[k^{\mathcal{X}}(X,X')\Big]\mathbb{E}_{P_Y,P_Y}\Big[k^{\mathcal{Y}}(Y,Y')\Big] \\
&= \left(w_P^2\,\mathbb{E}_{P,P}\Big[k(A,A')\Big] + w_Q^2\,\mathbb{E}_{Q,Q}\Big[k(B,B')\Big] + 2w_Pw_Q\,\mathbb{E}_{P,Q}\Big[k(A,B)\Big]\right)\left(w_P^2 + w_Q^2\right).
\end{aligned}$$

Combining these expressions, we obtain

$$\begin{aligned}
\mathrm{HSIC}^2_{k,\mathbf{1}}(P_{XY}) &= 2w_P^2w_Q^2\left(\mathbb{E}_{P,P}\Big[k(A,A')\Big] - 2\,\mathbb{E}_{P,Q}\Big[k(A,B)\Big] + \mathbb{E}_{Q,Q}\Big[k(B,B')\Big]\right) \\
&= 2w_P^2w_Q^2\,\mathrm{MMD}^2_k(P,Q).
\end{aligned} \tag{30}$$

In that setting, given m samples from P and n samples from Q, we have $w_P = m/(m+n), w_Q = n/(m+n)$, we similarly obtain

$$V_{\mathrm{HSIC}^2_{k,\mathbf{1}}} = \frac{2m^2n^2}{(m+n)^4}V_{\mathrm{MMD}^2_k}. \tag{31}$$

Noting that $k_\lambda(\cdot,\cdot) \to \mathbf{1}(\cdot = \cdot)$ as the bandwidth λ shrinks to 0, we also have

$$\mathrm{HSIC}^2_{k,k_\lambda}(P_{XY}) \to 2w_P^2w_Q^2\,\mathrm{MMD}^2_k(P,Q) \qquad \text{and} \qquad V_{\mathrm{HSIC}^2_{k,k_\lambda}} \to \frac{2m^2n^2}{(m+n)^4}V_{\mathrm{MMD}^2_k} \tag{32}$$

as $\lambda \to 0$.

2.3 KSD: Kernel Stein Discrepancy

SD measure. Stein's methods (Stein, 1972) have been widely used in the machine learning and statistics communities (see Anastasiou et al. (2021) for a recent review). At the heart of this field lies the concept of a Stein operator $\mathcal{A}_P \colon \mathcal{F} \to \mathcal{G}$ for function classes[6] $\mathcal{F} \subseteq \mathrm{Func}\big(\mathbb{R}^d \to \mathbb{R}^d\big)$ and $\mathcal{G} \subseteq \mathrm{Func}\big(\mathbb{R}^d \to \mathbb{R}\big)$, which is a linear operator satisfying Stein's identity (Stein, 1972; Stein et al., 2004)

$$P = Q \quad \Longleftrightarrow \quad \mathbb{E}_Q\big[(\mathcal{A}_P \boldsymbol{f})(X)\big] = 0 \;\; \text{for all} \;\; \boldsymbol{f} \in \mathcal{F}. \tag{33}$$

The Stein discrepancy (Gorham and Mackey, 2015) is then defined as the integral probability metric (using the range of $\mathcal{A}_P$ as the function class)

$$\mathrm{SD}_{\mathcal{A}_P}(P, Q) = \sup_{\boldsymbol{f} \in \mathcal{F}} \mathbb{E}_Q\big[(\mathcal{A}_P \boldsymbol{f})(X)\big] - \mathbb{E}_P\big[(\mathcal{A}_P \boldsymbol{f})(X)\big] = \sup_{\boldsymbol{f} \in \mathcal{F}} \mathbb{E}_Q\big[(\mathcal{A}_P \boldsymbol{f})(X)\big]. \tag{34}$$

Stein operators can be constructed from infinitesimal Markov process generators. In particular, assuming that the distribution P admits a density p with respect to the Lebesgue measure which is accessed only through the score function $\nabla \log p(x)$, starting from the overdamped Langevin equation leads to the (overdamped) Langevin Stein operator $\mathcal{A}_P^{\mathcal{L}}$ defined as (Gorham and Mackey, 2015, Equation 4)

$$(\mathcal{A}_P^{\mathcal{L}} \boldsymbol{f})(x) \coloneqq \boldsymbol{f}(x)^\top \nabla \log p(x) + \nabla^\top \boldsymbol{f}(x), \tag{35}$$

where $\nabla^\top \boldsymbol{f}(x) = \sum_{i=1}^d \frac{\partial}{\partial x_i} f_i(x)$ is the divergence of $\boldsymbol{f} = (f_1, \ldots, f_d)$ (*i.e.*, the trace of the Jacobian matrix of $\boldsymbol{f}$). The Langevin Stein operator can be expressed as a diffusion Stein operator (Gorham and Mackey, 2017, Section 3.1)

$$\begin{aligned}
(\mathcal{A}_P^{\mathcal{L}} \boldsymbol{f})(x) &= \boldsymbol{f}(x)^\top \nabla \log p(x) + \nabla^\top \boldsymbol{f}(x) \\
&= \boldsymbol{f}(x)^\top \left(\frac{\nabla p(x)}{p(x)} \right) + \nabla^\top \boldsymbol{f}(x) \\
&= \frac{1}{p(x)} \Big(\boldsymbol{f}(x)^\top \nabla p(x) + \Big(\nabla^\top \boldsymbol{f}(x) \Big) p(x) \Big) \\
&= \frac{1}{p(x)} \Big(\nabla^\top \big(\boldsymbol{f}(x) p(x) \big) \Big).
\end{aligned} \tag{36}$$

Using this expression, we can indeed verify that the Stein's identity holds

$$\mathbb{E}_P\big[(\mathcal{A}_P^{\mathcal{L}} \boldsymbol{f})(X)\big] = \int_{\mathbb{R}^d} (\mathcal{A}_P^{\mathcal{L}} \boldsymbol{f})(x) p(x) \, \mathrm{d}x = \int_{\mathbb{R}^d} \nabla^\top \big(\boldsymbol{f}(x) p(x) \big) \, \mathrm{d}x = \sum_{i=1}^d \int_{\mathbb{R}^d} \frac{\partial}{\partial x_i} \big(f_i(x) p(x) \big) \, \mathrm{d}x = 0 \tag{37}$$

for functions $\boldsymbol{f}$ such that $f_i(x)p(x)$ vanishes at the boundaries of the domain for $i = 1, \ldots, d$. Note also that (Ley and Swan, 2013)

$$\mathbb{E}_Q\big[(\mathcal{A}_P^{\mathcal{L}} \boldsymbol{f})(X)\big] = \mathbb{E}_Q\big[(\mathcal{A}_P^{\mathcal{L}} \boldsymbol{f})(X) - (\mathcal{A}_Q^{\mathcal{L}} \boldsymbol{f})(X)\big] = \mathbb{E}_Q\big[\boldsymbol{f}(X)^\top \big(\nabla \log p(X) - \nabla \log q(X) \big)\big]. \tag{38}$$

KSD measure. We present the KSD constructions of Chwialkowski et al. (2016) and Liu et al. (2016), more precisely, we follow the notation of the former. Let $\mathcal{H}$ be an RKHS in $\mathrm{Func}\big(\mathbb{R}^d \to \mathbb{R}\big)$ with reproducing kernel k. Denote by $\mathcal{H}^d$ the product RKHS consisting of elements of the form $\boldsymbol{f} = (f_1, \ldots, f_d)$ with $f_i \in \mathcal{H}$ for $i = 1, \ldots, d$, with the associated inner product $\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\mathcal{H}^d} = \sum_{i=1}^d \langle f_i, g_i \rangle_{\mathcal{H}}$ for all $\boldsymbol{f}, \boldsymbol{g} \in \mathcal{H}^d$. Note that elements of $\mathcal{H}^d$ can be seen as elements of $\mathrm{Func}\big(\mathbb{R}^d \to \mathbb{R}^d\big)$. The aim is to express the Stein operator with $\mathcal{F} \subseteq \mathcal{H}^d$ in

[6] The set $\mathrm{Func}(\mathcal{X} \to \mathcal{Y})$ consists of all functions from $\mathcal{X}$ to $\mathcal{Y}$.

terms of the kernel k, which will then give a closed-form expression to compute the Stein discrepancy. First, we recall the reproducing property of the kernel k which implies that (Equation 1)

$$f(x) = \langle f, k(x,\cdot)\rangle_{\mathcal{H}} \qquad \text{and} \qquad \tfrac{\partial}{\partial x_i} f(x) = \langle f, \tfrac{\partial}{\partial x_i} k(x,\cdot)\rangle_{\mathcal{H}}$$

for all $f \in \mathcal{H}$, $x \in \mathbb{R}^d$ and $i = 1, \dots, d$, under appropriate regularity conditions (Theorem 1, Zhou, 2008 and Lemma C.9, Barp et al., 2022). Using these properties, we can express the Stein operator in terms of the kernel k. First, note that

$$\nabla^\top \boldsymbol{f}(x) = \sum_{i=1}^d \tfrac{\partial}{\partial x_i} f_i(x) = \sum_{i=1}^d \langle f_i, \tfrac{\partial}{\partial x_i} k(x,\cdot)\rangle_{\mathcal{H}} = \langle \boldsymbol{f}, \nabla k(x,\cdot)\rangle_{\mathcal{H}^d}$$

and, with the notation $\boldsymbol{s}_P(x) \coloneqq \nabla \log p(x)$ for the score function, we have

$$\begin{aligned}
\boldsymbol{f}(x)^\top \nabla \log p(x) = \boldsymbol{f}(x)^\top \boldsymbol{s}_P(x) &= \sum_{i=1}^d f_i(x) s_i(x) = \sum_{i=1}^d \langle f_i, k(x,\cdot)\rangle_{\mathcal{H}} s_i(x) \\
&= \sum_{i=1}^d \langle f_i, k(x,\cdot) s_i(x)\rangle_{\mathcal{H}} = \langle \boldsymbol{f}, k(x,\cdot)\boldsymbol{s}_P(x)\rangle_{\mathcal{H}^d} \\
&= \langle \boldsymbol{f}, \nabla \log p(x) k(x,\cdot)\rangle_{\mathcal{H}^d}.
\end{aligned}$$

We conclude that the Stein operator can be expressed as

$$(\mathcal{A}_P^{\mathcal{L}} \boldsymbol{f})(x) = \boldsymbol{f}(x)^\top \nabla \log p(x) + \nabla^\top \boldsymbol{f}(x) = \langle \boldsymbol{f}, \nabla \log p(x) k(x,\cdot) + \nabla k(x,\cdot)\rangle_{\mathcal{H}^d}. \tag{39}$$

Writing $\boldsymbol{\xi}_P(x) \coloneqq \nabla \log p(x) k(x,\cdot) + \nabla k(x,\cdot)$ as in Chwialkowski et al. (2016, Equation 1), by properties of mean embeddings and linearity of expectation, we obtain that

$$\mathbb{E}_Q\big[(\mathcal{A}_P^{\mathcal{L}} \boldsymbol{f})(X)\big] = \mathbb{E}_Q[\langle \boldsymbol{f}, \boldsymbol{\xi}_P(X)\rangle_{\mathcal{H}^d}] = \big\langle \boldsymbol{f}, \mathbb{E}_Q[\boldsymbol{\xi}_P(X)]\big\rangle_{\mathcal{H}^d}. \tag{40}$$

The last equality holds under the Bochner integrability condition $\mathbb{E}_Q[\|\boldsymbol{\xi}_P(X)\|_{\mathcal{H}^d}] < \infty$ which itself holds provided by $\mathbb{E}_Q[h_P(X,X)] < \infty$ since $\mathbb{E}_Q[\|\boldsymbol{\xi}_P(X)\|_{\mathcal{H}^d}] \leq \sqrt{\mathbb{E}_Q\big[\|\boldsymbol{\xi}_P(X)\|_{\mathcal{H}^d}^2\big]} = \sqrt{\mathbb{E}_Q[h_P(X,X)]}$ as shown by Chwialkowski et al. (2016, Theorem 2.1). The Stein discrepancy with $\mathcal{F} = \{\boldsymbol{f} \in \mathcal{H}^d : \|\boldsymbol{f}\|_{\mathcal{H}^d} \leq 1\}$, which is referred to as KSD for Kernel Stein Discrepancy, is then equal to

$$\mathrm{KSD}_P(Q) = \sup_{\boldsymbol{f}\in\mathcal{F}} \mathbb{E}_Q\big[(\mathcal{A}_P^{\mathcal{L}} \boldsymbol{f})(X)\big] = \sup_{\boldsymbol{f}\in\mathcal{F}} \big\langle \boldsymbol{f}, \mathbb{E}_Q[\boldsymbol{\xi}_P(X)]\big\rangle_{\mathcal{H}^d} = \big\|\mathbb{E}_Q[\boldsymbol{\xi}_P(X)]\big\|_{\mathcal{H}^d}. \tag{41}$$

Hence, the squared KSD can be expressed as (Chwialkowski et al., 2016, Theorem 2.1)

$$\begin{aligned}
\mathrm{KSD}_P^2(Q) = \big\|\mathbb{E}_Q[\boldsymbol{\xi}_P(X)]\big\|_{\mathcal{H}^d}^2 &= \big\langle \mathbb{E}_Q[\boldsymbol{\xi}_P(X)], \mathbb{E}_Q[\boldsymbol{\xi}_P(Y)]\big\rangle_{\mathcal{H}^d} \\
&\overset{(\star)}{=} \mathbb{E}_{Q,Q}[\langle \boldsymbol{\xi}_P(X), \boldsymbol{\xi}_P(Y)\rangle_{\mathcal{H}^d}] = \mathbb{E}_{Q,Q}[h_P(X,Y)]
\end{aligned} \tag{42}$$

where the Stein kernel h_P is defined as

$$\begin{aligned}
h_P(x,y) = \langle \boldsymbol{\xi}_P(x), \boldsymbol{\xi}_P(y)\rangle_{\mathcal{H}^d} = &\Big(\nabla \log p(x)^\top \nabla \log p(y)\Big) k(x,y) + \nabla \log p(x)^\top \nabla_y k(x,y) \\
&+ \nabla \log p(y)^\top \nabla_x k(x,y) + \big\langle \nabla k(x,\cdot), \nabla k(y,\cdot)\big\rangle_{\mathcal{H}^d}
\end{aligned} \tag{43}$$

where $\langle \nabla k(x,\cdot), \nabla k(y,\cdot)\rangle_{\mathcal{H}^d} = \sum_{i=1}^d \left\langle \frac{\partial}{\partial x_i} k(x,\cdot), \frac{\partial}{\partial y_i} k(y,\cdot)\right\rangle_{\mathcal{H}} = \sum_{i=1}^d \frac{\partial^2}{\partial x_i \partial y_i} k(x,y)$ by the reproducing property with derivatives of Equation 1 (see also Steinwart and Christmann (2008, Lemma 4.34)). Again, the equation $(\star)$ holds under Bochner integrability which is satisfied when $\mathbb{E}_Q[h_P(X,X)] < \infty$. Stein's identity gives $\mathbb{E}_P\big[(\mathcal{A}_P^{\mathcal{L}} \boldsymbol{f})(x)\big] = 0$ for all $\boldsymbol{f} \in \mathcal{F}$, which implies that $\mathbb{E}_P[\boldsymbol{\xi}_P(X)] = 0$, and hence $\mathbb{E}_P[h_P(X,\cdot)] = 0$.

Writing $\boldsymbol{\delta}(x) = \boldsymbol{s}_P(x) - \boldsymbol{s}_Q(x)$ for the difference in score, we recall that Equation 38 gives that for any $\boldsymbol{f} \in \mathcal{F}$ we have

$$\mathbb{E}_Q\left[\boldsymbol{f}(X)^\top \boldsymbol{\delta}(X)\right] = \mathbb{E}_Q\left[(\mathcal{A}_P^{\mathcal{L}} \boldsymbol{f})(X)\right]. \tag{44}$$

As shown by Liu et al. (2016, Theorem 3.6), the KSD can also be expressed as

$$\begin{aligned}
\mathbb{E}_{Q,Q}\left[k(X,Y)\boldsymbol{\delta}(Y)^\top \boldsymbol{\delta}(X)\right] &= \mathbb{E}_{Q,Q}\left[\Big(k(X,Y)\boldsymbol{s}_P(X) + \nabla_X k(X,Y)\Big)^\top \boldsymbol{\delta}(Y)\right] \\
&= \mathbb{E}_{Q,Q}\left[h_P(X,Y)\right] \\
&= \mathrm{KSD}_P^2(Q)
\end{aligned} \tag{45}$$

where the first and second equalities hold by Equation 44 with $\boldsymbol{f}_1(X) = k(X,Y)\boldsymbol{\delta}(Y)$ for fixed Y and $\boldsymbol{f}_2(Y) = k(X,Y)\boldsymbol{s}_P(X) + \nabla_X k(X,Y)$ for fixed X, respectively, giving

$$\begin{aligned}
(\mathcal{A}_P^{\mathcal{L}} \boldsymbol{f}_1)(X) &= k(X,Y)\boldsymbol{\delta}(Y)^\top \boldsymbol{s}_P(X) + \nabla_X^\top\Big(k(X,Y)\boldsymbol{\delta}(Y)\Big) \\
&= \Big(k(X,Y)\boldsymbol{s}_P(X) + \nabla_X k(X,Y)\Big)^\top \boldsymbol{\delta}(Y)
\end{aligned}$$

as $\nabla_X^\top\big(k(X,Y)\boldsymbol{\delta}(Y)\big) = \sum_{i=1}^d \frac{\partial}{\partial X_i} k(X,Y)\delta_i(Y) = \big(\nabla_X k(X,Y)\big)^\top \boldsymbol{\delta}(Y)$, and

$$\begin{aligned}
(\mathcal{A}_P^{\mathcal{L}} \boldsymbol{f}_2)(Y) &= \Big(k(X,Y)\boldsymbol{s}_P(X) + \nabla_X k(X,Y)\Big)^\top \boldsymbol{s}_P(Y) + \nabla_Y^\top\Big(k(X,Y)\boldsymbol{s}_P(X) + \nabla_X k(X,Y)\Big) \\
&= k(X,Y)\boldsymbol{s}_P(X)^\top \boldsymbol{s}_P(Y) + \big(\nabla_X k(X,Y)\big)^\top \boldsymbol{s}_P(Y) + \big(\nabla_Y k(X,Y)\big)^\top \boldsymbol{s}_P(X) + \nabla_Y^\top\big(\nabla_X k(X,Y)\big) \\
&= h_P(X,Y)
\end{aligned}$$

as $\nabla_Y^\top\big(\nabla_X k(X,Y)\big) = \sum_{i=1}^d \frac{\partial}{\partial Y_i}\frac{\partial}{\partial X_i} k(X,Y) = \langle \nabla k(X,\cdot), \nabla k(Y,\cdot)\rangle_{\mathcal{H}^d}$. Using the Cauchy–Schwarz inequality as in Liu et al. (2016), the KSD can be upper bounded by the Fisher divergence as

$$\begin{aligned}
\mathrm{KSD}_P^2(Q) &= \mathbb{E}_{Q,Q}\Big[k(X,Y)\boldsymbol{\delta}(X)^\top \boldsymbol{\delta}(Y)\Big] \\
&\leq \sqrt{\mathbb{E}_{Q,Q}[k(X,Y)^2]\,\mathbb{E}_{Q,Q}\Big[\big(\boldsymbol{\delta}(X)^\top \boldsymbol{\delta}(Y)\big)^2\Big]} \\
&\leq \sqrt{\mathbb{E}_{Q,Q}[k(X,Y)^2]\,\mathbb{E}_{Q,Q}\big[\|\boldsymbol{\delta}(X)\|_2^2\,\|\boldsymbol{\delta}(Y)\|_2^2\big]} \\
&= \sqrt{\mathbb{E}_{Q,Q}[k(X,Y)^2]}\,\mathbb{E}_Q\big[\|\boldsymbol{\delta}(X)\|_2^2\big] \\
&= \sqrt{\mathbb{E}_{Q,Q}[k(X,Y)^2]}\,\mathrm{Fisher}(P,Q)
\end{aligned} \tag{46}$$

where the Fisher divergence (Johnson, 2004) is

$$\mathrm{Fisher}(P,Q) \coloneqq \mathbb{E}_Q\big[\|\nabla \log p(X) - \nabla \log q(X)\|_2^2\big]. \tag{47}$$

The definition of the Stein operator $\mathcal{A}_P^{\mathcal{L}}$ naturally extends to matrices $\boldsymbol{F} = (\boldsymbol{f}^{(1)}, \ldots, \boldsymbol{f}^{(d)})$ where each

$\boldsymbol{f}^{(i)} = (\boldsymbol{f}_1^{(i)}, \dots, \boldsymbol{f}_d^{(i)})$ is a vector for $i = 1, \dots, d$, as

$$\boldsymbol{\mathcal{A}}_P^{\mathcal{L}} \boldsymbol{F} = \left(\mathcal{A}_P^{\mathcal{L}} \boldsymbol{f}^{(1)}, \dots, \mathcal{A}_P^{\mathcal{L}} \boldsymbol{f}^{(d)} \right) \tag{48}$$

mapping functions from $\text{Func}\big(\mathbb{R}^d \to \mathbb{R}^{d \times d}\big)$ to $\text{Func}\big(\mathbb{R}^d \to \mathbb{R}^d\big)$. Let $\boldsymbol{K}(x, y) = k(x, y) I_{d \times d}$ which is equal to $\big(\boldsymbol{k}^{(1)}(x, y), \dots, \boldsymbol{k}^{(d)}(x, y)\big)$ where $\boldsymbol{k}^{(i)}(x, y)$ is d-dimensional vector of zeros with i-th entry $k(x, y)$, giving

$$(\boldsymbol{\mathcal{A}}_{P,x}^{\mathcal{L}} \boldsymbol{K})_i(x, y) = \boldsymbol{k}^{(i)}(x, y)^\top \boldsymbol{s}_P(x) + \nabla^\top \boldsymbol{k}^{(i)}(x, y) = k(x, y) \boldsymbol{s}_P(x_i) + \frac{\partial}{\partial x_i} k(x, y) \tag{49}$$

for $i = 1, \dots, d$, that is

$$(\boldsymbol{\mathcal{A}}_{P,x}^{\mathcal{L}} \boldsymbol{K})(x, y) = k(x, y) \boldsymbol{s}_P(x) + \nabla_x k(x, y) \tag{50}$$

where the subscripts x, y are used to specify which variable the Stein and gradient operators are operating on. Hence, the Stein kernel can be expressed as

$$h_P(x, y) = (\mathcal{A}_{P,y}^{\mathcal{L}} \, \boldsymbol{\mathcal{A}}_{P,x}^{\mathcal{L}} \, \boldsymbol{K})(x, y) \tag{51}$$

as shown above with $\mathcal{A}_P^{\mathcal{L}} \boldsymbol{f}_2$.

KSD V-statistic and U-statistic. We now present KSD estimators given a model distribution P and some samples $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} Q$, we denote the empirical distribution by $\widehat{Q}$. The Kernel Stein Discrepancy $\text{KSD}_P^2(Q)$ can be estimated using a one-sample second-order U- or V-statistic (Chwialkowski et al., 2016; Liu et al., 2016)

$$V_{\text{KSD}_k^2} \coloneqq \frac{1}{n^2} \sum_{1 \le i, i' \le n} h_P(X_i, X_{i'}) \qquad \text{and} \qquad U_{\text{KSD}_k^2} \coloneqq \frac{1}{n(n-1)} \sum_{1 \le i \ne i' \le n} h_P(X_i, X_{i'}) \tag{52}$$

which can be computed as

$$V_{\text{KSD}_k^2} \coloneqq \frac{1}{n^2} \mathbf{1}^\top H^{XX} \mathbf{1} \qquad \text{and} \qquad U_{\text{KSD}_k^2} \coloneqq \frac{1}{n(n-1)} \mathbf{1}^\top H^{XX} \mathbf{1} \tag{53}$$

where $H^{XX} = \big(h_P(X_i, X_j)\big)_{1 \le i, j \le n}$ with the Stein kernel h_P as in Equation 43, where $\bar{H}^{XX}$ is the matrix H^{XX} with diagonal entries set to zero, and where $\mathbf{1}$ is a vector of ones of size n. The V-statistic corresponds to the plugin estimator $\text{KSD}_P^2(\widehat{Q})$ which is strictly positive and can hence be used as an estimator for the (non-squared) KSD by taking its square root, unlike the U-statistic which is unbiased but can be negative. For consistency, the core KSD function h_k^{KSD} can be defined as h_P itself.

KSD kernel choice importance. The behaviour of the KSD when using a translation-invariant kernel k_λ, and letting the bandwidth λ tend to zero, is

$$\text{KSD}_\lambda^2 = \mathbb{E}_{Q,Q}\Big[k_\lambda(X, Y) \boldsymbol{\delta}(X)^\top \boldsymbol{\delta}(Y)\Big] \to \mathbb{E}_Q\Big[\boldsymbol{\delta}(X)^\top \boldsymbol{\delta}(X)\Big] = \text{Fisher}(P, Q)^2, \tag{54}$$

while, when the bandwidth λ tends to ∞, we have

$$\text{KSD}_\lambda^2 = \mathbb{E}_{Q,Q}\Big[k_\lambda(X, Y) \boldsymbol{\delta}(X)^\top \boldsymbol{\delta}(Y)\Big] \to \mathbb{E}_{Q,Q}\Big[\boldsymbol{\delta}(X)^\top \boldsymbol{\delta}(Y)\Big] = \|\mathbb{E}_Q[\boldsymbol{\delta}(X)]\|_2^2. \tag{55}$$

So, the KSD tends to the Fisher divergence when $\lambda \to 0$, and to the 2-norm of the expected difference in score when $\lambda \to \infty$, while the MMD and HSIC tend to zero in both of these regimes. The behaviour of the KSD U-stastistic and V-statistic when varying the bandwidth is difficult to characterise due to the complexity of

the Stein kernel h_P involving kernel derivatives which scale with the bandwidth. This is drastically different from the MMD and HSIC cases, and highlights even more the importance of the bandwidth choice for the KSD statistic computation in order to obtain meaningful results.

Fisher divergence as a KSD. With the kernel $k(x, y) = \mathbf{1}(x = y)$, the KSD is equal to the Fisher divergence

$$\mathrm{KSD}^2_\lambda = \mathbb{E}_{Q,Q}\Big[k_\lambda(X,Y)\boldsymbol{\delta}(X)^\top\boldsymbol{\delta}(Y)\Big] = \mathbb{E}_Q\Big[\boldsymbol{\delta}(X)^\top\boldsymbol{\delta}(X)\Big] = \mathrm{Fisher}(P,Q)^2. \tag{56}$$

From another point of view, the KSD can be seen as a kernelized Fisher divergence.

KSD as an MMD. The MMD with the Stein kernel h_P is equal to the KSD since

$$\mathrm{MMD}^2_{h_P}(P,Q) = \mathbb{E}_{Q,Q}[h_P(X,X')] - 2\,\mathbb{E}_{Q,P}[h_P(X,Y)] + \mathbb{E}_{P,P'}[h_P(Y,Y')] = \mathbb{E}_{Q,Q}[h_P(X,X')] = \mathrm{KSD}^2_P(Q) \tag{57}$$

using Stein's identity $\mathbb{E}_P[h_P(X,\cdot)] = 0$. We stress that the Stein kernel used in this MMD depends on the model distribution P.

MMD as a KSD. A simple operator can be defined as

$$(\mathcal{A}'_P\boldsymbol{f})(x) = f_1(x) - \mathbb{E}_P[f_1(X)] \tag{58}$$

for all $x \in \mathbb{R}^d$ and for $\boldsymbol{f} = (f_1, \ldots, f_d)$ where $f_i \colon \mathbb{R}^d \to \mathbb{R}$ for $i = 1, \ldots, d$. This is a Stein operator since $\mathbb{E}_P[(\mathcal{A}'_P\boldsymbol{f})(X)] = 0$ for all $\boldsymbol{f}$. Hence, we can define a Stein Discrepancy using this Stein operator, which we kernelise using an RKHS $\mathcal{H}$ with reproducing kernel k and unit ball $\mathcal{F} = \{\boldsymbol{f} \in \mathcal{H}^d : \|\boldsymbol{f}\|_{\mathcal{H}^d} \leq 1\}$,, as

$$\mathrm{KSD}_{\mathcal{A}'_P}(P,Q) = \sup_{\boldsymbol{f}\in\mathcal{F}} \mathbb{E}_Q\big[(\mathcal{A}'_P\boldsymbol{f})(X)\big] = \sup_{f_1\in\mathcal{H}\,:\,\|f_1\|_\mathcal{H}\leq 1} \mathbb{E}_Q[f_1] - \mathbb{E}_P[f_1] = \mathrm{MMD}_k(P,Q), \tag{59}$$

so the MMD itself can be seen as a KSD using a specific Stein operator.

3 Efficient kernel discrepancies estimators

Expectation. As seen in Equations 11, 21 and 52, MMD, HSIC and KSD can be estimated using one-sample second-order V-statistics, which are estimators of the quantity

$$\mathbb{E}\big[h(X,X')\big] \tag{60}$$

for some core function h, and where the expectation is over independent copies X and X'.

Statistics. Given i.i.d. variables $X_1, \ldots, X_n$, a class of estimators for this expected quantity takes the form

$$\frac{1}{|\mathcal{D}|}\sum_{(i,j)\in\mathcal{D}} h(X_i, X_j) \tag{61}$$

for some subset $\mathcal{D} \subseteq \{(i,j) : 1 \leq i,j \leq n\}$, often called the design, and can be computed in time $\mathcal{O}(|\mathcal{D}|)$, where $|\mathcal{D}|$ denotes the cardinality of $\mathcal{D}$.

V-statistic. The V-statistic (Mises, 1947) is defined by setting $\mathcal{D} = \{(i,j) : 1 \leq i,j \leq n\}$ giving

$$V = \frac{1}{n^2} \sum_{1 \leq i,j \leq n} h(X_i, X_j) \tag{62}$$

which can be computed in quadratic time $\mathcal{O}(n^2)$. Since the expectation is over independent copies, and that the V-statistic includes the terms $\{h(X_i, X_i) : i = 1, \ldots, n\}$, the V-statistic is biased.

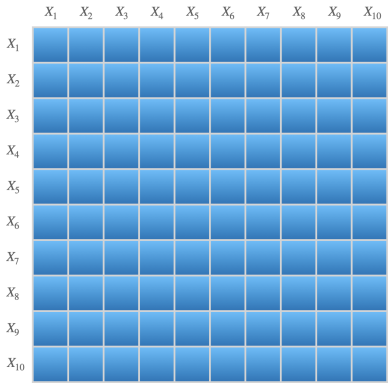

Figure 1: V-statistic. Visualisation of the core kernel matrix entries $h(X_i, X_j)$ considered (in blue) and ignored (in white) in the sum for the V-statistic computation with $n = 10$.

U-statistic. By not including these terms, *i.e.* by considering $\mathcal{D} = \{(i,j) : 1 \leq i \neq j \leq n\}$, we obtain the unbiased U-statistic (Hoeffding, 1948; Lee, 1990)

$$U = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} h(X_i, X_j), \tag{63}$$

also computable in quadratic time $\mathcal{O}(n^2)$. The U-statistic is known to be the minimum variance estimator of $\mathbb{E}[h(X, X')]$.

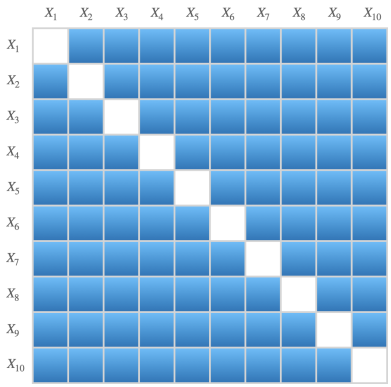

Figure 2: U-statistic. Visualisation of the core kernel matrix entries $h(X_i, X_j)$ considered (in blue) and ignored (in white) in the sum for the U-statistic computation with $n = 10$.

Incomplete statistic. The U-statistic and V-statistic are referred to as *complete*, unlike their *incomplete* counterparts which trade accuracy for computational efficiency (Blom, 1976; Janson, 1984; Lee, 1990) and

take the form

$$\frac{1}{|\mathcal{D}|}\sum_{(i,j)\in\mathcal{D}} h(X_i, X_j) \tag{64}$$

for some strictly smaller subset $\mathcal{D} \subset \{(i,j) : 1 \leq i \neq j \leq n\}$, and can be computed in time $\mathcal{O}(|\mathcal{D}|)$ which can be much faster than quadratic time. Incomplete statistics are unbiased, they are particularly useful when the number of samples is large, and the kernel function is computationally expensive to evaluate. The L-statistic, D-statistic, B-statistic, X-statistic and R-statistic, all introduced below, are examples of incomplete statistics.

Depending on the statistic, we sometimes define $\mathcal{D}$ as a subset of the upper triangular matrix entries $\{(i,j) : 1 \leq i < j \leq n\}$ and leverage the fact that the core h is symmetric. Nonethess, in the figures, we always provide illustrations considering the full core kernel matrix.

L-statistic. The linear L-statistic (Gretton et al., 2012, Lemma 14) is defined by considering the subset of the core kernel matrix entries $\mathcal{D} = \{(2i-1, 2i) : 1 \leq i \leq \lfloor n/2 \rfloor\}$, giving

$$L = \frac{1}{\lfloor n/2 \rfloor}\sum_{1\leq i\leq \lfloor n/2 \rfloor} h(X_{2i}, X_{2i-1}). \tag{65}$$

While this statistic can be computed in linear time $\mathcal{O}(n)$, it is rarely useful in practice as only very little information is captured when considering so few entries of the core kernel matrix.

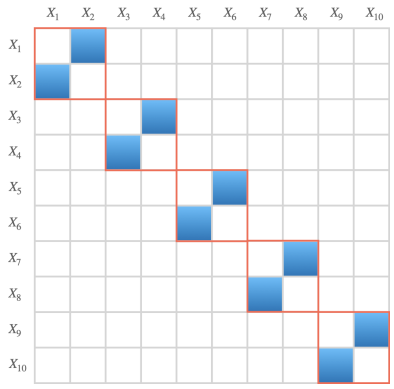

Figure 3: L-statistic. Visualisation of the core kernel matrix entries $h(X_i, X_j)$ considered (in blue) and ignored (in white) for the L-statistic computation with $n = 10$.

D-statistic. Another possibility is to include multiple subdiagonals of the core kernel matrix as done in Schrab et al. (2022b). Considering the first r subdiagonals, that is $\mathcal{D} \coloneqq \{(i, i+j) : i = 1, \ldots, n-j \text{ for } j = 1, \ldots, r\}$ with size $|\mathcal{D}| = rn - r(r+1)/2$, gives rise to a D-statistic

$$D = \frac{2}{r(2n-r-1)}\sum_{j=1}^{r}\sum_{i=1}^{n-j} h(X_i, X_{i+j}). \tag{66}$$

Its time complexity is $\mathcal{O}(rn)$, if r is set to a small fixed constant then this is linear, another common choice would be to set $r = \lfloor \sqrt{n} \rfloor$ to obtain an estimator computable in time $\mathcal{O}\big(n^{1.5}\big)$.

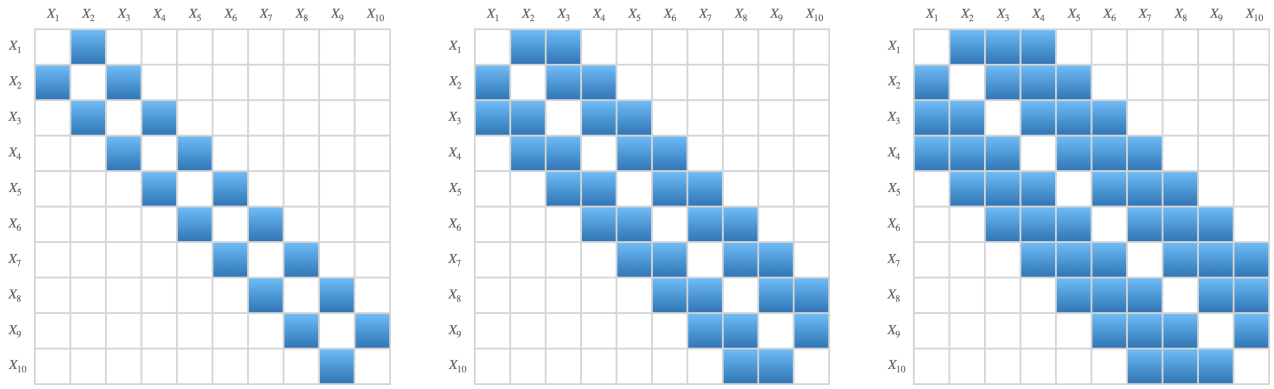

Figure 4: D-statistic. Visualisation of the core kernel matrix entries $h(X_i, X_j)$ considered (in blue) and ignored (in white) in the sum for the D-statistic computation with $n = 10$. *(Left)* $r = 1$. *(Centre)* $r = 2$. *(Right)* $r = 3$.

B-statistic (case $b = 2$). We now introduce block statistics (Ho and Shieh, 2006), also referred to as B-statistics. For illustration purposes, we start with the case of $b = 2$ blocks and consider some $n_1 + n_2 = n$. Then, letting $\mathcal{D} = \{(i,j) : 1 \leq i \neq j \leq n_1\} \cup \{(i,j) : n_1 + 1 \leq i \neq j \leq n\}$, we obtain

$$\begin{aligned} B &= U(X_1, \dots, X_{n_1}) + U(X_{n_1+1}, \dots, X_n) \\ &= \frac{1}{n_1(n_1-1)} \sum_{1 \leq i \neq j \leq n_1} h(X_i, X_j) + \frac{1}{n_2(n_2-1)} \sum_{1 \leq i \neq j \leq n_2} h(X_{n_1+i}, X_{n_1+j}) \end{aligned} \tag{67}$$

with time complexity $\mathcal{O}\big(n_1^2 + n_2^2\big)$. It is common to consider blocks of the same size, assuming n is even let $n_1 = n_2 = n/2$, then the time complexity becomes $2(n/2)^2$. In this example, the block statistic is composed of U-statistics, it can also be defined similarly using V-statistics instead (in which case it would lead to a biased statistic).

B-statistic (general case). We now consider the general case of b blocks of sizes $n_1, \dots, n_b$ where $\sum_{t=1}^{b} n_t = n$, and we let $n_0 = 0$. Then, considering

$$\mathcal{D} = \bigcup_{s=1}^{b} \left\{ (i,j) : 1 + \sum_{t=0}^{s-1} n_t \leq i \neq j \leq \sum_{t=0}^{s} n_t \right\} \tag{68}$$

gives the B-statistic (Ho and Shieh, 2006)

$$B = \frac{1}{|\mathcal{D}|} \sum_{s=1}^{b} n_t(n_t - 1) U\Big(X_{1+\sum_{t=0}^{s-1} n_t}, \dots, X_{\sum_{t=0}^{s} n_t}\Big) \tag{69}$$

where $n_t(n_t - 1)U\Big(X_{1+\sum_{t=0}^{s-1} n_t}, \dots, X_{\sum_{t=0}^{s} n_t}\Big)$ is an unscaled U-statistic, and where $|\mathcal{D}| = \sum_{s=1}^{b} n_t(n_t - 1)$. This B-statistic has time complexity $\mathcal{O}\big(n_1^2 + \dots + n_b^2\big)$. Assuming n is divisible by b and considering blocks of equal size $n_t = n/b$ for $t = 1, \dots, b$, we obtain

$$B = \frac{1}{b} \sum_{s=1}^{b} U\big(X_{1+(s-1)n/b}, \dots, X_{sn/b}\big) \tag{70}$$

with time complexity $\mathcal{O}\big(b(n/b)^2\big) = \mathcal{O}\big(n^2/b\big)$. In practice, it is common to set $b = \lfloor \sqrt{n} \rfloor$ (Zaremba et al., 2013; Zhang et al., 2018) and get an estimator with time complexity of the order $\mathcal{O}\big(n^{1.5}\big)$. When the sample size n is not divisible by the number of blocks b, we either have one block of size strictly less than n or even

ignore that smaller block for simplicity.

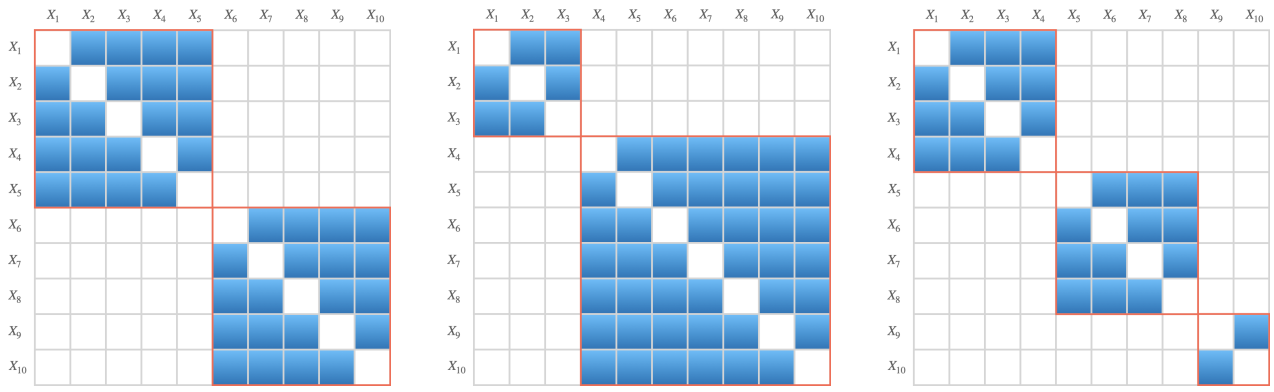

Figure 5: B-statistic. Visualisation of the core kernel matrix entries $h(X_i, X_j)$ considered (in blue) and ignored (in white) in the sum for the B-statistic computation with $n = 10$. *(Left)* $b = 2$, $n_1 = n_2 = 5$. *(Centre)* $b = 2$, $n_1 = 3$, $n_2 = 5$. *(Right)* $b = 3$, $n_1 = 4$, $n_2 = 4$, $n_3 = 2$.

X-statistic. The cross X-statistic, introduced by Kim and Ramdas, 2024 (see also Shekhar et al., 2022, 2023), considers the entries $\mathcal{D} \coloneqq \{(i, j) : i = 1, \dots, n_1 \text{ for } j = n_1 + 1, \dots, n\}$ for some $n_1 \in \{1, \dots, n-1\}$, giving

$$X = \frac{1}{n_1(n - n_1)} \sum_{i=1}^{n_1} \sum_{j=n_1+1}^{n} h(X_i, X_j) \tag{71}$$

computable in time complexity $\mathcal{O}(n_1(n - n_1))$. The main point of using this statistic is that the terms appearing in the first input of the core h and in its second input, are disjoint. Leveraging this fact, by scaling the statistic appropriately by some standard deviation (*i.e.* studentisation), asymptotic normality of the statistic can always be guaranteed (Kim and Ramdas, 2024). A typical choice for n_1 is simply to set it equal to $\lfloor n/2 \rfloor$, in which case the time complexity is still quadratic ($(n/2)^2$ rather than n^2) but this statistic can benefit from asymptotic normality.

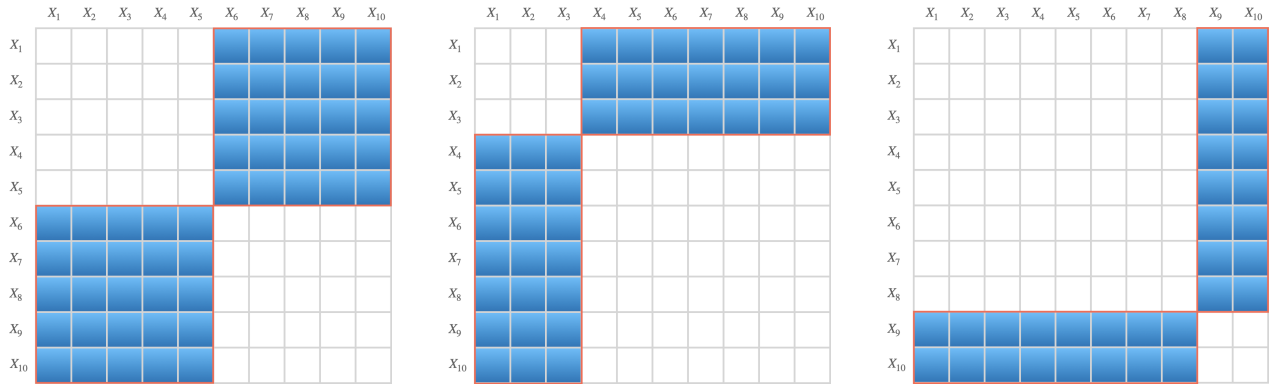

Figure 6: X-statistic. Visualisation of the core kernel matrix entries $h(X_i, X_j)$ considered (in blue) and ignored (in white) in the sum for the X-statistic computation with $n = 10$. *(Left)* $n_1 = 5$. *(Centre)* $n_1 = 3$. *(Right)* $n_1 = 8$.

R-statistic. Let $\mathcal{D}_r$ be a random subsample of $\{(i,j) : 1 \leq i < j \leq n\}$, either with or without replacement, of some prespecified size $|\mathcal{D}_r|$. Then, the random R-statistic (Lee, 1990) is defined as

$$R = \frac{1}{|\mathcal{D}_r|} \sum_{(i,j)\in\mathcal{D}_r} h(X_i, X_j) \tag{72}$$

with time complexity $\mathcal{O}(|\mathcal{D}_r|)$ chosen by the user. Given some fixed data $X_1, \ldots, X_n$, computing the R-statistic twice results in different values due to the additional source of randomness introduced in the statistic computation. This statistic has the benefit that in expectation it considers all non-diagonal entries of the core kernel matrix while being computationally faster to be evaluated.

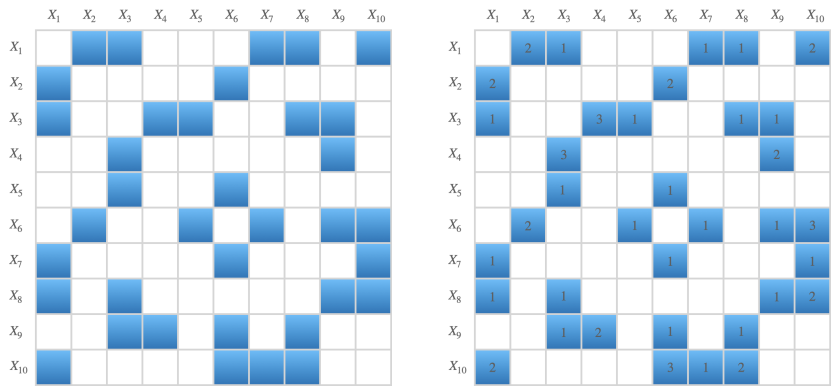

Figure 7: R-statistic. Random visualisation of the core kernel matrix entries $h(X_i, X_j)$ sampled (in blue) and not sampled (in white) in the sum for the R-statistic computation with $n = 10$. *(Left)* Without replacement, $|\mathcal{D}_r| = 18$. *(Right)* With replacement—the numbers represent how many times each entry has been sampled in the upper triangular matrix, $|\mathcal{D}_r| = 27$.

4 Kernel pooling: adaptive kernel discrepancies estimators

Adaptivity via kernel pooling. As aforementioned, the kernel choices for the MMD, HSIC and KSD estimators crucially affect their utility. To overcome this, we rely on kernel pooling, which consists in combining multiple statistics with different kernels, to construct estimators which are adaptive to the kernel selection. We first explain how the statistics can be normalised to be compared against each other, we then present three kernel pooling methods for combining them, and finally we propose a parameter-free method for constructing a collection of kernels. We note that kernel pooling can be used either with or without normalisation. In practice, we recommend using fuse kernel pooling (Equation 77) with normalisation (Equation 73), which is studied in depth in Biggs et al. (2023).

Normalisation. Consider a finite collection of kernels[7] $\mathcal{K}$ and their associated statistics[8]

$$S_k = \frac{1}{|\mathcal{D}_k|} \sum_{(i,j)\in\mathcal{D}_k} h_k(X_i, X_j).$$

In order to compare the statistics $S_1, \ldots, S_{|\mathcal{K}|}$, we need to ensure they are indeed comparable. For example, scaling the kernel by a constant trivially scales the MMD and HSIC by that factor. Moreover, some statistics might have much higher variance than others. These two facts illustrate that simply having a larger statistic

[7] For HSIC, this collection consists of product kernels $k = k^{\mathcal{X}} \times k^{\mathcal{Y}}$.

[8] For simplicity, we consider one-sample second-order statistics of this form, but the method holds more generally for any statistic (the normalisation needs to be adapted accordingly).

might not necessarily be significant. To account for this, we propose to normalise the estimators by some standard deviation term, that is, instead of considering S_k, we compute

$$S_k/\sigma_k \qquad \text{where} \qquad \sigma_k^2 \coloneqq \frac{4}{|\mathcal{D}_k^1|}\sum_{i\in\mathcal{D}_k^1}\left(\frac{1}{|\mathcal{D}_k^{2,i}|}\sum_{j\in\mathcal{D}_k^{2,i}} h_k(X_i,X_j)\right)^2 - \left(\frac{2}{|\mathcal{D}_k|}\sum_{(i,j)\in\mathcal{D}_k} h_k(X_i,X_j)\right)^2 \tag{73}$$

where $\mathcal{D}_k^1 \coloneqq \{i : (i,j) \in \mathcal{D}_k \text{ for some } j\}$ and $\mathcal{D}_k^{2,i} \coloneqq \{j : (i,j) \in \mathcal{D}_k\}$. This biased standard deviation estimator of the statistic under the alternative hypothesis in this incomplete form has been adapted from the complete variant proposed by Sutherland et al. (2017) and Liu et al. (2020, Equation 5), see Sutherland and Deka (2022) for unbiased standard deviation estimators.

While this appears to be similar to studentisation, we emphasise that the aim is different: we are not interested in obtaining asymptotic normality but in being able to compare all the normalised statistics $S_1/\sigma_1, \ldots, S_{|\mathcal{K}|}/\sigma_{|\mathcal{K}|}$ in a meaningful way. We note that, for studentisation, there is no real consensus in the literature on which form the estimated standard deviation should take. Here, we propose to use a simple one which aligns well with our study of different types of statistics in Section 3. The unnormalised case simply corresponds to using $\sigma_k = 1$.

We now present three methods for combining the (normalised) statistics, namely, mean, maximum and fuse kernel pooling. These run in time complexity $\mathcal{O}\big(\sum_{k\in\mathcal{K}}|\mathcal{D}_k|\big)$ which is $\mathcal{O}(|\mathcal{K}||\mathcal{D}|)$ if the same design $\mathcal{D}$ is used across all statistics.

Mean kernel pooling. One possibility is to take the mean (or sometimes the sum) of the normalised statistics, giving

$$\operatorname*{mean}_{k\in\mathcal{K}} S_k/\sigma_k \;=\; \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}} S_k/\sigma_k. \tag{74}$$

All normalised estimators are added up together, the intuition being that, as long as one statistic is 'large', then this will be captured in the sum.

Another common method is to take the mean (or the sum) of kernels, and then to simply compute one statistic with this mean kernel. For the case of MMD and HSIC, due to the linearity with respect to the kernel, we note that this is equivalent to taking the mean (or the sum) of the MMDs/HSICs without normalisation, that is

$$S_{\overline{k}} = \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}} S_k \qquad \text{where} \qquad \overline{k} \coloneqq \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}} k. \tag{75}$$

Maximum kernel pooling. In order to capture the discrepancy, another possibility is simply to take the largest of the normalised statistics, that is, to compute

$$\max_{k\in\mathcal{K}} S_k/\sigma_k. \tag{76}$$

Intuitively, if the maximum is 'large', the normalised statistic is 'large' for some kernel, meaning that the discrepancy can be detected for that kernel. If the maximum is 'small', we deduce that all statistics are 'small' and that there is no discrepancy detected by any of the kernels.

In this method, only one value is retained, which differs from the previous method in which all values are combined. The fact that many values are simply ignored, and that slightly modifying them might not change the maximum, can often not be desirable, both from a theoretical and a practical point of view. The unnormalized version of maximum kernel pooling is very closely related to the methods of Fukumizu et al. (2009) and Cárcamo et al. (2022). We next present a relaxed maximum which overcomes these issues.

Fuse kernel pooling. We can use a relaxed maximum of the normalised statistics which takes the form of a logsumexp expression

$$\operatorname*{fuse}_{k\in\mathcal{K}} S_k/\sigma_k \;=\; \frac{1}{\nu}\log\left(\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\exp\left(\nu S_k/\sigma_k\right)\right) \tag{77}$$

with fusing parameter $\nu > 0$. Noting that $\exp(\nu M)/|\mathcal{K}| \leq \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\exp(\nu S_k/\sigma_k) \leq \exp(\nu M)$ where $M \coloneqq \max_{k\in\mathcal{K}} S_k/\sigma_k$, we deduce that

$$\max_{k\in\mathcal{K}} S_k/\sigma_k - \frac{\log(|\mathcal{K}|)}{\nu} \;\leq\; \operatorname*{fuse}_{k\in\mathcal{K}} S_k/\sigma_k \;\leq\; \max_{k\in\mathcal{K}} S_k/\sigma_k \tag{78}$$

Hence, as ν tends to infinity, the estimator converges to the maximum of the $S_1/\sigma_1, \dots, S_{|\mathcal{K}|}/\sigma_{|\mathcal{K}|}$ values. Building upon the theory of Biggs et al. (2023), a typical choice of ν is to set it equal to $\max_{k\in\mathcal{K}}|\mathcal{D}_k|/N$ which increases with the sample size N for estimators computable in time longer than linear. For complete quadratic-time statistics, this gives $\nu = N$. The same intuition as for the true maximum holds, but having an estimator which changes with each normalised statistic can be beneficial for downstream tasks. As illustrated in Biggs et al. (2023, Appendix B), fuse kernel pooling also allows for (uncountable) distributions on the space of kernels instead of simply working with a finite collection of kernels (*i.e.*, uniform distribution on a discrete set of kernels). In practice, we recommend using the fuse variant of kernel pooling, which is analyzed in details in Biggs et al. (2023).

Kernel collection. Consider a radial kernel $k_\lambda(x, y) = \Psi(\|x - y\|_r/\lambda)$ for some $r \geq 1$, $\lambda > 0$ and $\Psi\colon \mathbb{R} \to \mathbb{R}$ normalised (either such that it integrates to 1, or such that $\psi(0) = 1$). Given some data $X_1, \dots, X_n$, consider the set of inter-sample distances

$$D = \left\{\|x - x'\|_r \colon x, x' \in \{X_1, \dots, X_n\},\, x \neq x'\right\} \setminus \{0\}. \tag{79}$$

A naive way of choosing the kernel bandwidth is simply to set it equal to the median of D (Gretton et al., 2012), while simple, this method fails to capture the discrepancy accurately in most cases and is not adaptive. However, the set D of distances remains very relevant as the kernel is evaluated at these values scaled by the inverse bandwidth. Hence, to construct a collection of bandwidths for k from D it makes sense to consider a discretisation of the interval between the minimum and maximum of D. In practice, to avoid numerical issues, we actually use the 5% and 95% quantiles of D instead, and discretise the interval between them linearly using 10 points (Biggs et al., 2023, Section 6). As noted in Schrab et al. (2023, Section 5.7), using only 10 bandwidths is sufficient to fully capture all the information, no advantage is observed for using more points in the discretisation. For the kernel k, this gives a bandwidth collection

$$\Lambda(k) = \left\{q_{5\%} + i(q_{95\%} - q_{5\%})/9 : i = 0, \dots, 9\right\}. \tag{80}$$

To construct the collection of kernels, we can then consider multiple kernel types and use for each the 10 bandwidths constructed above. In practice, as illustrated in Schrab et al. (2023, Section 5.7), we recommend combining Gaussian and Laplace kernels, with no advantage observed for including more types of kernels. The parameter-free kernel collection is then

$$\mathcal{K} = \left\{k_\lambda \colon k \in \{\text{Gaussian, Laplace}\}, \lambda \in \Lambda(k)\right\} \tag{81}$$

consisting of 20 kernels, which can then be used when computing an adaptive estimator through mean, maximum or fuse kernel pooling. In practice, when using these kernel metrics in general settings, we recommend using fuse pooling studied in details in Biggs et al. (2023). When using them for hypothesis testing, another powerful adaptive method is *aggregation* (Schrab et al., 2023, 2022a; Albert et al., 2022). See Schrab (2025b) for a unified view of hypothesis testing optimality results using these kernel discprepancies.

Bibliography

M. Albert, B. Laurent, A. Marrel, and A. Meynaoui. Adaptive test of independence based on HSIC measures. *The Annals of Statistics*, 50(2):858–879, 2022.

A. Anastasiou, A. Barp, F.-X. Briol, B. Ebner, R. E. Gaunt, F. Ghaderinezhad, J. Gorham, A. Gretton, C. Ley, Q. Liu, et al. Stein's Method Meets Statistics: A Review of Some Recent Developments. *arXiv preprint arXiv:2105.03481*, 2021.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3): 337–404, 1950.

A. Barp, C.-J. Simon-Gabriel, M. Girolami, and L. Mackey. Targeted separation and convergence with kernel discrepancies. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

F. Biggs, A. Schrab, and A. Gretton. MMD-FUSE: Learning and Combining Kernels for Two-Sample Testing Without Data Splitting. *arXiv preprint arXiv:2306.08777*, 2023.

G. Blom. Some properties of incomplete U-statistics. *Biometrika*, 63(3):573–580, 1976.

J. Cárcamo, A. Cuevas, and L.-A. Rodríguez. A uniform kernel trick for high-dimensional two-sample problems. *arXiv preprint arXiv:2210.02171*, 2022.

C. Carmeli, E. De Vito, A. Toigo, and V. Umanitá. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

S. L. Chau, A. Schrab, A. Gretton, D. Sejdinovic, and K. Muandet. Credal two-sample tests of epistemic ignorance. In *International Conference on Artificial Intelligence and Statistics*, 2025.

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning*, pages 2606–2615. PMLR, 2016.

K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99, 2004.

K. Fukumizu, A. Gretton, G. Lanckriet, B. Schölkopf, and B. K. Sriperumbudur. Kernel choice and classifiability for rkhs embeddings of probability distributions. *Advances in neural information processing systems*, 22, 2009.

J. Gorham and L. Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301. PMLR, 2017.

J. Gorham and L. W. Mackey. Measuring Sample Quality with Stein's Method. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 226–234, 2015.

A. Gretton. Introduction to RKHS, and some simple kernel algorithms. *Advanced Topopics Machine Learnearning Lecture Conducted from University College London*, 16(5-3):2, 2013.

A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A Kernel Method for the Two-Sample Problem. In *Advances in Neural Information Processing Systems*, pages 513–520, Cambridge, MA, 2007. MIT Press.

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, volume 1, pages 585–592, 2008.

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

H.-C. Ho and G. S. Shieh. Two-stage U-statistics for Hypothesis Testing. *Scandinavian journal of statistics*, 33(4):861–873, 2006.

W. Hoeffding. A class of statistics with asymptotically normal distribution. 19(3):293–325, 1948.

S. Janson. The asymptotic distributions of incomplete U-statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 66(4):495–505, 1984.

O. Johnson. *Information theory and the central limit theorem*, volume 8. World Scientific, 2004.

I. Kim and A. Ramdas. Dimension-agnostic inference using cross U-statistics. *Bernoulli*, 30(1):683–711, 2024.

I. Kim and A. Schrab. Differentially private permutation tests: Applications to kernel methods. Arxiv preprint 2310.19043., 2023.

J. Lee. *U-statistics: Theory and Practice*. Citeseer, 1990.

C. Ley and Y. Swan. Stein's density approach and information inequalities. *Electronic Communications in Probability*, 18:1–14, 2013.

F. Liu, W. Xu, J. Lu, G. Zhang, A. Gretton, and D. J. Sutherland. Learning Deep Kernels for Non-Parametric Two-Sample Tests. In *International Conference on Machine Learning*, 2020.

Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, pages 276–284. PMLR, 2016.

R. v. Mises. On the asymptotic distribution of differentiable statistical functions. *The annals of mathematical statistics*, 18(3):309–348, 1947.

K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 1:429–443, 1997.

R. Pogodin, N. Deka, Y. Li, D. J. Sutherland, V. Veitch, and A. Gretton. Efficient conditionally invariant representation learning. *International Conference on Learning Representations, ICLR*, 2022.

R. Pogodin, A. Schrab, Y. Li, D. J. Sutherland, and A. Gretton. Practical kernel tests of conditional independence. 2024.

W. Rudin. *Real and complex analysis*. Third edition, McGraw Hill, 1987.

A. Schrab. *Optimal Kernel Hypothesis Testing*. PhD thesis, UCL (University College London), 2025a.

A. Schrab. A Unified View of Optimal Kernel Hypothesis Testing. *arXiv preprint arXiv:2503.07084*, 2025b.

A. Schrab and I. Kim. Robust kernel hypothesis testing under data corruption. In *International Conference on Artificial Intelligence and Statistics*, 2025.

A. Schrab, B. Guedj, and A. Gretton. KSD Aggregated Goodness-of-fit Test. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, 2022a.

A. Schrab, I. Kim, B. Guedj, and A. Gretton. Efficient Aggregated Kernel Tests using Incomplete U-statistics. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, volume 35, pages 18793–18807, 2022b.

A. Schrab, I. Kim, M. Albert, B. Laurent, B. Guedj, and A. Gretton. MMD Aggregated Two-Sample Test. *Journal of Machine Learning Research*, 24(194):1–81, 2023.

D. Sejdinovic and A. Gretton. What is an RKHS? *Lecture Notes*, 25, 2012.

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

S. Shekhar, I. Kim, and A. Ramdas. A permutation-free kernel two-sample test. *Advances in Neural Information Processing Systems*, 35:18168–18180, 2022.

S. Shekhar, I. Kim, and A. Ramdas. A permutation-free kernel independence test. *Journal of Machine Learning Research*, 24(369):1–68, 2023.

L. Song, A. J. Smola, A. Gretton, J. Bedo, and K. M. Borgwardt. Feature Selection via Dependence Maximization. *Journal of Machine Learning Research*, 13(5):1393–1434, 2012.

B. Sriperumbudur, K. Fukumizu, and G. Lanckriet. On the relation between universality, characteristic kernels and RKHS embedding of measures. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 773–780. JMLR Workshop and Conference Proceedings, 2010a.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective Hilbert space embeddings of probability measures. In *21st Annual Conference on Learning Theory (COLT 2008)*, pages 111–122. Omnipress, 2008.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010b.

B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, Characteristic Kernels and RKHS Embedding of Measures. *Journal of Machine Learning Research*, 12(7):2389–2410, 2011.

C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*, volume 6, pages 583–603. University of California Press, 1972.

C. Stein, P. Diaconis, S. Holmes, and G. Reinert. Use of exchangeable pairs in the analysis of simulations. *Lecture Notes-Monograph Series*, pages 1–26, 2004.

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

D. J. Sutherland and N. Deka. Unbiased estimators for the variance of mmd estimators. 2022.

D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations*, 2017.

W. Zaremba, A. Gretton, and M. Blaschko. B-test: A non-parametric, low variance kernel two-sample test. *Advances in neural information processing systems*, 26, 2013.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. *Conference on Uncertainty in Artificial Intelligence*, 2011.

Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130, 2018. doi: 10.1007/s11222-016-9721-7.

D.-X. Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of computational and Applied Mathematics*, 220(1-2):456–463, 2008.

Acknowledgements

I, Antonin Schrab, acknowledge support from the U.K. Research and Innovation under grant number EP/S021566/1.