# Kernel-based information theoretical measures: accelerations and limits

Zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

von der KIT-Fakultät für Informatik des
Karlsruher Instituts für Technologie (KIT)

genehmigte
Dissertation

von

Florian Kalinke

Tag der mündlichen Prüfung: 19. Mai 2025
1. Referent: Jun.-Prof. Dr. Jan Stühmer
2. Referent: Prof. Zoltán Szabó
Betreuer: Prof. Dr.-Ing. Klemens Böhm

# Acknowledgements

First and foremost, I would like to thank my doctoral adviser, Prof. Klemens Böhm, for instilling in me the idea of pursuing fundamental research and for his continuous support throughout, whether financially, structurally, or by offering intellectual freedom and discussing projects and articles. He influenced this journey in more ways than I can list. I am also grateful to Edouard Fouché for advising me in numerous ways and for critically evaluating, dissecting, and refining ideas through discussions and text.

This thesis largely took its current form beginning with a research stay with Zoltán Szabó in London, which led to numerous collaborations and subsequent visits. Thank you for always welcoming me, answering all my (often naïve) questions, and teaching me so much. I thank Bharath K. Sriperumbudur for sharing his love of mathematics and southern Indian food with us. My gratitude also belongs to Jan Stühmer for his advice and guidance on the final meters of this project.

I want to thank all my colleagues from the Chair of Information Systems for being part of this journey and for keeping up with my notation-heavy presentations. Special thanks to Daniel Betsche for always having an open ear and sharing an office with me, and to Tobias Fuchs for our collaborative work.

Last but not least, thank you, Mum and Dad, for setting me on the way. Thank you, Melanie, for sharing your life with me and always trusting in my abilities when proofs become challenging.

# Abstract

Kernel methods have been at the forefront of data science for several decades and provide the basis of some of the most powerful and principled machine learning algorithms currently known. The key properties rendering kernel methods ubiquitous are the number of domains they have been designed for, the Hilbert structure of the function class associated with kernels facilitating their statistical analysis, and their ability to represent probability measures as elements in a reproducing kernel Hilbert space without loss of information under very mild assumptions. These properties have led to the invention of many kernel-based information theoretical measures such as the maximum mean discrepancy (MMD; also known as energy distance in the statistics literature), quantifying the difference of two distributions, the Hilbert-Schmidt independence criterion (HSIC; also known as distance covariance in the statistics literature), quantifying the dependence of a distribution, or kernel Stein discrepancies (KSD), quantifying the difference of a distribution to a given target. These measures have found numerous applications, most prominently in designing two-sample, independence, and goodness-of-fit tests. However, while powerful, their classical U- and V-statistic-based estimators have a runtime complexity that scales quadratically with the number of samples $n$, prohibiting their application to large-sample settings. To tackle this severe limitation, this dissertation makes the following contributions.

We propose the first accelerated Nyström-based HSIC estimator capable of handling more than two random variables, prove its $\sqrt{n}$-consistency, and evaluate its performance on synthetic data, dependency testing of media annotations, and causal discovery. Further, we establish the minimax optimal rate of HSIC estimation for continuous bounded translation-invariant kernels on $\mathbb{R}^d$ for Borel measures containing the Gaussians to be $O\left(n^{-1/2}\right)$, settling a question that has been open since the introduction of HSIC 20 years ago. In this setting, the result also implies the minimax optimality of our proposed HSIC acceleration. Regarding KSD, we propose a Nyström-based acceleration, prove its $\sqrt{n}$-consistency with a classical sub-Gaussian assumption, and show its state-of-the-art performance in goodness-of-fit testing on a suite of benchmarks. Last, we design an efficient online approximation of MMD that allows its computation on data streams and gives rise to a powerful change detection algorithm. Extensive experiments show that the proposed change detector achieves state-of-the-art performance on synthetic and real-world data.

Overall, this dissertation advances our understanding of estimating kernel-based information theoretical measures and establishes fundamental tools for analyzing their accelerations. All code replicating the experiments is made openly available.

# Zusammenfassung

Kernelmethoden bilden die Grundlage für einige der leistungsstärksten und fundiertesten Algorithmen für maschinelles Lernen. Die Eigenschaften, welche Kernelmethoden omnipräsent machen, sind die Anzahl der Domänen, für die sie entwickelt wurden, die Hilbert-Struktur der mit Kerneln verbundenen Funktionsklasse, die ihre statistische Analyse erlaubt, und die Möglichkeit, Wahrscheinlichkeitsmaße als Elemente in einem reproduzierenden Kernel-Hilbert-Raum ohne Informationsverlust und unter sehr milden Annahmen abzubilden. All diese Eigenschaften haben zur Entwicklung zahlreicher kernelbasierter informationstheoretischer Maße geführt, wie zum Beispiel der Maximum Mean Discrepancy (MMD; in der Statistikliteratur auch als "energy distance" bezeichnet), die den Unterschied zwischen zwei Wahrscheinlichkeitsmaßen quantifiziert; dem Hilbert-Schmidt Independence Criterion (HSIC; in der Statistikliteratur auch als "distance covariance" bezeichnet), welches die (Un-)abhängigkeit einer Verteilung quantifiziert; oder der Kernel Stein Discrepancy (KSD), die den Unterschied einer Verteilung zu einem gegebenen Ziel quantifiziert. Diese Maße haben zahlreiche Anwendungen gefunden, vor allem bei der Entwicklung von Zweistichproben-, Unabhängigkeits- und Anpassungsgütetests. Die existierenden U- und V-Statistik-basierten Schätzer sind zwar leistungsfähig, haben aber eine Laufzeitkomplexität, die quadratisch mit der Stichprobengröße $n$ wächst, was ihre Anwendung auf große Stichproben stark beeinträchtigt. Um dieser schwerwiegenden Einschränkung zu begegnen, leistet diese Dissertation die folgenden Beiträge.

Wir schlagen den ersten beschleunigten Nyström-basierten HSIC-Schätzer vor, der mehr als zwei Zufallsvariablen verarbeiten kann, beweisen seine $\sqrt{n}$-Konsistenz und evaluieren seine Leistung auf synthetischen Daten, Abhängigkeitstests von Medienannotationen und dem Finden von kausalen Zusammenhängen. Darüber hinaus zeigen wir, dass die minimax-optimale Rate der HSIC-Schätzung für kontinuierliche, beschränkte, translationsinvariante Kernel auf $\mathbb{R}^d$ für Borel-Maße, die die Normalverteilungen enthalten, $O\left(n^{-1/2}\right)$ ist. Damit beantworten wir eine Frage, die seit der Einführung von HSIC vor mehr als 20 Jahren unbeantwortet war. Unser Ergebnis impliziert auch die Minimax-Optimalität der von uns vorgeschlagenen HSIC-Beschleunigung. In Bezug auf KSD schlagen wir ebenfalls eine Nyström-basierte Beschleunigung vor, beweisen ihre $\sqrt{n}$-Konsistenz mit einer klassischen sub-Gaußschen Annahme und zeigen mit einer Reihe von Anpassungsgütetest-Benchmarks, dass diese den bisherigen Stand der Forschung übertrifft. Schließlich entwerfen wir eine effiziente Online-Approximation von MMD, die deren Berechnung auf Datenströmen ermöglicht und die Basis für einen leistungsstarken Change Detection Algorithmus bietet. Umfangreiche Experimente zeigen, dass der vorgeschlagene Algorithmus sowohl auf synthetischen als auch auf realen Daten eine herausragende Leistung erzielt.

Insgesamt leistet diese Dissertation einen wissenschaftlichen Beitrag, indem sie beschleunigte Schätzer für kernelbasierte informationstheoretische Maße vorstellt und Werkzeuge für deren Analyse einführt. Unsere theoretischen und experimentellen Ergebnisse zeigen die hervorragenden Eigenschaften dieser Schätzer. Sämtlicher Code für die Replikation der Experimente ist frei verfügbar.

# A Guide to Notation

Below is a summary of the fundamental definitions used throughout this work. All kernel-related quantities are introduced in Chapter 2 and indexed at the end of this guide.

**General.** For a positive integer $M$, $[M] := \{1, \dots, M\}$. For positive sequences $(a_n)_{n=1}^{\infty}$ and $(b_n)_{n=1}^{\infty}$, $b_n = O(a_n)$ if there exist constants $C > 0$ and $n_0 \in \mathbb{N}_{>0}$ such that $b_n \le C a_n$ for all $n \ge n_0$; $a_n = \Theta(b_n)$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. $a_n = \tilde{O}(b_n)$ means $a_n = O(b_n)$ up to logarithmic terms; $a_n = \tilde{\Theta}(b_n)$ is defined likewise. For $a_1, a_2 \ge 0$, $a_1 \lesssim a_2$ (resp. $a_1 \gtrsim a_2$) means that $a_1 \le C a_2$ (resp. $a_1 \ge C' a_2$) for an absolute constant $C > 0$ (resp. $C' > 0$), and we write $a_1 \asymp a_2$ iff. $a_1 \lesssim a_2$ and $a_1 \gtrsim a_2$. We write $\mathbb{1}_A$ for the indicator function of a set $A$, $\{\{\cdot\}\}$ for a multiset, and $\times_{m=1}^{M} A_m$ for the Cartesian product of sets $(A_m)_{m=1}^{M}$.

**Linear algebra.** The Euclidean inner product of vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ is denoted by $\langle \mathbf{v}, \mathbf{w} \rangle$; the Euclidean norm is $\|\mathbf{v}\|_{\mathbb{R}^d} := \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$. The Hadamard product of matrices $\mathbf{A}_m \in \mathbb{R}^{d_1 \times d_2}$ of equal size ($m \in [M]$) is $\circ_{m \in [M]} \mathbf{A}_m := \left[ \prod_{m \in [M]} (\mathbf{A}_m)_{i,j} \right]_{i,j=1}^{d_1, d_2}$. Matrix multiplication takes precedence over the Hadamard one. For a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, $\mathrm{tr}(\mathbf{A}) := \sum_{i \in [d]} A_{i,i}$ denotes its trace, $\mathbf{A}^{-1}$ is its inverse (assuming that $\mathbf{A}$ is non-singular), and $\mathbf{A}^{-}$ is its Moore–Penrose inverse. The transpose of a matrix $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}$ is denoted by $\mathbf{A}^{\mathsf{T}} \in \mathbb{R}^{d_2 \times d_1}$. The Frobenius norm of a matrix $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}$ is $\|\mathbf{A}\|_{\mathrm{F}} := \sqrt{\sum_{i \in [d_1], j \in [d_2]} (A_{i,j})^2}$. The $d$-dimensional vector of ones is $\mathbf{1}_d$. The $d \times d$-sized identity matrix is denoted by $\mathbf{I}_d$. $\mathrm{bdiag}(\mathbf{M}_1, \dots, \mathbf{M}_N)$ forms a block-diagonal matrix from its arguments $(\mathbf{M}_n)_{n=1}^{N}$ ($\mathbf{M}_n \in \mathbb{R}^{d_n \times d_n}$, $n \in [N]$) and $|\mathbf{A}|$ denotes the determinant of a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$. For a set $S$ in a vector space, $\mathrm{span}(S)$ denotes the linear hull of $S$.

**Functional analysis.** Let $\mathcal{H}$ be a separable Hilbert space. A linear operator $A : \mathcal{H} \to \mathcal{H}$ is called bounded if $\|A\|_{\mathrm{op}} := \sup_{\|h\|_{\mathcal{H}}=1} \|Ah\|_{\mathcal{H}} < \infty$; the set of $\mathcal{H} \to \mathcal{H}$ bounded linear operators is denoted by $\mathcal{L}(\mathcal{H})$. Let $A \in \mathcal{L}(\mathcal{H})$. If it exists, $A^{-1}$ denotes the inverse of $A$; it is also bounded linear. $A$ is called positive (shortly $A \succcurlyeq 0$) if it is self-adjoint ($A^* = A$, where $A^* \in \mathcal{L}(\mathcal{H})$ is defined by $\langle Af, g \rangle_{\mathcal{H}} = \langle f, A^* g \rangle_{\mathcal{H}}$ for all $f, g \in \mathcal{H}$), and $\langle Ah, h \rangle_{\mathcal{H}} \ge 0$ for all $h \in \mathcal{H}$. If $A \succcurlyeq 0$, then there exists a unique $B \succcurlyeq 0$ such that $B^2 = A$; we write $B = A^{\frac{1}{2}}$ and call $B$ the square root of $A$. An $A \in \mathcal{L}(\mathcal{H})$ is called trace-class if $\sum_{i \in I} \langle (A^* A)^{\frac{1}{2}} e_i, e_i \rangle_{\mathcal{H}} < \infty$ for some countable orthonormal basis (ONB) $(e_i)_{i \in I}$ of $\mathcal{H}$, and in this case $\mathrm{tr}(A) := \sum_{i \in I} \langle Ae_i, e_i \rangle_{\mathcal{H}} < \infty$.[1] For a self-adjoint trace-class operator $A$ with eigenvalues $(\lambda_i)_{i \in I}$, $\mathrm{tr}(A) = \sum_{i \in I} \lambda_i$. An operator $A \in \mathcal{L}(\mathcal{H})$ is called compact if $\overline{\{Ah \mid h \in \mathcal{H}, \|h\|_{\mathcal{H}} \le 1\}}$ is compact, where $\overline{\{\cdot\}}$ denotes the closure. A trace class operator is compact, and a compact positive operator $A$ has largest eigenvalue $\|A\|_{\mathrm{op}}$. For any $A \in \mathcal{L}(\mathcal{H})$, it holds that $\|A^* A\|_{\mathrm{op}} = \|A\|_{\mathrm{op}}^2$ (which is called the $C^*$ property). Given a closed linear subspace $U \subseteq \mathcal{H}$, the (orthogonal) projection of $h \in \mathcal{H}$ on $U$ is denoted by $P_U h \in U$; $u = P_U h$ is the unique vector such that $h - u \perp U$. For any $u \in U$, $\|h - P_U h\|_{\mathcal{H}} \le \|h - u\|_{\mathcal{H}}$, that is, $P_U h$ is the closest element in $U$ to $h$.

[1] The trace-class property and the value of $\mathrm{tr}(A)$ is independent of the specific ONB chosen. The separability of $\mathcal{H}$ implies the existence of a countable ONB in it.

**Probability and measure.** Let $(\mathcal{X}, \tau_\mathcal{X})$ be a topological space and $\mathcal{B}(\tau_\mathcal{X})$ the Borel sigma-algebra induced by the topology $\tau_\mathcal{X}$. All probability measures in this thesis are meant with respect to the measurable space $(\mathcal{X}, \mathcal{B}(\tau_\mathcal{X}))$, and they are denoted by $\mathcal{M}_1^+(\mathcal{X})$. For $x \in \mathcal{X}$ and $B \in \mathcal{B}(\tau_\mathcal{X})$, the Dirac measure $\delta_x(B) := \mathbb{1}_B(x)$. The $n$-fold product measure of $\mathbb{P} \in \mathcal{M}_1^+(\mathcal{X})$ is denoted by $\mathbb{P}^n \in \mathcal{M}_1^+(\mathcal{X}^n)$. The product of $\mathbb{P}_1 \in \mathcal{M}_1^+(\mathcal{X}_1)$ and $\mathbb{P}_2 \in \mathcal{M}_1^+(\mathcal{X}_2)$ is written as $\mathbb{P}_1 \otimes \mathbb{P}_2$ $(\in \mathcal{M}_1^+(\mathcal{X}_1 \times \mathcal{X}_2))$, where $\left(\mathcal{X}_1, \tau_{\mathcal{X}_1}\right)$ and $\left(\mathcal{X}_2, \tau_{\mathcal{X}_2}\right)$ are topological spaces. For $\mathbb{P} \in \mathcal{M}_1^+(\mathcal{X})$ absolutely continuous w.r.t. $\mathbb{Q} \in \mathcal{M}_1^+(\mathcal{X})$ (iff. for any $B \in \mathcal{B}(\tau_\mathcal{X})$ for which $\mathbb{Q}(B) = 0$, $\mathbb{P}(B) = 0$), we denote the Radon-Nikodym derivative by $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}$. For a sequence of $r_n > 0$-s and a sequence of real-valued random variables $X_n$, $X_n = O_P(r_n)$ means that $\frac{X_n}{r_n}$ is bounded in probability. For $r \in \{1, 2\}$, let $\psi_r(u) = e^{u^r} - 1$ and $\|X\|_{\psi_r} := \inf\left\{C > 0 \mid \mathbb{E}_{X\sim\mathbb{P}}\psi_r\left(\frac{|X|}{C}\right) \leq 1\right\}$. A real-valued random variable $X \sim \mathbb{P} \in \mathcal{M}_1^+(\mathbb{R})$ is called sub-exponential if $\|X\|_{\psi_1} < \infty$ and sub-Gaussian if $\|X\|_{\psi_2} < \infty$. For $r \geq 1$, a measure space $(\mathcal{X}, \mathcal{B}(\tau_\mathcal{X}), \Lambda)$, and a real-valued measurable function $f : (\mathcal{X}, \mathcal{B}(\tau_\mathcal{X})) \to (\mathbb{R}, \mathcal{B}(\tau_\mathbb{R}))$, $\|f\|_{L^r(\mathcal{X},\mathcal{B}(\tau_\mathcal{X}),\Lambda)} := \left(\int_\mathcal{X} |f(x)|^r \mathrm{d}\Lambda(x)\right)^{\frac{1}{r}}$. We write $L^2\left(\mathbb{R}^d, \Lambda\right) := L^2\left(\mathbb{R}^d, \mathcal{B}(\tau_{\mathbb{R}^d}), \Lambda\right)$ for the Hilbert space of (equivalence classes of) measurable functions $f : \left(\mathbb{R}^d, \mathcal{B}(\tau_{\mathbb{R}^d})\right) \to (\mathbb{R}, \mathcal{B}(\tau_\mathbb{R}))$ for which $\|f\|_{L^2\left(\mathbb{R}^d,\Lambda\right)} < \infty$. For a measure space $\left(\mathbb{R}^d, \mathcal{B}(\tau_{\mathbb{R}^d}), \Lambda\right)$, the support of $\Lambda$, written as $\mathrm{supp}(\Lambda)$, is the subset of $\mathbb{R}^d$ for which every open neighborhood of $\mathbf{x} \in \mathbb{R}^d$ has positive measure [Cohn, 2013, p. 207].

We introduce the following notations in the main text and collect them here for easy reference.

| Notation | Description | Page |
|---|---|---|
| $\mathcal{H}_k$ | reproducing kernel Hilbert space | 5 |
| $k$ | kernel function | 5 |
| $\phi_k(x)$ | $= k(\cdot, x)$, canonical feature map | 5 |
| $\otimes_{m=1}^M k_m$ | tensor product of kernels | 6 |
| $\partial_i f(\mathbf{x})$ | partial derivative | 7 |
| $\nabla_\mathbf{x} f(\mathbf{x})$ | vector of partial derivatives | 7 |
| $\partial_i \partial_j f(\mathbf{x})$ | mixed partial derivative | 7 |
| $D_f(\mathbb{P}, \mathbb{Q})$ | $f$-divergence | 7 |
| $\gamma_\mathcal{F}(\mathbb{P}, \mathbb{Q})$ | integral probability metric | 8 |
| $\mu_k(\mathbb{P})$ | kernel mean embedding | 8 |
| $C_{\mathbb{P},k}$ | covariance operator | 9 |
| $C_{\mathbb{P},k,\lambda}$ | regularized covariance operator | 9 |
| $\mathcal{N}_{\mathbb{P},k}(\lambda)$ | effective dimension | 9 |
| $\tilde{C}_{\mathbb{P},k}$ | cross-covariance operator | 10 |
| $\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q})$ | maximum mean discrepancy | 10 |
| $\mathrm{HSIC}_k(\mathbb{P})$ | Hilbert-Schmidt independence criterion | 11 |
| $S_p(\mathbb{Q})$ | kernel Stein discrepancy | 13 |

# Contents

# 1. Introduction

Kernel methods [Aronszajn, 1950] have been on the forefront of data science for over two decades [Schölkopf and Smola, 2002, Steinwart and Christmann, 2008], and they underpin some of the most powerful and principled machine learning techniques currently known. The key idea of kernels is to map the data into a (possibly infinite-dimensional) feature space—a reproducing kernel Hilbert space (RKHS)—in which one computes the inner product implicitly through a symmetric, positive definite function, the so-called kernel function.

Kernel methods are widely applicable as kernel functions have been designed for strings [Watkins, 1999, Lodhi et al., 2002, Cuturi and Vert, 2005], time series and sequences [Cuturi et al., 2007, Cuturi, 2011, Király and Oberhauser, 2019], sets [Haussler, 1999, Gärtner et al., 2002], rankings [Jiao and Vert, 2016], fuzzy domains [Guevara et al., 2017], and graphs [Gärtner et al., 2003, Vishwanathan et al., 2010, Borgwardt et al., 2020, Nikolentzos and Vazirgiannis, 2023]. In particular, their extension to the space of probability measures [Berlinet and Thomas-Agnan, 2004, Hein and Bousquet, 2005, Smola et al., 2007, Sriperumbudur et al., 2010] allows to represent distributions in an RKHS by the so-called mean embedding.

Such embeddings form the main building block of maximum mean discrepancy (MMD; Smola et al. [2007], Gretton et al. [2012]), which quantifies the discrepancy of two distributions as the RKHS distance of their respective mean embeddings. MMD is (i) a semi-metric on probability measures, (ii) a metric iff. the kernel is characteristic [Fukumizu et al., 2007, Sriperumbudur et al., 2010], (iii) an instance of an integral probability metric (IPM; Zolotarev 1983, Müller 1997) when the underlying function class of the IPM is chosen to be the unit ball in an RKHS. MMD also allows capturing the (in)dependence of a distribution by considering the discrepancy of a joint distribution to the product of its marginals, which then gives rise to the Hilbert-Schmidt independence criterion (HSIC; Gretton et al. 2005, Quadrianto et al. 2009, Pfister et al. 2018). Further, given a fixed target measure and using Stein's method [Stein, 1972, Chen, 2021, Anastasiou et al., 2023] to design a specific RKHS, so-called kernel Stein discrepancies (KSD; Liu et al. 2016, Chwialkowski et al. 2016) capture the discrepancy of a distribution to the target. In the latter approach, even partial knowledge of the target, for example, knowledge of the probability density function up to the normalizing constant, can suffice, which renders the approach particularly appealing in Bayesian contexts.

Like classical information-theoretical quantities, for example, the Kullback-Leibler divergence [Kullback and Leibler, 1951] or the Kolmogorov-Smirnov distance [Kolmogorov, 1933, Smirnov, 1948], MMD underpins many data science tasks—but, in contrast to most classical quantities, MMD is easy to estimate on any kernel-enriched domain as it permits closed-form estimators, typically as a combination of U-statistics or V-statistics [Hoeffding, 1948, v. Mises, 1947], which have been well-studied [Serfling, 1980] independently from kernel methods.

With their broad applicability, closed-form estimators, and solid theoretical foundations, kernel-based information theoretical measures have found numerous applications. Examples include two-sample testing [Baringhaus and Franz, 2004, Székely and Rizzo, 2004, 2005, Harchaoui and Cappé, 2007, Gretton et al., 2012], independence testing in batch [Gretton et al., 2008, Wehbe and Ramdas, 2015, Bilodeau and Nangue, 2017, Górecki et al., 2018, Pfister et al., 2018, Albert et al., 2022, Shekhar et al., 2023] and

streaming [Podkopaev et al., 2023] settings, feature selection [Camps-Valls et al., 2010, Song et al., 2012, Yamada et al., 2014, Wang et al., 2022] with applications in biomarker detection [Climente-González et al., 2019] and wind power prediction [Bouche et al., 2023], clustering [Song et al., 2007, Climente-González et al., 2019], and causal discovery [Mooij et al., 2016, Pfister et al., 2018, Chakraborty and Zhang, 2019, Schölkopf et al., 2021]. In addition, recent successful applications are in sensitivity analysis [Veiga, 2015, Freitas Gustavo et al., 2023, Fellmann et al., 2024, Herrando-Pérez and Saltré, 2024], in the context of uncertainty quantification [Stenger et al., 2020], for the analysis of data augmentation methods for brain tumor detection [Anaya-Isaza and Mera-Jiménez, 2022], and that of multimodal neural networks trained on neuroimaging data [Fedorov et al., 2024]. Further applications are goodness-of-fit testing [Liu et al., 2016, Chwialkowski et al., 2016, Balasubramanian et al., 2021, Hagrass et al., 2025], change point detection in the offline [Harchaoui and Cappé, 2007] and online [Li et al., 2019, Keriven et al., 2020, Wei and Xie, 2022] setting, regression [Szabó et al., 2016, Law et al., 2018], and training of generative adversarial neural networks [Dziugaite et al., 2015, Li et al., 2015, Binkowski et al., 2018], among many others.

However, the classical U-statistic and V-statistic-based estimators for the kernel-based information theoretical measures have a runtime complexity that is quadratic in the number of observations, hindering their application in large-scale or online settings. While building blocks for accelerating the computations exist, typically relying on random Fourier features (RFF; Rahimi and Recht 2007, Sriperumbudur and Szabó 2015), the incomplete Cholesky factorization [Wright, 1999, Bach and Jordan, 2002], or the Nyström method [Nyström, 1930, Williams and Seeger, 2001], these leave two questions open. The first is the application of the method to specific tasks like HSIC or KSD estimation. The second, more fundamental question is understanding the computational-statistical trade-off: the price that one pays w.r.t. statistical accuracy for the gain in computational efficiency. Answering the latter sheds light on the interplay between the computational and statistical requirements of learning.

For RFF-based approaches, which approximate the feature map of a translation-invariant kernel by a finite-dimensional vector, both questions have been tackled, for example, in the case of kernel principal component analysis (PCA; Sriperumbudur and Sterge 2022) and kernel ridge regression [Li et al., 2021]. The Nyström-based acceleration, which relies on a subsample of the data and places no assumptions on the kernel function, has also been analyzed in the case of kernel PCA [Sterge and Sriperumbudur, 2022], kernel ridge regression [Rudi et al., 2015], and computation of the kernel mean embedding and MMD [Chatalic et al., 2022].

In a similar spirit, this thesis tackles the quadratic runtime bottlenecks of MMD, HSIC, and KSD, respectively. In particular, we introduce efficient Nyström-based HSIC and KSD estimators, and analyze their computational-statistical trade-off. We establish that, given a suitably chosen subsample size, the proposed algorithms preserve the statistical accuracy of the quadratic-time estimators while reducing their runtime requirements. Further, we show the minimax lower-bound of HSIC estimation for a rich class of kernel- and distribution pairs, which, while also of independent interest, nicely settles the optimality of our proposed HSIC approximation. Last, we design an online MMD approximation scheme and validate its performance on change detection tasks. The following section gives an overview of the contributions.

## 1.1. Contributions

The contributions of this dissertation can be coarsely summarized as follows; we provide a more nuanced account of the respective contributions at the beginnings of Chapters 3–6, respectively.

1. We propose a Nyström-based accelerated HSIC estimator that can handle $M \geq 2$ random variables, show its $\sqrt{n}$-consistency (where $n$ is the sample size), and validate the proposed method via experiments on synthetic data, dependency testing of media annotations, and causal discovery.
2. We establish the minimax lower bound $O\left(n^{-1/2}\right)$ of HSIC estimation with $M \geq 2$ components on $\mathbb{R}^d$ with continuous bounded translation-invariant characteristic kernels. As this lower bound matches the known upper bounds of the existing "classical" U-statistic and V-statistic-based estimators, and that of our Nyström HSIC estimator, this result settles their minimax optimality.
3. We apply the Nyström-based acceleration to KSD and propose an estimator with a runtime of $O\left(mn + m^3\right)$, with $n$ samples and $m \ll n$ Nyström points, along with an accelerated wild bootstrap for goodness-of-fit testing. We lift the usual boundedness assumption on the feature map by considering a sub-Gaussian assumption instead, allowing us to establish the $\sqrt{n}$-consistency of our proposed acceleration. Our method achieves state-of-the-art results on a suite of benchmarks.
4. Our last contribution is a change detection algorithm, MMDEW, which builds on our novel online approximation of MMD. When considering a data stream with $t$ observations, MMDEW has a memory requirement of $O(\log t)$ and a runtime complexity of $O\left(\log^2 t\right)$ for each new observation. Extensive experiments show that MMDEW achieves state-of-the-art change detection performance across a range of synthetic and real-world data sets.

All code reproducing our results is openly available, with links in the chapters corresponding to the following publications.

## 1.2. Publications

The work within this dissertation appears in the following publications.

- F. Kalinke and Z. Szabó. Nyström M-Hilbert-Schmidt independence criterion. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 1005–1015, 2023. (Accepted for poster spotlight). Parts of the work were carried out while the author was a research associate at the Department of Statistics, London School of Economics.
- F. Kalinke and Z. Szabó. The minimax rate of HSIC estimation for translation-invariant kernels. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 108468–108489, 2024.
- F. Kalinke, Z. Szabó, and B. K. Sriperumbudur. Nyström kernel Stein discrepancy. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 388–396, 2025b.
- F. Kalinke, M. Heyden, G. Gntuni, E. Fouché, and K. Böhm. Maximum mean discrepancy on exponential windows for online change detection. *Transactions on Machine Learning Research*, 2025a.

We have reused these prior works' content but restructured and rephrased it to form a coherent monograph. In particular, we unified all notations and introductions. Chapters 3–6, which are the main part of this dissertation, explicitly state which prior work they are based on. All other chapters may reuse content from each of these prior works.

## 1.3. Thesis outline

The remainder of this work is structured as follows. We introduce our notations and the information-theoretical quantities that we consider in Chapter 2. Chapter 3 details our proposed HSIC acceleration, its analysis, and our experiments. In Chapter 4, we derive the minimax lower bound of HSIC estimation. The Nyström-based KSD acceleration, with guarantees and experiments, is in Chapter 5. Chapter 6 details our online change point detection algorithm. Conclusions and future work are in Chapter 7. We collect the external results that we use in the appendices.

# 2. Background

In this chapter, we recall a few facts on reproducing kernel Hilbert spaces (Section 2.1) before we introduce information theoretical measures in general (Section 2.2) and, in particular, our quantities of interest (Section 2.3).

## 2.1. Reproducing kernel Hilbert space

A reproducing kernel Hilbert space (RKHS; Aronszajn 1950) is a function space, which, in a sense made precise in a second, contains very well-behaved functions. The structure of RKHSs is well understood [Berlinet and Thomas-Agnan, 2004, Steinwart and Christmann, 2008, Saitoh and Sawano, 2016, Paulsen and Raghupathi, 2016], which allows the statistical analysis of RKHS-based approaches, and these spaces are well-suited for computations. Both properties will be evident throughout this thesis.

**Definition 2.1.1** (Reproducing kernel Hilbert space; RKHS)**.** *Let $\mathcal{X}$ be a nonempty set. The RKHS $\mathcal{H}_k$ on $\mathcal{X}$ associated with a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the Hilbert space of functions $h : \mathcal{X} \to \mathbb{R}$ such that*

*(i) $k(\cdot, x) \in \mathcal{H}_k$ for all $x \in \mathcal{X}$, and*

*(ii) $\langle h, k(\cdot, x)\rangle_{\mathcal{H}_k} = h(x)$ for all $x \in \mathcal{X}$ and $h \in \mathcal{H}_k$,*

*where $k(\cdot, x)$ stands for $x' \in \mathcal{X} \mapsto k(x', x) \in \mathbb{R}$ with $x \in \mathcal{X}$ fixed.*

Definition 2.1.1(i) and (ii) directly yield that

$$k(x, x') = \langle k(\cdot, x), k(\cdot, x')\rangle_{\mathcal{H}_k} =: \langle \phi_k(x), \phi_k(x')\rangle_{\mathcal{H}_k} \text{ for all } x, x' \in \mathcal{X}, \tag{2.1}$$

and $\phi_k$ is referred to as the canonical feature map.

To allow the definition of the kernel mean embedding—one of the key ingredients for the quantities in which we are interested—, we make the following assumption throughout this work.

**Assumption 2.1.1.** *We assume that $(\mathcal{X}, \tau_{\mathcal{X}})$ is a topological space. All kernels are assumed to be measurable and $\mathcal{H}_k$ is assumed to be separable.*

The separability of $\mathcal{H}_k$ can be guaranteed on a separable topological space $(\mathcal{X}, \tau_{\mathcal{X}})$ by taking a continuous kernel $k$ [Steinwart and Christmann, 2008, Lemma 4.33].

In the remainder of this section, we explore a few properties of kernels, setting the stage for later chapters.

Kernel functions permit certain algebraic operations. For $m \in [M]$ with $M \geq 2$, let $k_m : \mathcal{X}_m \times \mathcal{X}_m \to \mathbb{R}$ be a kernel on a space $\mathcal{X}_m$ with associated RKHS $\mathcal{H}_{k_m}$. The tensor product of the kernels $(k_m)_{m=1}^M$

$$\otimes_{m=1}^M k_m \left( (x_m)_{m=1}^M, (x'_m)_{m=1}^M \right) := \prod_{m \in [M]} k_m(x_m, x'_m),$$

with $x_m, x'_m \in \mathcal{X}_m$ ($m \in [M]$), is also a kernel; we will use the shorthand $k = \otimes_{m=1}^M k_m$. The associated RKHS has a simple structure $\mathcal{H}_k = \otimes_{m=1}^M \mathcal{H}_{k_m}$ [Berlinet and Thomas-Agnan, 2004] with the r.h.s. denoting the tensor product of the RKHSs $(\mathcal{H}_{k_m})_{m=1}^M$. Indeed, for $h_m, v_m \in \mathcal{H}_{k_m}$, the multi-linear operator $\otimes_{m=1}^M h_m \in \otimes_{m=1}^M \mathcal{H}_{k_m}$ acts as $\otimes_{m=1}^M h_m(v_1, \ldots, v_M) = \prod_{m \in [M]} \langle h_m, v_m \rangle_{\mathcal{H}_{k_m}}$. The space $\otimes_{m=1}^M \mathcal{H}_{k_m}$ is the closure of the linear combination of such $\otimes_{m=1}^M h_m$-s:

$$\otimes_{m=1}^M \mathcal{H}_{k_m} = \overline{\text{span}} \left( \otimes_{m=1}^M h_m \ : \ h_m \in \mathcal{H}_{k_m}, m \in [M] \right),$$

where the closure is meant w.r.t. to the (linear extension of the) inner product defined as

$$\left\langle \otimes_{m=1}^M a_m, \otimes_{m=1}^M b_m \right\rangle_{\otimes_{m=1}^M \mathcal{H}_{k_m}} := \prod_{m \in [M]} \langle a_m, b_m \rangle_{\mathcal{H}_{k_m}}, \quad a_m, b_m \in \mathcal{H}_{k_m}. \tag{2.2}$$

Specifically, (2.2) implies that

$$\left\| \otimes_{m=1}^M a_m \right\|_{\otimes_{m=1}^M \mathcal{H}_{k_m}} = \prod_{m \in [M]} \|a_m\|_{\mathcal{H}_{k_m}}. \tag{2.3}$$

Next, we introduce a common assumption that we make in Chapter 3, Chapter 4, and Chapter 6.

**Assumption 2.1.2.** *A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is bounded iff. there exists $B \geq 0$ such that*

$$\sup_{x, x' \in \mathcal{X}} \sqrt{k(x, x')} \leq B.$$

A kernel $k$ is bounded iff. the associated feature map $\phi_k$ is bounded, that is, $\sup_{x \in \mathcal{X}} \|\phi_k(x)\|_{\mathcal{H}_k} \overset{(2.1)}{=} \sup_{x \in \mathcal{X}} \|k(\cdot, x)\|_{\mathcal{H}_k} = \sup_{x \in \mathcal{X}} \sqrt{k(x, x)} \leq B$. We note that, in some cases, Assumption 2.1.2 is too strict. We tackle one particular case in Chapter 5.

If one restricts the domain of the kernel to be $\mathcal{X} = \mathbb{R}^d$, continuous bounded translation-invariant kernels permit an alternative representation, which we detail after introducing the necessary definition. Recall that a function $\kappa : \mathbb{R}^d \to \mathbb{R}$ is called positive definite if $\sum_{i,j \in [n]} c_i c_j \kappa(\mathbf{x}_i - \mathbf{x}_j) \geq 0$ for all $n \in \mathbb{N}_{>0}$, $\mathbf{c} = (c_i)_{i=1}^n \in \mathbb{R}^n$, and $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$.

**Definition 2.1.2** (Translation-invariant kernel)**.** *A kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is said to be translation-invariant if there exists a positive definite function $\kappa : \mathbb{R}^d \to \mathbb{R}$ such that $k(\mathbf{x}, \mathbf{x}') = \kappa(\mathbf{x} - \mathbf{x}')$ for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$.*

By Bochner's theorem [Wendland, 2005, Theorem 6.6] (recalled in Theorem A.2.1) for a continuous bounded translation-invariant kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ there exists a finite non-negative Borel measure $\Lambda_k$ such that

$$k(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x} - \mathbf{y}, \boldsymbol{\omega} \rangle} \mathrm{d}\Lambda_k(\boldsymbol{\omega}) \tag{2.4}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, with $i := \sqrt{-1}$. We use representation (2.4) in Chapter 4.

The RKHS of a continuously differentiable kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} \subset \mathbb{R}^d$ is open, contains the derivatives of the feature maps [Steinwart and Christmann, 2008, Lemma 4.34]. The details are as follows; see Section 2.3.3 and Chapter 5 for an application.

For a (twice) differentiable $f : \mathbb{R}^d \to \mathbb{R}$ write

$$\partial_i f(\mathbf{x}) := \frac{\partial f(\mathbf{x})}{\partial x_i}, \qquad \nabla_{\mathbf{x}} f(\mathbf{x}) := (\partial_i f(\mathbf{x}))_{i=1}^d \in \mathbb{R}^d, \qquad \text{and} \qquad \partial_i \partial_j f(\mathbf{x}) := \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}.$$

We may interpret $k$ as $\tilde{k} : \mathbb{R}^{2d} \to \mathbb{R}$. As it is known that the mixed partial derivative $\partial_i \partial_{i+d} k(\mathbf{x}, \mathbf{x}')$ is equivalent to $\partial_i \partial_{i+d} \tilde{k}(\mathbf{x}, \mathbf{x}')$, let $\partial_i \partial_{i+d} k(\mathbf{x}, \mathbf{x}') := \partial_i \partial_{i+d} \tilde{k}(\mathbf{x}, \mathbf{x}')$.[1] Assume that, for $i \in [d]$, $\partial_i \partial_{i+d} k(\mathbf{x}, \mathbf{x}')$ exists and is continuous. Then $\partial_i k(\cdot, \mathbf{x})$ exists and for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$

$$\langle \partial_i k(\cdot, \mathbf{x}), \partial_i k(\cdot, \mathbf{x}') \rangle_{\mathcal{H}_k} = \partial_i \partial_{i+d} k(\mathbf{x}, \mathbf{x}') = \partial_{i+d} \partial_i k(\mathbf{x}, \mathbf{x}').$$

In particular, for $h \in \mathcal{H}_k$, one has the derivative-reproducing property [Zhou, 2008, Theorem 1]

$$\partial_i h(\mathbf{x}) = \langle h, \partial_i k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}_k}, \tag{2.5}$$

which directly implies that

$$\nabla_{\mathbf{x}} h(\mathbf{x}) = \left( \langle h, \partial_i k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}_k} \right)_{i=1}^d \in \mathbb{R}^d.$$

We introduce a few information theoretical measures in the next section.

## 2.2. Information theoretical measures

Information theoretical measures, such as Kullback-Leibler divergence [Kullback and Leibler, 1951] or Kolmogorov-Smirnov distance [Kolmogorov, 1933, Smirnov, 1948], are functionals on probability measures and provide the basis for foundational tasks in statistics and data science like two-sample testing, independence testing, or goodness-of-fit testing. The quantities that we consider are particular instances of the wider class of so-called integral probability metrics (IPMs; Zolotarev 1983, Müller 1997). $f$-divergences [Ali and Silvey, 1966, Csiszár, 1967] are another pillar for constructing such measures. We quickly recall the abstract setting, before we give a detailed account of our quantities of interest.

**Definition 2.2.1** ($f$-divergence)**.** *Let $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_1^+(\mathcal{X})$ and $f : [0, \infty) \to (-\infty, \infty]$ a convex function.[2] The $f$-divergence of $\mathbb{P}$ and $\mathbb{Q}$ is*

$$D_f(\mathbb{P}, \mathbb{Q}) = \int_{\mathcal{X}} f\left( \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x) \right) \mathrm{d}\mathbb{Q}(x) \tag{2.6}$$

*if $\mathbb{P}$ is absolutely continuous w.r.t. $\mathbb{Q}$ and $+\infty$ otherwise.*

[1] We emphasize that in the former expression $(\cdot, \cdot)$ denotes application while in the latter $(\cdot, \cdot)$ denotes concatenation.
[2] Following Sriperumbudur et al. [2012], we do not require the usual $f(1) = 0$ condition.

Well-known instances of (2.6) are the Kullback-Leibler divergence ($f(t) = t \log t$), Hellinger distance ($f(t) = (\sqrt{t} - 1)^2$), $\chi^2$-divergence ($f(t) = (t-1)^2$), or total variation distance ($f(t) = |t - 1|$) [Sriperumbudur et al., 2012]. But, in practice, expression (2.6) can be challenging to estimate from samples, especially if one puts no additional assumptions on $\mathbb{P}$ and $\mathbb{Q}$ [Rubenstein et al., 2019].

Integral probability metrics take the following form.

**Definition 2.2.2** (Integral probability metric; IPM)**.** *Let $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_1^+(\mathcal{X})$ and $\mathcal{F}$ a class of real-valued measurable functions. The IPM of $\mathbb{P}$ and $\mathbb{Q}$ is*

$$\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{X}} f(x) \mathrm{d}\mathbb{P}(x) - \int_{\mathcal{X}} f(x) \mathrm{d}\mathbb{Q}(x) \right|. \tag{2.7}$$

By appropriately choosing $\mathcal{F}$, (2.7) specializes to the Kantorovich metric ($\mathcal{F} = \{f : \|f\|_L \leq 1\}$, with $\|\cdot\|_L$ the Lipschitz-seminorm of bounded continuous functions on the metric space $(\mathcal{X}, d)$) and known to be the dual [Dudley, 2002, Theorem 11.8.2] of the Wasserstein distance [Dudley, 2002, p. 420] if $(\mathcal{X}, d)$ is separable, the total variation distance ($\mathcal{F} = \{f : \|f\|_\infty := \sup_{x \in \mathcal{X}} |f(x)| \leq 1\}$), the Kolmogorov-Smirnov distance ($\mathcal{F} = \{\mathbb{1}_{\{(-\infty, t]\}} : t \in \mathbb{R}^d\}$), and the maximum mean discrepancy ($\mathcal{F} = \{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1\}$), among others [Sriperumbudur et al., 2012]. It is known that the total variation is the only non-trivial distance that is both an $f$-divergence and an IPM [Sriperumbudur et al., 2012, Appendix A].

Like $f$-divergences, most of the mentioned IPMs are difficult to estimate in practice[3] but they are important tools in probability theory. In contrast, estimators for the kernel-based maximum mean discrepancy can easily be constructed in closed-form on any kernel-enriched domain, which makes kernel-based approaches particularly appealing. We detail kernel-based methods in the following and refer to Sriperumbudur et al. [2012] for a detailed comparison of estimators of different IPMs w.r.t. their consistency and convergence rates.

## 2.3. Kernel-based information theoretical measures

In this section, we collect fundamental definitions and define our quantities of main interest. The maximum mean discrepancy is in Section 2.3.1, the Hilbert-Schmidt independence criterion is in Section 2.3.2, and the kernel Stein discrepancy is in Section 2.3.3.

The key quantity for defining information theoretical measures using kernels is the kernel mean embedding [Smola et al., 2007, Berlinet and Thomas-Agnan, 2004, Gretton et al., 2012]. We refer to Muandet et al. [2017] for a detailed overview of kernel mean embeddings and also note that the following quantities are instances of the more general kernelized cumulants [Bonnier et al., 2023].

**Definition 2.3.1** (Kernel mean embedding)**.** *Let $\mathcal{H}_k$ denote the RKHS associated to kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The mean embedding $\mu_k$ of a probability measure $\mathbb{P} \in \mathcal{M}_1^+(\mathcal{X})$ is*

$$\mu_k(\mathbb{P}) = \int_{\mathcal{X}} k(\cdot, x) \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k, \tag{2.8}$$

*where the integral is meant in Bochner's sense [Diestel and Uhl, 1977, Chapter II.2].*

[3] For example, estimation of the Kantorovich metric is accomplished by solving a linear program [Sriperumbudur et al., 2012, Theorem 2.1].

The mean element $\mu_k(\mathbb{P})$ exists iff. $\int_{\mathcal{X}} \|k(\cdot, x)\|_{\mathcal{H}_k} \mathrm{d}\mathbb{P}(x) < \infty$ [Diestel and Uhl, 1977, p. 45; Theorem 2]. This condition is satisfied for instance when $k(\cdot, x)$ is bounded [Sriperumbudur et al., 2010, Proposition 2]. If $k(\cdot, x)$ is unbounded, one restricts the class $\mathcal{M}_1^+(\mathcal{X})$ to ensure the existence of the mean embedding.

To gain intuition, note that the mean embedding can be thought of as a generalization of the moment-generating function (if it exists) or the characteristic function, as it can fully characterize a distribution. In fact, the mean embedding recovers the moment-generating function for a specific choice of $k$; see Muandet et al. [2017, Example 3.1–3.2]. Further, given $h \in \mathcal{H}_k$, the mean element $\mu_k(\mathbb{P})$ acts as

$$\langle h, \mu_k(\mathbb{P})\rangle_{\mathcal{H}_k} \overset{(a)}{=} \int_{\mathcal{X}} \langle h, k(\cdot, x)\rangle_{\mathcal{H}_k} \mathrm{d}\mathbb{P}(x) \overset{(b)}{=} \mathbb{E}_{X\sim\mathbb{P}} h(X),$$

where Steinwart and Christmann [2008, (A.32)] allows exchanging integration and the inner product in (a), and the reproducing property gives (b).

Similar to the mean element (2.8), the definition of the covariance operator [Baker, 1970, 1973] with kernels is as follows.

**Definition 2.3.2** (Covariance operator)**.** *In the setting of Definition 2.3.1, the covariance operator $C_{\mathbb{P},k} : \mathcal{H}_k \to \mathcal{H}_k$ of $\mathbb{P} \in \mathcal{M}_1^+(\mathcal{X})$ w.r.t. $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is*

$$C_{\mathbb{P},k} := \int_{\mathcal{X}} k(\cdot, x) \otimes k(\cdot, x) \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k \otimes \mathcal{H}_k. \tag{2.9}$$

The operator exists iff. $\int_{\mathcal{X}} \|k(\cdot, x)\|^2_{\mathcal{H}_k} \mathrm{d}\mathbb{P}(x) < \infty$; $C_{\mathbb{P},k}$ is a positive trace class operator.

Let $g, h \in \mathcal{H}_k$. Then

$$\begin{aligned}\left\langle g, C_{\mathbb{P},k} h\right\rangle_{\mathcal{H}_k} &\overset{(a)}{=} \int_{\mathcal{X}} \langle g, k(\cdot, x) \otimes k(\cdot, x) h\rangle_{\mathcal{H}_k} \mathrm{d}\mathbb{P}(x) \overset{(b)}{=} \int_{\mathcal{X}} \langle g, k(\cdot, x)\rangle_{\mathcal{H}_k} \langle h, k(\cdot, x)\rangle_{\mathcal{H}_k} \mathrm{d}\mathbb{P}(x) \\ &\overset{(c)}{=} \mathbb{E}_{X\sim\mathbb{P}} g(X) h(X),\end{aligned}$$

and hence $\left\langle g, C_{\mathbb{P},k} h\right\rangle_{\mathcal{H}_k}$ is equivalent to the covariance of centered $g$ and $h$. (a) is by Steinwart and Christmann [2008, (A.32)], (b) uses that $k(\cdot, x) \otimes k(\cdot, x) : \mathcal{H}_k \to \mathcal{H}_k$ acts as $h \mapsto k(\cdot, x) \langle k(\cdot, x), h\rangle_{\mathcal{H}_k}$, and (c) follows from the reproducing property.

We define $C_{\mathbb{P},k,\lambda} := C_{\mathbb{P},k} + I\lambda$ for $\lambda > 0$, where $I$ denotes the identity operator. Further, the effective dimension is[4]

$$\mathcal{N}_{\mathbb{P},k}(\lambda) := \mathrm{tr}\left(C_{\mathbb{P},k} C^{-1}_{\mathbb{P},k,\lambda}\right) \leq \frac{\mathrm{tr}\left(C_{\mathbb{P},k}\right)}{\lambda}. \tag{2.10}$$

$\mathcal{N}_{\mathbb{P},k}(\lambda)$ can be seen as a measure of the capacity of the hypothesis space [Zhang, 2002, Caponnetto and De Vito, 2007] and, in that sense, captures the difficulty of the learning problem. The effective dimension plays a key role in the analysis of the proposed Nyström $M$-HSIC in Chapter 3 and our Nyström KSD in Chapter 5.

The cross-covariance of $M \geq 2$ random variables also has an RKHS-valued analogue.

[4] The inequality is implied by $\mathrm{tr}\left(C_{\mathbb{P},k} C^{-1}_{\mathbb{P},k,\lambda}\right) = \sum_{i\in I} \frac{\lambda_i}{\lambda_i+\lambda} \leq \frac{1}{\lambda} \sum_{i\in I} \lambda_i = \frac{\mathrm{tr}(C_{\mathbb{P},k})}{\lambda}$, where $(\lambda_i)_{i\in I}$ are the eigenvalues of $C_{\mathbb{P},k}$.

**Definition 2.3.3** (Cross-covariance operator)**.** *Let $X = (X_m)_{m=1}^M$ denote a random variable with distribution $\mathbb{P} \in \mathcal{M}_1^+(\mathcal{X})$ on the product space $\mathcal{X} = \times_{m=1}^M \mathcal{X}_m$, where $\mathcal{X}_m$ is enriched with kernel $k_m : \mathcal{X}_m \times \mathcal{X}_m \to \mathbb{R}$. The distribution of the m-th marginal $X_m$ of $X$ is denoted by $\mathbb{P}_m \in \mathcal{M}_1^+(\mathcal{X}_m)$; the product of these $M$ marginals is $\otimes_{m=1}^M \mathbb{P}_m \in \mathcal{M}_1^+(\mathcal{X})$. The cross-covariance operator is*

$$\tilde{C}_{\mathbb{P},k} := \int_{\mathcal{X}} \otimes_{m=1}^M k_m(\cdot, x_m) \mathrm{d}\mathbb{P}(x_1, \ldots, x_M) \in \otimes_{m=1}^M \mathcal{H}_{k_m} = \mathcal{H}_k, \tag{2.11}$$

*where $k = \otimes_{m=1}^M k_m$.*

Similar to before, $\tilde{C}_{\mathbb{P},k}$ exists iff.

$$\int_{\mathcal{X}} \left\| \otimes_{m=1}^M k_m(\cdot, x_m) \right\|_{\mathcal{H}_k} \mathrm{d}\mathbb{P}(x_1, \ldots, x_M) \overset{(2.3)}{=} \int_{\mathcal{X}} \prod_{m \in [M]} \|k_m(\cdot, x_m)\|_{\mathcal{H}_{k_m}} \mathrm{d}\mathbb{P}(x_1, \ldots, x_M) < \infty.$$

Clearly, $\tilde{C}_{\mathbb{P},k\otimes k} = C_{\mathbb{P},k}$, implying that (2.9) is a special case of (2.11).

In the next sections, we detail information theoretical measures based on the introduced quantities.

### 2.3.1. Maximum mean discrepancy

The existence of (2.8) gives rise to a (semi-)metric on $\mathcal{M}_1^+(\mathcal{X})$, also obtained by considering the IPM $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$, with $\mathcal{F} = \{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1\}$ and using the reproducing property [Gretton et al., 2012, Lemma 4].

**Definition 2.3.4** (Maximum mean discrepancy; MMD)**.** *In the setting of Definition 2.3.1, with $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_1^+(\mathcal{X})$ and assuming that $\mathbb{E}_{X\sim\mathbb{P}} \|k(\cdot, X)\|_{\mathcal{H}_k} < \infty$ and $\mathbb{E}_{Y\sim\mathbb{Q}} \|k(\cdot, Y)\|_{\mathcal{H}_k} < \infty$,*

$$\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) = \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k}. \tag{2.12}$$

The injectivity of the mean embedding $\mu_k$ is equivalent to $\mathrm{MMD}_k$ being a metric [Fukumizu et al., 2007, Sriperumbudur et al., 2010]; in this case the kernel $k$ is called characteristic. On Euclidean domains, the characteristic property is satisfied for instance if $k$ is continuous bounded translation-invariant and (i) $k(\mathbf{x}, \mathbf{x}') = \kappa(\mathbf{x} - \mathbf{x}')$ coincides with the probability density function of a symmetric infinitely divisible distribution [Nishiyama and Fukumizu, 2016], or (ii) the Borel measure $\Lambda_k$ in (2.4) has support $\mathbb{R}^d$ [Sriperumbudur et al., 2010]. In general, to ensure the characteristic property, universality [Steinwart, 2001, Micchelli et al., 2006] is sufficient on compact metric domains. With $k$ characteristic and closed-form estimators, MMD allows powerful two-sample tests on any kernel-enriched domain; this property underpins our change detection algorithm detailed in Chapter 6. In particular, we will base our proposed online approximation scheme of MMD on the following known estimator.

Given samples $x_1, \ldots, x_m \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ and $y_1, \ldots, y_n \overset{\text{i.i.d.}}{\sim} \mathbb{Q}$, define the empirical measures $\hat{\mathbb{P}}_n = \frac{1}{m} \sum_{i=1}^m \delta_{x_i}$ and $\hat{\mathbb{Q}}_n = \frac{1}{n} \sum_{i=1}^n \delta_{y_i}$. A closed-form biased estimator of (2.12) can then be obtained by considering the plug-in estimator [Gretton et al., 2012, (5)]

$$\mathrm{MMD}_k^2\left(\hat{\mathbb{P}}_m, \hat{\mathbb{Q}}_n\right) = \frac{1}{m^2} \sum_{i,j=1}^m k(x_i, x_j) + \frac{1}{n^2} \sum_{i,j=1}^n k(y_i, y_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(x_i, y_j), \tag{2.13}$$

where the equality holds by the linearity of the inner product and the reproducing property. The runtime complexity of (2.13) is in $O\left(m^2 + n^2\right)$.

Expression 2.12 naturally leads to a two-sample test. The details are as follows.

To decide whether the value of $\mathrm{MMD}\left(\hat{\mathbb{P}}_m, \hat{\mathbb{Q}}_n\right)$ indicates a significant difference between $\mathbb{P}$ and $\mathbb{Q}$, one tests the null hypothesis $H_0 : \mathbb{P} = \mathbb{Q}$ versus its alternative $H_1 : \mathbb{P} \neq \mathbb{Q}$ by defining an acceptance region for a given level $\alpha \in (0, 1)$, which takes the form $\mathrm{MMD}\left(\hat{\mathbb{P}}_m, \hat{\mathbb{Q}}_n\right) < \epsilon_\alpha$. One rejects $H_0$ if the test statistic exceeds the threshold $\epsilon_\alpha$. The level $\alpha$ is a bound for the probability that the tests rejects $H_0$ incorrectly [Casella and Berger, 1990]. Assuming that $k$ is nonnegative and bounded by $K > 0$, that is, $0 \leq k(x, y) \leq K$ for all $x, y \in \mathcal{X}$, Gretton et al. [2012, Corollary 9] provides the distribution-free threshold $\epsilon_\alpha$ for the case that both samples $\hat{\mathbb{P}}_m$ and $\hat{\mathbb{Q}}_m$ have the same size ($m = n$) as

$$\mathrm{MMD}_k\left(\hat{\mathbb{P}}_m, \hat{\mathbb{Q}}_m\right) < \sqrt{\frac{2K}{m}}\left(1 + \sqrt{2\log\frac{1}{\alpha}}\right) =: \epsilon_\alpha. \tag{2.14}$$

Computing (2.14) costs $O(1)$. As the change detection setting in Chapter 6 requires the case that $m \neq n$, we will extend (2.14) in Section 6.3.

### 2.3.2. Hilbert-Schmidt independence criterion

By considering the distance of a joint distribution to the product of its marginals, MMD also allows quantifying dependence of $M = 2$ components [Gretton et al., 2005] and of $M \geq 2$ components [Quadrianto et al., 2009, Sejdinovic et al., 2013a, Pfister et al., 2018]; it is then called Hilbert-Schmidt independence criterion.

**Definition 2.3.5** (Hilbert-Schmidt independence criterion; HSIC)**.** *In the setting of Definition 2.3.3,*

$$\mathrm{HSIC}_k(\mathbb{P}) := \mathrm{MMD}_k\left(\mathbb{P}, \otimes_{m=1}^M \mathbb{P}_m\right), \tag{2.15}$$

*where $k = \otimes_{m=1}^M k_m$ and given that the respective mean embeddings exist.*

HSIC is seen to be equivalent to the RKHS-norm of the centered cross-covariance operator (2.11)

$$\mathrm{HSIC}_k(\mathbb{P}) = \left\|\tilde{C}_{\mathbb{P},k} - \mu_k\left(\otimes_{m=1}^M \mathbb{P}_m\right)\right\|_{\mathcal{H}_k} =: \left\|\tilde{C}^c_{\mathbb{P},k}\right\|_{\mathcal{H}_k}, \tag{2.16}$$

where, for $M = 2$, the centered cross-covariance operator is equivalent to subtracting the mean element from the respective feature maps

$$\tilde{C}_{\mathbb{P},k_1\otimes k_2} - \mu_{k_1}(\mathbb{P}_1) \otimes \mu_{k_2}(\mathbb{P}_2) = \int_{\mathcal{X}} \left(k_1(\cdot, x_1) - \mu_{k_1}(\mathbb{P}_1)\right) \otimes \left(k_2(\cdot, x_2) - \mu_{k_2}(\mathbb{P}_2)\right) \mathrm{d}\mathbb{P}(x_1, x_2).$$

The equivalence does not hold in the general case ($M > 2$) [Sejdinovic et al., 2013a].

Further, one of the most widely-used independence measures in statistics, distance covariance [Székely et al., 2007, Székely and Rizzo, 2009, Lyons, 2013], was shown to be equivalent to HSIC [Sejdinovic et al., 2013b] when the latter is specialized to $M = 2$ components; Sheng and Sriperumbudur [2023] proved a similar result for the conditional case. HSIC is known to capture the independence of $M = 2$ random variables with characteristic $(k_m)_{m=1}^2$ kernels (on the respective domains) as proved by Lyons [2013]; for more than two components ($M > 2$) universality [Steinwart, 2001, Micchelli et al., 2006] of $(k_m)_{m=1}^M$-s is sufficient [Szabó and Sriperumbudur, 2018].

The classical HSIC estimator takes the following form. Given an i.i.d. sample of $M$-tuples of size $n$

$$\hat{\mathbb{P}}_n := \left\{ \left(x_1^1, \ldots, x_M^1\right), \ldots, \left(x_1^n, \ldots, x_M^n\right) \right\} \subset \mathcal{X} \tag{2.17}$$

drawn from $\mathbb{P}$, the corresponding empirical estimate of the squared HSIC, obtained by replacing the population means with the sample means, gives rise to the V-statistic based estimator [Quadrianto et al., 2009, Pfister et al., 2018], which we call V-HSIC,

$$0 \leq \mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) = \frac{1}{n^2}\, \mathbf{1}_n^\mathsf{T} \left(\circ_{m\in[M]}\, \mathbf{K}_{k_m}\right) \mathbf{1}_n + \frac{1}{n^{2M}} \prod_{m\in[M]} \mathbf{1}_n^\mathsf{T} \mathbf{K}_{k_m} \mathbf{1}_n - \frac{2}{n^{M+1}} \mathbf{1}_n^\mathsf{T} \left(\circ_{m\in[M]}\, \mathbf{K}_{k_m} \mathbf{1}_n\right) \tag{2.18}$$

with Gram matrices

$$\mathbf{K}_{k_m} = \left[ k_m\left(x_m^i, x_m^j\right) \right]_{i,j\in[n]} \in \mathbb{R}^{n\times n}, \tag{2.19}$$

which can be computed in $O(n^2 M)$. This prohibitive runtime inspired the development of HSIC approximations [Zhang et al., 2018] using the Nyström method and random Fourier features. We review the Nyström-based construction in Section 3.2 and explain why the technique is restricted to $M = 2$ components, before presenting our alternative approximation scheme of HSIC, capable of handling $M \geq 2$ components, in Section 3.3.

### 2.3.3. Kernel Stein discrepancy

We now introduce our third quantity of interest, the kernel Stein discrepancy (KSD; Liu et al. 2016, Chwialkowski et al. 2016), which allows quantifying goodness of fit. We restrict our attention to the Langevin-Stein operator-based KSD (KSD in the following) for quantifying goodness of fit on $\mathcal{X} = \mathbb{R}^d$, employed in Chapter 5, and refer to Hagrass et al. [2025] for the general construction and examples on non-Euclidean domains.

We make the following assumption.

**Assumption 2.3.1.** *Let $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_1^+\left(\mathbb{R}^d\right)$ be fixed. Assume that*

*(i)* *$\mathbb{P}$ is absolutely continuous w.r.t. the Lebesgue measure with corresponding density $p$,*

*(ii)* *$p$ is continuously differentiable with support $\mathbb{R}^d$,*

*(iii)* *$p(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathbb{R}^d$,*

*(iv)* *$\lim_{\|\mathbf{x}\|\to\infty} f(\mathbf{x})p(\mathbf{x}) = 0$ for all $f \in \mathcal{H}_k$,[5] and*

*(v)* *$k$ is continuously differentiable in both arguments.*

We refer to $\mathbb{P}$ as the target distribution and to $\mathbb{Q}$ as the sampling distribution. Property (iv) holds for instance if $p$ is bounded and $\lim_{\|\mathbf{x}\|\to\infty} f(\mathbf{x}) = 0$ for all $f \in \mathcal{H}_k$; the latter is guaranteed if $k(\cdot, \mathbf{x}) \in C_0\left(\mathbb{R}^d\right)$ for all $\mathbf{x} \in \mathbb{R}^d$, where $C_0\left(\mathbb{R}^d\right)$ denotes the space of continous functions on $\mathbb{R}^d$ vanishing at infinity. Condition (v) will imply the measurability of $h_p$ and the separability of $\mathcal{H}_{h_p}$, both quantities defined below.

[5] On finite-dimensional vector spaces all norms are equivalent; hence the choice of $\|\cdot\|$ does not matter.

Let $\mathcal{H}_k^d := \times_{i=1}^d \mathcal{H}_k$ be the product RKHS with inner product defined by $\langle \mathbf{f}, \mathbf{g} \rangle_{\mathcal{H}_k^d} := \sum_{i=1}^d \langle f_i, g_i \rangle_{\mathcal{H}_k}$ for $\mathbf{f} = (f_i)_{i=1}^d, \mathbf{g} = (g_i)_{i=1}^d \in \mathcal{H}_k^d$. The (Langevin-)Stein operator [Gorham and Mackey, 2015, (4)] is defined as $\left(\mathcal{T}_p \mathbf{f}\right)(\mathbf{x}) = \langle \nabla_{\mathbf{x}} [\log p(\mathbf{x})], \mathbf{f}(\mathbf{x}) \rangle + \sum_{i=1}^d \partial_i f_i(\mathbf{x}) \left(\mathbf{f} \in \mathcal{H}_k^d,\ \mathbf{x} \in \mathbb{R}^d\right)$. With this definition at hand and using (2.5),

$$\begin{aligned} \left(\mathcal{T}_p \mathbf{f}\right)(\mathbf{x}) &= \left\langle \mathbf{f}, \xi_p(\mathbf{x}) \right\rangle_{\mathcal{H}_k^d}, \\ \xi_p(\mathbf{x}) &= \left[\nabla_{\mathbf{x}} \left(\log p(\mathbf{x})\right) k\left(\cdot, \mathbf{x}\right) + \nabla_{\mathbf{x}} k\left(\cdot, \mathbf{x}\right)\right] \in \mathcal{H}_k^d \end{aligned} \tag{2.20}$$

for all $\mathbf{f} \in \mathcal{H}_k^d$ and $\mathbf{x} \in \mathbb{R}^d$, with kernel (for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$)

$$h_p(\mathbf{x}, \mathbf{y}) = \left\langle \xi_p(\mathbf{x}), \xi_p(\mathbf{y}) \right\rangle_{\mathcal{H}_k^d} = \langle h_p(\cdot, \mathbf{x}), h_p(\cdot, \mathbf{y}) \rangle_{\mathcal{H}_{h_p}}. \tag{2.21}$$

Notice that $\xi_p(\mathbf{x})$ and $h_p(\cdot, \mathbf{x})$ map to different feature spaces ($\mathcal{H}_k^d$ and $\mathcal{H}_{h_p}$, respectively) but yield the same kernel $h_p$, which, with (2.20), takes the explicit form

$$\begin{aligned} h_p(\mathbf{x}, \mathbf{y}) = &\left\langle \nabla_{\mathbf{x}} \log p(\mathbf{x}), \nabla_{\mathbf{y}} \log p(\mathbf{y}) \right\rangle_{\mathbb{R}^d} k(\mathbf{x}, \mathbf{y}) + \left\langle \nabla_{\mathbf{y}} \log p(\mathbf{y}), \nabla_{\mathbf{x}} k(\mathbf{x}, \mathbf{y}) \right\rangle_{\mathbb{R}^d} \\ &+ \left\langle \nabla_{\mathbf{x}} \log p(\mathbf{x}), \nabla_{\mathbf{y}} k(\mathbf{x}, \mathbf{y}) \right\rangle_{\mathbb{R}^d} + \sum_{i=1}^d \frac{\partial^2 k(\mathbf{x}, \mathbf{y})}{\partial x_i \partial y_i}. \end{aligned}$$

KSD then is defined as the IPM $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$, with $\mathcal{F} = \left\{\mathbf{f} \in \mathcal{H}_k^d : \|\mathbf{f}\|_{\mathcal{H}_k^d} \leq 1\right\}$, and given by

$$\begin{aligned} S_p(\mathbb{Q}) &= \sup_{\mathbf{f} \in \mathcal{F}} \underbrace{\mathbb{E}_{X \sim \mathbb{P}} \left[\mathcal{T}_p \mathbf{f}(X)\right]}_{\overset{(a)}{=} 0} - \mathbb{E}_{X \sim \mathbb{Q}} \left[\mathcal{T}_p \mathbf{f}(X)\right] = \sup_{\mathbf{f} \in \mathcal{F}} \left\langle \mathbf{f}, \mathbb{E}_{X \sim \mathbb{Q}} \xi_p(X) \right\rangle_{\mathcal{H}_k^d} = \left\| \mathbb{E}_{X \sim \mathbb{Q}} \xi_p(X) \right\|_{\mathcal{H}_k^d} \\ &\overset{(b)}{=} \left\| \mathbb{E}_{X \sim \mathbb{Q}} h_p\left(\cdot, X\right) \right\|_{\mathcal{H}_{h_p}}, \end{aligned} \tag{2.22}$$

where (a) holds by the construction of KSD and (b) follows from (2.21).

In particular, (2.22) motivates the following definition.

**Definition 2.3.6** (Kernel Stein discrepancy; KSD)**.** *In the setting of Assumption 2.3.1, define the kernel Stein discrepancy as*

$$S_p(\mathbb{Q}) = \left\| \mathbb{E}_{X \sim \mathbb{Q}} h_p\left(\cdot, X\right) \right\|_{\mathcal{H}_{h_p}}, \tag{2.23}$$

*with $h_p$ as in* (2.21) *and given that $\mathbb{E}_{X \sim \mathbb{Q}} \left\| h_p(\cdot, X) \right\|_{\mathcal{H}_{h_p}} < \infty$.*

We note that (2.23) corresponds to the RKHS norm of the mean embedding of $\mathbb{Q}$ with kernel $h_p$.

Given a sample $\hat{\mathbb{Q}}_n = \{\mathbf{x}_i\}_{i=1}^n \sim \mathbb{Q}^n$, the popular V-statistic-based estimator [Chwialkowski et al., 2016, Section 2.2] is obtained by replacing $\mathbb{Q}$ with the empirical measure $\hat{\mathbb{Q}}_n$, similar to the biased MMD (2.13) and HSIC (2.18) estimators; it takes the form

$$S_p^2\left(\hat{\mathbb{Q}}_n\right) = \frac{1}{n^2} \sum_{i,j=1}^n h_p(\mathbf{x}_i, \mathbf{x}_j), \tag{2.24}$$

and can be computed in $O(n^2)$ time. The corresponding U-statistic-based estimator [Liu et al., 2016, (14)] has a similar expression but omits the diagonal terms, that is,

$$S^2_{p,u}\left(\hat{\mathbb{Q}}_n\right) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n}^{n} h_p(\mathbf{x}_i, \mathbf{x}_j); \tag{2.25}$$

it also has a runtime cost of $O(n^2)$. For large-scale applications, the quadratic runtime is a significant bottleneck; this is the shortcoming we tackle in Chapter 5.

Now, we are ready to detail our contributions in the following chapters.

# 3. Nyström $M$-Hilbert-Schmidt Independence Criterion

The content of this chapter is based on the following publication.

- F. Kalinke and Z. Szabó. Nyström M-Hilbert-Schmidt independence criterion. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 1005–1015, 2023. PMLR.

The code replicating all experiments is available at github.com/flopska/nystroem-mhsic.

## 3.1. Introduction

Recall from Chapter 1 that HSIC has been deployed successfully across a wide range of domains and while its V-statistic-based estimator is powerful, its runtime increases quadratically with the number of samples, which limits its applicability in large-scale settings. To tackle this severe computational bottleneck, approximations of HSIC (N-HSIC, RFF-HSIC) have been proposed [Zhang et al., 2018], relying on the Nyström [Williams and Seeger, 2001] and the random Fourier feature (RFF; Rahimi and Recht 2007) method, respectively. However, these estimators (i) are limited to two components, (ii) their extension to more than two components is not straightforward, and (iii) they lack theoretical guarantees. The RFF-based approach is further restricted to finite-dimensional Euclidean domains and to translation-invariant kernels. The normalized finite set independence criterion (NFSIC; Jitkrittum et al. 2017a) replaces the RKHS norm of HSIC with an $L^2$ one, which allows the construction of linear-time estimators. However, NFSIC requires $\mathbb{R}^d$-valued input and analytic kernels [Chwialkowski et al., 2015]. Novel complementary approaches are the kernel partial correlation coefficient (KPCC; Huang et al. 2022), and tests based on incomplete U-statistics [Schrab et al., 2022c]. One drawback of KPCC is its cubic runtime complexity w.r.t. the sample size when applied to kernel-enriched domains. Schrab et al. [2022c]'s approach can run in linear time, but it is limited to $M = 2$ components. We note that all approaches require choosing an appropriate kernel: Here, one can optimize over various parametric families of kernels for increasing a proxy of test power in case of MMD [Jitkrittum et al., 2016, Liu et al., 2020], and in case of HSIC [Jitkrittum et al., 2017a]. One can also design (almost) minimax-optimal MMD-based two-sample tests using spectral regularization [Hagrass et al., 2024a]. We summarize the estimators most closely related to our work in Table 3.1.

The restriction of existing HSIC approximations to two components is a severe limitation in recent applications like causal discovery which require independence tests capable of handling more than two components. Furthermore, the emergence of large-scale data sets necessitates algorithms that scale well in the sample size. To alleviate these bottlenecks, we make the following **contributions**.

- We propose Nyström $M$-HSIC, an efficient HSIC estimator, which can handle more than two components and has runtime $O\left(Mn'^3 + Mn'n\right)$, where $n$ denotes the number of samples, $n' \ll n$ stands for the number of Nyström points, and $M$ is the number of random variables whose independence is measured.

**Table 3.1.:** Comparison of kernel independence measures: $n$ – number of samples, $M$ – number of components, $n'$ – number of Nyström samples, $s$ – number of random Fourier features, $d$ – data dimensionality.

| Independence Measure | Runtime Complexity | $M$ | Domain | Admissible Kernels |
|---|---|---|---|---|
| V-HSIC [Pfister et al., 2018] | $O(Mn^2)$ | $M \geq 2$ | any | universal |
| NFSIC [Jitkrittum et al., 2017a] | $O(n)$ | $M = 2$ | $\mathbb{R}^d$ | analytic, characteristic |
| N-HSIC [Zhang et al., 2018] | $O(n'^3 + nn'^2)$ | $M = 2$ | any | characteristic |
| RFF-HSIC [Zhang et al., 2018] | $O(s^2 n)$ | $M = 2$ | $\mathbb{R}^d$ | translation-invariant, characteristic |
| KPCC [Huang et al., 2022] | $O(n^3)$ | $M = 2$ | any | characteristic |
| **Nyström $M$-HSIC** (N-MHSIC) | $O(Mn'^3 + Mn'n)$ | $M \geq 2$ | any | universal |

- We provide theoretical guarantees for Nyström $M$-HSIC: we prove that our estimator converges with rate $O(n^{-1/2})$ for $n' = \tilde{\Theta}(\sqrt{n})$, which matches the convergence of the quadratic-time estimator.
- We perform an extensive suite of experiments to demonstrate the efficiency of Nyström $M$-HSIC. These applications include dependency testing of media annotations and causal discovery. In the former, we achieve similar runtime and power as existing HSIC approximations. The latter requires testing joint independence of more than two components, which is beyond the capabilities of existing HSIC accelerations. Here, the proposed algorithm achieves the same performance as the quadratic-time HSIC estimator V-HSIC with a significantly reduced runtime.

The remainder of this chapter is structured as follows. We recall the existing Nyström-based HSIC approximation for $M = 2$ components in Section 3.2 before we detail our proposed estimator for $M \geq 2$ components in Section 3.3. Experiments are in Section 3.4. We collect all our proofs in Section 3.5.

## 3.2. Existing Nyström approximation

In this section, we recall the existing Nyström approximation, which can handle $M = 2$ components.

Recall from Section 2.3.2 that the classical quadratic-time (V-HSIC) estimator (2.18) takes the form

$$\mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) = \frac{1}{n^2}\,\mathbf{1}_n^\mathsf{T}\left(\circ_{m\in[M]}\,\mathbf{K}_{k_m}\right)\mathbf{1}_n + \frac{1}{n^{2M}}\prod_{m\in[M]}\mathbf{1}_n^\mathsf{T}\mathbf{K}_{k_m}\mathbf{1}_n - \frac{2}{n^{M+1}}\mathbf{1}_n^\mathsf{T}\left(\circ_{m\in[M]}\,\mathbf{K}_{k_m}\mathbf{1}_n\right)$$

which, for $M = 2$ components [Gretton et al., 2005] can be written as

$$\mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) = \frac{1}{n^2}\,\mathrm{tr}\left(\mathbf{H}\mathbf{K}_{k_1}\mathbf{H}\mathbf{K}_{k_2}\right), \tag{3.1}$$

with the centering matrix $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T} \in \mathbb{R}^{n\times n}$, Gram matrices $\mathbf{K}_{k_1}$, $\mathbf{K}_{k_2}$ defined in (2.19), and sample $\hat{\mathbb{P}}_n := \left\{(x_1^1, x_2^1), \ldots, (x_1^n, x_2^n)\right\}$ as in (2.17) with $M = 2$. The naive computation of (3.1) costs $O(n^3)$. However, noticing that $\mathrm{tr}(\mathbf{A}^\mathsf{T}\mathbf{B}) = \sum_{i,j\in[n]} A_{i,j}B_{i,j}$, the computational complexity reduces to $O(n^2)$. The quadratic complexity can be reduced by the Nyström approximation[1] [Zhang et al., 2018]

$$\mathrm{HSIC}_{k,\mathrm{N}_0}^2\left(\hat{\mathbb{P}}_n\right) = \frac{1}{n^2}\,\mathrm{tr}\left(\mathbf{H}\mathbf{K}_{k_1}^{\mathrm{Nys}}\mathbf{H}\mathbf{K}_{k_2}^{\mathrm{Nys}}\right) \overset{(*)}{=} \frac{1}{n^2}\left\|\left(\mathbf{H}\phi_{k_1}^{\mathrm{Nys}}\right)^\mathsf{T}\mathbf{H}\phi_{k_2}^{\mathrm{Nys}}\right\|_\mathrm{F}^2, \tag{3.2}$$

[1] $\mathrm{HSIC}_k^2(\hat{\mathbb{P}}_n)$ denotes the application of $\mathrm{HSIC}_k^2$ to the empirical measure $\hat{\mathbb{P}}_n$. $\mathrm{HSIC}_{k,\mathrm{N}_0}^2(\hat{\mathbb{P}}_n)$ and $\mathrm{HSIC}_{k,\mathrm{N}}^2(\hat{\mathbb{P}}_n)$ indicate dependence on $\hat{\mathbb{P}}_n$. Similarly, $\mu_\ell(\hat{\mathbb{Q}}_n)$ stands for application, $\mu_\ell(\tilde{\mathbb{Q}}_{n'})$, $\mu_{k_m}(\tilde{\mathbb{P}}_{m,n'})$ and $\mu_k(\tilde{\mathbb{P}}_{n'})$ indicate dependence on the argument.

which we detail in the following. The Nyström approximation relies on a subsample of size $n' \leq n$ of $\hat{\mathbb{P}}_n$, which we denote by $\tilde{\mathbb{P}}_{n'} := \left\{\left(\tilde{x}_1^1, \tilde{x}_2^1\right), \ldots, \left(\tilde{x}_1^{n'}, \tilde{x}_2^{n'}\right)\right\}$; the tilde indicates a relabeling. The subsample allows to define three matrices

$$\begin{aligned} \mathbf{K}_{k_m,n',n'} &= \left[k_m\left(\tilde{x}_m^i, \tilde{x}_m^j\right)\right]_{i,j\in[n']} \in \mathbb{R}^{n'\times n'}, \\ \mathbf{K}_{k_m,n,n} &= \mathbf{K}_{k_m} \in \mathbb{R}^{n\times n}, \\ \mathbf{K}_{k_m,n',n} &= \left[k_m(\tilde{x}_m^i, x_m^j)\right]_{i\in[n'],j\in[n]} \in \mathbb{R}^{n'\times n}, \end{aligned} \tag{3.3}$$

where $m \in [2]$ and $\mathbf{K}_{k_m}$ is defined in (2.19); we let $\mathbf{K}_{k_m,n,n'} = \mathbf{K}_{k_m,n',n}^\mathsf{T} \in \mathbb{R}^{n\times n'}$. The matrices $\mathbf{K}_{k_m}^{\mathrm{Nys}}$ ($m \in [2]$) as used in (3.2) are

$$\mathbf{K}_{k_m}^{\mathrm{Nys}} := \mathbf{K}_{k_m,n,n'}\mathbf{K}_{k_m,n',n'}^{-1}\mathbf{K}_{k_m,n',n} = \underbrace{\mathbf{K}_{k_m,n,n'}\mathbf{K}_{k_m,n',n'}^{-\frac{1}{2}}}_{=:\phi_{k_m}^{\mathrm{Nys}}\in\mathbb{R}^{n\times n'}} \left(\underbrace{\mathbf{K}_{k_m,n,n'}\mathbf{K}_{k_m,n',n'}^{-\frac{1}{2}}}_{=\phi_{k_m}^{\mathrm{Nys}}}\right)^\mathsf{T} \in \mathbb{R}^{n\times n},$$

provided that the inverse $\mathbf{K}_{k_m,n',n'}^{-1}$ exists. In (3.2) the r.h.s. of $(*)$ has a computational complexity of $O\left(n'^3 + nn'^2\right)$,[2,3] which is smaller than $O\left(n^2\right)$ of (3.1), provided that $n' = o\left(\sqrt{n}\right)$; this speeds up the computation. $(*)$ relies on the cyclic invariance property of the trace, and the idempotence of $\mathbf{H}$ (in other words, $\mathbf{HH} = \mathbf{H}$), limiting the above derivation to $M = 2$ components; the approach does not extend naturally to the case of $M > 2$.

## 3.3. Proposed Nyström $M$-HSIC estimator

We now elaborate the proposed Nyström HSIC approximation for $M \geq 2$ components.

Recall from (2.16) that the centered cross-covariance operator takes the form

$$\tilde{C}_{\mathbb{P},k}^c = \mu_k(\mathbb{P}) - \mu_k\left(\otimes_{m=1}^M \mathbb{P}_m\right) = \mu_k(\mathbb{P}) - \otimes_{m=1}^M \mu_{k_m}(\mathbb{P}_m)\,.$$

There are $M+1$ expectations in this expression; we estimate these mean embeddings separately. This conceptually simple construction, is to the best of our knowledge, the first that handles $M \geq 2$ components, and it allows to leverage recent bounds on mean estimators (Lemma 3.3.1). We first detail the general Nyström method for approximating expectations $\int_{\mathcal{Y}} \phi_\ell(y)\mathrm{d}\mathbb{Q}(y)$ associated to a kernel $\ell : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ and probability distribution $\mathbb{Q} \in \mathcal{M}_1^+(\mathcal{Y})$. One can then choose

$$(\mathcal{Y}, \ell, \mathbb{Q}) = (\mathcal{X}, k, \mathbb{P}), \qquad \text{and} \qquad (\mathcal{Y}, \ell, \mathbb{Q}) = (\mathcal{X}_m, k_m, \mathbb{P}_m), \quad m \in [M], \tag{3.4}$$

to achieve our goal.

Let $\tilde{\mathbb{Q}}_{n'} = \left\{\tilde{y}^1, \ldots, \tilde{y}^{n'}\right\}$ be a subsample (with replacement) of $\hat{\mathbb{Q}}_n = \left\{y^1, \ldots, y^n\right\} \overset{\text{i.i.d.}}{\sim} \mathbb{Q}$ referred to as Nyström points; the tilde again indicates relabeling. The usual estimator of the mean embedding replaces the population mean with its empirical counterpart over $n$ samples[1]

$$\mu_\ell(\mathbb{Q}) = \int_{\mathcal{Y}} \phi_\ell(y)\mathrm{d}\mathbb{Q}(y) \approx \frac{1}{n}\sum_{i\in[n]} \phi_\ell(y^i) = \mu_\ell\left(\hat{\mathbb{Q}}_n\right).$$

[2] This follows from the complexity of $O\left(n'^3\right)$ of inverting an $n'\times n'$ matrix and the complexity of multiplying both feature representations [Zhang et al., 2018].

[3] While asymptotically faster algorithms for matrix inversion exist, we consider the cost that one typically encounters in practice.

Instead, the Nyström approximation uses a weighted sum with weights $\alpha_i \in \mathbb{R}$ ($i \in [n']$): given $n'$ Nyström points, the estimator takes the form[1]

$$\mu_\ell(\mathbb{Q}) \approx \sum_{i\in[n']} \alpha_i \phi_\ell(\tilde{y}^i) = \mu_\ell\left(\tilde{\mathbb{Q}}_{n'}\right) \in \mathcal{H}_\ell^{\mathrm{Nys}},$$

where $\mathcal{H}_\ell^{\mathrm{Nys}} := \operatorname{span}\left(\phi_\ell(\tilde{y}^i) \,:\, i \in [n']\right) \subset \mathcal{H}_\ell$. The coefficients $\boldsymbol{\alpha}_\ell = (\alpha_\ell^1, \ldots, \alpha_\ell^{n'})^\mathsf{T} \in \mathbb{R}^{n'}$ are obtained by the minimum norm solution of

$$\min_{\boldsymbol{\alpha}_\ell \in \mathbb{R}^{n'}} \left\| \mu_\ell\left(\hat{\mathbb{Q}}_n\right) - \sum_{i\in[n']} \alpha_i \phi_\ell(\tilde{y}^i) \right\|_{\mathcal{H}_\ell}^2 . \tag{3.5}$$

The following lemma describes the solution of (3.5).

**Lemma 3.3.1** (Nyström mean embedding, Chatalic et al. [2022])**.** *For a kernel $\ell$ with corresponding feature map $\phi_\ell$, an i.i.d. sample $\hat{\mathbb{Q}}_n$ of distribution $\mathbb{Q}$, and a subsample $\tilde{\mathbb{Q}}_{n'}$ of $\hat{\mathbb{Q}}_n$, the Nyström estimate of $\mu_\ell(\mathbb{Q})$ is given by*

$$\mu_\ell\left(\tilde{\mathbb{Q}}_{n'}\right) = \sum_{i\in[n']} \alpha_\ell^i \phi_\ell(\tilde{y}^i),$$
$$\boldsymbol{\alpha}_\ell = \frac{1}{n} \left(\mathbf{K}_{\ell,n',n'}\right)^- \mathbf{K}_{\ell,n',n} \mathbf{1}_n,$$

*with Gram matrix $\mathbf{K}_{\ell,n',n'} = \left[\ell(\tilde{x}^i, \tilde{x}^j)\right]_{i,j\in[n']} \in \mathbb{R}^{n'\times n'}$, and $\mathbf{K}_{\ell,n',n} = \left[\ell(\tilde{x}^i, x^j)\right]_{i\in[n'],j\in[n]} \in \mathbb{R}^{n'\times n}$.*

Let

$$\tilde{\mathbb{P}}_{n'} = \left\{\left(\tilde{x}_1^1, \ldots, \tilde{x}_M^1\right), \ldots, \left(\tilde{x}_1^{n'}, \ldots, \tilde{x}_M^{n'}\right)\right\} \tag{3.6}$$

be a subsample (with replacement) of $\hat{\mathbb{P}}_n = \left\{\left(x_1^1, \ldots, x_M^1\right), \ldots, \left(x_1^n, \ldots, x_M^n\right)\right\}$ defined in (2.17), and

$$\tilde{\mathbb{P}}_{m,n'} = \left\{\tilde{x}_m^1, \ldots, \tilde{x}_m^{n'}\right\} \tag{3.7}$$

be the corresponding subsample of the $m$-th marginal ($m \in [M]$). Using our choice (3.4) with Lemma 3.3.1, the estimators for the embeddings of marginal distributions take the form[1]

$$\mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right) = \sum_{i\in[n']} \alpha_{k_m}^i \phi_{k_m}(\tilde{x}_m^i),$$
$$\boldsymbol{\alpha}_{k_m} = \frac{1}{n} \left(\mathbf{K}_{k_m,n',n'}\right)^- \mathbf{K}_{k_m,n',n} \mathbf{1}_n, \tag{3.8}$$

and the estimator of the mean embedding of the joint distribution is[1]

$$\mu_k\left(\tilde{\mathbb{P}}_{n'}\right) = \sum_{i\in[n']} \alpha_k^i \otimes_{m=1}^M \phi_{k_m}(\tilde{x}_m^i),$$

$$\boldsymbol{\alpha}_k = \frac{1}{n} \left(\mathbf{K}_{k,n',n'}\right)^- \left(\mathbf{K}_{k,n',n}\right) \mathbf{1}_n \overset{(*)}{=} \frac{1}{n} \overbrace{\Big(\underbrace{\circ_{m\in[M]} \mathbf{K}_{k_m,n',n'}}_{(a)}\Big)^-}^{(c)} \times \Big(\underbrace{\circ_{m\in[M]} \mathbf{K}_{k_m,n',n}}_{(b)}\Big)\mathbf{1}_n, \tag{3.9}$$

where $(*)$ holds as for the Gram matrix $\mathbf{K}_{k,n',n'}$ associated with the product kernel $k = \otimes_{m=1}^{M} k_m$ one has

$$\mathbf{K}_{k,n',n'} = \left[k\left((\tilde{x}_1^i, \ldots, \tilde{x}_M^i), (\tilde{x}_1^j, \ldots, \tilde{x}_M^j)\right)\right]_{i,j\in[n']} = \left[\prod_{m\in[M]} k_m(\tilde{x}_m^i, \tilde{x}_m^j)\right]_{i,j\in[n']} = \circ_{m\in[M]} \mathbf{K}_{k_m,n',n'},$$

and similarly $\mathbf{K}_{k,n',n} = \circ_{m\in[M]} \mathbf{K}_{k_m,n',n}$, with $\mathbf{K}_{k_m,n',n'}$ and $\mathbf{K}_{k_m,n',n}$ defined in (3.3).

Combining the $M+1$ Nyström estimators in (3.8) and in (3.9) gives rise to the overall Nyström HSIC estimator, which is elaborated in the following lemma.

**Lemma 3.3.2** (Computation of Nyström $M$-HSIC)**.** *The Nyström estimator for HSIC can be expressed as*[1]

$$\mathrm{HSIC}^2_{k,N}\left(\hat{\mathbb{P}}_n\right) = \boldsymbol{\alpha}_k^\mathsf{T}\left(\circ_{m\in[M]} \mathbf{K}_{k_m,n',n'}\right)\boldsymbol{\alpha}_k + \prod_{m\in[M]} \boldsymbol{\alpha}_{k_m}^\mathsf{T}\mathbf{K}_{k_m,n',n'}\boldsymbol{\alpha}_{k_m} - 2\boldsymbol{\alpha}_k^\mathsf{T}\left(\circ_{m\in[M]} \mathbf{K}_{k_m,n',n'}\boldsymbol{\alpha}_{k_m}\right), \quad (3.10)$$

*with $\boldsymbol{\alpha}_{k_m}$ and $\boldsymbol{\alpha}_k$ defined in* (3.8) *and* (3.9)*, respectively, $\mathbf{K}_{k_m,n',n'}$ is defined in* (3.3)*, and $N$ in the subscript of the estimator refers to Nyström. Note that* (3.10) *depends on $\hat{\mathbb{P}}_n$ as one must solve* (3.5)*.*

**Remark 3.3.1.**

- ***Uniform weights, no subsampling.*** *The estimator* (3.10) *gives back* (2.18) *when $\boldsymbol{\alpha}_k := \boldsymbol{\alpha}_{k_m} := \frac{1}{n}\mathbf{1}_n$ for all $m \in [M]$, and when there is no subsampling applied.*
- ***Runtime complexity.*** *In order to determine the computational complexity of* (3.10) *one has to find that of* (3.9)*; that of* (3.8) *follows by choosing $M = 1$ in* (3.9)*. $(a)$ and $(b)$ in* (3.9) *are Hadamard products; hence their computational complexity is $O(Mn'^2)$ and $O(Mnn')$. $(c)$ in* (3.9) *is the Moore-Penrose inverse of an $n' \times n'$ matrix; thus its complexity is $O(n'^3)$. Hence, the computation of $\boldsymbol{\alpha}_k$ costs $O(Mn'^2 + n'^3 + Mn'n)$, and that of $(\boldsymbol{\alpha}_{k_m})_{m=1}^M$ is $O(n'^2 + n'^3 + n'n)$ for each $m \in [M]$. In* (3.10) *each term can be computed in $O(Mn'^2)$. Overall the Nyström $M$-HSIC estimator has complexity $O(Mn'^2 + Mn'^3 + Mn'n) = O(Mn'^3 + Mn'n)$.*
- ***Difference compared to the estimator by Zhang et al. [2018].*** *For $M = 2$,* (3.10) *reduces to*

$$\mathrm{HSIC}^2_{k,N}\left(\hat{\mathbb{P}}_n\right) = \boldsymbol{\alpha}_k^\mathsf{T}\left(\circ_{i\in[2]}\mathbf{K}_{k_i,n',n'}\right)\boldsymbol{\alpha}_k + \prod_{i\in[2]} \alpha_{k_i}^\mathsf{T}\mathbf{K}_{k_i,n',n'}\boldsymbol{\alpha}_{k_i} - 2\boldsymbol{\alpha}_k^\mathsf{T}\left(\circ_{i\in[2]}\mathbf{K}_{k_i,n',n'}\boldsymbol{\alpha}_{k_i}\right). \quad (3.11)$$

*Using the equivalence of* (2.18) *and* (3.1) *in case $M = 2$ gives*

$$\mathrm{tr}\left(\mathbf{H}\mathbf{K}_{k_1}\mathbf{H}\mathbf{K}_{k_2}\right) = \frac{1}{n^2}\mathbf{1}_n^\mathsf{T}\left(\mathbf{K}_{k_1} \circ \mathbf{K}_{k_2}\right)\mathbf{1}_n + \frac{1}{n^4}\prod_{i\in[2]} \mathbf{1}_n^\mathsf{T}\mathbf{K}_{k_i}\mathbf{1}_n - \frac{2}{n^3}\mathbf{1}_n^\mathsf{T}\left(\mathbf{K}_{k_1}\mathbf{1}_n \circ \mathbf{K}_{k_2}\mathbf{1}_n\right),$$

*hence* (3.2) *becomes*

$$\mathrm{HSIC}^2_{k,N_0}\left(\hat{\mathbb{P}}_n\right) = \frac{1}{n^2}\mathbf{1}_n^\mathsf{T}\left(\mathbf{K}_{k_1}^{Nys} \circ \mathbf{K}_{k_2}^{Nys}\right)\mathbf{1}_n + \frac{1}{n^4}\prod_{i\in[2]} \mathbf{1}_n^\mathsf{T}\mathbf{K}_{k_i}^{Nys}\mathbf{1}_n - \frac{2}{n^3}\mathbf{1}_n^\mathsf{T}\left(\mathbf{K}_{k_1}^{Nys}\mathbf{1}_n \circ \mathbf{K}_{k_2}^{Nys}\mathbf{1}_n\right). \quad (3.12)$$

*The estimators* (3.11) *and* (3.12) *are identical if $\boldsymbol{\alpha}_k = \boldsymbol{\alpha}_{k_m} = \frac{1}{n}\mathbf{1}_n$ for all $m \in [2]$ and when there is no subsampling; in the general case they do not coincide. In* (3.2) *the dominant term in the complexity is $(n')^2 n$ (since $n' < n$), this reduces to $n'n$ in our proposed estimator* (3.10)*.*

Key to showing the consistency of the proposed Nyström $M$-HSIC estimator (3.10) (Proposition 3.3.1) is our next lemma, which describes how the Nyström approximation error of the mean embeddings of the components ($d_{k_m}$ below) can be propagated through tensor products.

**Lemma 3.3.3** (Error propagation on tensor products). *Let $X = (X_m)_{m=1}^M \in \mathcal{X} = \times_{m=1}^M \mathcal{X}_m$, $k_m : \mathcal{X}_m \times \mathcal{X}_m \to \mathbb{R}$ bounded kernels ($\exists a_{k_m} \in (0,\infty)$ such that $\sup_{x_m \in \mathcal{X}_m} \sqrt{k_m(x_m, x_m)} \leq a_{k_m}$, $m \in [M]$), $k = \otimes_{m=1}^M k_m$, $\mathcal{H}_k$ the RKHS associated to $k$, $X \sim \mathbb{P} \in \mathcal{M}_1^+(\mathcal{X})$, $\mathbb{P}_m$ the $m$-th marginal of $\mathbb{P}$ ($m \in [M]$), $n' \leq n$, and $\tilde{\mathbb{P}}_{m,n'}$ defined according to* (3.7). *Then*

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

*where* $d_{k_m} = \left\| \mu_{k_m}(\mathbb{P}_m) - \mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right) \right\|_{\mathcal{H}_{k_m}}$.

Our resulting Nyström $M$-HSIC performance guarantee is as follows.

**Proposition 3.3.1** (Error bound for Nyström $M$-HSIC). *Let $X = (X_m)_{m=1}^M \in \mathcal{X} = \times_{m=1}^M \mathcal{X}_m$, $X \sim \mathbb{P} \in \mathcal{M}_1^+(\mathcal{X})$, $(\mathcal{X}_m)_{m\in[M]}$ locally compact, second-countable topological spaces, $k_m : \mathcal{X}_m \times \mathcal{X}_m \to \mathbb{R}$ bounded kernels, i.e., $\exists a_{k_m} \in (0,\infty)$ such that $\sup_{x_m \in \mathcal{X}_m} \sqrt{k_m(x_m, x_m)} \leq a_{k_m}$ for all $m \in [M]$, $k = \otimes_{m=1}^M k_m$, $a_k = \prod_{m=1}^M a_{k_m}$, $\phi_{k_m}(x_m) = k_m(\cdot, x_m)$ for all $x_m \in \mathcal{X}_m$, $\phi_k = \otimes_{m=1}^M \phi_{k_m}$, $C_{\mathbb{P},k}$ and $C_{\mathbb{P}_m,k_m}$ as in* (2.9), *$\mathcal{N}_{\mathbb{P},k}$ and $\mathcal{N}_{\mathbb{P}_m,k_m}$ as in* (2.10), *the number of Nyström points $n' \leq n$, and $\hat{\mathbb{P}}_n$ defined according to* (2.17). *Then, for any $\delta \in \left(0, \frac{1}{M+1}\right)$*

$$\begin{aligned}
\left|\mathrm{HSIC}_k(\mathbb{P}) - \mathrm{HSIC}_{k,N}\left(\hat{\mathbb{P}}_n\right)\right| \leq\ & \underbrace{\frac{c_{k,1}}{\sqrt{n}}}_{t_{k,1}} + \underbrace{\frac{c_{k,2}}{n'}}_{t_{k,2}} + \underbrace{\frac{c_{k,3}\sqrt{\log(n'/\delta)}}{n'}\sqrt{\mathcal{N}_{\mathbb{P},k}\left(\frac{12a_k^2\log(n'/\delta)}{n'}\right)}}_{t_{k,3}} \\
& + \prod_{m\in[M]} \Bigg[ a_{k_m} + \underbrace{\frac{c_{k_m,1}}{\sqrt{n}}}_{t_{k_m,1}} + \underbrace{\frac{c_{k_m,2}}{n'}}_{t_{k_m,2}} \\
& + \underbrace{\frac{c_{k_m,3}\sqrt{\log(n'/\delta)}}{n'}\sqrt{\mathcal{N}_{\mathbb{P}_m,k_m}\left(\frac{12a_{k_m}^2\log(n'/\delta)}{n'}\right)}}_{t_{k_m,3}} \Bigg] \\
& - \prod_{m\in[M]} a_{k_m}
\end{aligned}$$

*holds with probability at least $1 - (M+1)\delta$, provided that*

$$n' \geq \max_{m\in[M]} \left(67, 12a_k^2 \left\|C_{\mathbb{P},k}\right\|_{\mathrm{op}}^{-1}, 12a_{k_m}^2 \left\|C_{\mathbb{P}_m,k_m}\right\|_{\mathrm{op}}^{-1}\right) \log\left(\frac{n'}{\delta}\right),$$

*with $c_{k,1} = 2a_k\sqrt{2\log(6/\delta)}$, $c_{k,2} = 4\sqrt{3}a_k\log(12/\delta)$, $c_{k,3} = 12\sqrt{3\log(12/\delta)}a_k$, $c_{k_m,1} = 2a_{k_m}\sqrt{2\log(6/\delta)}$, $c_{k_m,2} = 4\sqrt{3}a_{k_m}\log(12/\delta)$, and $c_{k_m,3} = 12\sqrt{3\log(12/\delta)}a_{k_m}$ for $m \in [M]$.*

As a baseline, to interpret the result (see the second bullet point in Remark 3.3.2), one could consider the V-statistic based HSIC estimator (2.18) for $M \geq 2$, which according to our following lemma has a convergence rate of $O_P\left(\frac{1}{\sqrt{n}}\right)$.

**Lemma 3.3.4** (Convergence rate of V-statistic based HSIC estimator)**.** *Let* $\mathrm{HSIC}_k\left(\hat{\mathbb{P}}_n\right)$ *be as in* (2.18) *on a metric space* $\mathcal{X} = \times_{m=1}^{M} \mathcal{X}_m$*, and* $\mathrm{HSIC}_k\left(\mathbb{P}\right) > 0$*. Then*

$$\left|\mathrm{HSIC}_k(\mathbb{P}) - \mathrm{HSIC}_k\left(\hat{\mathbb{P}}_n\right)\right| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

**Remark 3.3.2.**

- *From the terms* $t_{k,1}, t_{k,2}, t_{k_m,1}, t_{k_m,2}, m \in [M]$ *it follows that for* $n' < \sqrt{n}$ *the respective second term dominates, thus increasing the error; for* $n' > \sqrt{n}$ *the respective first term dominates and the computational complexity increases. The effective dimension* $\left(t_{k,3}, t_{k_m,3}\right)$ *controls the trade off between the two terms and can be related [Chatalic et al., 2022] to the decay of the eigenvalues of the respective covariance operator. A convergence rate of* $n^{-1/2}$ *for the sums* $t_{k,1} + t_{k,2} + t_{k,3}$ *and* $t_{k_m,1} + t_{k_m,2} + t_{k_m,3}$ *can be achieved if*
  - $\max_{m\in[M]} \left(\mathcal{N}_{\mathbb{P},k}(\lambda), \mathcal{N}_{\mathbb{P}_m,k_m}(\lambda)\right) \leq c\lambda^{-\gamma}$ *for some* $c > 0$ *and* $\gamma \in (0, 1]$ *with*

$$n' = n^{1/(2-\gamma)} \log(n/\delta), \text{ or}$$

  - $\max_{m\in[M]} \left(\mathcal{N}_{\mathbb{P},k}(\lambda), \mathcal{N}_{\mathbb{P}_m,k_m}(\lambda)\right) \leq \log(1 + c/\lambda)/\beta$ *for some* $c > 0$, $\beta > 0$, *and*

$$n' = \sqrt{n} \log\left(\sqrt{n} \max_{m\in[M]} \left(\frac{1}{\delta}, \frac{c}{6a_k^2}, \frac{c}{6a_{k_m}^2}\right)\right).$$

  *This rate of convergence propagates through the product.*
- *Lemma 3.3.4 establishes that the V-statistic based estimator of HSIC converges with rate* $n^{-1/2}$*. Recalling the last line of Table 3.1, setting* $n' = o\left(n^{2/3}\right)$*, the proposed estimator yields an asymptotic speedup over V-HSIC. Hence, setting* $n' = \tilde{\Theta}\left(\sqrt{n}\right)$ *allows to obtain the same rate of convergence while decreasing runtime. Assumption* $\mathrm{HSIC}_k(\mathbb{P}) > 0$ *in Lemma 3.3.4 protects one from attaining a convergence rate of* $n^{-1}$ *of* $\mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right)$.

## 3.4. Experiments

In this section, we demonstrate the efficiency of the proposed method (N-MHSIC) against the baselines NFSIC, RFF-HSIC, N-HSIC, and the quadratic-time V-statistic based HSIC estimator (V-HSIC) in the context of independence testing. Hence, the null hypothesis $H_0$ is that the joint distribution factorizes to the product of the marginals, the alternative $H_1$ is that this is not the case. The experiments study both synthetic (Section 3.4.1) and real-world (Section 3.4.2) examples, in terms of power and runtime.

We use the Gaussian kernel $k_m(\mathbf{x}_m, \mathbf{x}'_m) = \exp\left(-\gamma_{k_m} \left\|\mathbf{x}_m - \mathbf{x}'_m\right\|_2^2\right)$ for all experiments, with $\gamma_{k_m}$ chosen according to the median heuristic [Garreau et al., 2018]. For a fair comparison of the test power, we approximate the null distribution of each test statistic by the permutation approach with 250 samples. We then perform a one-sided test with an acceptance region of 5% ($\alpha = 0.05$), which we repeat, for all

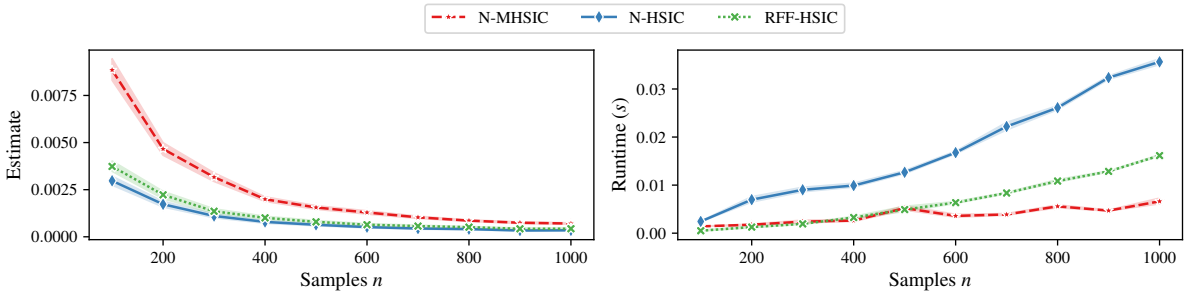


**Figure 3.1.:** Estimation accuracy for $M = 2$ components; the theoretical HSIC value is zero.

power experiments, on 100 independent draws of the data; the runtime results include these. We set each algorithm's parameters as recommended by the respective authors: For NFSIC, we set the number of test locations $J = 5$; the number of Fourier features (RFF-HSIC) and Nyström samples (N-HSIC) is set to $\sqrt{n}$. The number of Nyström samples of N-MHSIC is indicated within the experiment description. The opaque area in the figures indicates the 0.95-quantile obtained over 5 runs. All experiments were performed on a PC with Ubuntu 20.04, 124GB RAM, and 32 cores with 2GHz each.

### 3.4.1. Synthetic data

We examine three toy problems in the following, illustrating runtime and statistical power.

**Comparison of HSIC approximations under $H_0$.** First, for $M = 2$ components, we compare our proposed method to the existing accelerated HSIC estimators (N-HSIC, RFF-HSIC) on independent data to assess convergence w.r.t. runtime. Specifically, we set $X_1, X_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. The theoretical value of HSIC is thus zero. Figure 3.1 shows the estimates for sample sizes from 100 to 1000; the number of Nyström samples for N-MHSIC is set to $n' = 2\sqrt{n}$. All approaches converge to zero, with N-MHSIC converging a bit slower than the existing HSIC approximations. However, we note that the gap is on the order of $10^{-3}$ so it is close to the theoretical value also for small sample sizes. The runtime scales as predicted by the complexity analysis, with the proposed approach running faster than both N-HSIC and RFF-HSIC starting from $n = 500$ samples.

**Dependent data ($H_1$ holds).** To evaluate the statistical power on $M = 2$ components, we set $X_1 \sim \mathcal{N}(0, 1)$, $X_2 = X_1 + \epsilon$, and $\epsilon \sim \mathcal{N}(0, 1)$, with $n'$ set as before. Figure 3.2 shows that N-MHSIC achieves a power of one for $n \approx 100$ and that it is slightly worse than the existing HSIC approximations for small sample sizes. V-HSIC has the highest power but also the highest runtime. Even though NFSIC has linear runtime complexity it is slower than all other statistics on small sample sizes.

**Causal discovery.** The experiments until now considered $M = 2$ components. However, N-MHSIC allows for handling $M \geq 2$ components and thus can estimate the directed acyclic graph (DAG) governing causality if one assumes an additive noise model.

Specifically, we sample from the structural equations $X_i = \sum_{j \in \mathrm{PA}_i} f^{i,j}(X_j) + \epsilon_i$ for $i \in [M]$, of a randomly selected fully connected DAG with four nodes ($M = 4$), of which there are 24. In the equation,

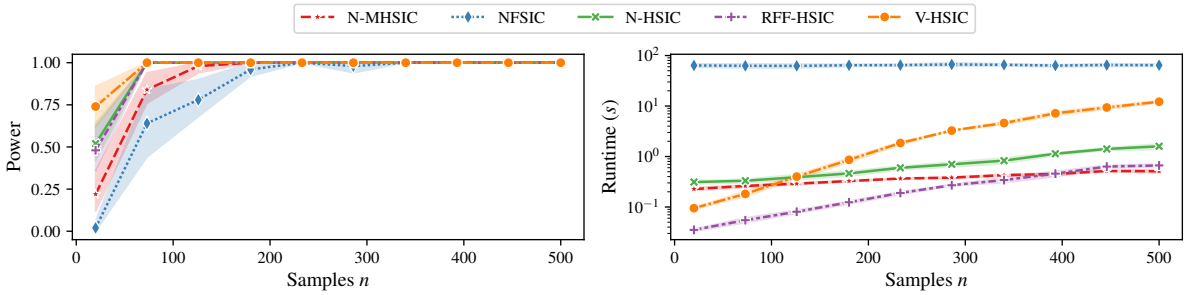


**Figure 3.2.:** Power on dependent data. Runtime on log scale.

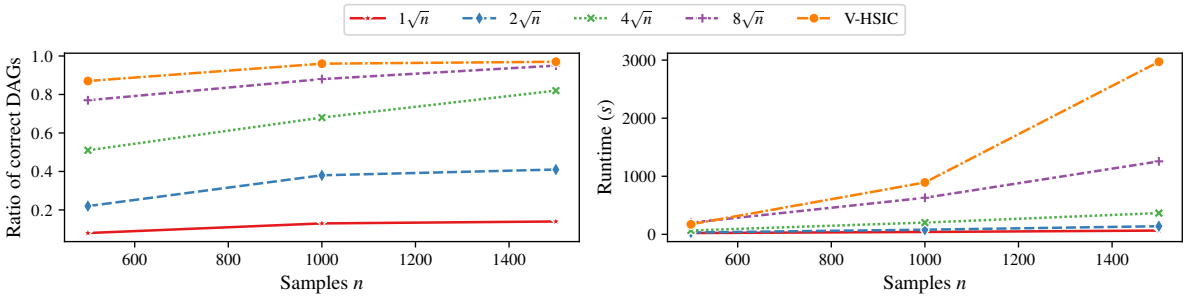


**Figure 3.3.:** Ratio of correctly identified DAGs with 4 nodes.

$\mathrm{PA}_i$ denotes the parents of $i$ in the associated DAG, and the $\epsilon_i$ are normally distributed and jointly independent, with a variance sampled independently from the uniform distribution $\mathrm{Unif}\left(1, \sqrt{2}\right)$.

To now test whether a particular DAG fits the data, Pfister et al. [2018] propose to use generalized additive model regression to find the residuals when regressing each node onto all its parents and to reject the DAG if the residuals are not jointly independent. If these are independent, we accept the causal structure. In this application, one is only interested in the relative $p$-values when performing the procedure for all possible DAGs with the correct number of nodes.

V-HSIC has the best performance in [Pfister et al., 2018], so we only compare against V-HSIC; it is also the only other approach which allows testing joint independence of more than two components. Figure 3.3 shows how often N-MHSIC and V-HSIC identify the correct DAG in 100 samples. V-HSIC has higher power than N-MHSIC and more often identifies the correct DAG for small sample sizes. However, as the r.h.s. of Figure 3.3 shows, the proposed algorithm runs twice as fast as V-HSIC even for $n' = 8\sqrt{n}$ and $n = 1500$ while producing the same result quality. Due to their different runtime complexities, the gap in runtime widens further with increasing sample size.

### 3.4.2. Real-world data

This section is dedicated to benchmarks on real-world data.

**Million Song Data.** The Million Song Data [Bertin-Mahieux et al., 2011] contains approximately 500,000 songs. Each has 90 features ($X$) together with its year of release, which ranges from 1922 to 2011 ($Y$). The algorithms must detect the dependence between the features and the year of release. To approximate the power, we draw 100 independent samples of the whole data set. Figure 3.4 shows the results, for level $\alpha = 0.01$; the different ranges of $n$ highlight the asymptotic runtime gains. In

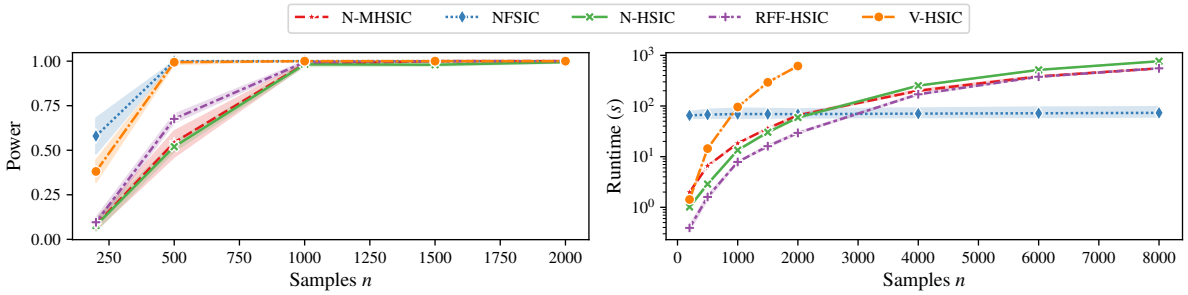


**Figure 3.4.:** Test power vs. runtime on the Million Song Data.

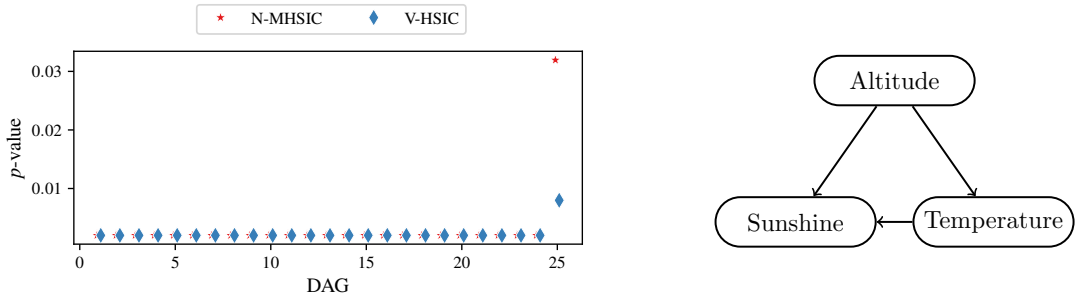


**Figure 3.5.:** Testing for joint independence on the residuals of DAGs with three nodes (left) and the DAG with the largest $p$-value (right). The $p$-values agree on DAGs 1 to 24.

contrast to a similar experiment of Jitkrittum et al. [2017a], we use a permutation approach for all two-sample tests and increase the number of Nyström samples (random Fourier features) as a function of $n$, obtaining higher power throughout. The problem is sufficiently challenging, so that we set the number of Nyström samples to $8\sqrt{n}$ for N-MHSIC. V-HSIC and NFSIC achieve maximum power from $n = 500$. N-MHSIC features similar runtime and power as the existing HSIC approximations N-HSIC and RFF-HSIC but can handle more than two components. The runtime plot illustrates that the lower asymptotic complexity of N-MHSIC compared to V-HSIC also holds in practice.

**Weather Causal Discovery.** Here, we aim to infer the correct causality DAG from real-world data, namely the data set of Mooij et al. [2016] which contains 349 measurements consisting of altitude, temperature and sunshine. The goal is to infer the most plausible DAG with three nodes ($d = 3$) out of the 25 possible DAGs ($3^3 - 2 = 25$; two graphs have a cycle). We assume the structural equations discussed before. Figure 3.5 shows the $p$-values with the estimated DAG (with index 25) having the largest $p$-value. Again, we compare our results to V-HSIC and find that both successfully identify the most plausible DAG [Pfister et al., 2018].

These experiments demonstrate the efficiency of the proposed Nyström $M$-HSIC method.

The next section contains our proofs.

## 3.5. Proofs

This section is dedicated to proofs. Lemma 3.3.2 is derived in Section 3.5.1. Proposition 3.3.1 is proved in Section 3.5.4 relying on an auxiliary lemma shown in Section 3.5.2 and Lemma 3.3.3, proved in Section 3.5.3. Lemma 3.3.4 is proved in Section 3.5.6, with an auxiliary result in Section 3.5.5.

### 3.5.1. Proof of Lemma 3.3.2

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

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

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

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

Similarly, we obtain for term $B$ that

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

where in $(*)$ we used (2.2), the linearity of the inner product, and the reproducing property.

Last, we express term $C$ as

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

where (a) follows from the linearity of the inner product, (b) holds by (2.2), (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 $\mathbf{K}_{k_m,n',n'}$ as $\left(\mathbf{K}_{k_m,n',n'}\right)_i$.

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

### 3.5.2. Lemma to the Proof of Proposition 3.3.1

Our main result relies on Lemma 3.3.3 and on the following result.

**Lemma 3.5.1** (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 $\sup_{x_m\in\mathcal{X}_m} \sqrt{k_m(x_m,x_m)} \leq a_{k_m}$ 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 $k$, $\phi_k = \otimes_{m=1}^M \phi_{k_m}$, $C_{\mathbb{P},k}$ as in (2.9), $n' \leq n$, and $\tilde{\mathbb{P}}_{n'}$ defined according to (3.6). Then for any $\delta \in (0,1)$ it holds that*

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

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

$$
n' \geq \max\left(67, 12a_k^2 \left\|C_{\mathbb{P},k}\right\|_{\mathrm{op}}^{-1}\right) \log\left(\frac{n'}{\delta}\right),
$$

*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 $\sqrt{k(x,x)} = \prod_{m=1}^M \sqrt{k_m(x_m,x_m)} \leq a_k$, the statement is implied by Theorem A.1.1. □

### 3.5.3. Proof of Lemma 3.3.3

To simplify notation, we define $\mu_{k_m} = \mu_{k_m}(\mathbb{P}_m)$, $\tilde{\mu}_{k_m} = \mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right)$, $\mathcal{H}_k = \otimes_{m=1}^{M}\mathcal{H}_{k_m}$, 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. = $\left\|\mu_{k_1}(\mathbb{P}_1) - \mu_{k_1}\left(\tilde{\mathbb{P}}_{1,n'}\right)\right\|_{\mathcal{H}_k}$ is satisfied, and we assume that the statement holds for $M = M - 1$, to obtain

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

where (a) holds by the triangle inequality, (b) is implied by (2.3) and the definition of $d_{k_M}$, (c) follows from

$$
\left\|\mu_{k_m}\right\|_{\mathcal{H}_{k_m}} = \left\|\int_{\mathcal{X}_m} k_m(\cdot, x_m)\mathrm{d}\mathbb{P}_m(x_m)\right\|_{\mathcal{H}_{k_m}} \overset{(e)}{\leq} \int_{\mathcal{X}_m}\underbrace{\left\|k_m(\cdot, x_m)\right\|_{\mathcal{H}_{k_m}}}_{\overset{(f)}{=}\sqrt{k_m(x_m,x_m)}\overset{(g)}{\leq}a_{k_m}}\mathrm{d}\mathbb{P}_m(x_m) \leq a_{k_m}, \quad (3.13)
$$

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

(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 (3.13) and the definition of $d_{k_M}$.

### 3.5.4. Proof of Proposition 3.3.1

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

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

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 3.5.1; in other words, for any $\delta \in (0,1)$ with probability at least $(1-\delta)$ it holds that

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

provided that $n' \geq \max\left(67, 12a_k^2\left\|C_{\mathbb{P},k}\right\|_{\mathrm{op}}^{-1}\right)\log\left(\frac{n'}{\delta}\right)$, 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 3.3.3 to the second term gives

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

We now bound the error of each of the $M$ factors by Theorem A.1.1, i.e., for fixed $m \in [M]$; particularly, we get that for any $\delta \in (0,1)$ with probability at least $1-\delta$

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

Hence,

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

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

$$\begin{aligned}&\prod_{m\in[M]}\left[a_{k_m} + \left\|\mu_{k_m}(\mathbb{P}_m) - \mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right)\right\|_{\mathcal{H}_{k_m}}\right] \leq \\ &\quad\leq \prod_{m\in[M]}\left[a_{k_m} + \frac{c_{k_m,1}}{\sqrt{n}} + \frac{c_{k_m,2}}{n'} + \frac{c_{k_m,3}\sqrt{\log(n'/\delta)}}{n'}\sqrt{\mathcal{N}_{\mathbb{P}_m,k_m}\left(\frac{12a_{k_m}^2\log(n'/\delta)}{n'}\right)}\right],\end{aligned}$$

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

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

provided that

$$n'\geq\max\left(67,12a_{k_m}^2\left\|C_{\mathbb{P}_m,k_m}\right\|_{\mathrm{op}}^{-1}\right)\log\left(\frac{n'}{\delta}\right)$$

for all $m\in[M]$, with $C_{\mathbb{P}_m,k_m}=\mathbb{E}_{\mathbb{P}_m}\left[\phi_{k_m}(X_m)\otimes\phi_{k_m}(X_m)\right]$ 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.

### 3.5.5. Lemma to the Proof of Lemma 3.3.4

**Lemma 3.5.2** (Deviation bound for U-statistics based HSIC estimator)**.** *It holds that*

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

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

*Proof.* We show that (2.15) can be expressed as a sum of U-statistics and then bound the terms individually. First, square (2.15) 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 $\mathrm{HSIC}_{k,u}^2\left(\hat{\mathbb{P}}_n\right)=A_n'+B_n'-2C_n'$, and split $t$ as $\alpha t+\beta t+(1-\alpha-\beta)t$, with $\alpha,\beta>0$ and $\alpha+\beta<1$. One obtains

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

Doubling and rewriting Theorem A.1.2, we have that for U-statistics and any $\delta\in(0,1)$

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

Now, choosing the $(\theta,U_n,u)$ triplet to be $\left(A,A_n',\alpha t\right)$, $\left(B,B_n',\beta t\right)$, $\left(C,C_n',\frac{(1-\alpha-\beta)t}{2}\right)$, respectively, setting $m=2M$, and observing that $a\leq k(x,y)\leq b$ 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. □

### 3.5.6. Proof of Lemma 3.3.4

We consider the decomposition

$$\left|\mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}_k^2\left(\mathbb{P}\right)\right| \leq \underbrace{\left|\mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}_{k,u}^2\left(\hat{\mathbb{P}}_n\right)\right|}_{t_1} + \underbrace{\left|\mathrm{HSIC}_{k,u}^2\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}_k^2\left(\mathbb{P}\right)\right|}_{t_2}, \quad (3.14)$$

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

**Second term ($t_2$)**: Lemma 3.5.2 establishes that $t_2 = O_P\left(\frac{1}{\sqrt{n}}\right)$.

**First term ($t_1$)**: To bound $t_1$, first, by Markov's inequality (Lemma A.1.2) observe that, for $\epsilon := \frac{\mathbb{E}\left(\left|\mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}_{k,u}^2\left(\hat{\mathbb{P}}_n\right)\right|\right)}{a}$,

$$\begin{aligned}
\mathbb{P}\left(\left|\mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}_{k,u}^2\left(\hat{\mathbb{P}}_n\right)\right| \geq \frac{\mathbb{E}\left(\left|\mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}_{k,u}^2\left(\hat{\mathbb{P}}_n\right)\right|\right)}{\epsilon}\right) \leq \epsilon, \\
\mathbb{P}\left(\left|\mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}_{k,u}^2\left(\hat{\mathbb{P}}_n\right)\right| < \frac{\mathbb{E}\left(\left|\mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}_{k,u}^2\left(\hat{\mathbb{P}}_n\right)\right|\right)}{\epsilon}\right) \geq 1 - \epsilon, \\
\mathbb{P}\left(\left|\mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}_{k,u}^2\left(\hat{\mathbb{P}}_n\right)\right| < \frac{C}{n\epsilon}\right) \overset{(*)}{\geq} 1 - \epsilon, \\
\mathbb{P}\left(\left|\mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}_{k,u}^2\left(\hat{\mathbb{P}}_n\right)\right| \geq \frac{C}{n\epsilon}\right) \leq \epsilon, \quad (3.15)
\end{aligned}$$

for constant $C > 0$ and $n$ large enough, where $(*)$ follows from Lemma A.1.1 (with $r = 1$). (3.15) implies that

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

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

$$\begin{aligned}
\left|\mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}_k^2\left(\mathbb{P}\right)\right| &\overset{(3.14)}{\leq} \left|\mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}_{k,u}^2\left(\hat{\mathbb{P}}_n\right)\right| + \left|\mathrm{HSIC}_{k,u}^2\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}_k^2\left(\mathbb{P}\right)\right| \\
&= O_P\left(\frac{1}{n}\right) + O_P\left(\frac{1}{\sqrt{n}}\right) = O_P\left(\frac{1}{\sqrt{n}}\right). \quad (3.16)
\end{aligned}$$

Hence

$$\begin{aligned}
O_P\left(\frac{1}{\sqrt{n}}\right) &\overset{(3.16)}{\geq} \left|\mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}_k^2\left(\mathbb{P}\right)\right| = \left|\mathrm{HSIC}_k\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}_k\left(\mathbb{P}\right)\right| \left|\underbrace{\mathrm{HSIC}_k\left(\hat{\mathbb{P}}_n\right)}_{\overset{(2.18)}{\geq} 0} + \underbrace{\mathrm{HSIC}_k\left(\mathbb{P}\right)}_{\geq 0}\right| \\
&\geq \left|\mathrm{HSIC}_k\left(\hat{\mathbb{P}}_n\right) - \mathrm{HSIC}_k\left(\mathbb{P}\right)\right| \mathrm{HSIC}_k\left(\mathbb{P}\right),
\end{aligned}$$

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

# 4. The Minimax Rate of HSIC Estimation for Translation-Invariant Kernels

The content of this chapter is based on the following publication.

- F. Kalinke and Z. Szabó. The minimax rate of HSIC estimation for translation-invariant kernels. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 108468–108489, 2024. neurips.cc.

## 4.1. Introduction

Recall from Section 2.3.2 and Chapter 3 that HSIC has found numerous applications and that many estimators for HSIC exist. The classical ones, detailed in Section 2.3.2, rely on U-statistics or V-statistics [Gretton et al., 2005, Quadrianto et al., 2009, Pfister et al., 2018] and are known to converge at a rate of $O_P\left(n^{-1/2}\right)$. In fact, the V-statistic-based estimator is obtained by replacing the population kernel mean embedding with its empirical counterpart; estimating the mean embedding can be carried out at a speed of $O_P\left(n^{-1/2}\right)$ [Smola et al., 2007, Theorem 2], which implies that HSIC can be estimated at the same rate. Existing approximations such as Nyström $M$-HSIC [Kalinke and Szabó, 2023] (recalled in Chapter 3), also achieve this rate under the assumption of an appropriate rate of decay of the effective dimension. While all of these upper bounds match asymptotically, it is not known whether HSIC can be estimated at a faster rate, that is, whether the upper bound of $O_P\left(n^{-1/2}\right)$ is optimal in the minimax sense, or if designing estimators achieving better rates is possible. Lower bounds for the related MMD are known [Tolstikhin et al., 2016], but the existing analysis considers radial kernels and relies on independent Gaussian distributions. Radial kernels are a special case of the more general class of translation-invariant kernels that we consider.[1] The reliance on independent Gaussian distributions renders the analysis of Tolstikhin et al. [2016] inapplicable for HSIC estimation.

We tackle both of these severe restrictions with the following **contributions**.

- We establish the minimax lower bound $O\left(n^{-1/2}\right)$ of HSIC estimation with $M \geq 2$ components on $\mathbb{R}^d$ with continuous bounded translation-invariant characteristic kernels. As this lower bound matches the known upper bounds of the existing "classical" U-statistic and V-statistic-based estimators, and that of the Nyström $M$-HSIC estimator (Chapter 3), our result settles their minimax optimality.
- Specifically, our result also implies the minimax lower bound of $O\left(n^{-1/2}\right)$ for the estimation of the cross-covariance operator, which can be further specialized to get back the minimax result [Zhou et al., 2019, Theorem 5] on the estimation of the covariance operator.

The remainder of this chapter is structured as follows. Our results are in Section 4.2 and we collect their proofs in Section 4.3.

[1] The family of radial kernels encompasses, for example, Gaussians, mixtures of Gaussians, inverse multiquadratics, and Matérn kernels; the Laplace kernel is translation-invariant but not radial (with respect to the traditionally-chosen Euclidean norm $\|\cdot\|_{\mathbb{R}^d}$).

## 4.2. Results

This section is dedicated to our results: The minimax lower bound for the estimation of $\mathrm{HSIC}_k(\mathbb{P})$, where $k$ is a product of continuous bounded translation-invariant characteristic kernels is given in Theorem 4.2.1(ii). For the specific case where $k$ is a product of Gaussian kernels (stated in Theorem 4.2.1(i)), the constant in the lower bound is made explicit. Theorem 4.2.1(ii) also helps to establish a lower bound on the estimation of the cross-covariance operator (Corollary 4.2.1).

Before presenting our results, we recall the framework of minimax estimation [Tsybakov, 2009] adapted to our setting. Let $\hat{F}_n$ denote any estimator of $\mathrm{HSIC}_k(\mathbb{P})$ based on $n$ i.i.d. samples from $\mathbb{P}$. A sequence $(\xi_n)_{n=1}^{\infty}$ ($\xi_n > 0$ for all $n$) is said to be a lower bound of HSIC estimation w.r.t. a class $\mathcal{P}$ of Borel probability measures on $\mathbb{R}^d$ if there exists a constant $c > 0$ such that

$$\inf_{\hat{F}_n} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}^n \left\{ \xi_n^{-1} \left| \mathrm{HSIC}_k(\mathbb{P}) - \hat{F}_n \right| \geq c \right\} > 0. \tag{4.1}$$

If a specific estimator of HSIC $\tilde{F}_n$ has an upper bound that matches $(\xi_n)_{n=1}^{\infty}$ up to constants, that is,

$$\left| \mathrm{HSIC}_k(\mathbb{P}) - \tilde{F}_n \right| = O_P(\xi_n), \tag{4.2}$$

then $\tilde{F}_n$ is called minimax optimal.

We use Le Cam's method [Cam, 1973, Tsybakov, 2009] (recalled in Theorem A.2.5) to obtain bounds as in (4.1); estimators of HSIC achieving the bounds in (4.2) with $\xi_n = n^{-1/2}$ are quoted in the introduction to this chapter. The key to the application of the method is to show that there exist $\alpha > 0$ and $n_0 \in \mathbb{N}_{>0}$ such that for all $n \geq n_0$ one can find an adversarial pair of distributions $\left(\mathbb{P}_{\theta_0}, \mathbb{P}_{\theta_1}\right) = \left(\mathbb{P}_{\theta_0}(n), \mathbb{P}_{\theta_1}(n)\right) \in \mathcal{P} \times \mathcal{P}$ and $s_n > 0$ for which

1. $\mathrm{KL}\left(\mathbb{P}_{\theta_1}^n || \mathbb{P}_{\theta_0}^n\right) \leq \alpha$, in other words, the corresponding $n$-fold product measures must be similar in the sense of Kullback-Leibler divergence, but
2. $\left|\mathrm{HSIC}_k(\mathbb{P}_{\theta_1}) - \mathrm{HSIC}_k(\mathbb{P}_{\theta_0})\right| \geq 2s_n$, that is, their corresponding values of HSIC must be dissimilar.

In this case, $\inf_{\hat{F}_n} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}^n \left\{ \left| \mathrm{HSIC}_k(\mathbb{P}) - \hat{F}_n \right| \geq s_n \right\} \geq \max\left( \frac{e^{-\alpha}}{4}, \frac{1-\sqrt{\alpha/2}}{2} \right)$ for all $n \geq n_0$; hence to establish the minimax optimality of existing estimators w.r.t. their known upper bounds, it is sufficient to find adversarial pairs $\left\{ \left(\mathbb{P}_{\theta_0}(n), \mathbb{P}_{\theta_1}(n)\right) \right\}_{n \geq n_0}$ that satisfy 1. for some positive constant $\alpha$ and also fulfill 2. with $s_n = \Theta\left(n^{-1/2}\right)$.

The proof of the first part of our statement relies on the following Lemma 4.2.1 which yields the analytical value of $\mathrm{HSIC}_k(\mathcal{N}(\boldsymbol{\mu}, \Sigma))$, where $k = \otimes_{m=1}^{M} k_m$ is the product of Gaussian kernels $k_m$ ($m \in [M]$) and $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ denotes the multivariate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$.

**Lemma 4.2.1** (Analytical value of HSIC for the Gaussian setting)**.** *Let us consider the Gaussian kernel $k(\mathbf{x}, \mathbf{y}) = e^{-\frac{\gamma}{2}\|\mathbf{x}-\mathbf{y}\|_{\mathbb{R}^d}^2}$ ($\gamma > 0$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$) and Gaussian random variable $X = (X_m)_{m=1}^{M} \sim \mathcal{N}(\mathbf{m}, \Sigma) =: \mathbb{P}$, where $X_m \in \mathbb{R}^{d_m}$ ($m \in [M]$), $\mathbf{m} = (\mathbf{m}_m)_{m=1}^{M} \in \mathbb{R}^d$, $\Sigma = [\Sigma_{i,j}]_{i,j\in[M]} \in \mathbb{R}^{d\times d}$, $\Sigma_{i,j} \in \mathbb{R}^{d_i \times d_j}$, and $d = \sum_{m\in[M]} d_m$. In this case, with $\Sigma_1 = \Sigma$ and $\Sigma_2 = \mathrm{bdiag}(\Sigma_{1,1}, \ldots, \Sigma_{M,M})$, we have*

$$\mathrm{HSIC}_k^2(\mathbb{P}) = \frac{1}{|2\gamma\Sigma_1 + \mathbf{I}_d|^{\frac{1}{2}}} + \frac{1}{|2\gamma\Sigma_2 + \mathbf{I}_d|^{\frac{1}{2}}} - \frac{2}{|\gamma\Sigma_1 + \gamma\Sigma_2 + \mathbf{I}_d|^{\frac{1}{2}}}.$$

In this work, we focus on continuous bounded translation-invariant kernels, which are fully characterized by Bochner's theorem [Wendland, 2005, Theorem 6.6] (recalled in Theorem A.2.1); the theorem states that a function on $\mathbb{R}^d$ is positive definite if and only if it is the Fourier transform of a finite nonnegative measure.[2] We use this description to obtain our main result, which is as follows.

**Theorem 4.2.1** (Lower bound for HSIC estimation on $\mathbb{R}^d$)**.** *Let $\mathcal{P}$ be a class of Borel probability measures over $\mathbb{R}^d$ containing the $d$-dimensional Gaussian distributions. Let $d = \sum_{m\in[M]} d_m$ and $\hat{F}_n$ denote any estimator of $\mathrm{HSIC}_k(\mathbb{P})$ with $n \geq 2 =: n_0$ i.i.d. samples from $\mathbb{P} \in \mathcal{P}$. Assume further that $k = \otimes_{m=1}^M k_m$ where either, for $m \in [M]$,*

*(i) the kernels $k_m : \mathbb{R}^{d_m} \times \mathbb{R}^{d_m} \to \mathbb{R}$ are Gaussian with common bandwidth parameter $\gamma > 0$ defined by $(\mathbf{x}_m, \mathbf{x}'_m) \mapsto e^{-\frac{\gamma}{2}\left\|\mathbf{x}_m - \mathbf{x}'_m\right\|^2_{\mathbb{R}^{d_m}}}$ $(\mathbf{x}_m, \mathbf{x}'_m \in \mathbb{R}^{d_m})$, or*

*(ii) the kernels $k_m : \mathbb{R}^{d_m} \times \mathbb{R}^{d_m} \to \mathbb{R}$ are continuous bounded translation-invariant characteristic kernels.*

*Then, for any $n \geq n_0$, it holds that*

$$\inf_{\hat{F}_n} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}^n \left\{ \left| \mathrm{HSIC}_k\left(\mathbb{P}\right) - \hat{F}_n \right| \geq \frac{c}{\sqrt{n}} \right\} \geq \frac{1 - \sqrt{\frac{5}{8}}}{2},$$

*with (i) the constant $c = \frac{\gamma}{2(2\gamma+1)^{\frac{d}{4}+1}} > 0$ (depending on $\gamma$ and $d$ only) in the first case, or (ii) some constant $c > 0$ in the second case.*

We note that while Theorem 4.2.1(ii) applies to the more general class of translation-invariant kernels, we include Theorem 4.2.1(i) as it makes the constant $c$ explicit.

The following corollary allows to recover the recent lower bound on the estimation of the covariance operator by Zhou et al. [2019, Theorem 5] as a special case that we detail in Remark 4.2.1(e).

**Corollary 4.2.1** (Lower bound on cross-covariance operator estimation)**.** *In the setting of Theorem 4.2.1(ii), let $\hat{F}_n$ denote any estimator of the centered cross-covariance operator $\tilde{C}^c_{\mathbb{P},k} \in \mathcal{H}_k$ defined in* (2.16) *with $n \geq 2 =: n_0$ i.i.d. samples from $\mathbb{P} \in \mathcal{P}$. Then, for any $n \geq n_0$, it holds that*

$$\inf_{\hat{F}_n} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}^n \left\{ \left\| \tilde{C}^c_{\mathbb{P},k} - \hat{F}_n \right\|_{\mathcal{H}_k} \geq \frac{c}{\sqrt{n}} \right\} \geq \frac{1 - \sqrt{\frac{5}{8}}}{2},$$

*for some constant $c > 0$.*

**Remark 4.2.1.**

*(a)* ***Validness of HSIC.*** *Though generally the characteristic property of $(k_m)_{m=1}^M$-s is not enough [Szabó and Sriperumbudur, 2018, Example 2] for $M > 2$ to ensure the $\mathcal{I}$-characteristic property of $k = \otimes_{m=1}^M k_m$ (in other words, that $\mathrm{HSIC}_k(\mathbb{P}) = 0$ iff. $\mathbb{P} = \otimes_{m=1}^M \mathbb{P}_m$), on $\mathbb{R}^d$ under the imposed continuous bounded translation-invariant assumption (i) $k$ being characteristic, (ii) $k$ being $\mathcal{I}$-characteristic, and (iii) $(k_m)_{m=1}^M$-s being characteristic are equivalent (Theorem A.2.4).*

[2] We note that for many translation-invariant kernels, the corresponding spectral measures are known [Sriperumbudur et al., 2010, Table 2].

(b) ***Minimax optimality of existing HSIC estimators.*** *The lower bounds in Theorem 4.2.1 asymptotically match the known upper bounds of the U-statistic and V-statistic-based estimators of $\xi_n = n^{-1/2}$. The Nyström-based HSIC estimator achieves the same rate under an appropriate decay of the eigenspectrum of the respective covariance operator. Hence, Theorem 4.2.1 implies the optimality of these estimators on $\mathbb{R}^d$ with continuous bounded translation-invariant characteristic kernels in the minimax sense.*

(c) ***Difference compared to Tolstikhin et al. [2016] (minimax MMD estimation).*** *We note that a lower bound for the related $\mathrm{MMD}_k$ exists. However, the adversarial distribution pair $(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_0})$ constructed by Tolstikhin et al. [2016, Theorem 1] to obtain the lower bound on MMD estimation has a product structure which implies that $\left|\mathrm{HSIC}_k(\mathbb{P}_{\theta_1}) - \mathrm{HSIC}_k(\mathbb{P}_{\theta_0})\right| = 0$ and hence it is not applicable in our case of HSIC; Tolstikhin et al. [2016, Theorem 2] with radial kernels has the same restriction.*

(d) ***Difference compared to Tolstikhin et al. [2017] (minimax mean embedding estimation).*** *The estimation of the mean embedding $\mu_k(\mathbb{P})$ is known to have a minimax rate of $O\left(n^{-1/2}\right)$. But, this rate does not imply an optimal lower bound for the estimation of MMD as is evident from the two works [Tolstikhin et al., 2016, 2017]. The same conclusion holds for HSIC estimation.*

(e) ***Difference compared to Zhou et al. [2019] (minimax covariance operator estimation).*** *For the related problem of estimating the centered covariance operator*

$$C^c_{\mathbb{P},k} := \int_{\mathbb{R}^d} (\phi_k(x) - \mu_k(\mathbb{P})) \otimes (\phi_k(x) - \mu_k(\mathbb{P})) \, \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k \otimes \mathcal{H}_k,$$

*Zhou et al. [2019, Theorem 5] give the lower bound*

$$\inf_{\hat{F}_n} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}^n \left\{ \left\| C^c_{\mathbb{P},k} - \hat{F}_n \right\|_{\mathcal{H}_k \otimes \mathcal{H}_k} \geq \frac{c}{\sqrt{n}} \right\} \geq 1/8$$

*in the same setting as in Theorem 4.2.1(ii), where $\hat{F}_n$ is any estimator of the centered covariance $C^c_{\mathbb{P},k}$, and $c$ is a positive constant. By noting that the centered covariance is the centered cross-covariance of a random variable with itself, Corollary 4.2.1 recovers their result.*

The next section contains our proofs.

## 4.3. Proofs

This section is dedicated to our proofs. We present the proof of Lemma 4.2.1 in Section 4.3.1, an auxiliary result in Section 4.3.2, the proof of Theorem 4.2.1 in Section 4.3.3, and that of Corollary 4.2.1 in Section 4.3.4.

### 4.3.1. Proof of Lemma 4.2.1

As

$$\begin{aligned}\mathrm{HSIC}^2_k(\mathbb{P}) = \mathrm{MMD}^2_k(\mathbb{P}, \mathbb{Q}) &= \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|^2_{\mathcal{H}_k} \\ &= \langle \mu_k(\mathbb{P}), \mu_k(\mathbb{P}) \rangle_{\mathcal{H}_k} + \langle \mu_k(\mathbb{Q}), \mu_k(\mathbb{Q}) \rangle_{\mathcal{H}_k} - 2\langle \mu_k(\mathbb{P}), \mu_k(\mathbb{Q}) \rangle_{\mathcal{H}_k}\end{aligned}$$

with $\mathbb{Q} = \otimes_{m=1}^M \mathbb{P}_m = \mathcal{N}(\mathbf{m}, \mathrm{bdiag}(\Sigma_{1,1}, \ldots, \Sigma_{M,M}))$, $\mathbb{P}_m = \mathcal{N}(\mathbf{m}_m, \Sigma_{m,m})$, it is sufficient to be able to compute $\langle \mu_k(\mathbb{P}), \mu_k(\mathbb{Q}) \rangle_{\mathcal{H}_k}$-type quantities with $\mathbb{P} = \mathcal{N}(\mathbf{m}_1, \Sigma_1)$ and $\mathbb{Q} = \mathcal{N}(\mathbf{m}_2, \Sigma_2)$. One can show

[Muandet et al., 2011, Table 1] that $\langle \mu_k(\mathbb{P}), \mu_k(\mathbb{Q}) \rangle_{\mathcal{H}_k} = \frac{e^{-\frac{1}{2}(\mathbf{m}_1-\mathbf{m}_2)^\top \left(\Sigma_1+\Sigma_2+\gamma^{-1}\mathbf{I}_d\right)^{-1}(\mathbf{m}_1-\mathbf{m}_2)}}{|\gamma\Sigma_1+\gamma\Sigma_2+\mathbf{I}_d|^{\frac{1}{2}}}$. Using this fact and that $\mathbf{m} = \mathbf{m}_1 = \mathbf{m}_2$, the result follows.

### 4.3.2. Auxiliary result

In this section, we collect an auxiliary result. Lemma 4.3.1 presents an upper bound on the Kullback-Leibler divergence between multivariate normal distributions.

**Lemma 4.3.1** (Upper bound on KL divergence). *Let $d = \sum_{m=1}^{M} d_m$, with $d_m \in \mathbb{N}_{>0}$ ($m \in [M]$). Fix $i \in [d_1]$. Let $j = i + 1$, $\mathbb{P}_{\theta_0} = \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$, and $\mathbb{P}_{\theta_1} = \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$, with $\boldsymbol{\mu}_1 = \frac{1}{\sqrt{dn}}\mathbf{1}_d \in \mathbb{R}^d$, and $\Sigma_1 = \Sigma(i, j, \rho_n) \in \mathbb{R}^{d\times d}$ defined as in* (4.3) *($\rho_n \in (0, 1)$). Then, for $2 \leq n \in \mathbb{N}$,*

$$\mathrm{KL}(\mathbb{P}_{\theta_1}^n || \mathbb{P}_{\theta_0}^n) \leq \frac{1}{2n} + \frac{n}{2}\frac{\rho_n^2}{1-\rho_n^2}.$$

*In particular, for $\rho_n^2 = 1/n$, it holds that $\mathrm{KL}(\mathbb{P}_{\theta_1}^n || \mathbb{P}_{\theta_0}^n) \leq \frac{5}{4}$.*

*Proof.* With $\boldsymbol{\mu}_0 = \mathbf{0}_d$ and $\Sigma_0 = \mathbf{I}_d$, we obtain that

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\theta_1}^n || \mathbb{P}_{\theta_0}^n) &\overset{(a)}{=} \sum_{i\in[n]} \mathrm{KL}(\mathbb{P}_{\theta_1} || \mathbb{P}_{\theta_0}) \overset{(b)}{=} \frac{n}{2}\left[\mathrm{tr}(\Sigma_0^{-1}\Sigma_1) + (\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^\top\Sigma_0^{-1}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1) - d + \ln\left(\frac{|\Sigma_0|}{|\Sigma_1|}\right)\right] \\ &= \frac{n}{2}\left[\underbrace{\mathrm{tr}(\Sigma_1)}_{=d} + \underbrace{\|\boldsymbol{\mu}_1\|_{\mathbb{R}^d}^2}_{=\frac{1}{n^2}} - d + \ln\left(\underbrace{\frac{1}{|\Sigma_1|}}_{\overset{(c)}{=}\frac{1}{1-\rho_n^2}}\right)\right] = \frac{1}{2n} + \frac{n}{2}\ln\left(\frac{1}{1-\rho_n^2}\right) \overset{(d)}{\leq} \frac{1}{2n} + \frac{n}{2}\frac{\rho_n^2}{1-\rho_n^2} \overset{(e)}{\leq} \frac{5}{4},\end{aligned}$$

where (a) is implied by Lemma A.2.1, (b) follows from Lemma A.2.2, (c) follows from the definition of the determinant, (d) is the consequence of the inequality $\ln(x) \leq x - 1$ holding for $x > 0$, and (e) holds for $n \geq 2$ and $\rho_n^2 = 1/n$ as

$$\frac{n}{2}\underbrace{\frac{1/n}{1-1/n}}_{\frac{1}{n-1}} \leq 1 \iff \frac{n}{2}\frac{1}{n-1} \leq 1 \iff n \leq 2(n-1) \iff n \geq 2,$$

and in this case (for $n \geq 2$) one has that $\frac{1}{2n} \leq \frac{1}{4}$. □

### 4.3.3. Proof of Theorem 4.2.1

The setup and the upper bound on $\mathrm{KL}(\mathbb{P}_{\theta_1}^n || \mathbb{P}_{\theta_0}^n)$ agree for (i) and (ii) but the methods that we use to lower bound $\left|\mathrm{HSIC}_k(\mathbb{P}_{\theta_1}) - \mathrm{HSIC}_k(\mathbb{P}_{\theta_0})\right|$ differ. We structure the proof accordingly and present the overlapping part before we branch out into (i) and (ii). Both parts of the statement rely on Le Cam's method, which we state as Theorem A.2.5 for self-completeness.

To construct the adversarial pair, we consider a class $\mathcal{G}$ of Gaussian distributions over $\mathbb{R}^d$ such that every element $\mathcal{N}(\boldsymbol{\mu}, \Sigma) \in \mathcal{G}$, with

$$\Sigma = \Sigma(i, j, \rho) = \begin{bmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 1 & \rho & \cdots & 0 \\ 0 & \cdots & \rho & 1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{d \times d}, \tag{4.3}$$

and (fixed) $i = d_1$, $j = d_1 + 1$, $\rho \in (-1, 1)$. In other words, $\Sigma$ is essentially the $d$-dimensional matrix $\boldsymbol{I}_d$ except for the $(i, j)$ and $(j, i)$ entry; both entries are identical to $\rho$, and they specify the correlation of the respective coordinates. This family of distributions is indexed by a tuple $(\boldsymbol{\mu}, \rho) \in \mathbb{R}^d \times (-1, 1) =: \mathcal{A}$ and, for $a \in \mathcal{A}$, we write $\mathbb{P}_a$ for the associated distribution. To bring ourselves into the setting of Theorem A.2.5, we fix $n \in \mathbb{N}_{>0}$, choose $\mathcal{X} = \left(\mathbb{R}^d\right)^n$, set $\Theta = \{\theta_a := \mathrm{HSIC}_k(\mathbb{P}_a) : a \in \mathcal{A}\}$, $\mathcal{P}_\Theta = \{\mathbb{P}_a^n : a \in \mathcal{A}\} = \{\mathbb{P}_a^n : \theta_a \in \Theta\}$, and use the metric $(x, y) \mapsto |x - y|$ for $x, y \in \mathbb{R}$. Hence, the data $D \sim \mathbb{P}_\theta \in \mathcal{P}_\Theta$. For brevity, let $F : \mathcal{A} \to \mathbb{R}$ stand for $a \mapsto \mathrm{HSIC}_k(\mathbb{P}_a)$, and let $\hat{F}_n$ stand for the corresponding estimator based on $n$ samples.

As $\mathcal{G} \subseteq \mathcal{P}$, it holds for every positive $s$ that

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}^n \left\{ \left| \mathrm{HSIC}_k(\mathbb{P}) - \hat{F}_n \right| \geq s \right\} \geq \sup_{\mathbb{P} \in \mathcal{G}} \mathbb{P}^n \left\{ \left| \mathrm{HSIC}_k(\mathbb{P}) - \hat{F}_n \right| \geq s \right\}.$$

Let $\mathbb{P}_{\theta_0} = \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$ and $\mathbb{P}_{\theta_1} = \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ with

$$\boldsymbol{\mu}_0 = \mathbf{0}_d \in \mathbb{R}^d, \qquad \Sigma_0 = \Sigma(d_1, d_1 + 1, 0) = \mathbf{I}_d \in \mathbb{R}^{d \times d},$$
$$\boldsymbol{\mu}_1 = \frac{1}{\sqrt{dn}} \mathbf{1}_d \in \mathbb{R}^d, \qquad \Sigma_1 = \Sigma(d_1, d_1 + 1, \rho_n) \in \mathbb{R}^{d \times d},$$

where $\rho_n \in (-1, 1)$ will be chosen appropriately later.[3] We now proceed to upper bound $\mathrm{KL}\left(\mathbb{P}_{\theta_1}^n || \mathbb{P}_{\theta_0}^n\right)$ and lower bound $|F(\theta_1) - F(\theta_0)|$.

**Upper bound for KL divergence** Lemma 4.3.1 implies that with $\rho_n^2 = \frac{1}{n}$, one has the bound

$$\mathrm{KL}\left(\mathbb{P}_{\theta_1}^n || \mathbb{P}_{\theta_0}^n\right) \leq \alpha := \frac{5}{4}$$

for $n \geq 2 =: n_0$.

**Lower bound (i): Gaussian kernels.** Recall that the considered kernel is $k(\mathbf{x}, \mathbf{y}) = e^{-\frac{\gamma}{2}\|\mathbf{x}-\mathbf{y}\|_{\mathbb{R}^d}^2}$ ($\gamma > 0$). The idea of the proof is as follows.

1. We express $|F(\theta_1) - F(\theta_0)|$ in closed form as a function of $\gamma$, $\rho_n$, and $d$.
2. Using the analytical form obtained in the 1st step, we construct the lower bound.

This is what we detail next.

[3] Notice the dependence of $\mathbb{P}_{\theta_1}$ on $n$.

- **Analytical form of** $|F(\theta_1) - F(\theta_0)|$: Using the fact that $\mathrm{HSIC}_k(\mathbb{P}_{\theta_0}) = 0$, we have that

$$\begin{aligned}\left|F(\theta_1) - \underbrace{F(\theta_0)}_{=0}\right|^2 = F^2(\theta_1) &= \mathrm{HSIC}_k^2\left(\mathbb{P}_{\theta_1}\right) = \mathrm{MMD}_k^2\left(\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1), \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I}_d)\right) \\ &= \left\|\mu_k\left(\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)\right) - \mu_k\left(\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I}_d)\right)\right\|_{\mathcal{H}_k}^2 \\ &= \underbrace{\left\langle \mu_k\left(\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)\right), \mu_k\left(\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)\right)\right\rangle_{\mathcal{H}_k}}_{(i)} + \underbrace{\left\langle \mu_k\left(\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I}_d)\right), \mu_k\left(\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I}_d)\right)\right\rangle_{\mathcal{H}_k}}_{(ii)} \\ &\quad - 2\underbrace{\left\langle \mu_k\left(\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)\right), \mu_k\left(\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I}_d)\right)\right\rangle_{\mathcal{H}_k}}_{(iii)},\end{aligned}$$

  which we compute term-by-term with Lemma 4.2.1, and obtain

$$\begin{aligned}(i) &= |2\gamma\Sigma_1 + \mathbf{I}_d|^{-1/2} = \left[(2\gamma+1)^{d-2}\left((2\gamma+1)^2 - (2\gamma\rho_n)^2\right)\right]^{-1/2}, \\ (ii) &= |2\gamma\mathbf{I}_d + \mathbf{I}_d|^{-1/2} = \left[(2\gamma+1)^d\right]^{-1/2}, \\ (iii) &= |\gamma\Sigma_1 + \gamma\mathbf{I}_d + \mathbf{I}_d|^{-1/2} = \left[(2\gamma+1)^{d-2}\left((2\gamma+1)^2 - (\gamma\rho_n)^2\right)\right]^{-1/2}.\end{aligned}$$

  Combining (i), (ii), and (iii) yields that

$$\begin{aligned}\mathrm{HSIC}_k^2\left(\mathbb{P}_{\theta_1}\right) &= (i) + (ii) - 2(iii) \\ &= \left[(2\gamma+1)^{d-2}\left((2\gamma+1)^2 - (2\gamma\rho_n)^2\right)\right]^{-1/2} + \left[(2\gamma+1)^d\right]^{-1/2} \\ &\quad - 2\left[(2\gamma+1)^{d-2}\left((2\gamma+1)^2 - (\gamma\rho_n)^2\right)\right]^{-1/2}.\end{aligned}$$

- **Lower bound on** $|F(\theta_1) - F(\theta_0)|$: Next, we show that there exists $c > 0$ such that for any $n \in \mathbb{N}_{>0}$ it holds that $\mathrm{HSIC}_k^2\left(\mathbb{P}_{\theta_1}\right) \geq \frac{c}{n}$.

  For $0 < x < \left(1 + \frac{1}{2\gamma}\right)^2$, let us consider the function

$$\begin{aligned}f_c(x) &= \left[(2\gamma+1)^{d-2}\left((2\gamma+1)^2 - 4\gamma^2 x\right)\right]^{-1/2} + \left[(2\gamma+1)^d\right]^{-1/2} \\ &\quad - 2\left[(2\gamma+1)^{d-2}\left((2\gamma+1)^2 - \gamma^2 x\right)\right]^{-1/2} - cx \\ &= \left[z^{d-2}\left(z^2 - 4\gamma^2 x\right)\right]^{-1/2} + \left(z^d\right)^{-1/2} - 2\left[z^{d-2}\left(z^2 - \gamma^2 x\right)\right]^{-1/2} - cx,\end{aligned}$$

with the shorthand $z := 2\gamma + 1$.[4] With this notation, $f_c(1/n) = \mathrm{HSIC}_k^2\left(\mathbb{P}_{\theta_1}\right) - c/n$; our aim is to determine $c > 0$ such that $f_c(1/n) \geq 0$ for any positive integer $n$. To achieve this goal, notice that $f_c(0) = 0$, and

$$\begin{aligned} f_c'(x) &= \frac{2\gamma^2 z^{d-2}}{\left[z^{d-2}\left(z^2 - 4x\gamma^2\right)\right]^{3/2}} - \frac{\gamma^2 z^{d-2}}{\left[z^{d-2}\left(z^2 - x\gamma^2\right)\right]^{3/2}} - c \\ &> \frac{2\gamma^2 z^{d-2}}{\left[z^{d-2}\left(z^2 - x\gamma^2\right)\right]^{3/2}} - \frac{\gamma^2 z^{d-2}}{\left[z^{d-2}\left(z^2 - x\gamma^2\right)\right]^{3/2}} - c = \frac{\gamma^2 z^{d-2}}{\left[z^{d-2}\left(z^2 - x\gamma^2\right)\right]^{3/2}} - c \\ &> \frac{\gamma^2 z^{d-2}}{\left(z^{d-2} z^2\right)^{3/2}} - c = \frac{\gamma^2}{z^2\sqrt{z^d}} - c = \frac{\gamma^2}{(2\gamma+1)^2\sqrt{(2\gamma+1)^d}} - c. \end{aligned}$$

Choosing now $c = \frac{\gamma^2}{(2\gamma+1)^2\sqrt{(2\gamma+1)^d}} > 0$, we have $f_c'(x) \geq 0$, so $f$ is a nondecreasing function. Note that $f_c(1/n) = \mathrm{HSIC}_k^2\left(\mathbb{P}_{\theta_1}\right) - c/n \geq 0$, with $x = 1/n$ and $\left(1 + \frac{1}{2\gamma}\right)^{-2} < 1 \leq n < \infty$. By taking the positive square root, this means that

$$\mathrm{HSIC}_k\left(\mathbb{P}_{\theta_1}\right) \geq \frac{\gamma}{(2\gamma+1)\left((2\gamma+1)^d\right)^{1/4}\sqrt{n}} =: 2s$$

holds for $n \geq 1$, implying that $|F(\theta_1) - F(\theta_0)| \geq 2s > 0$.

We conclude the proof by Theorem A.2.5 using that $\alpha = \frac{5}{4}$ and $\max\left(\frac{e^{-\frac{5}{4}}}{4}, \frac{1-\sqrt{\frac{5}{8}}}{2}\right) = \frac{1-\sqrt{\frac{5}{8}}}{2}$.

**Lower bound (ii): translation-invariant kernels.** Let $\Lambda_k$ denote the spectral measure associated to the kernel $k$ according to (2.4). Using the fact that $\mathrm{HSIC}_k(\mathbb{P}_{\theta_0}) = 0$, we have for $|F(\theta_1) - F(\theta_0)|$ that

$$\begin{aligned} \left|F(\theta_1) - \underbrace{F(\theta_0)}_{=0}\right|^2 &= F^2(\theta_1) = \mathrm{HSIC}_k^2\left(\mathbb{P}_{\theta_1}\right) = \mathrm{MMD}_k^2\left(\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1), \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_0)\right) \\ &\overset{(i)}{=} \left\|\psi_{\mathcal{N}(\boldsymbol{\mu}_1,\Sigma_1)} - \psi_{\mathcal{N}(\boldsymbol{\mu}_1,\Sigma_0)}\right\|_{L^2\left(\mathbb{R}^d,\Lambda_k\right)}^2 \\ &\overset{(ii)}{=} \int_{\mathbb{R}^d} \left|e^{i\langle\boldsymbol{\mu}_1,\boldsymbol{\omega}\rangle - \frac{1}{2}\langle\boldsymbol{\omega},\Sigma_1\boldsymbol{\omega}\rangle} - e^{i\langle\boldsymbol{\mu}_1,\boldsymbol{\omega}\rangle - \frac{1}{2}\langle\boldsymbol{\omega},\Sigma_0\boldsymbol{\omega}\rangle}\right|^2 \mathrm{d}\Lambda_k(\boldsymbol{\omega}) \\ &= \int_{\mathbb{R}^d} \underbrace{\left|e^{i\langle\boldsymbol{\mu}_1,\boldsymbol{\omega}\rangle}\right|^2}_{=1} \left|e^{-\frac{1}{2}\langle\boldsymbol{\omega},\Sigma_1\boldsymbol{\omega}\rangle} - e^{-\frac{1}{2}\langle\boldsymbol{\omega},\Sigma_0\boldsymbol{\omega}\rangle}\right|^2 \mathrm{d}\Lambda_k(\boldsymbol{\omega}) \\ &\overset{(iii)}{\geq} \int_A \left|e^{-\frac{1}{2}\langle\boldsymbol{\omega},\Sigma_1\boldsymbol{\omega}\rangle} - e^{-\frac{1}{2}\langle\boldsymbol{\omega},\Sigma_0\boldsymbol{\omega}\rangle}\right|^2 \mathrm{d}\Lambda_k(\boldsymbol{\omega}) \overset{(iv)}{\geq} \rho_n^2 \underbrace{\int_A \left[h_{\boldsymbol{\omega}}'(0)\right]^2 \mathrm{d}\Lambda_k(\boldsymbol{\omega})}_{=:(2c)^2} \overset{(v)}{=} \underbrace{\frac{(2c)^2}{n}}_{=:(2s)^2>0}, \end{aligned}$$

where $(i)$ holds by Sriperumbudur et al. [2010, Corollary 4(i)] (recalled in Theorem A.2.2). $(ii)$ follows from the analytical form $\psi_{\mathcal{N}(\boldsymbol{\mu},\Sigma)}(\mathbf{t}) = e^{i\langle\boldsymbol{\mu},\boldsymbol{t}\rangle - \frac{1}{2}\langle\mathbf{t},\Sigma\mathbf{t}\rangle}$ of the characteristic function of a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$. For $(iii)$, we define the non-empty open set

$$A = \left\{\boldsymbol{\omega} = (\omega_1, \ldots, \omega_d)^\mathsf{T} \in \mathbb{R}^d \,:\, \omega_{d_1}\omega_{d_1+1} < 0\right\} \subset \mathbb{R}^d,$$

[4] Notice that $(2\gamma+1)^2 - \gamma^2 x > (2\gamma+1)^2 - 4\gamma^2 x$, and $(2\gamma+1)^2 - 4\gamma^2 x > 0 \Leftrightarrow x < \left(1 + \frac{1}{2\gamma}\right)^2$ for a positive $x$; hence the imposed assumption on $x$ ensures that the function $f_c$ is well-defined.

and use that the integration of a non-negative function over a subset yields a lower bound. In $(iv)$, fix $\boldsymbol{\omega} \in A$ and let

$$h_{\boldsymbol{\omega}} : \rho \in [0,1] \mapsto e^{-\frac{1}{2}\langle \boldsymbol{\omega}, \Sigma(d_1, d_1+1, \rho)\boldsymbol{\omega}\rangle} \in (0,1].$$

Note that $h_{\boldsymbol{\omega}}(\rho) = e^{-\frac{1}{2}\left(\boldsymbol{\omega}^\mathsf{T}\boldsymbol{\omega} + 2\rho\omega_{d_1}\omega_{d_1+1}\right)}$; $h_{\boldsymbol{\omega}}$ is continuous on $[0,1]$ and differentiable on $(0,1)$. Hence for any $\rho \in (0,1)$, by the mean value theorem, there exists $\tilde{\rho} \in (0,1)$ such that

$$h_{\boldsymbol{\omega}}(\rho) - h_{\boldsymbol{\omega}}(0) = \rho h'_{\boldsymbol{\omega}}(\tilde{\rho}) \geq \rho \min_{c\in[0,1]} h'_{\boldsymbol{\omega}}(c).$$

We have the first and second derivatives

$$h'_{\boldsymbol{\omega}}(c) = -\omega_{d_1}\omega_{d_1+1} e^{-\frac{1}{2}\left(\boldsymbol{\omega}^\mathsf{T}\boldsymbol{\omega} + 2c\omega_{d_1}\omega_{d_1+1}\right)}, \qquad h''_{\boldsymbol{\omega}}(c) = \omega_{d_1}^2\omega_{d_1+1}^2 e^{-\frac{1}{2}\left(\boldsymbol{\omega}^\mathsf{T}\boldsymbol{\omega} + 2c\omega_{d_1}\omega_{d_1+1}\right)} > 0,$$

which implies that $c \mapsto h'_{\boldsymbol{\omega}}(c)$ is a strictly increasing function of $c$ and that it attains its minimum at $c = 0$, that is,

$$h_{\boldsymbol{\omega}}(\rho) - h_{\boldsymbol{\omega}}(0) \geq \rho h'_{\boldsymbol{\omega}}(0) > 0,$$

where the 2nd inequality holds by $\rho > 0$ and $\boldsymbol{\omega} \in A$. This shows that

$$\left[h_{\boldsymbol{\omega}}(\rho) - h_{\boldsymbol{\omega}}(0)\right]^2 \geq \left[\rho h'_{\boldsymbol{\omega}}(0)\right]^2,$$

and the monotonicity of integration gives $(iv)$. For $(v)$, we note that the kernel $k = \otimes_{m=1}^M k_m$ is characteristic [Szabó and Sriperumbudur, 2018, Theorem 4] (recalled in Theorem A.2.4) as the $(k_m)_{m=1}^M$-s are characteristic. Thus, $\operatorname{supp}(\Lambda_k) = \mathbb{R}^d$ (see Sriperumbudur et al. [2010, Theorem 9]; recalled in Theorem A.2.3), implying that $\Lambda_k(A) > 0$. $(v)$ follows from the positivity of $h'_{\boldsymbol{\omega}}(0)$ (for any $\boldsymbol{\omega} \in A$), from the fact that the integral of a positive function on a set with positive measure is positive, and from our choice of $\rho_n = n^{-1/2}$.

Now, by taking the positive square root, we have

$$|F(\theta_1) - F(\theta_0)| \geq \frac{2c}{\sqrt{n}} =: 2s. \tag{4.4}$$

We conclude by the application of Theorem A.2.5 using that $\alpha = \frac{5}{4}$ and $\max\left(\frac{e^{-\frac{5}{4}}}{4}, \frac{1-\sqrt{\frac{5}{8}}}{2}\right) = \frac{1-\sqrt{\frac{5}{8}}}{2}$.

### 4.3.4. Proof of Corollary 4.2.1

We use the same argument as in the beginning of the proof of Theorem 4.2.1 in Section 4.3.3 but adjust the setting in which we apply Theorem A.2.5. Specifically, we now let $\Theta = \{\theta_a := \tilde{C}^c_{\mathbb{P}_{a},k} : a \in \mathcal{A}\}$, with $\tilde{C}^c_{\mathbb{P},k}$ defined as in (2.16), be the set of centered cross-covariance operators, use the metric $(x,y) \mapsto \|x-y\|_{\mathcal{H}_k}$ for $x, y \in \mathcal{H}_k$, and keep the remaining part of the setup the same. Hence, it remains to lower bound $\left\|\tilde{C}^c_{\mathbb{P}_{\theta_1},k} - \tilde{C}^c_{\mathbb{P}_{\theta_0},k}\right\|_{\mathcal{H}_k}$. By using that HSIC is the RKHS norm of the centered cross-covariance operator, we obtain that

$$\left\|\tilde{C}^c_{\mathbb{P}_{\theta_1},k} - \tilde{C}^c_{\mathbb{P}_{\theta_0},k}\right\|_{\mathcal{H}_k} \overset{(i)}{\geq} \Big|\underbrace{\left\|\tilde{C}^c_{\mathbb{P}_{\theta_1},k}\right\|_{\mathcal{H}_k}}_{=\mathrm{HSIC}_k\left(\mathbb{P}_{\theta_1}\right)} - \underbrace{\left\|\tilde{C}_{\mathbb{P}_{\theta_0},k}\right\|_{\mathcal{H}_k}}_{=\mathrm{HSIC}_k\left(\mathbb{P}_{\theta_0}\right)}\Big| = |F(\theta_1) - F(\theta_0)| \overset{(ii)}{\geq} 2s = \frac{2c}{\sqrt{n}},$$

where $(i)$ holds by the reverse triangle inequality, $F$ is defined as in Section 4.3.3, and $(ii)$ is guaranteed by (4.4) for $c > 0$. We conclude as in the proof of Theorem 4.2.1(ii) to obtain the stated result.

# 5. Nyström Kernel Stein Discrepancy

The content of this chapter is based on the following publication.

- F. Kalinke, Z. Szabó, and B. K. Sriperumbudur. Nyström kernel Stein discrepancy. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 388–396, 2025b. PMLR.

The code replicating all experiments is available at github.com/flopska/nystroem-ksd.

## 5.1. Introduction

In addition to two-sample and independence tests, testing for goodness-of-fit (GoF; Ingster and Suslina 2003, Lehmann and Romano 2021) is also of central importance in data science and statistics, which involves testing $H_0 : \mathbb{Q} = \mathbb{P}$ vs. $H_1 : \mathbb{Q} \neq \mathbb{P}$ based on samples from an unknown sampling distribution $\mathbb{Q}$ and a (fixed known) target distribution $\mathbb{P}$. Classical GoF tests, e.g., the Kolmogorov-Smirnov test [Kolmogorov, 1933, Smirnov, 1948], or the test for normality by Baringhaus and Henze [1988], usually require explicit knowledge of the target distribution. However, in practical applications, the target distribution is frequently only known up to a normalizing constant. Examples include validating the output of Markov Chain Monte Carlo (MCMC) samplers [Welling and Teh, 2011, Bardenet et al., 2014, Korattikara et al., 2014], or assessing deep generative models [Koller and Friedman, 2009, Salakhutdinov, 2015]. In all these examples, one desires a powerful test, even though the normalization constant might be difficult to obtain.

A recent approach to tackle GoF testing, detailed in Section 2.3.3, involves applying a Stein operator [Stein, 1972, Chen, 2021, Anastasiou et al., 2023] to functions in an RKHS and using them as test functions to measure the discrepancy between distributions, referred to as kernel Stein discrepancies (KSD; Chwialkowski et al. 2016, Liu et al. 2016). An empirical estimator of KSD can be used as a test statistic to address the GoF problem. In particular, the Langevin-Stein operator [Gorham and Mackey, 2015, Chwialkowski et al., 2016, Liu et al., 2016, Oates et al., 2017, Gorham and Mackey, 2017] in combination with the kernel mean embedding gives rise to a KSD on the Euclidean space $\mathbb{R}^d$, which we consider in this chapter. As a test statistic, KSD has many desirable properties. In particular, KSD requires only knowledge of the derivative of the score function of the target distribution — implying that KSD is agnostic to the normalization of the target and therefore does not require solving, either analytically or numerically, complex normalization integrals in Bayesian settings. This property has led to its widespread use, e.g., for assessing and improving sample quality [Gorham and Mackey, 2015, Chen et al., 2018, 2019, Futami et al., 2019, Gorham et al., 2020], validating MCMC methods [Coullon et al., 2023], comparing deep generative models [Lim et al., 2019], detecting out-of-distribution inputs [Nalisnick et al., 2019], assessing Bayesian seismic inversion [Izzatullah et al., 2020], modeling counterfactuals [Martinez-Taboada and Kennedy, 2024], and explaining predictions [Sarvmaili et al., 2025]. GoF testing with KSDs has been explored on Euclidean data [Liu et al., 2016, Chwialkowski et al., 2016], discrete data [Yang et al., 2018], point processes [Yang et al., 2019], time-to-event data [Fernandez et al., 2020], graph data [Xu and Reinert, 2021], sequential models [Baum et al., 2023], and functional

data [Wynne et al., 2024]. The KSD statistic has also been extended to the conditional case [Jitkrittum et al., 2020].

Recall from Section 2.3.1 that estimators for the Langevin-Stein operator-based KSD exist. But, the classical V-statistic- [Chwialkowski et al., 2016] and U-statistic-based [Liu et al., 2016] estimators, (2.24) and (2.25), respectively, have a runtime complexity that scales quadratically with the number of samples of the sampling distribution, which limits their deployment to large-scale settings. To address this bottleneck, Chwialkowski et al. [2016] introduced a linear-time statistic that suffers from low statistical power compared to its quadratic-time counterpart. Jitkrittum et al. [2017b] proposed the finite set Stein discrepancy (FSSD), a linear-time approach that replaces the RKHS-norm by the $L^2$-norm approximated by sampling; the sampling can either be random (FSSD-rand) or optimized w.r.t. a power proxy (FSSD-opt). Another approach [Huggins and Mackey, 2018] is employing the random Fourier feature (RFF; Rahimi and Recht 2007, Sriperumbudur and Szabó 2015) method to accelerate the KSD estimation. However, it is known [Chwialkowski et al., 2015, Proposition 1] that the resulting statistic fails to distinguish a large class of measures. Huggins and Mackey [2018] generalize the idea of replacing the RKHS-norm by going from $L^2$-norms to $L^p$ ones, to obtain feature Stein discrepancies. They present an efficient approximation, random feature Stein discrepancies (RFSD), which is a (near-)linear time estimator. However, successful deployment of the method depends on a good choice of parameters, which, while the authors provide guidelines, can be challenging to select and tune in practice.

Our work alleviates these severe bottlenecks. We employ the Nyström method [Williams and Seeger, 2001] to accelerate KSD estimation and show the $\sqrt{n}$-consistency of our proposed estimator. The main technical challenge is that the Stein kernel (induced by the Langevin-Stein operator and the original kernel) is typically unbounded while existing statistical Nyström analysis [Rudi et al., 2015, Chatalic et al., 2022, Sterge and Sriperumbudur, 2022, Kalinke and Szabó, 2023, Chatalic et al., 2025] usually considers bounded kernels. To tackle unbounded kernels, we select a classical sub-Gaussian assumption, which we impose on the feature map associated to the kernel, and show that existing methods of analysis can successfully be extended to handle this novel case. In this sense, our work, besides Della Vecchia et al. [2021], which requires a similar sub-Gaussian condition for analyzing empirical risk minimization on random subspaces, is a first step in analyzing the consistency of the unbounded case in the Nyström setting.

Specifically, we make the following **contributions**.

- We introduce a Nyström-based acceleration of kernel Stein discrepancy. The proposed estimator runs in $O\left(mn + m^3\right)$ time, with $n$ samples and $m \ll n$ Nyström points.
- We prove the $\sqrt{n}$-consistency of our estimator in a classical sub-Gaussian setting, which extends (in a non-trivial fashion) existing results for Nyström-based methods [Rudi et al., 2015, Chatalic et al., 2022, Sterge and Sriperumbudur, 2022, Kalinke and Szabó, 2023] focusing on bounded kernels.
- We perform an extensive suite of experiments to demonstrate the applicability of the proposed method. Our proposed approach achieves competitive results throughout all experiments.

This chapter is structured as follows. The proposed Nyström-based acceleration with guarantees is in Section 5.2. Section 5.3 details our experiments and Section 5.4 discusses limitations. All proofs are in Section 5.5.

## 5.2. Proposed Nyström KSD

To enable the efficient estimation of (2.22), we propose a Nyström technique-based estimator in Section 5.2.1 and an accelerated wild bootstrap test in Section 5.2.2. In Section 5.2.3, our consistency results are collected.

### 5.2.1. The Nyström KSD estimator

We consider a subsample $\tilde{\mathbb{Q}}_m = \{\{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_m\}\}$ of size $m$ (sampled with replacement), the so-called Nyström sample, of the original sample $\hat{\mathbb{Q}}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$; the tilde indicates a relabeling. The best approximation of $S_p(\mathbb{Q})$ in RKHS-norm-sense, when using $m$ Nyström samples, can be obtained by considering the orthogonal projection of $\mathbb{E}_{X\sim\mathbb{Q}} h_p(\cdot, X)$ onto $\mathcal{H}_{h_p,m} := \operatorname{span}\left\{h_p(\cdot, \tilde{\mathbf{x}}_i) \mid i \in [m]\right\} \subset \mathcal{H}_{h_p}$, with feature map $h_p(\cdot, \tilde{\mathbf{x}}_i)$ and associated kernel $h_p$ defined in (2.21). As $\mathbb{Q}$ is unknown in practice and only available via samples $\hat{\mathbb{Q}}_n \sim \mathbb{Q}^n$, we consider the orthogonal projection of $\mathbb{E}_{X\sim\hat{\mathbb{Q}}_n} h_p(\cdot, X)$ onto $\mathcal{H}_{h_p,m}$ instead. In other words, we aim to find the weights $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^m \in \mathbb{R}^m$ that correspond to the minimum norm solution of the cost function

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^m} \left\| \underbrace{\frac{1}{n}\sum_{i=1}^n h_p(\cdot, \mathbf{x}_i)}_{=\mathbb{E}_{X\sim\hat{\mathbb{Q}}_n} h_p(\cdot,X)} - \sum_{i=1}^m \alpha_i h_p(\cdot, \tilde{\mathbf{x}}_i) \right\|_{\mathcal{H}_{h_p}}, \tag{5.1}$$

which gives rise to the squared KSD estimator[1]

$$\tilde{S}_p^2\left(\hat{\mathbb{Q}}_n\right) := \left\| \sum_{i=1}^m \alpha_i h_p(\cdot, \tilde{\mathbf{x}}_i) \right\|_{\mathcal{H}_{h_p,m}}^2 = \left\| P_{\mathcal{H}_{h_p,m}} \mathbb{E}_{X\sim\hat{\mathbb{Q}}_n} h_p(\cdot, X) \right\|_{\mathcal{H}_{h_p,m}}^2. \tag{5.2}$$

**Lemma 5.2.1** (Nyström-KSD Estimator)**.** *The squared KSD estimator* (5.2) *takes the form*

$$\tilde{S}_p^2\left(\hat{\mathbb{Q}}_n\right) = \boldsymbol{\beta}_p^\mathsf{T} \mathbf{K}_{h_p,m,m}^- \boldsymbol{\beta}_p, \tag{5.3}$$

*with* $\boldsymbol{\beta}_p = \frac{1}{n}\mathbf{K}_{h_p,m,n}\mathbf{1}_n \in \mathbb{R}^m$*, Gram matrix* $\mathbf{K}_{h_p,m,m} = \left[h_p\left(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j\right)\right]_{i,j=1}^m \in \mathbb{R}^{m\times m}$*, and* $\mathbf{K}_{h_p,m,n} = \left[h_p\left(\tilde{\mathbf{x}}_i, \mathbf{x}_j\right)\right]_{i,j=1}^{m,n} \in \mathbb{R}^{m\times n}$*.*

**Remark 5.2.1.**

(a) ***Runtime complexity.*** *The computation of* (5.3) *consists of calculating* $\boldsymbol{\beta}_p$*, pseudo-inverting* $\boldsymbol{K}_{h_p,m,m}$*, and obtaining the quadratic form* $\boldsymbol{\beta}_p^\mathsf{T}\mathbf{K}_{h_p,m,m}^-\boldsymbol{\beta}_p$*. The calculation of* $\boldsymbol{\beta}_p$ *requires* $O(mn)$ *operations, due to the multiplication of an* $m \times n$ *matrix with a vector of length* $n$*. Inverting the* $m \times m$ *matrix* $\boldsymbol{K}_{h_p,m,m}$ *costs* $O(m^3)$*,*[3] *dominating the cost of* $O(m^2)$ *needed for the computation of* $\boldsymbol{K}_{h_p,m,m}$*. The quadratic form* $\boldsymbol{\beta}_p^\mathsf{T}\mathbf{K}_{h_p,m,m}^-\boldsymbol{\beta}_p$ *has a computational cost of* $O\left(m^2\right)$*. Hence,* (5.3) *can be computed in* $O\left(mn + m^3\right)$*, which means that for* $m = o\left(n^{2/3}\right)$*, our proposed Nyström-KSD estimator guarantees an asymptotic speedup.*

[1] $\tilde{S}_p^2\left(\hat{\mathbb{Q}}_n\right)$ indicates dependence on $\hat{\mathbb{Q}}_n$.

(b) ***Comparison of*** (2.24) ***and*** (5.3)**.** *The Nyström estimator* (5.3) *recovers the V-statistic-based estimator* (2.24) *when no subsampling is performed and provided that* $\mathbf{K}_{h_p,n,n}$ *is invertible.*

(c) ***Comparison to Chatalic et al. [2022].*** *We note that the estimator* (5.3) *corresponds precisely to Chatalic et al. [2022, (5)]. We consider the analysis of this known estimator in the case of unbounded feature maps—which arise in the KSD setting—as one of our core contributions, which we detail in Section 5.2.3.*

### 5.2.2. Nyström bootstrap testing

In this section, we discuss how one can use (5.3) for accelerated goodness-of-fit testing. We recall that the goal of goodness-of-fit testing is to test $H_0 : \mathbb{Q} = \mathbb{P}$ versus $H_1 : \mathbb{Q} \neq \mathbb{P}$, given samples $\hat{\mathbb{Q}}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and target distribution $\mathbb{P}$. Recall that KSD relies on score functions ($\nabla_{\mathbf{x}}[\log p(\mathbf{x})]$); hence knowing $\mathbb{P}$ up to a multiplicative constant is enough. To use the Nyström-based estimator (5.3) for goodness-of-fit testing, we propose to obtain its null distribution by a Nyström-based bootstrap procedure. Our method builds on the existing test for the V-statistic-based KSD, detailed in Chwialkowski et al. [2016, Section 2.2], which we quote in the following. Define the bootstrapped statistic by

$$B_n = \frac{1}{n^2} \sum_{i,j=1}^{n} w_i w_j h_p\left(\mathbf{x}_i, \mathbf{x}_j\right), \tag{5.4}$$

with $w_i \in \{-1, 1\}$ an auxiliary Markov chain defined by

$$w_i = \mathbb{1}_{(U_i > 0.5)} w_{i-1} - \mathbb{1}_{(U_i \leq 0.5)} w_{i-1}, \tag{5.5}$$

where $U_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(0, 1)$, that is, $w_i$ changes sign with probability 0.5. The test procedure is as follows.

1. Calculate the test statistic (2.24).
2. Obtain $D$ wild bootstrap samples $\{B_{n,i}\}_{i=1}^{D}$ with (5.4) and estimate the $1 - \alpha$ empirical quantile of these samples.
3. Reject the null hypothesis if the test statistic (2.24) exceeds the quantile.

(5.4) requires $O(n^2)$ computations, which yields a total cost of $O(Dn^2)$ for obtaining $D$ bootstrap samples, rendering large-scale goodness-of-fit tests unpractical.

We propose the Nyström-based bootstrap

$$B_n^{\text{Nys}} = \frac{1}{n^2} \mathbf{w}^{\mathsf{T}} \mathbf{K}_{h_p,n,m} \mathbf{K}^{-}_{h_p,m,m} \mathbf{K}_{h_p,m,n} \mathbf{w}, \tag{5.6}$$

with $\mathbf{w} = (w_i)_{i=1}^{n} \in \mathbb{R}^n$ collecting the $w_i$-s ($i \in [n]$) defined in (5.5); $\mathbf{K}_{h_p,n,m}$ ($= \mathbf{K}^{\mathsf{T}}_{h_p,m,n}$) and $\mathbf{K}_{h_p,m,m}$ are defined as in Lemma 5.2.1. The approximation is based on the fact [Williams and Seeger, 2001] that $\mathbf{K}_{h_p,n,m} \mathbf{K}^{-}_{h_p,m,m} \mathbf{K}_{h_p,m,n}$ is a low-rank approximation of $\mathbf{K}_{h_p,n,n}$, that is, $\mathbf{K}_{h_p,n,m} \mathbf{K}^{-}_{h_p,m,m} \mathbf{K}_{h_p,m,n} \approx \mathbf{K}_{h_p,n,n}$. Hence, our proposed procedure (5.6) approximates (5.4) but reduces the cost from $O(n^2)$ to $O(nm + m^3)$ if one computes from left to right (also refer to Remark 5.2.1(a)); in the case of $m = o(n^{2/3})$ this guarantees an asymptotic speedup. We obtain a total cost of $O(D(nm + m^3))$ for obtaining the wild bootstrap samples. This acceleration allows KSD-based goodness-of-fit tests to be applied on large data sets.

### 5.2.3. Guarantees

This section is dedicated to the statistical behavior of the proposed Nyström-KSD estimator (5.3).

The existing analysis of Nyström estimators [Rudi et al., 2015, Chatalic et al., 2022, Sterge and Sriperumbudur, 2022, Kalinke and Szabó, 2023] considers bounded kernels only. Indeed, if one has that $\sup_{\mathbf{x}\in\mathbb{R}^d} \left\|h_p\left(\cdot,\mathbf{x}\right)\right\|_{\mathcal{H}_{h_p}} < \infty$, the consistency of (5.3) is implied by Chatalic et al. [2022, Theorem 4.1], which we include here for convenience of comparison. In the following, we denote the randomness in the choice of Nyström samples by $(i_j)_{j=1}^m \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([n]) =: \Lambda$, which means that $\tilde{\mathbf{x}}_j = \mathbf{x}_{i_j}$ $(j \in [m])$.

**Theorem 5.2.1** (Bounded case)**.** *Assume the setting of Lemma 5.2.1, $C_{\mathbb{Q},h_p} \neq 0$, $m \geq 4$ Nyström samples, and a bounded Stein feature map ($\sup_{\mathbf{x}\in\mathbb{R}^d} \left\|h_p\left(\cdot,\mathbf{x}\right)\right\|_{\mathcal{H}_{h_p}} =: K < \infty$). Then, for any $\delta \in (0,1)$, it holds with $(\mathbb{Q}^n \otimes \Lambda^m)$-probability of at least $1-\delta$ that*

$$\left|S_p(\mathbb{Q}) - \tilde{S}_p\left(\hat{\mathbb{Q}}_n\right)\right| \leq \frac{c_1}{\sqrt{n}} + \frac{c_2}{m} + \frac{c_3\sqrt{\log\frac{m}{\delta}}}{m}\sqrt{\mathcal{N}_{\mathbb{Q},h_p}\left(\frac{12K^2\log\frac{m}{\delta}}{m}\right)},$$

*when $m \geq \max\left(67, 12K^2\left\|C_{\mathbb{Q},h_p}\right\|_{\mathrm{op}}^{-1}\right)\log(m/\delta)$, where $c_1$, $c_2$, and $c_3$ are positive constants.*

However, in practice, the feature map of KSD is typically unbounded and Theorem 5.2.1 is not applicable, as it is illustrated in the following example with the frequently-used Gaussian kernel.

**Example 5.2.1** (KSD yields unbounded kernel)**.** *Consider univariate data ($d=1$), unnormalized target density $p(x) = e^{-x^2/2}$ (corresponding to $\mathbb{P} = \mathcal{N}(0,1)$), and (i) the RBF kernel $k(x,y) = \exp\left(-\gamma(x-y)^2\right)$ with $\gamma > 0$, or (ii) the IMQ kernel $k(x,y) = \left(c^2 + (x-y)^2\right)^{-\beta}$ with $\beta, c > 0$. By using (2.21), direct calculation yields (i) $\left\|\xi_p(\cdot,x)\right\|_{\mathcal{H}_k}^2 = x^2 + 2\gamma \overset{x\to\infty}{\to} \infty$ in the first, and (ii) $\left\|\xi_p(\cdot,x)\right\|_{\mathcal{H}_k}^2 = x^2c^{2\beta} - 2\beta c^{2(\beta-1)} \overset{x\to\infty}{\to} \infty$ in the second case.*

**Remark 5.2.2.** *In fact, a more general result holds: If one considers a bounded continuously differentiable translation-invariant kernel $k$, the induced Stein kernel is only bounded provided that the target density $p(\mathbf{x})$ has tails that are no thinner than $e^{-\sum_{i=1}^d |x_i|}$ [Hagrass et al., 2025, Remark 2], which clearly rules out Gaussian targets.*

For analyzing the setting of unbounded feature maps, we make the following assumption.[2]

**Assumption 5.2.1.** *The centered Stein feature map $\bar{h}_p\left(\cdot,X\right) = h_p\left(\cdot,X\right) - \mathbb{E}_{X\sim\mathbb{Q}}h_p\left(\cdot,X\right)$ with the sampling distribution $\mathbb{Q} \in \mathcal{M}_1^+\left(\mathbb{R}^d\right)$ is sub-Gaussian, that is,*

$$\left\|\left\langle \bar{h}_p\left(\cdot,X\right), u\right\rangle_{\mathcal{H}_{h_p}}\right\|_{\psi_2} \lesssim \left\|\left\langle \bar{h}_p\left(\cdot,X\right), u\right\rangle_{\mathcal{H}_{h_p}}\right\|_{L^2(\mathbb{Q})} < \infty$$

*holds for all $u \in \mathcal{H}_{h_p}$, with a $u$-independent absolute constant in $\lesssim$.*

[2] We specialize Definition 2 by Koltchinskii and Lounici [2017] stated for Banach spaces to (reproducing kernel) Hilbert spaces by using the Riesz representation theorem.

**Example 5.2.2** (Applicability of Assumption 5.2.1)**.** *In the simple case $d = 1$, $k(x, y) = xy$ ($\mathcal{H}_k = \mathbb{R}$), and target measure $\mathbb{P} = \mathcal{N}(0, 1)$, Assumption 5.2.1 is satisfied, for instance, for $\mathbb{Q} = \text{Unif}\left(-\sqrt{3}, \sqrt{3}\right)$. The details are as follows. From (2.20), $\xi_p(\cdot, x) = h_p(\cdot, x) = 1 - x^2$ ($x \in \mathbb{R}$). We note that $\mathbb{E}_{X\sim\mathbb{Q}} h_p(\cdot, X) = 0$ implies that $\bar{h}_p(\cdot, x) = h_p(\cdot, x)$ and we obtain $\left\|\left\langle h_p(\cdot, X), u\right\rangle_{\mathbb{R}}\right\|_{\psi_2} = |u| \left\|1 - X^2\right\|_{\psi_2} \overset{(a)}{\leq} |u|c_1 \overset{(b)}{=} |u|c_1c_2 \left\|1 - X^2\right\|_{L_2(\mathbb{Q})} \lesssim \left\|\left\langle h_p(\cdot, X), u\right\rangle_{\mathbb{R}}\right\|_{L^2(\mathbb{Q})}$. The boundedness of $X$ implies the sub-Gaussianity (in the real-valued sense) of $1 - X^2$ in (a); hence, $\left\|1 - X^2\right\|_{\psi_2} \leq c_1$. In (b), we let $c_2 = \left\|1 - X^2\right\|^{-1}_{L^2(\mathbb{Q})}$.*

We elaborate further on Assumption 5.2.1 in Remark 5.2.3(c), after we state our following main result.

**Theorem 5.2.2** (Consistency of Nyström-KSD)**.** *Let Assumption 5.2.1 hold, $C_{\mathbb{Q},\bar{h}_p} \neq 0$, and assume the setting of Lemma 5.2.1. Then, for any $\delta \in (0, 1)$ with $(\mathbb{Q}^n \otimes \Lambda^m)$-probability of at least $1 - \delta$ it holds that*

$$\left|S_p(\mathbb{Q}) - \tilde{S}_p\left(\hat{\mathbb{Q}}_n\right)\right| \lesssim \frac{\sqrt{\text{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)\log(6/\delta)}}{n} + \sqrt{\frac{\text{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)\log(6/\delta)}{n}} + \frac{\sqrt{\text{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)\log(12n/\delta)\log(12/\delta)}}{m}\sqrt{\mathcal{N}_{\mathbb{Q},\bar{h}_p}\left(\frac{c\,\text{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}{m}\right)}$$

*when $m \gtrsim \max\left\{\left\|C_{\mathbb{Q},\bar{h}_p}\right\|^{-1}_{\text{op}} \text{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right), \log(12/\delta)\right\}$, where $c > 1$ is a constant.*

To interpret the consistency guarantee of Theorem 5.2.2, we consider the three terms on the r.h.s. w.r.t. the magnitude of $m$. The first two terms converge with $O\left(n^{-1/2}\right)$, independent of the choice of $m$. By using the universal upper bound $\mathcal{N}_{\mathbb{Q},\bar{h}_p}\left(\frac{c\,\text{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}{m}\right) \lesssim m$ on the effective dimension, the last term reveals that an overall rate of $O\left(n^{-1/2}\right)$ can only be achieved with further assumptions regarding the rate of decay of the effective dimension if one also requires $m = o\left(n^{2/3}\right)$ — as is necessary for a speed-up, see Remark 5.2.1(a). Indeed, the rate of decay of the effective dimension can be linked to the rate of decay of the eigenvalues of the covariance operator [Della Vecchia et al., 2021, Proposition 4, 5], which is known to frequently decay exponentially, or, at least, polynomially. In this sense, the last term acts as a balance, which takes the characteristics of the data and of the kernel into account.

The next corollary shows that an overall rate of $O\left(n^{-1/2}\right)$ can be achieved, depending on the properties of the covariance operator.

**Corollary 5.2.1.** *In the setting of Theorem 5.2.2, assume that the spectrum of the covariance operator $C_{\mathbb{Q},\bar{h}_p}$ decays either (i) polynomially, implying that $\mathcal{N}_{\mathbb{Q},\bar{h}_p}(\lambda) \lesssim \lambda^{-\gamma}$ for some $\gamma \in (0, 1]$, or (ii) exponentially, implying that, $\mathcal{N}_{\mathbb{Q},\bar{h}_p}(\lambda) \lesssim \log(1 + \frac{c_1}{\lambda})$ for some $c_1 > 0$. Then it holds that*

$$\left|S_p(\mathbb{Q}) - \tilde{S}_p\left(\hat{\mathbb{Q}}_n\right)\right| = O_{\mathbb{Q}^n\otimes\Lambda^m}\left(\frac{1}{\sqrt{n}}\right),$$

*assuming that the number of Nyström points satisfies (i) $m \gtrsim n^{\frac{1}{2-\gamma}} \log^{\frac{1}{2-\gamma}}(12n/\delta) \log^{\frac{1}{2-\gamma}}(12/\delta)$ in the first case, or (ii) $m \gtrsim \sqrt{n}\left(\log\left(1 + \frac{c_1 n}{c\,\text{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}\right)\log(12n/\delta)\log(12/\delta)\right)^{1/2}$ in the second case.*

To interpret these rates—see Remark 5.2.3(d)—, we obtain the (matching) $\sqrt{n}$-consistency of the quadratic-time estimator (2.24) in our following result.

**Theorem 5.2.3** (Consistency of KSD)**.** *Assume that* $\left\|\left\|h_p(\cdot, X)\right\|_{\mathcal{H}_{h_p}}\right\|_{\psi_2} < \infty$ *and let* $\hat{\mathbb{Q}}_n = \{X_1, \ldots, X_n\}$, *where* $\{X_i\}_{i\in[n]} \overset{i.i.d.}{\sim} \mathbb{Q}$. *Then it holds that*

$$\left|S_p(\mathbb{Q}) - S_p\left(\hat{\mathbb{Q}}_n\right)\right| = O_{\mathbb{Q}^n}\left(\frac{1}{\sqrt{n}}\right).$$

The following example illustrates that, in some cases, the assumption $\left\|\left\|h_p(\cdot, X)\right\|_{\mathcal{H}_{h_p}}\right\|_{\psi_2} < \infty$ can be verified analytically.

**Example 5.2.3** (Assumption $\left\|\left\|h_p(\cdot, X)\right\|_{\mathcal{H}_{h_p}}\right\|_{\psi_2} < \infty$)**.** *Assume that* $d = 1$, $k = \exp\left(-\gamma(x-y)^2\right)$ $(\gamma > 0)$, *target measure* $\mathbb{P} = \mathcal{N}(0,1)$, *and samples* $X, X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{Q}$ *with* $\|X\|_{\psi_2} < \infty$. *Then*

$$\left\|\left\|h_p(\cdot, X)\right\|_{\mathcal{H}_{h_p}}\right\|_{\psi_2}^2 \overset{(a)}{=} \left\|h_p(X, X)\right\|_{\psi_1} \overset{(b)}{=} \left\|X^2 + 2\gamma\right\|_{\psi_1} \overset{(c)}{\leq} \left\|X^2\right\|_{\psi_1} + \|2\gamma\|_{\psi_2} \overset{(d)}{=} \|X\|_{\psi_2} + \frac{2\gamma}{\sqrt{\log 2}} < \infty,$$

*with the following details. Lemma A.3.2(iv) and the reproducing property yield (a). (b) follows from the explicit form of* $h_p$ *given in Example 5.2.1(i). The triangle inequality gives (c) and (d) follows from the definition of the* $\psi_2$*-norm using that* $2\gamma$ *is non-random.*

*In this setting, similar computations using Example 5.2.1(ii) show that the assumption is also satisfied with the IMQ kernel.*

A few remarks are in order.

**Remark 5.2.3.**

(a) ***Runtime benefit.*** *Recall that — see Remark 5.2.1(a) —, our proposed Nyström estimator* (5.3) *requires* $m = o\left(n^{2/3}\right)$ *Nyström samples to achieve a speed-up. Hence, in the case of polynomial decay, an asymptotic speed-up with a statistical accuracy that matches the quadratic-time estimator* (2.24) *is guaranteed for* $\gamma < 1/2$*; in the case of exponential decay, large enough* $n$ *always suffices.*

(b) ***Comparison of Theorem 5.2.1 and Theorem 5.2.2.*** *Recall that both theorems target precisely the same estimators, Chatalic et al. [2022, (5)] and* (5.3)*, respectively. We note that in the finite-dimensional case, every bounded random variable is also sub-Gaussian. This property does not carry over to sub-Gaussianity in the infinite-dimensional case; see the remark after Della Vecchia et al. [2021, Definition 1]. In this sense, the assumptions of both statements are not directly comparable. Still, the takeaway of both results—with these different sets of conditions—is the same.*

(c) ***Sub-Gaussian assumption.*** *Key to the proof of Theorem 5.2.2 is having an adequate notion of non-boundedness of the feature map. One approach—common for controlling unbounded real-valued random variables— is to impose a sub-Gaussian assumption. In Hilbert spaces, various definitions of sub-Gaussian behavior have been investigated [Talagrand, 1987, Fukuda, 1990, Antonini, 1997]; see Giorgobiani et al. [2020] for a recent survey. Among the definitions of sub-Gaussianity, we carefully*

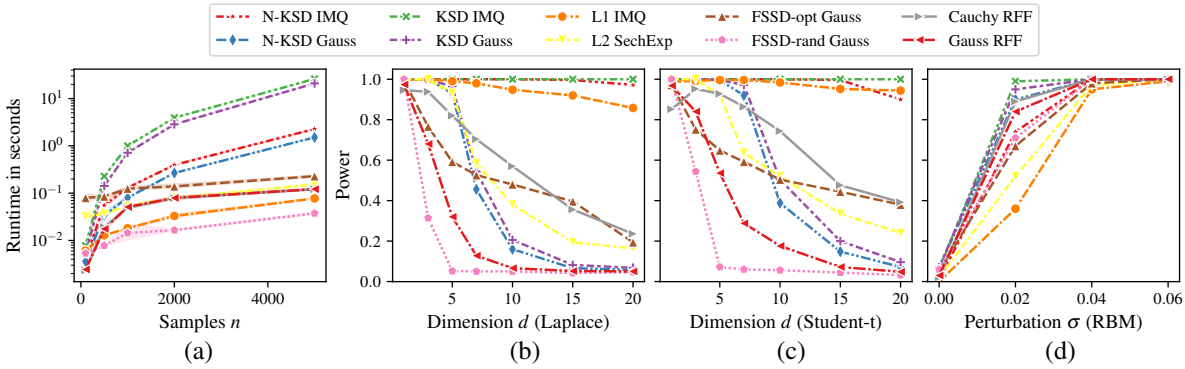


**Figure 5.1.:** Comparison of goodness-of-fit tests w.r.t. their runtime and their power.

*selected Koltchinskii and Lounici [2017, Def. 2].*[3] *Specifically, this assumption allows us to derive our key Lemma 5.5.1 and Lemma 5.5.3. The former is similar to Rudi et al. [2015, Lemma 6], which is typically employed for Nyström analysis in the bounded case [Chatalic et al., 2022, Sterge and Sriperumbudur, 2022, Kalinke and Szabó, 2023], but our result applies to the sub-Gaussian setting. The main technical challenge we resolve is transforming our setting to a form in which existing concentration results can be leveraged. Especially the case of* $\mathbb{P} \neq \mathbb{Q}$ *requires special care, which we tackle by systematically using the centered covariance operator* $C_{\mathbb{Q},\bar{h}_p}$*; we refer to the respective proof for details.*[4] *The latter, Lemma 5.5.3, intuitively states that norms of sub-Gaussian vectors whitened by* $C^{-1/2}_{\mathbb{Q},\bar{h}_p,\lambda}$ *inherit the sub-Gaussian property. Together, these lemmas open the door to proving Theorem 5.2.2.*

(d) ***Comparison of Theorem 5.2.2 and Theorem 5.2.3.*** *With the weaker condition on the RKHS norm of the feature map, that is,* $\left\| \left\| h_p(\cdot, X) \right\|_{\mathcal{H}_{h_p}} \right\|_{\psi_2} < \infty$ *(implied by Assumption 5.2.1, see Lemma 5.5.3), Theorem 5.2.3 shows that the quadratic-time estimator* (2.24) *converges with rate* $O\left(n^{-1/2}\right)$*. Our Nyström result, Theorem 5.2.2 with Corollary 5.2.1, shows that a matching rate can be achieved (given an appropriate decay of the effective dimension) with* $m = \tilde{\Theta}\left(\sqrt{n}\right)$*; this choice of* $m$ *satisfies* $m = o\left(n^{2/3}\right)$ *and thus implies an asymptotic speedup by (a).*

(e) ***General KSD framework.*** *We note that our results also hold in the general KSD framework [Hagrass et al., 2025] but we present them on* $\mathbb{R}^d$*, which one arguably most frequently encounters in practice, to simplify exposition.*

## 5.3. Experiments

We verify the viability of our proposed method, abbreviated as N-KSD in this section, by comparing its runtime and its power to existing methods: the quadratic-time KSD [Liu et al., 2016, Chwialkowski et al., 2016], the linear-time goodness-of-fit test finite set Stein discrepancy (FSSD; Jitkrittum et al. 2017b), RFF-based KSD approximations [Huggins and Mackey, 2018], and the linear-time goodness-of-fit test using random feature Stein discrepancy (L1 IMQ, L2 SechExp; Huggins and Mackey 2018). For FSSD, we consider randomized test locations (FSSD-rand) and optimized test locations (FSSD-opt); the optimality is meant w.r.t. a power proxy detailed in the cited work. For all competitors, we use the settings

[3] The condition is also referred to as *sub-Gaussian in Fukuda's sense* [Giorgobiani et al., 2020, Def. 1].

[4] We note that an analysis of the centered setting is also challenging in the bounded case; for instance, Sterge and Sriperumbudur [2022] tackle the resulting difficulties (in case of kernel PCA) with U-statistics, of which our method is independent.

and implementations provided by the respective authors. We use the well-known Gaussian kernel $k(\mathbf{x}, \mathbf{y}) = \exp\left(-\gamma \|\mathbf{x} - \mathbf{y}\|^2_{\mathbb{R}^d}\right)$ ($\gamma > 0$) with the median heuristic [Garreau et al., 2018], and the IMQ kernel $k(\mathbf{x}, \mathbf{y}) = \left(c^2 + \|\mathbf{x} - \mathbf{y}\|^2_{\mathbb{R}^d}\right)^{-\beta}$ [Gorham and Mackey, 2017], with the choices of $\beta, c > 0$ detailed in the respective experiment description. To approximate the null distribution of N-KSD, we perform a bootstrap with (5.6), setting $D = 500$. To allow an easy comparison, our experiments in Section 5.3.1 replicate goodness-of-fit testing experiments from Chwialkowski et al. [2016], Jitkrittum et al. [2017b] and Huggins and Mackey [2018]. Additionally, in Section 5.3.2, we investigate the trade-off between power and runtime of the tested approaches. Section 5.3.3 shows the impact of the size of the Nyström sample. We ran all experiments on a PC with Ubuntu 20.04, 124GB RAM, and 32 cores with 2GHz each.

### 5.3.1. Goodness-of-fit testing benchmarks

This section replicates the GoF testing experiments from Chwialkowski et al. [2016], Jitkrittum et al. [2017b] and Huggins and Mackey [2018], taking our proposed method into account. The details are as follows.

**Runtime.** We set $m = 4\sqrt{n}$ for N-KSD to match the settings in our later experiments. As per recommendation, we fix the number of test locations $J = 10$ for L1 IMQ, L2 SechExp, Cauchy RFF, Gauss RFF, and both FSSD methods. The data is randomly generated with $d = 10$ dimensions. We note that the dimensionality enters the complexity only through the kernel evaluation; the dependence is linear in our case. The runtime, see Figure 5.1(a) for the average over 10 repetitions (the error bars indicate the estimated 95% quantile), behaves as predicted by the complexity analysis. The proposed approach runs orders of magnitudes faster than the quadratic-time KSD estimator (2.24). From $n = 1500$, all (near-)linear-time approaches are faster (excluding FSSD-opt, which has a relatively large fixed cost). Still, N-KSD achieves competitive runtime results even for $n = 5000$.

**Laplace vs. standard normal.** We fix the target distribution $\mathbb{P} = \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ and obtain $n = 1000$ samples from the alternative $\mathbb{Q} = \mathrm{Lap}\left(0, \frac{1}{\sqrt{2}}\right)^d$, a product of $d$ Laplace distributions. We test $H_0 : \mathbb{Q} = \mathbb{P}$ vs. $H_1 : \mathbb{Q} \neq \mathbb{P}$ with a level of $\alpha = 0.05$. We set the kernel parameters $c$ and $\beta$ for KSD IMQ and N-KSD IMQ as per the recommendation for L1 IMQ in the corresponding experiment by Huggins and Mackey [2018]. Figure 5.1(b) reports the power (obtained over 500 draws of the data) of the different approaches. KSD Gauss and its approximation N-KSD Gauss perform similarly but their power diminishes from $d = 3$. KSD IMQ achieves full power for all tested dimensions and performs best overall. N-KSD IMQ ($m = 4\sqrt{n}$) achieves comparable results, with a small decline from $D = 15$. Our proposed method outperforms all existing KSD accelerations.

**Student-t vs. standard normal.** The setup is similar to that of the previous experiment, but we consider samples from $\mathbb{Q}$ a multivariate student-t distribution with 5 degrees of freedom, set $n = 2000$, and repeat the experiment 250 times to estimate the power. We show the results in Figure 5.1(c). All approaches employing the Gaussian kernel quickly loose in power; all techniques utilizing the IMQ kernel, including N-KSD IMQ, achieve comparably high power throughout.

**Restricted Boltzmann machine (RBM).** Similar to Liu et al. [2016], Jitkrittum et al. [2017b], we consider the case where the target $\mathbb{P}$ is the non-normalized density of an RBM with 50 visible and 40 hidden dimensions; the samples $\hat{\mathbb{Q}}_n$ are obtained from the same RBM perturbed by independent Gaussian noise with variance $\sigma^2$. For $\sigma^2 = 0$, $H_0 : \mathbb{Q} = \mathbb{P}$ holds, and for $\sigma^2 > 0$, implying that the alternative $H_1 : \mathbb{Q} \neq \mathbb{P}$ holds, the goal is to detect that the $n = 1000$ samples come from a forged RBM.

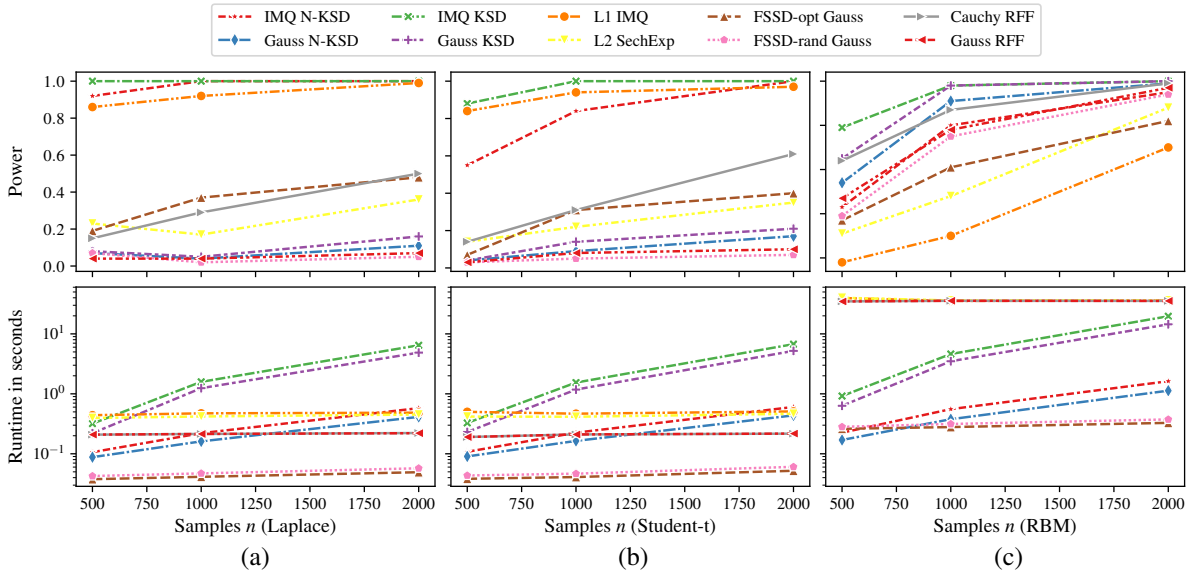


**Figure 5.2.:** Runtime and power trade-off of the tested approximations.

For the IMQ kernel (L1 IMQ, N-KSD IMQ, KSD IMQ), we set $c = 1$ and $\beta = -1/2$. We show the results in Figure 5.1(d), using 100 repetitions to obtain the power. KSD with the IMQ and with the Gaussian kernel performs best. Our proposed Nyström-based method ($m = 4\sqrt{n}$) nearly matches its performance with the IMQ kernel while requiring only a fraction of the runtime. Besides Cauchy RFF and Gauss RFF, all other approaches achieve less power for $\sigma \in \{0.02, 0.04\}$.

These experiments demonstrate the efficiency of the proposed Nyström-KSD method.

### 5.3.2. Runtime vs. power trade-off

In this section, we perform an additional set of experiments to contrast runtime and power; the setups match those of Section 5.3.1. We repeated each setup for 100 rounds to obtain the given power and average runtime. The quadratic-time approaches are considered as baseline.

**Laplace vs. standard normal.** We fix $d = 15$, $m = 4\sqrt{n}$, and vary $n \in \{500, 1000, 2000\}$. The remaining parameters match the ones stated in Section 5.3.1.

Figure 5.2(a) summarizes our results regarding power and runtime. The results show that the proposed IMQ N-KSD approach has the highest power of all approximations across all tested $n$. Second best is L1 IMQ. W.r.t. runtime, the proposed method is faster than L1 IMQ for $n \in \{500, 1000\}$. For $n = 2000$, IMQ N-KSD has a similar runtime but still features better power. The FSSD approaches are the fastest but do not have a high power in this experiment.

**Student-t vs. standard normal.** Again, we fix $d = 15$, set $m = 4\sqrt{n}$, and vary $n \in \{500, 1000, 2000\}$. The other parameters are the sames as the ones stated in Section 5.3.1.

Figure 5.2(b) shows that L1 IMQ achieves higher power than the proposed IMQ N-KSD for $n \in \{500, 1000\}$ but at the price of a larger runtime. For $n = 2000$, the performance of IMQ N-KSD is slightly

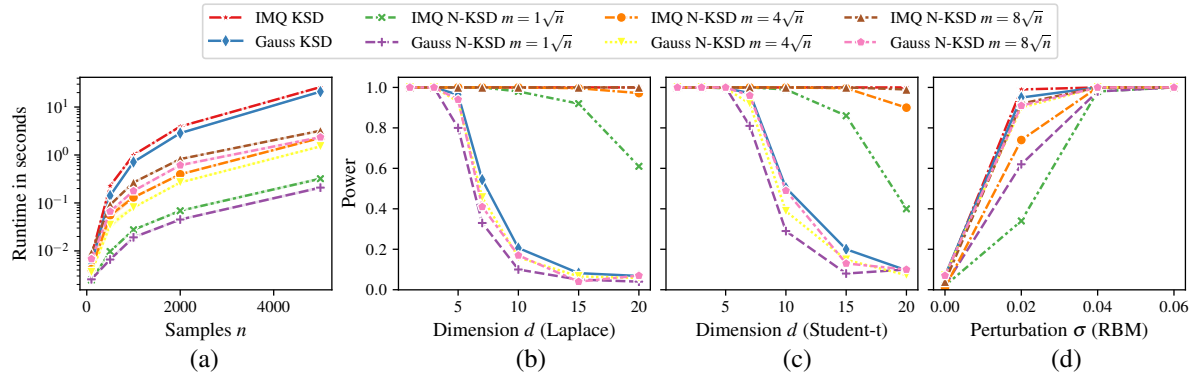


**Figure 5.3.:** Impact of different choices of factor $c$ for the number of Nyström samples $m = c\sqrt{n}$.

better than that of L1 IMQ while both approaches have a similar runtime. The remaining approaches perform worse in terms of power.

**Restricted Boltzmann machine (RBM).** For the RBM experiment, we set $\sigma = 0.02$, $m = 4\sqrt{n}$, and select $n \in \{500, 1000, 2000\}$; all other parameters match the ones detailed in Section 5.3.1.

We summarize the results in Figure 5.2(c). While both random feature Stein discrepancies (L1 IMQ, L2 SechExp) scale linearly in $n$, the higher dimensionality and difficulty of this problem result in a runtime that is orders of magnitude larger than that of all other approximations; the same observation w.r.t. runtime applies to the RFF approaches. We also observe that the runtimes of the related FSSD approaches increase compared to their runtime results in the Laplace and Student-t experiments.

Regarding power, the proposed Gauss N-KSD achieves the best result of all approximations from $n \geq 1000$ while being among the fastest methods. While, for $n \in \{1000, 2000\}$, it is a bit slower than the FSSD approaches, the proposed method achieves higher power across all choices of $n$.

From these results, we conclude that the proposed N-KSD has a very good runtime/power trade-off.

### 5.3.3. Impact of the size of the Nyström sample

Figure 5.3(a–d) captures the impact of the choice of the Nyström sample size $m = c\sqrt{n}$ for $c \in \{1, 4, 8\}$; the $\sqrt{n}$ dependence follows from Corollary 5.2.1(ii), where we neglect the logarithmic terms due to their small contribution. We include the quadratic-time approaches as baselines; the experimental setup matches the experiments detailed in Section 5.3.1. Generally, as one expects, both runtime and power increase for larger $c$. Still, even for $c = 8$, where the power of the proposed approximation is hardly discernible from the baselines across all experiments, its runtime is an order of magnitude lower, which further strengthens the benefit of employing our proposed method.

## 5.4. Limitations

Assumption 5.2.1, which underpins our main result (Theorem 5.2.2), can be difficult to verify in some cases. We refer to Example 5.2.2 for a case where the analytical verification is possible. The weaker assumption $\left\| \left\| h_p(\cdot, X) \right\|_{\mathcal{H}_{h_p}} \right\|_{\psi_2} < \infty$ of Theorem 5.2.3 is usually easier to verify analytically, as we show

in Example 5.2.3. We note that, as with all kernel-based tests, the choice of the kernel, corresponding to the setting of $\gamma$ for the Gaussian kernel (resp. the setting of $\beta, c$ for the IMQ kernel), has an impact on the power of the test. While optimizing kernel parameters is not the focus of this work, there exist methods in the literature to (approximately) achieve this goal [Jitkrittum et al., 2016, 2017a,b, Liu et al., 2020, Schrab et al., 2022a,b, Hagrass et al., 2024a,b, 2025].

## 5.5. Proofs

This section is dedicated to our proofs. We collect auxiliary results in Section 5.5.1. Lemma 5.2.1 is proved in Section 5.5.2. We prove our main result (Theorem 5.2.2) in Section 5.5.3; Corollary 5.2.1 is shown in Section 5.5.4. The proof of Theorem 5.2.3 is in Section 5.5.5.

### 5.5.1. Auxiliary results

This section collects our auxiliary results. Lemma 5.5.1 builds on Rudi et al. [2015, Lemma 6], which assumes bounded feature maps, and on Della Vecchia et al. [2021, Lemma 5], which is stated in the context of leverage scores. The main technical challenge that we resolve lies in introducing and handling the centered covariance operator that allows us to make use of existing concentration results. Lemma 5.5.2 states that a sub-exponential random variable satisfies Bernstein's conditions, and Lemma 5.5.3 is about the sub-Gaussianity of norms of Hilbert space-valued random variables. In Lemma 5.5.4, we show how tensor products interplay with linearly transformed vectors. Lemma 5.5.5 is about the maximum of real-valued sub-Gaussian random variables; it is a slightly altered restatement of Canonne [2021]. In Lemma 5.5.6 and Lemma 5.5.7, we collect inequalities of positive operators and of norms of covariance operators, respectively.

**Lemma 5.5.1** (Projected covariance operator bound)**.** *Let Assumption 1 hold, and assume* $0 < \lambda \leq \left\| C_{\mathbb{Q},\bar{h}_p} \right\|_{\mathrm{op}}$*. Then, for any* $\delta \in (0,1)$*, it holds that*

$$(\mathbb{P}^n \otimes \Lambda^m)\left( \left\| \left( I - P_{\mathcal{H}_{h_p,m}} \right) C^{1/2}_{\mathbb{Q},\bar{h}_p,\lambda} \right\|^2_{\mathrm{op}} \lesssim \lambda \right) \geq 1 - \delta,$$

*provided that* $m \gtrsim \max\left\{ \frac{\mathrm{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}{\lambda}, 1 \right\} \log(4/\delta)$.

*Proof.* The proof proceeds in two steps: First, we show that $\left\| \left( I - P_{\mathcal{H}_{h_p,m}} \right) C^{1/2}_{\mathbb{Q},\bar{h}_p,\lambda} \right\|^2_{\mathrm{op}} \leq \frac{\lambda}{1-\beta(\lambda)}$, when $\beta(\lambda) := \lambda_{\max}\left( C^{-1/2}_{\mathbb{Q},\bar{h}_p,\lambda} \left( C_{\mathbb{Q},\bar{h}_p} - C_{\tilde{\mathbb{Q}}_m,\tilde{h}_p} \right) C^{-1/2}_{\mathbb{Q},\bar{h}_p,\lambda} \right) < 1$, where

$$\tilde{h}_p(\cdot,\mathbf{x}) := h_p(\cdot,\mathbf{x}) - \frac{1}{m}\sum_{i\in[m]} h_p(\cdot,\tilde{\mathbf{x}}_i) \quad (\mathbf{x} \in \mathbb{R}^d),$$

$$\begin{aligned} C_{\tilde{\mathbb{Q}}_m,\tilde{h}_p} &= \frac{1}{m}\sum_{i\in[m]} \tilde{h}_p(\cdot,\tilde{\mathbf{x}}_i) \otimes \tilde{h}_p(\cdot,\tilde{\mathbf{x}}_i) \\ &= \frac{1}{m}\sum_{i\in[m]} h_p(\cdot,\tilde{\mathbf{x}}_i) \otimes h_p(\cdot,\tilde{\mathbf{x}}_i) - \left( \frac{1}{m}\sum_{i\in[m]} h_p(\cdot,\tilde{\mathbf{x}}_i) \right) \otimes \left( \frac{1}{m}\sum_{i\in[m]} h_p(\cdot,\tilde{\mathbf{x}}_i) \right). \end{aligned}$$

In the second step, we show that $\beta(\lambda) < 1$ (with high probability) for $m$ large enough.

**Step 1.** Define the sampling operator $Z_m : \mathcal{H}_{h_p} \to \mathbb{R}^m$ by $f \mapsto \frac{1}{\sqrt{m}} \left(f(\tilde{\mathbf{x}}_i)\right)_{i=1}^m$. Its adjoint $Z_m^* : \mathbb{R}^m \to \mathcal{H}_{h_p}$ (see Sterge and Sriperumbudur [2022, Lemma A.7(i)] is given by $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^m \mapsto \frac{1}{\sqrt{m}} \sum_{i=1}^m \alpha_i h_p(\cdot, \tilde{\mathbf{x}}_i)$. Recall that $\mathcal{H}_{h_p,m} = \operatorname{span}\left\{h_p(\cdot, \tilde{\mathbf{x}}_i) \mid i \in [m]\right\}$ and notice that $\operatorname{range} P_{\mathcal{H}_{h_p,m}} = \overline{\operatorname{range} Z_m^*}$. We obtain

$$\begin{aligned}\left\|\left(I - P_{\mathcal{H}_{h_p,m}}\right) C_{\mathbb{Q},\bar{h}_p,\lambda}^{1/2}\right\|_{\mathrm{op}}^2 &\overset{(a)}{\leq} \lambda \left\|\left(Z_m^* Z_m + \lambda I\right)^{-1/2} C_{\mathbb{Q},\bar{h}_p,\lambda}^{1/2}\right\|_{\mathrm{op}}^2 \overset{(b)}{=} \lambda \left\|C_{\tilde{\mathbb{Q}}_m,h_p,\lambda}^{-1/2} C_{\mathbb{Q},\bar{h}_p,\lambda}^{1/2}\right\|_{\mathrm{op}}^2 \\ &\overset{(c)}{\leq} \lambda \left\|C_{\tilde{\mathbb{Q}}_m,\tilde{h}_p,\lambda}^{-1/2} C_{\mathbb{Q},\bar{h}_p,\lambda}^{1/2}\right\|_{\mathrm{op}}^2\end{aligned} \tag{5.7}$$

where (a) follows from Rudi et al. [2015, Proposition 3] with $X := C_{\mathbb{Q},\bar{h}_p,\lambda}^{1/2}$ therein. (b) is by Sterge and Sriperumbudur [2022, Lemma A.7(iv)]. Lemma 5.5.6(5) with $C := C_{\tilde{\mathbb{Q}}_m,h_p,\lambda}^{-1/2}$, $D := C_{\tilde{\mathbb{Q}}_m,\tilde{h}_p,\lambda}^{-1/2}$, and $X := C_{\mathbb{Q},\bar{h}_p,\lambda}^{1/2}$ yields (c), as we obtain $C^*C = C_{\tilde{\mathbb{Q}}_m,h_p,\lambda}^{-1} \preccurlyeq C_{\tilde{\mathbb{Q}}_m,\tilde{h}_p,\lambda}^{-1} = D^*D$; the positive definite relationship holding by the following chain of inequalities

$$\begin{aligned}C_{\tilde{\mathbb{Q}}_m,h_p,\lambda}^{-1} \preccurlyeq C_{\tilde{\mathbb{Q}}_m,\tilde{h}_p,\lambda}^{-1} &\overset{\text{Lemma 5.5.6(4)}}{\iff} C_{\tilde{\mathbb{Q}}_m,h_p,\lambda} \succcurlyeq C_{\tilde{\mathbb{Q}}_m,\tilde{h}_p,\lambda} \overset{(d)}{\iff} C_{\tilde{\mathbb{Q}}_m,h_p} \succcurlyeq C_{\tilde{\mathbb{Q}}_m,\tilde{h}_p} \\ &\overset{(e)}{\iff} 0 \preccurlyeq \mu_{h_p}\left(\tilde{\mathbb{Q}}_m\right) \otimes \mu_{h_p}\left(\tilde{\mathbb{Q}}_m\right),\end{aligned}$$

which is true as the r.h.s. is a positive operator. In (d), we subtract $\lambda I$ from both sides. (e) follows from subtracting $C_{\tilde{\mathbb{Q}}_m,h_p}$ and by multiplying with $-1$.

Applying the second inequality in the statement of Rudi et al. [2015, Proposition 7] to (5.7) (we specialize $A := C_{\tilde{\mathbb{Q}}_m,\tilde{h}_p}$ and $B := C_{\mathbb{Q},\bar{h}_p}$ therein), we obtain

$$\lambda \left\|C_{\tilde{\mathbb{Q}}_m,\tilde{h}_p,\lambda}^{-1/2} C_{\mathbb{Q},\bar{h}_p,\lambda}^{1/2}\right\|_{\mathrm{op}}^2 \leq \frac{\lambda}{1 - \beta(\lambda)}, \tag{5.8}$$

when $\beta(\lambda) < 1$. The combination of (5.7) and (5.8) yields the first stated claim.

**Step 2.** It remains to show that $\beta(\lambda) < 1$ holds with high probability. Let us introduce the shorthands $\tilde{\mu}_{h_p} = \mu_{h_p}\left(\tilde{\mathbb{Q}}_m\right) = \frac{1}{m} \sum_{i \in [m]} h_p(\cdot, \tilde{\mathbf{x}}_i)$ and $\mu_{h_p} = \mu_{h_p}(\mathbb{Q})$. Notice that we have

$$C_{\tilde{\mathbb{Q}}_m,\tilde{h}_p} = C_{\tilde{\mathbb{Q}}_m,\bar{h}_p} - \left[\tilde{\mu}_{h_p} - \mu_{h_p}\right] \otimes \left[\tilde{\mu}_{h_p} - \mu_{h_p}\right], \tag{5.9}$$

which is verified by using the linearity of tensor products and by using that

$$C_{\tilde{\mathbb{Q}}_m,\bar{h}_p} = \frac{1}{m} \sum_{i \in [m]} \bar{h}_p(\cdot, \tilde{\mathbf{x}}_i) \otimes \bar{h}_p(\cdot, \tilde{\mathbf{x}}_i).$$

Instead of showing that $\beta(\lambda) < 1$, we will show that the following stronger requirement can be satisfied:

$$\begin{aligned}
\beta(\lambda) &\overset{(a)}{\leq} \left\| C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} \left( C_{\mathbb{Q},\bar{h}_p} - C_{\tilde{\mathbb{Q}}_m,\bar{h}_p} \right) C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} \right\|_{\mathrm{op}} \\
&\overset{(b)}{=} \left\| C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} \left( C_{\mathbb{Q},\bar{h}_p} - C_{\tilde{\mathbb{Q}}_m,\bar{h}_p} + \left[\tilde{\mu}_{h_p} - \mu_{h_p}\right] \otimes \left[\tilde{\mu}_{h_p} - \mu_{h_p}\right] \right) C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} \right\|_{\mathrm{op}} \\
&\overset{(c)}{\leq} \left\| C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} \left( C_{\mathbb{Q},\bar{h}_p} - C_{\tilde{\mathbb{Q}}_m,\bar{h}_p} \right) C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} \right\|_{\mathrm{op}} \\
&\quad + \left\| C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} \left( \left[\tilde{\mu}_{h_p} - \mu_{h_p}\right] \otimes \left[\tilde{\mu}_{h_p} - \mu_{h_p}\right] \right) C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} \right\|_{\mathrm{op}} \\
&\overset{(d)}{=} \underbrace{\left\| C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} \left( C_{\mathbb{Q},\bar{h}_p} - C_{\tilde{\mathbb{Q}}_m,\bar{h}_p} \right) C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} \right\|_{\mathrm{op}}}_{=:T_1} + \underbrace{\left\| C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} \left( \tilde{\mu}_{h_p} - \mu_{h_p} \right) \right\|_{\mathcal{H}_{h_p}}^2}_{=:T_2} < 1.
\end{aligned}$$

In (a), we use that the spectral radius is bounded by the operator norm. (b) uses (5.9) and (c) holds by the triangle inequality. Lemma 5.5.4 and Lemma A.3.1 applied to the second term yield (d).

- **First term ($T_1$).** We will bring ourselves into the setting of Koltchinskii and Lounici [2017, Theorem 9] (recalled in Theorem A.3.1). First, we condition on the Nyström selection and define the centered random variables $\eta_{i_j} = C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} \left( h_p\left(\cdot, \tilde{\mathbf{x}}_j\right) - \mathbb{E}_{X\sim\mathbb{Q}} h_p\left(\cdot, X\right) \right) \left( = C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} \bar{h}_p(\cdot, \tilde{\mathbf{x}}_j) \right)$ $(j \in [m])$, which satisfy the sub-Gaussian assumption. Indeed, let $u \in \mathcal{H}_{h_p}$ be arbitrary, and $v = C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} u \in \mathcal{H}_{h_p}$; the invertibility of $C_{\mathbb{Q},\bar{h}_p,\lambda}$ guarantees the well-definedness of $v$. With this notation, for any $j \in [m]$,

$$\begin{aligned}
\left\| \left\langle \eta_{i_j}, u \right\rangle_{\mathcal{H}_{h_p}} \right\|_{\psi_2} &\overset{(a)}{=} \left\| \left\langle C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} \bar{h}_p(\cdot, \tilde{\mathbf{x}}_j), u \right\rangle_{\mathcal{H}_{h_p}} \right\|_{\psi_2} \overset{(b)}{=} \left\| \left\langle \bar{h}_p(\cdot, \tilde{\mathbf{x}}_j), C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} u \right\rangle_{\mathcal{H}_{h_p}} \right\|_{\psi_2} \\
&\overset{(c)}{=} \left\| \left\langle \bar{h}_p(\cdot, \tilde{\mathbf{x}}_j), v \right\rangle_{\mathcal{H}_{h_p}} \right\|_{\psi_2} \overset{(d)}{\lesssim} \underbrace{\left\| \left\langle \bar{h}_p(\cdot, \tilde{\mathbf{x}}_j), v \right\rangle_{\mathcal{H}_{h_p}} \right\|_{L^2(\mathbb{Q})}}_{(\dagger)} \overset{(e)}{=} \left\| \left\langle \bar{h}_p(\cdot, \tilde{\mathbf{x}}_j), C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} u \right\rangle_{\mathcal{H}_{h_p}} \right\|_{L^2(\mathbb{Q})} \\
&\overset{(f)}{=} \left\| \left\langle C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} \bar{h}_p(\cdot, \tilde{\mathbf{x}}_{i_j}), u \right\rangle_{\mathcal{H}_{h_p}} \right\|_{L^2(\mathbb{Q})} \overset{(g)}{=} \left\| \left\langle \eta_{i_j}, u \right\rangle_{\mathcal{H}_{h_p}} \right\|_{L^2(\mathbb{Q})} < \infty.
\end{aligned}$$

(a) is the definition of the $\eta_{i_j}$-s, (b) uses the self-adjointness of $C_{\mathbb{Q},\bar{h}_p,\lambda}$, and (c) follows from the definition of $v$. The sub-Gaussian assumption implies (d), (e) again follows from the definition of $v$, and (f) is implied by the self-adjointness of $C_{\mathbb{Q},\bar{h}_p,\lambda}$. Inserting the definition of $\eta_{i_j}$ in (g) proves their sub-Gaussianity by using that $(\dagger) < \infty$ according to Assumption 5.2.1 and as the derivation afterwards only involved equalities.
Let $A = C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} C_{\mathbb{Q},\bar{h}_p} C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2}$. Theorem A.3.1 yields that, conditioned on the Nyström selection, it holds with probability at least $1 - \delta/2$ that

$$\underbrace{\left\| C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} \left( C_{\mathbb{Q},\bar{h}_p} - C_{\tilde{\mathbb{Q}}_m,\bar{h}_p} \right) C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} \right\|_{\mathrm{op}}}_{= \frac{1}{m}\sum_{j=1}^m \eta_{i_j} \otimes \eta_{i_j} - \mathbb{E}\left[\eta_{i_j} \otimes \eta_{i_j}\right]} \lesssim \|A\|_{\mathrm{op}} \max\left( \sqrt{\frac{r(A)}{m}}, \sqrt{\frac{\log(2/\delta)}{m}} \right),$$

provided that $m \geq \max\{r(A), \log(2/\delta)\}$, with $r(A) = \frac{\operatorname{tr}(A)}{\|A\|_{\mathrm{op}}}$. Using Lemma 5.5.6(2), $A \preccurlyeq I$, hence $\|A\|_{\mathrm{op}} \leq 1$. Moreover by Lemma 5.5.7(3), $r(A) \leq \frac{2\operatorname{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}{\lambda}$, which implies that, with the same probability,

$$\left\| C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} \left( C_{\mathbb{Q},\bar{h}_p} - C_{\tilde{\mathbb{Q}}_m,\bar{h}_p} \right) C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} \right\|_{\mathrm{op}} \lesssim \max\left( \sqrt{\frac{\operatorname{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}{\lambda m}}, \sqrt{\frac{\log(2/\delta)}{m}} \right),$$

holds when $m \geq \max\left\{ \frac{2\operatorname{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}{\lambda}, \log(2/\delta) \right\}$. Therefore, one can take $m \gtrsim \max\left\{ \frac{\operatorname{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}{\lambda}, \log(2/\delta) \right\}$ to get $\left\| C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} \left( C_{\mathbb{Q},\bar{h}_p} - C_{\tilde{\mathbb{Q}}_m,\bar{h}_p} \right) C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} \right\|_{\mathrm{op}} < \frac{1}{2}$ holding with probability at least $1 - \delta/2$.

- **Second term ($T_2$).** We condition again on the Nyström selection, let $\eta_{i_j} = C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2} \bar{h}_p\left(\cdot, \mathbf{x}_{i_j}\right)$ for $j \in [m]$, and observe that $\frac{1}{m}\sum_{j\in[m]} \eta_{i_j} = C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2}\left(\tilde{\mu}_{h_p} - \mu_{h_p}\right)$. The $\eta_{i_j}$-s are centered, and, by Lemma 5.5.3, it holds for any $j \in [m]$ that

$$\left\| \left\| \eta_{i_j} \right\|_{\mathcal{H}_{h_p}} \right\|_{\psi_2}^2 \lesssim \operatorname{tr}\left( C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1} C_{\mathbb{Q},\bar{h}_p} \right),$$

that is, the $\left\|\eta_{i_j}\right\|_{\mathcal{H}_{h_p}}$-s are sub-Gaussian. Hence, by Lemma A.3.2(3), they are sub-exponential, and, by Lemma 5.5.2, they satisfy the Bernstein condition with $\sigma, B \lesssim \sqrt{\operatorname{tr}\left( C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1} C_{\mathbb{Q},\bar{h}_p} \right)}$. Therefore, application of Theorem A.3.2 yields that, conditioned on the Nyström choice, it holds with probability at least $1 - \delta/2$ that

$$\begin{aligned} \left\| \frac{1}{m} \sum_{j=1}^{m} \eta_{i_j} \right\|_{\mathcal{H}_{h_p}} &\lesssim \frac{\sqrt{\operatorname{tr}\left( C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1} C_{\mathbb{Q},\bar{h}_p} \right)} \log(4/\delta)}{m} + \sqrt{\frac{\operatorname{tr}\left( C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1} C_{\mathbb{Q},\bar{h}_p} \right) \log(4/\delta)}{m}} \\ &\overset{(a)}{\lesssim} \sqrt{\frac{\operatorname{tr}\left( C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1} C_{\mathbb{Q},\bar{h}_p} \right) \log(4/\delta)}{m}} \overset{(b)}{\leq} \sqrt{\frac{\operatorname{tr}\left( C_{\mathbb{Q},\bar{h}_p} \right) \log(4/\delta)}{\lambda m}} \end{aligned}$$

where in (a), we assume that $m \geq \log(4/\delta)$ and notice that this condition implies that the first term is smaller than the second term. Lemma 5.5.7(1) yields (b). The obtained bound means that choosing $m \gtrsim \max\left\{ \frac{\operatorname{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}{\lambda}, 1 \right\} \log(4/\delta)$ guarantees that $\left\| \frac{1}{m}\sum_{j=1}^{m} \eta_{i_j} \right\|_{\mathcal{H}_{h_p}}^2 < \frac{1}{2}$ holds with probability at least $1 - \delta/2$.

As a final step, we observe that $\log(2/\delta) < \log(4/\delta)$ and $\log(4/\delta) > 1$, which, by union bound, shows that, for $m \gtrsim \max\left\{ \frac{\operatorname{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}{\lambda}, 1 \right\} \log(4/\delta)$, it holds with probability at least $1 - \delta$ that $\beta(\lambda) < 1$. We lift the conditioning by integrating over all Nyström selections. □

**Lemma 5.5.2** (Sub-exponential satisfies Bernstein conditions)**.** *Let $Y$ be a real-valued random variable which is sub-exponential, i.e. $\|Y\|_{\psi_1} < \infty$. Let $\sigma := \sqrt{2}\|Y\|_{\psi_1}$, $B := \|Y\|_{\psi_1} > 0$. Then the Bernstein condition*

$$\mathbb{E}|Y|^p \leq \frac{1}{2} p! \sigma^2 B^{p-2} < \infty$$

*holds for any $p \geq 2$.*

*Proof.* For any $p \geq 2$, we have

$$\mathbb{E}|Y|^p = p! B^p \mathbb{E}\frac{|Y|^p}{B^p p!} \overset{(a)}{<} p! B^p \underbrace{\left[\mathbb{E}\exp\left(\frac{|Y|}{B}\right) - 1\right]}_{\overset{(b)}{\leq} 1} = \frac{1}{2} p! B^{p-2} \left(\sqrt{2}B\right)^2,$$

where in (a) we use that $\frac{x^n}{n!} < e^x - 1$ holds for all $n, x > 0$, and (b) follows from the definition of the sub-exponential Orlicz norm. □

The next lemma shows that $\bar{h}_p(\cdot, X)$ and the "whitened" random variable $C^{-1/2}_{\mathbb{Q},\bar{h}_p,\lambda}\bar{h}_p(\cdot, X)$ enjoy sub-Gaussian properties in terms of their respective $\mathcal{H}_{h_p}$ norms.

**Lemma 5.5.3** (Sub-Gaussianity of norm of Hilbert space-valued random variables)**.** *Let $\mathcal{H}$ be a separable Hilbert space, $Y \sim \mathbb{Q} \in \mathcal{M}_1^+(\mathcal{H})$, and $A \in \mathcal{L}(\mathcal{H})$ invertible, and positive. Assume that $Y$ is sub-Gaussian, in other words $\left\|\langle Y, u\rangle_{\mathcal{H}}\right\|_{\psi_2} \lesssim \left\|\langle Y, u\rangle_{\mathcal{H}}\right\|_{L^2(\mathbb{Q})}$ holds for all $u \in \mathcal{H}$. Then*

$$\left\|\left\|A^{1/2}Y\right\|_{\mathcal{H}}\right\|_{\psi_2}^2 \lesssim \operatorname{tr}\left(A\mathbb{E}_{Y\sim\mathbb{Q}}\left(Y \otimes Y\right)\right).$$

*Specifically, with Assumption 5.2.1, choosing $A := I$ and $Y := \bar{h}_p(\cdot, X)$, and $A := C^{-1}_{\mathbb{Q},\bar{h}_p,\lambda}$ $(\lambda > 0)$ and $Y := \bar{h}_p(\cdot, X)$, respectively, it holds that*

$$\left\|\left\|\bar{h}_p(\cdot, X)\right\|_{\mathcal{H}_{h_p}}\right\|_{\psi_2} < \infty, \qquad \text{and} \qquad \left\|\left\|C^{-1/2}_{\mathbb{Q},\bar{h}_p,\lambda}\bar{h}_p(\cdot, X)\right\|_{\mathcal{H}_{h_p}}\right\|_{\psi_2}^2 \lesssim \operatorname{tr}\left(C^{-1}_{\mathbb{Q},\bar{h}_p,\lambda}C_{\mathbb{Q},\bar{h}_p}\right) < \infty,$$

*that is, both $\left\|\bar{h}_p(\cdot, X)\right\|_{\mathcal{H}_{h_p}}$ and $\left\|C^{-1/2}_{\mathbb{Q},\bar{h}_p,\lambda}\bar{h}_p(\cdot, X)\right\|_{\mathcal{H}_{h_p}}$ are sub-Gaussian.*

*Proof.* Let $(e_i)_{i\in I}$ be a countable ONB of the separable $\mathcal{H}$. We obtain

$$\begin{aligned}
\left\|\left\|A^{1/2}Y\right\|_{\mathcal{H}}\right\|_{\psi_2}^2 &\overset{(a)}{=} \left\|\left\|A^{1/2}Y\right\|_{\mathcal{H}}^2\right\|_{\psi_1} \overset{(b)}{=} \left\|\sum_{i\in I}\left\langle A^{1/2}Y, e_i\right\rangle_{\mathcal{H}}^2\right\|_{\psi_1} \overset{(c)}{\leq} \sum_{i\in I}\left\|\left\langle A^{1/2}Y, e_i\right\rangle_{\mathcal{H}}^2\right\|_{\psi_1} \\
&\overset{(d)}{=} \sum_{i\in I}\left\|\left\langle A^{1/2}Y, e_i\right\rangle_{\mathcal{H}}\right\|_{\psi_2}^2 \overset{(e)}{\lesssim} \sum_{i\in I}\left\|\left\langle A^{1/2}Y, e_i\right\rangle_{\mathcal{H}}\right\|_{L^2(\mathbb{Q})}^2 \overset{(f)}{=} \sum_{i\in I}\mathbb{E}_{Y\sim\mathbb{Q}}\left\langle A^{1/2}Y, e_i\right\rangle_{\mathcal{H}}^2 \\
&\overset{(g)}{=} \sum_{i\in I}\mathbb{E}_{Y\sim\mathbb{Q}}\left\langle\left(A^{1/2}Y\right)\otimes\left(A^{1/2}Y\right), e_i \otimes e_i\right\rangle_{\mathcal{H}\otimes\mathcal{H}} \\
&\overset{(h)}{=} \sum_{i\in I}\mathbb{E}_{Y\sim\mathbb{Q}}\left\langle A^{1/2}\left(Y\otimes Y\right)A^{1/2}, e_i\otimes e_i\right\rangle_{\mathcal{H}\otimes\mathcal{H}} \\
&\overset{(i)}{=} \sum_{i\in I}\left\langle A^{1/2}\mathbb{E}_{Y\sim\mathbb{Q}}\left(Y\otimes Y\right)A^{1/2}, e_i\otimes e_i\right\rangle_{\mathcal{H}\otimes\mathcal{H}} \overset{(j)}{=} \sum_{i\in I}\left\langle A^{1/2}\mathbb{E}_{Y\sim\mathbb{Q}}\left(Y\otimes Y\right)A^{1/2}e_i, e_i\right\rangle_{\mathcal{H}} \\
&\overset{(k)}{=} \operatorname{tr}\left(A^{1/2}\mathbb{E}_{Y\sim\mathbb{Q}}\left(Y\otimes Y\right)A^{1/2}\right) \overset{(l)}{=} \operatorname{tr}\left(A\mathbb{E}_{Y\sim\mathbb{Q}}\left(Y\otimes Y\right)\right).
\end{aligned}$$

The details are as follows. (a) uses Lemma A.3.2(4), Parseval's identity yields (b), and the triangle inequality implies (c). (d) holds by Lemma A.3.2(4). For (e), let $u_i = A^{1/2}e_i$ and observe that

$$\left\|\left\langle A^{1/2}Y, e_i\right\rangle_{\mathcal{H}}\right\|_{\psi_2}^2 \overset{(m)}{=} \left\|\langle Y, u_i\rangle_{\mathcal{H}}\right\|_{\psi_2}^2 \overset{(n)}{\lesssim} \left\|\langle Y, u_i\rangle_{\mathcal{H}}\right\|_{L^2(\mathbb{Q})}^2 \overset{(o)}{=} \left\|\left\langle A^{1/2}Y, e_i\right\rangle_{\mathcal{H}}\right\|_{L^2(\mathbb{Q})}^2,$$

where (m) uses the self-adjointness of $A^{1/2}$ (implied by the positivity of $A$), (n) follows from the sub-Gaussian assumption on $Y$ holding for arbitrary $u_i \in \mathcal{H}$, and (o), again, uses the self-adjointness of $A$. (f) is the definition of the $L^2(\mathbb{Q})$-norm, (g) holds by the definition of the tensor product, and Lemma 5.5.4 yields (h). (i) integral and bounded linear operators are swapped by Steinwart and Christmann [2008, (A.32)], (j) is a property of Hilbert-Schmidt operators, and (k) uses the definition of the trace of a linear operator w.r.t. an ONB. The cyclic invariance property of the trace yields (l) and concludes the proof of the first statement.

With $A := I$ and $Y := \bar{h}_p(\cdot, X)$, we have $\left\|\left\|\bar{h}_p(\cdot, X)\right\|_{\mathcal{H}_{h_p}}\right\|_{\psi_2} \lesssim \operatorname{tr}\left(\mathbb{E}\left(\bar{h}_p(\cdot, X) \otimes \bar{h}_p(\cdot, X)\right)\right) = \operatorname{tr}(C_{\mathbb{Q},\bar{h}_p}) < \infty$, which is the second statement. The last part follows from considering $A := C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1}$ and $Y := \bar{h}_p(\cdot, X)$; the invertibility of $C_{\mathbb{Q},\bar{h}_p,\lambda}$ guarantees the well-definedness of the $u_i$-s ($i \in I$). □

The following lemma is a natural generalization of the property $(\mathbf{Ca})(\mathbf{Db})^\mathsf{T} = \mathbf{C}\left(\mathbf{ab}^\mathsf{T}\right)\mathbf{D}^\mathsf{T}$, where $\mathbf{C}, \mathbf{D} \in \mathbb{R}^{d\times d}$ and $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$.

**Lemma 5.5.4** (Tensor product of linearly transformed vectors)**.** *Let $\mathcal{H}$ be a Hilbert space and $C, D \in \mathcal{L}(\mathcal{H})$. Then for any $a, b \in \mathcal{H}$, $(Ca) \otimes (Db) = C(a \otimes b)D^*$. Specifically, when $D$ is self-adjoint, it holds that $(Ca) \otimes (Db) = C(a \otimes b)D$.*

*Proof.* Let $h \in \mathcal{H}$ be arbitrary and fixed. Then,

$$[(Ca) \otimes (Db)](h) \overset{(a)}{=} Ca\langle Db, h\rangle_{\mathcal{H}},$$
$$[C(a \otimes b)D^*](h) = C(a \otimes b)(D^*h) \overset{(b)}{=} Ca\langle b, D^*h\rangle_{\mathcal{H}} \overset{(c)}{=} Ca\langle Db, h\rangle_{\mathcal{H}}.$$

In (a) and (b), we used the definition of $\otimes$, (c) follows from the definition of the adjoint and by the property $(D^*)^* = D$. The shown equality of $[(Ca) \otimes (Db)](h) = [C(a \otimes b)D^*](h)$ for any $h \in \mathcal{H}$ proves the claimed statement. □

**Lemma 5.5.5** (Maximum of sub-Gaussian random variables)**.** *Let $(X_i)_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$ be real-valued sub-Gaussian random variables. Then $\mathbb{P}\left(\max_{i\in[n]} |X_i| \lesssim \sqrt{\|X_1\|_{\psi_2}^2 \log(2n/\delta)}\right) \geq 1-\delta$ holds for any $\delta \in (0,1)$.*

*Proof.* Let $c > 0$ be an absolute constant. As $X_1$ is sub-Gaussian, by Vershynin [2018, Proposition 2.5.2], there exists $K_1 \leq c\|X_1\|_{\psi_2}$ such that $\mathbb{P}(|X_1| \geq t) \leq 2e^{-\frac{t^2}{K_1^2}}$ for all $t \geq 0$. Let $u = \sqrt{K_1^2(\log(2n) + t)}$. Then

$$\mathbb{P}\left(\max_{i\in[n]} |X_i| \geq u\right) \overset{(a)}{\leq} \sum_{i=1}^n \mathbb{P}(|X_i| \geq u) \overset{(b)}{\leq} 2ne^{-\frac{u^2}{K_1^2}} \overset{(c)}{=} e^{-t},$$

where (a) uses that the probability of a maximum exceeding a value is less than the sum of the probability of its arguments exceeding the value, (b) uses the mentioned property of sub-Gaussian random variables, and (c) is our definition of $u$. Solving for $\delta := e^{-t}$ gives $t = \log(1/\delta)$, and considering the complement yields $\mathbb{P}\left(\max_{i\in[n]} |X_i| \leq \sqrt{K_1^2 \log(2n/\delta)}\right) \geq 1-\delta$. Using that $K_1 \leq c\|X_1\|_{\psi_2}$ concludes the proof. □

The following result shows that positive operators share some well-known properties of positive (semi-)definite matrices; we refer to Bhatia [2007] for the related matrix cases.

**Lemma 5.5.6** (Properties of positive operators). *Let $\mathcal{H}$ be a Hilbert space and assume $A, B \in \mathcal{L}(\mathcal{H})$ are positive and invertible. Then, the following hold.*

1. *If $A \preccurlyeq B$, then $X^*AX \preccurlyeq X^*BX$ for any $X \in \mathcal{L}(\mathcal{H})$.*
2. *If $A \preccurlyeq B$, then $B^{-1/2}AB^{-1/2} \preccurlyeq I$.*
3. *If $B \preccurlyeq I$, then $B^{-1} \succcurlyeq I$.*
4. *If $A \preccurlyeq B$, then $A^{-1} \succcurlyeq B^{-1}$.*
5. *If $C^*C \preccurlyeq D^*D$, then $\|CX\|_{\mathrm{op}} \leq \|DX\|_{\mathrm{op}}$ for any $C, D, X \in \mathcal{L}(\mathcal{H})$.*

*Proof.*

1. For any $x \in \mathcal{H}$, it holds that $\langle x, X^*AXx\rangle_{\mathcal{H}} = \langle Xx, AXx\rangle_{\mathcal{H}} \overset{(\dagger)}{\leq} \langle Xx, BXx\rangle_{\mathcal{H}} = \langle x, X^*BXx\rangle_{\mathcal{H}}$; $(\dagger)$ follows from $A \preccurlyeq B$ applied to $Xx$.
2. We apply (1.) with $X = B^{-1/2}$.
3. We have $B^{-1} = B^{-1/2}IB^{-1/2} \succcurlyeq B^{-1/2}BB^{-1/2} = I$, where we used (1.) in the second step.
4. By (2.), it holds that $B^{-1/2}AB^{-1/2} \preccurlyeq I$, from which (3.) implies that $B^{1/2}A^{-1}B^{1/2} \succcurlyeq I$. Now apply (1.) with $X = B^{-1/2}$ to obtain the stated result.
5. The $C^*$-property, the definition of the adjoint and that of the operator norm yield

$$\begin{aligned}\|CX\|_{\mathrm{op}}^2 = \|X^*C^*CX\|_{\mathrm{op}} &= \sup_{\|x\|_{\mathcal{H}}=1} \langle x, X^*C^*CXx\rangle_{\mathcal{H}} = \sup_{\|x\|_{\mathcal{H}}=1} \langle Xx, C^*CXx\rangle_{\mathcal{H}} \\ &\leq \sup_{\|x\|_{\mathcal{H}}=1} \langle Xx, D^*DXx\rangle_{\mathcal{H}} = \sup_{\|x\|_{\mathcal{H}}=1} \langle x, X^*D^*DXx\rangle_{\mathcal{H}} = \|X^*D^*DX\|_{\mathrm{op}} = \|DX\|_{\mathrm{op}}^2,\end{aligned}$$

   which, after taking the positive square root, proves the claim. □

The following lemma collects some inequalities for the trace and operator norms of covariance operators. Many of these are known and frequently employed without proof; we provide proofs here for completeness.

**Lemma 5.5.7** (Covariance operator inequalities). *Let $\mathcal{H}$ be a separable Hilbert space, $X \sim \mathbb{Q} \in \mathcal{M}_1^+(\mathcal{H})$, $C_{\mathbb{Q}} = \mathbb{E}[X \otimes X]$, $C_{\mathbb{Q},\lambda} = C_{\mathbb{Q}} + \lambda I$, and let $r(\cdot) = \frac{\mathrm{tr}(\cdot)}{\|\cdot\|_{\mathrm{op}}}$ be defined on trace-class operators. Assume that $0 < \lambda \leq \left\|C_{\mathbb{Q}}\right\|_{\mathrm{op}}$. Then, the following hold.*

1. $\frac{1}{2}r\left(C_{\mathbb{Q}}\right) \leq \mathrm{tr}\left(C_{\mathbb{Q},\lambda}^{-1}C_{\mathbb{Q}}\right) \leq \frac{\mathrm{tr}\left(C_{\mathbb{Q}}\right)}{\lambda}$,
2. $\frac{1}{2} \leq \left\|C_{\mathbb{Q},\lambda}^{-1/2}C_{\mathbb{Q}}C_{\mathbb{Q},\lambda}^{-1/2}\right\|_{\mathrm{op}} < 1$, *and*
3. $r\left(C_{\mathbb{Q},\lambda}^{-1/2}C_{\mathbb{Q}}C_{\mathbb{Q},\lambda}^{-1/2}\right) \leq \frac{2\,\mathrm{tr}\left(C_{\mathbb{Q}}\right)}{\lambda}$.

*Proof.* Let $(\lambda_i)_{i\in I}$ denote the eigenvalues of $C_{\mathbb{Q}}$, with $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$.

1. The first inequality follows from $\operatorname{tr}\left(C_{\mathbb{Q},\lambda}^{-1} C_{\mathbb{Q}}\right) = \sum_{i\in I} \frac{\lambda_i}{\lambda_i+\lambda} \geq \sum_{i\in I} \frac{\lambda_i}{2\left\|C_{\mathbb{Q}}\right\|_{\mathrm{op}}} = \frac{\operatorname{tr}(C_{\mathbb{Q}})}{2\left\|C_{\mathbb{Q}}\right\|_{\mathrm{op}}}$. The second one was shown with (2.10).
2. For the first inequality, observe that $\left\|C_{\mathbb{Q},\lambda}^{-1/2} C_{\mathbb{Q}} C_{\mathbb{Q},\lambda}^{-1/2}\right\|_{\mathrm{op}} = \frac{\lambda_1}{\lambda_1+\lambda} \overset{(\dagger)}{\geq} \frac{1}{2}$, where $(\dagger) \Leftrightarrow 2\lambda_1 \geq \lambda_1 + \lambda \Leftrightarrow (\left\|C_{\mathbb{Q}}\right\|_{\mathrm{op}} =) \lambda_1 \geq \lambda$, which holds by assumption. The second one is implied as $\frac{\lambda_1}{\lambda_1+\lambda} \overset{(\dagger)}{<} 1$, where $(\dagger) \Leftrightarrow \lambda_1 < \lambda_1 + \lambda \Leftrightarrow 0 < \lambda$; this condition was again assumed.
3. We upper bound the numerator of $r(C_{\mathbb{Q},\lambda}^{-1/2} C_{\mathbb{Q}} C_{\mathbb{Q},\lambda}^{-1/2})$ by using result (1.) after we rewrite it as $\operatorname{tr}\left(C_{\mathbb{Q},\lambda}^{-1/2} C_{\mathbb{Q}} C_{\mathbb{Q},\lambda}^{-1/2}\right) = \operatorname{tr}\left(C_{\mathbb{Q},\lambda}^{-1} C_{\mathbb{Q}}\right)$ using the cyclic invariance of the trace, and lower bound the denominator by (2.). □

### 5.5.2. Proof of Lemma 5.2.1

By (2.22), KSD is the norm of the mean embedding of $\mathbb{Q}$ under $h_p(\cdot,\cdot)$, that is,

$$S_p(\mathbb{Q}) = \left\| \int_{\mathbb{R}^d} h_p(\cdot, \mathbf{x}) \, \mathrm{d}\mathbb{Q}(\mathbf{x}) \right\|_{\mathcal{H}_{h_p}} = \left\| \mu_{h_p}(\mathbb{Q}) \right\|_{\mathcal{H}_{h_p}}. \tag{5.10}$$

Hence, with Chatalic et al. [2022, (5)], the optimization problem (5.1) has the solution $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^m = \frac{1}{n} \mathbf{K}_{h_p,m,m}^{-} \mathbf{K}_{h_p,m,n} \mathbf{1}_n \in \mathbb{R}^m$. Now, using (5.10), we have

$$\begin{aligned}
\left\| \sum_{i=1}^m \alpha_i h_p(\cdot, \tilde{\mathbf{x}}_i) \right\|_{\mathcal{H}_{h_p,m}}^2 &\overset{(a)}{=} \left\langle \sum_{i=1}^m \alpha_i h_p(\cdot, \tilde{\mathbf{x}}_i), \sum_{i=1}^m \alpha_i h_p(\cdot, \tilde{\mathbf{x}}_i) \right\rangle_{\mathcal{H}_{h_p,m}} \\
&\overset{(b)}{=} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j \left\langle h_p(\cdot, \tilde{\mathbf{x}}_i), h_p(\cdot, \tilde{\mathbf{x}}_j) \right\rangle_{\mathcal{H}_{h_p,m}} \overset{(c)}{=} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j h_p(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \overset{(d)}{=} \boldsymbol{\alpha}^{\mathsf{T}} \mathbf{K}_{h_p,m,m} \boldsymbol{\alpha} \\
&\overset{(e)}{=} \frac{1}{n^2} \mathbf{1}_n^{\mathsf{T}} \mathbf{K}_{h_p,n,m} \underbrace{\mathbf{K}_{h_p,m,m}^{-} \mathbf{K}_{h_p,m,m} \mathbf{K}_{h_p,m,m}^{-}}_{=\mathbf{K}_{h_p,m,m}^{-}} \mathbf{K}_{h_p,m,n} \mathbf{1}_n = \boldsymbol{\beta}_p^{\mathsf{T}} \mathbf{K}_{h_p,m,m}^{-} \boldsymbol{\beta}_p.
\end{aligned}$$

In (a) we used that $\|\cdot\|_{\mathcal{H}_{h_p,m}}$ is inner product induced, (b) follows from the linearity of the inner product, (c) is implied by the reproducing property, (d) is by the definition of the Gram matrix, in (e) we made use of the explicit form of $\boldsymbol{\alpha}$, the symmetry of Gram matrices, the property $\mathbf{K}_{h_p,m,n}^{\mathsf{T}} = \mathbf{K}_{h_p,n,m}$, and that the Moore-Penrose inverse satisfies $\mathbf{A}^{-}\mathbf{A}\mathbf{A}^{-} = \mathbf{A}^{-}$ for any matrix $\mathbf{A}$.

### 5.5.3. Proof of Theorem 5.2.2

Contrasting the existing related work [Rudi et al., 2015, Chatalic et al., 2022, Sterge and Sriperumbudur, 2022, Kalinke and Szabó, 2023], we do not impose a boundedness assumption on the feature map. This relaxation leads to new technical difficulties that we resolve in the following. We start our analysis from a decomposition similar to Chatalic et al. [2022, Lemma 4.1]; the difference is that we introduce the centered covariance operator $C_{\mathbb{Q},\bar{h}_p,\lambda}$ which allows us to handle both $\mathbb{P} = \mathbb{Q}$ and the challenging case of $\mathbb{P} \neq \mathbb{Q}$ in a unified fashion.

To simplify notation, let $\mu_{h_p} := \mu_{h_p}(\mathbb{Q})$, $\hat{\mu}_{h_p} := \mu_{h_p}\left(\hat{\mathbb{Q}}_n\right)$, and $\hat{\mu}_{h_p}^{\mathrm{Nys}} := P_{\mathcal{H}_{h_p,m}} \mu_{h_p}\left(\hat{\mathbb{Q}}_n\right)$. First, we decompose the error as follows.

$$\begin{aligned}
\left|S_p(\mathbb{Q}) - \tilde{S}_p\left(\hat{\mathbb{Q}}_n\right)\right| &\overset{(a)}{=} \left|\left\|\mu_{h_p}\right\|_{\mathcal{H}_{h_p}} - \left\|\hat{\mu}^{\text{Nys}}_{h_p}\right\|_{\mathcal{H}_{h_p}}\right| \overset{(b)}{\leq} \left\|\mu_{h_p} - \hat{\mu}^{\text{Nys}}_{h_p}\right\|_{\mathcal{H}_{h_p}} \\
&\overset{(c)}{\leq} \left\|\mu_{h_p} - \hat{\mu}_{h_p}\right\|_{\mathcal{H}_{h_p}} + \left\|\hat{\mu}_{h_p} - \hat{\mu}^{\text{Nys}}_{h_p}\right\|_{\mathcal{H}_{h_p}} \\
&\overset{(d)}{=} \left\|\mu_{h_p} - \hat{\mu}_{h_p}\right\|_{\mathcal{H}_{h_p}} + \left\|\left(I - P_{\mathcal{H}_{h_p,m}}\right)\left(\hat{\mu}_{h_p} - \frac{1}{m}\sum_{i=1}^{m} h_p\left(\cdot, \tilde{\mathbf{x}}_i\right)\right)\right\|_{\mathcal{H}_{h_p}} \\
&\overset{(e)}{\leq} \underbrace{\left\|\mu_{h_p} - \hat{\mu}_{h_p}\right\|_{\mathcal{H}_{h_p}}}_{=:t_1} + \underbrace{\left\|\left(I - P_{\mathcal{H}_{h_p,m}}\right) C^{1/2}_{\mathbb{Q},\bar{h}_p,\lambda}\right\|_{\text{op}}}_{=:t_2} \underbrace{\left\|C^{-1/2}_{\mathbb{Q},\bar{h}_p,\lambda}\left(\hat{\mu}_{h_p} - \frac{1}{m}\sum_{i=1}^{m} h_p\left(\cdot, \tilde{\mathbf{x}}_i\right)\right)\right\|_{\mathcal{H}_{h_p}}}_{=:t_3}. \qquad (5.11)
\end{aligned}$$

(a) is implied by (5.10) and (5.2); (b) follows from the reverse triangle inequality; $\pm\hat{\mu}_{h_p}$ and the triangle inequality yield (c); in (d), we use the distributive property and introduce $\frac{1}{m}\sum_{i=1}^{m} h_p\left(\cdot, \tilde{\mathbf{x}}_i\right) \in \mathcal{H}_{h_p,m}$ whose projection onto the orthogonal complement of $\mathcal{H}_{h_p,m}$ is zero; in (e) $I = C^{1/2}_{\mathbb{Q},\bar{h}_p,\lambda} C^{-1/2}_{\mathbb{Q},\bar{h}_p,\lambda}$ was introduced and we used the definition of the operator norm.

We next obtain individual probabilistic bounds for the three terms $t_1$, $t_2$, and $t_3$, which we subsequently combine by union bound. We will then conclude the proof by showing that all assumptions that we imposed along the way are satisfied.

- **Term $t_1$.** The first term measures the deviation of an empirical mean $\hat{\mu}_{h_p}$ to its population counterpart $\mu_{h_p}$. To bound this deviation $\left\|\hat{\mu}_{h_p} - \mu_{h_p}\right\|_{\mathcal{H}_{h_p}} = \left\|\frac{1}{n}\sum_{i=1}^{n} \bar{h}_p\left(\cdot, \mathbf{x}_i\right)\right\|_{\mathcal{H}_{h_p}}$, we will use the Bernstein inequality (Theorem A.3.2) with the $\eta_i := \bar{h}_p\left(\cdot, \mathbf{x}_i\right) \in \mathcal{H}_{h_p}$ $(i \in [n])$ centered random variables, by gaining control on the moments of $Y := \left\|\bar{h}_p\left(\cdot, X\right)\right\|_{\mathcal{H}_{h_p}}$. This is what we elaborate next.
  By Assumption 5.2.1 and Lemma 5.5.3, $Y$ is sub-Gaussian; hence, by Lemma A.3.2(3) it is also sub-exponential, and therefore (Lemma 5.5.2) it satisfies the Bernstein condition

$$\mathbb{E}|Y|^p \leq \frac{1}{2} p! \sigma^2 B^{p-2} < \infty, \quad \text{with} \qquad \sigma = \sqrt{2}\left\|Y\right\|_{\psi_1}, \qquad B = \left\|Y\right\|_{\psi_1},$$

  for any $p \geq 2$. Notice that $B = \left\|Y\right\|_{\psi_1} \overset{(a)}{\lesssim} \left\|Y\right\|_{\psi_2} \overset{(b)}{\lesssim} \sqrt{\text{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}$. (a) follows from Lemma A.3.2(3) and (b) is implied by Lemma 5.5.3. As $\sigma \asymp B$, we also got that $\sigma \lesssim \sqrt{\text{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}$.
  Having obtained a bound on the moments, we can apply Bernstein's inequality for separable Hilbert spaces (Yurinsky 1995; recalled in Theorem A.3.2) to the centered $\eta_i = \bar{h}_p\left(\cdot, \mathbf{x}_i\right) \in \mathcal{H}_{h_p}$-s $(i \in [n])$, and obtain that for any $\delta \in (0,1)$ it holds that

$$\mathbb{Q}^n\left(\underbrace{\left\|\mu_{h_p} - \hat{\mu}_{h_p}\right\|_{\mathcal{H}_{h_p}}}_{\left(=\left\|\frac{1}{n}\sum_{i=1}^{n}\eta_i\right\|_{\mathcal{H}_{h_p}}\right)} \lesssim \frac{\sqrt{\text{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}\log(6/\delta)}{n} + \sqrt{\frac{\text{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)\log(6/\delta)}{n}}\right) \geq 1 - \delta/3. \qquad (5.12)$$

  Note that (5.12) also holds with the measure $\mathbb{Q}^n \otimes \Lambda^m$, since the event considered in (5.12) has no randomness w.r.t. $\Lambda^m$.

- **Term $t_2$.** Assume that $0 < \lambda \leq \left\| C_{\mathbb{Q},\bar{h}_p} \right\|_{\mathrm{op}}$. Then, we can handle the second term with Lemma 5.5.1 and obtain that for any $\delta \in (0,1)$ it holds that

$$(\mathbb{Q}^n \otimes \Lambda^m)\left( \left\| \left(I - P_{\mathcal{H}_{h_p,m}}\right) C^{1/2}_{\mathbb{Q},\bar{h}_p,\lambda} \right\|_{\mathrm{op}} \lesssim \sqrt{\lambda} \right) \geq 1 - \delta/3 \tag{5.13}$$

provided that $m \gtrsim \max\left\{ \frac{\mathrm{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}{\lambda}, 1 \right\} \log(12/\delta)$.

- **Term $t_3$.** The third term depends on the sample $(\mathbf{x}_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{Q}$ and on the Nyström selection $(i_j)_{j=1}^m \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([n]) =: \Lambda$; with this notation $\tilde{\mathbf{x}}_j = \mathbf{x}_{i_j}$ $(j \in [m])$. Our goal is to take both sources of randomness into account.
**Fixed $\mathbf{x}_i$-s, randomness in $i_j$-s:** Let the sample $(\mathbf{x}_i)_{i=1}^n$ be fixed. As the $(\mathbf{x}_{i_j})_{j=1}^m$-s are i.i.d.,

$$t_3 = \left\| C^{-1/2}_{\mathbb{Q},\bar{h}_p,\lambda} \left( \hat{\mu}_{h_p} - \frac{1}{m} \sum_{i=1}^m h_p\left(\cdot, \tilde{\mathbf{x}}_i\right) \right) \right\|_{\mathcal{H}_{h_p}} = \left\| \frac{1}{m} \sum_{i=1}^m \underbrace{\left[ C^{-1/2}_{\mathbb{Q},\bar{h}_p,\lambda} \left( h_p\left(\cdot, \tilde{\mathbf{x}}_i\right) - \hat{\mu}_{h_p} \right) \right]}_{=: Y_i} \right\|_{\mathcal{H}_{h_p}}$$

measures the concentration of the sum $\frac{1}{m}\sum_{i=1}^m Y_i$ around its expectation, which is zero as one has that $\mathbb{E}_J\left[h_p(\cdot, \mathbf{x}_J)\right] = \hat{\mu}_{h_p}$ with $J \sim \Lambda$. Notice that

$$\begin{aligned}
\max_{i\in[m]} \|Y_i\|_{\mathcal{H}_{h_p}} &= \max_{i\in[m]} \left\| C^{-1/2}_{\mathbb{Q},\bar{h}_p,\lambda} \left( h_p\left(\cdot, \tilde{\mathbf{x}}_i\right) - \hat{\mu}_{h_p} \right) \right\|_{\mathcal{H}_{h_p}} \\
&= \max_{i\in[m]} \left\| C^{-1/2}_{\mathbb{Q},\bar{h}_p,\lambda} \left( h_p\left(\cdot, \tilde{\mathbf{x}}_i\right) - \mathbb{E}_{X\sim\mathbb{Q}} h_p(\cdot, X) + \mathbb{E}_{X\sim\mathbb{Q}} h_p(\cdot, X) - \hat{\mu}_{h_p} \right) \right\|_{\mathcal{H}_{h_p}} \\
&\leq \max_{i\in[m]} \left\| C^{-1/2}_{\mathbb{Q},\bar{h}_p,\lambda} \Big( \underbrace{h_p\left(\cdot, \tilde{\mathbf{x}}_i\right) - \mathbb{E}_{X\sim\mathbb{Q}} h_p(\cdot, X)}_{=\bar{h}_p(\cdot, \tilde{\mathbf{x}}_i)} \Big) \right\|_{\mathcal{H}_{h_p}} + \left\| C^{-1/2}_{\mathbb{Q},\bar{h}_p,\lambda} \left( \hat{\mu}_{h_p} - \mathbb{E}_{X\sim\mathbb{Q}} h_p(\cdot, X) \right) \right\|_{\mathcal{H}_{h_p}} \\
&\leq \max_{i\in[n]} \left\| C^{-1/2}_{\mathbb{Q},\bar{h}_p,\lambda} \bar{h}_p(\cdot, \mathbf{x}_i) \right\|_{\mathcal{H}_{h_p}} + \left\| C^{-1/2}_{\mathbb{Q},\bar{h}_p,\lambda} \left( \hat{\mu}_{h_p} - \mathbb{E}_{X\sim\mathbb{Q}} h_p(\cdot, X) \right) \right\|_{\mathcal{H}_{h_p}} \\
&\leq \max_{i\in[n]} \left\| C^{-1/2}_{\mathbb{Q},\bar{h}_p,\lambda} \bar{h}_p(\cdot, \mathbf{x}_i) \right\|_{\mathcal{H}_{h_p}} + \frac{1}{n} \sum_{i\in[n]} \left\| C^{-1/2}_{\mathbb{Q},\bar{h}_p,\lambda} \Big( \underbrace{h_p(\cdot, \mathbf{x}_i) - \mathbb{E}_{X\sim\mathbb{Q}} h_p(\cdot, X)}_{=\bar{h}_p(\cdot, \mathbf{x}_i)} \Big) \right\|_{\mathcal{H}_{h_p}} \\
&\leq 2 \max_{i\in[n]} \left\| C^{-1/2}_{\mathbb{Q},\bar{h}_p,\lambda} \bar{h}_p(\cdot, \mathbf{x}_i) \right\|_{\mathcal{H}_{h_p}} =: K = K(\mathbf{x}_1, \ldots, \mathbf{x}_n),
\end{aligned}$$

where we used that $\pm\mathbb{E}_{X\sim\mathbb{Q}} h_p(\cdot, X) = 0$, the triangle inequality, and the homogeneity of the norm. An application of Theorem A.3.3 yields that, conditioned on the sample $(\mathbf{x}_i)_{i=1}^n$, it holds that

$$\Lambda^m\left( (i_j)_{j=1}^m \,:\, t_3 \leq K \frac{\sqrt{2\log(12/\delta)}}{\sqrt{m}} \;\middle|\; (\mathbf{x}_i)_{i=1}^n \right) \geq 1 - \frac{\delta}{6}. \tag{5.14}$$

**Randomness in $\mathbf{x}_i$-s:** Let $Z_i := \left\| C^{-1/2}_{\mathbb{Q},\bar{h}_p,\lambda} \bar{h}_p(\cdot, \mathbf{x}_i) \right\|_{\mathcal{H}_{h_p}}$ $(i \in [n])$ with $(\mathbf{x}_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{Q}$. By Assumption 5.2.1 and Lemma 5.5.3, the $Z_i$-s are sub-Gaussian random variables. Hence, by Lemma 5.5.5, with probability at least $1 - \delta/6$, it holds that

$$K = 2 \max_{i\in[n]} |Z_i| \lesssim \sqrt{\|Z_1\|^2_{\psi_2} \log(12n/\delta)}.$$

By Lemma 5.5.3, $\|Z_1\|_{\psi_2}^2 \lesssim \operatorname{tr}\left(C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1} C_{\mathbb{Q},\bar{h}_p}\right)$. We have shown that

$$\mathbb{Q}^n\left((\mathbf{x}_i)_{i=1}^n \,:\, K \lesssim \sqrt{\operatorname{tr}\left(C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1} C_{\mathbb{Q},\bar{h}_p}\right)\log(12n/\delta)}\right) \geq 1 - \frac{\delta}{6}. \tag{5.15}$$

**Combination:** We now combine the intermediate results. Let

$$A = \left\{ \left((\mathbf{x}_i)_{i=1}^n, (i_j)_{j=1}^m\right) \,:\, t_3 \lesssim \frac{\sqrt{\operatorname{tr}\left(C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1} C_{\mathbb{Q},\bar{h}_p}\right)\log(12n/\delta)\log(12/\delta)}}{\sqrt{m}} \right\},$$

$$B = \left\{ (\mathbf{x}_i)_{i=1}^n \,:\, K \lesssim \sqrt{\operatorname{tr}\left(C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1} C_{\mathbb{Q},\bar{h}_p}\right)\log(12n/\delta)} \right\},$$

$$C = \left\{ \left((\mathbf{x}_i)_{i=1}^n, (i_j)_{j=1}^m\right) \,:\, t_3 \leq K\frac{\sqrt{2\log(12/\delta)}}{\sqrt{m}}, (\mathbf{x}_i)_{i=1}^n \in B \right\} \subseteq A.$$

Then, with $\mathbb{Q}^n \otimes \Lambda^m$ denoting the product measure of $\mathbb{Q}^n$ and $\Lambda^m$, we obtain

$$\begin{aligned}
(\mathbb{Q}^n \otimes \Lambda^m)(A) &= \mathbb{E}_{\mathbb{Q}^n}\left[\Lambda^m\left(A \mid (\mathbf{x}_i)_{i=1}^n\right)\right] = \int_{\left(\mathbb{R}^d\right)^n} \Lambda^m\left(A \mid (\mathbf{x}_i)_{i=1}^n\right) \mathrm{d}\mathbb{Q}^n(\mathbf{x}_1, \ldots, \mathbf{x}_n) \\
&\geq \int_B \Lambda^m\left(A \mid (\mathbf{x}_i)_{i=1}^n\right) \mathrm{d}\mathbb{Q}^n(\mathbf{x}_1, \ldots, \mathbf{x}_n) \geq \int_B \Lambda^m\left(C \mid (\mathbf{x}_i)_{i=1}^n\right) \mathrm{d}\mathbb{Q}^n(\mathbf{x}_1, \ldots, \mathbf{x}_n) \\
&\overset{(a)}{\geq} \left(1 - \frac{\delta}{6}\right)\mathbb{Q}^n(B) \overset{(b)}{\geq} (1 - \delta/6)^2 = 1 - \delta/3 + \delta^2/6^2 > 1 - \delta/3.
\end{aligned} \tag{5.16}$$

(a) is implied by the uniform lower bound established in (5.14). (b) was shown in (5.15).

**Combination of $t_1$, $t_2$, and $t_3$.** To conclude, we use decomposition (5.11), and union bound (5.12), (5.13), and (5.16). Further, we observe that $\operatorname{tr}\left(C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1} C_{\mathbb{Q},\bar{h}_p}\right) = \mathcal{N}_{\mathbb{Q},\bar{h}_p}(\lambda)$, and obtain that

$$\begin{aligned}
(\mathbb{Q}^n \otimes \Lambda^m)\Bigg(\left|S_p(\mathbb{Q}) - \tilde{S}_p\left(\hat{\mathbb{Q}}_n\right)\right| \lesssim \frac{\sqrt{\operatorname{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)\log(6/\delta)}}{n} + \sqrt{\frac{\operatorname{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)\log(6/\delta)}{n}} + \\
+ \sqrt{\frac{\lambda\mathcal{N}_{\mathbb{Q},\bar{h}_p}(\lambda)\log(12n/\delta)\log(12/\delta)}{m}}\Bigg) \geq 1 - \delta
\end{aligned}$$

provided that $m \gtrsim \max\left\{\frac{\operatorname{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}{\lambda}, 1\right\}\log(12/\delta)$ and $0 < \lambda \leq \left\|C_{\mathbb{Q},\bar{h}_p}\right\|_{\mathrm{op}}$ both hold. Now, specializing $\lambda = \frac{c\operatorname{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}{m}$ for some absolute constant $c > 1$, all constraints are satisfied for

$$m \gtrsim \max\left\{\log(12/\delta), \operatorname{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)\left\|C_{\mathbb{Q},\bar{h}_p}\right\|_{\mathrm{op}}^{-1}\right\}.$$

Using our choice of $\lambda$, after rearranging, we get the stated claim.

### 5.5.4. Proof of Corollary 5.2.1

The proof is split into two parts. The first one considers the polynomial decay assumption, the second one is about the exponential decay assumption.

- **Polynomial decay.** The $\sqrt{n}$-consistency of the first two addends in Theorem 5.2.2 was established in the discussion following the statement. Hence, we limit our considerations to the last addend. Assume that $\mathcal{N}_{\mathbb{Q},\bar{h}_p}(\lambda) \lesssim \lambda^{-\gamma}$ for $\gamma \in (0,1]$. Observing that the trace expression is constant, the last addend in Theorem 5.2.2 yields that

$$\sqrt{\frac{\log(12/\delta)\log(12n/\delta)\mathcal{N}_{\mathbb{Q},h_p}\left(\frac{c\,\mathrm{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}{m}\right)}{m^2}} \overset{(a)}{\lesssim} \sqrt{\frac{\log(12/\delta)\log(12n/\delta)}{m^{2-\gamma}}} \overset{(b)}{=} O\left(\frac{1}{\sqrt{n}}\right),$$

  with (a) implied by the polynomial decay assumption and (b) follows from our choice of $m \gtrsim n^{\frac{1}{2-\gamma}}\log^{\frac{1}{2-\gamma}}(12n/\delta)\log^{\frac{1}{2-\gamma}}(12/\delta)$. This derivation confirms the first stated result.

- **Exponential decay.** Assume it holds that $\mathcal{N}_{\mathbb{Q},\bar{h}_p}(\lambda) \lesssim \log(1 + c_1/\lambda)$. Observe that as per the discussion following Theorem 5.2.2, the first two addends are $O\left(n^{-1/2}\right)$. For the last addend, again noticing that the trace is constant, we have

$$\sqrt{\frac{\log(12/\delta)\log(12n/\delta)\mathcal{N}_{\mathbb{Q},\bar{h}_p}\left(\frac{c\,\mathrm{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}{m}\right)}{m^2}} \overset{(a)}{\lesssim} \sqrt{\frac{\log(12/\delta)\log(12n/\delta)\log\left(1+\frac{c_1 m}{c\,\mathrm{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}\right)}{m^2}}$$
$$\overset{(b)}{\lesssim} \sqrt{\frac{\log(12/\delta)\log(12n/\delta)\log\left(1+\frac{c_1 n}{c\,\mathrm{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}\right)}{m^2}} \overset{(c)}{=} O\left(\frac{1}{\sqrt{n}}\right),$$

  where (a) uses the exponential decay assumption. (b) uses that $n \geq m$ and that the logarithm is a monotonically increasing function. (c) follows from our choice of

$$m \gtrsim \sqrt{n}\sqrt{\log\left(1+\frac{c_1 n}{c\,\mathrm{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}\right)\log(12n/\delta)\log(12/\delta)},$$

  finishing the proof of the corollary.

### 5.5.5. Proof of Theorem 5.2.3

By the reverse triangle inequality, we obtain

$$\left|S_p(\mathbb{Q}) - S_p\left(\hat{\mathbb{Q}}_n\right)\right| \leq \left\|\mu_{h_p}(\mathbb{Q}) - \mu_{h_p}\left(\hat{\mathbb{Q}}_n\right)\right\|_{\mathcal{H}_{h_{\mathbb{P}}}} = \left\|\frac{1}{n}\sum_{i=1}^{n}\underbrace{\left[h_p(\cdot,X_i) - \mathbb{E}_{X\sim\mathbb{Q}}h_p(\cdot,X)\right]}_{=:\eta_i}\right\|_{\mathcal{H}_{h_p}},$$

which measures the concentration of i.i.d. centered random variables. To obtain the bound, we will use Bernstein's inequality (recalled in Theorem A.3.2) by gaining control on the moments of $\|\eta_i\|_{\mathcal{H}_{h_p}}$ with Lemma 5.5.2.

First, note that the $\|\eta_i\|_{\mathcal{H}_{h_p}}$-s ($i \in [n]$) are sub-Gaussian as

$$\left\|\|\eta_i\|_{\mathcal{H}_{h_p}}\right\|_{\psi_2} \overset{(a)}{=} \left\|\left\|h_p(\cdot, X_i) - \mathbb{E}_{X\sim\mathbb{Q}} h_p(\cdot, X)\right\|_{\mathcal{H}_{h_p}}\right\|_{\psi_2} \overset{(b)}{\leq} \left\|\left\|h_p(\cdot, X_i)\right\|_{\mathcal{H}_{h_p}} + \left\|\mathbb{E}_{X\sim\mathbb{Q}} h_p(\cdot, X)\right\|_{\mathcal{H}_{h_p}}\right\|_{\psi_2}$$
$$\overset{(c)}{\leq} \left\|\left\|h_p(\cdot, X_i)\right\|_{\mathcal{H}_{h_p}} + \mathbb{E}_{X\sim\mathbb{Q}}\left\|h_p(\cdot, X)\right\|_{\mathcal{H}_{h_p}}\right\|_{\psi_2} \overset{(d)}{\lesssim} \left\|\left\|h_p(\cdot, X_i)\right\|_{\mathcal{H}_{h_p}}\right\|_{\psi_2} < \infty.$$

We use the definition of $\eta_i$ in (a). (b) is implied by the triangle inequality and the monotonicity of the norm. (c) is by Jensen's inequality holding for Bochner integrals, and (d) follows from Lemma A.3.2(1); finiteness is due to the imposed assumption.

Hence, $\|\eta_i\|_{\mathcal{H}_{h_p}}$ is sub-exponential (Lemma A.3.2(3)), and, by Lemma 5.5.2, it holds for any $p \geq 2$ that

$$\mathbb{E}_{X\sim\mathbb{Q}} \|\eta_i\|^p_{\mathcal{H}_{h_p}} \leq \frac{1}{2} p! \sigma^2 B^{p-2},$$

with $\sigma, B \lesssim \left\|\|\eta_i\|_{\mathcal{H}_{h_p}}\right\|_{\psi_1} =: K$. Now, applying Theorem A.3.2 yields that, for any $\delta \in (0,1)$, it holds with probability at least $1-\delta$ that

$$\left\|\frac{1}{n}\sum_{i=1}^n \eta_i\right\|_{\mathcal{H}_{h_p}} \lesssim \frac{2K\log(2/\delta)}{n} + \sqrt{\frac{2K^2\log(2/\delta)}{n}},$$

which is the stated claim.

# 6. Maximum Mean Discrepancy on Exponential Windows for Online Change Detection

The content of this chapter is based on the following publication.

- F. Kalinke, M. Heyden, G. Gntuni, E. Fouché, and K. Böhm. Maximum mean discrepancy on exponential windows for online change detection. *Transactions on Machine Learning Research*, 2025a. TMLR.

The code replicating all experiments is available at github.com/flopska/mmdew-change-detector.

## 6.1. Introduction

In this chapter, we shift from the offline to the online setting. In particular, we consider data streams, which are possibly infinite sequences of observations that arrive over time. Such streams can have different sources: sensors in industrial settings, online transactions from financial institutions, click monitoring on websites, or online feeds, among others. Quickly detecting when a change takes place can yield useful insights, for example, about machine failure, malicious financial transactions, changes in customer preferences, and public opinions.

A change occurs if the underlying distribution of the data stream changes at a certain point in time. We call this moment change point [Gama, 2010]; it is sometimes also referred to as concept drift. Change detection is an unsupervised task that has received and still is receiving a lot of interest. The earliest approaches, for example, Shewhart [1925], Page [1954], originated from quality control and require strong parametric assumptions on the pre and post-change distributions. More recent work in the parametric regime weakens these assumptions by allowing post-change distributions from a parametric family with an unknown parameter [Lorden, 1970, Siegmund and Venkatraman, 1995] or by allowing any post-change distribution [Sparks, 2000, Lorden and Pollak, 2005, Abbasi and Haq, 2019, Xie et al., 2023].

In our work, both the pre and post-change distribution are assumed to be unknown, which is a challenging setting that can be tackled with non-parametric approaches. We detail the approaches most related to our proposed method in the following and refer to Wang and Xie [2024] for a recent more extensive survey on parametric and non-parametric change detection methods.

A principled and widely-used approach to detect changes in a non-parametric fashion is to use two-sample tests. The null hypothesis of such tests is that the data before and after the potential change point follow the same distribution. If the test rejects the hypothesis, one assumes that a change occurred. One way to construct two-sample tests, applicable on many domains, is using MMD, as detailed in Section 2.3.1. Recall that estimating MMD with classical estimators for two data sets of sizes $m$ and $n$, respectively, costs $O(m^2 + n^2)$, with a memory complexity in $O(m + n)$. Naively computing MMD for each possible change point on a data stream with $t = m + n$ observations has a complexity in $O(t^3)$ for each new observation. These properties render the direct application of MMD to change detection in

**Table 6.1.:** Comparison of change detectors. Complexity — runtime complexity per new observation, ARL / MTD — type of known results, domain — data types, $t$ — total number of observations, $d$ — dimensionality (for Euclidean spaces), $k$ — parameter, $W$ — window length / block size, $N$ — number of windows.

| Change Detection Algorithm | Runtime Complexity | ARL / MTD | Domain |
|---|---|---|---|
| ADWINK [Faithfull et al., 2019] | $O(dk \log W)$ | empirical | $\mathbb{R}^d$ |
| WATCH [Faber et al., 2021] | unknown[a] | empirical | $\mathbb{R}^d$ |
| Scan $B$-statistics [Li et al., 2019] | $O(NW^2)$ | analytical | any |
| NEWMA [Keriven et al., 2020] | $O(md)$[b] | analytical | $\mathbb{R}^d$ |
| D3 [Gözüaçik et al., 2019] | $O(W^3)$[c] | none | $\mathbb{R}^d$ |
| IBDD [de Souza et al., 2021] | $O(pq)$[d] | none | $\mathbb{R}^d$ |
| **MMDEW** | $O(\log^2 t)$ | empirical | any |

data streams impractical. However, there exist methods to compute MMD in the streaming setting, for example, linear time tests [Gretton et al., 2012], but their statistical power is low. Zaremba et al. [2013] introduce $B$-tests, which have higher power. However, both can not directly be used for change detection. Li et al. [2019] enable the estimation of MMD on data streams for change detection by introducing Scan $B$-statistics. Wei and Xie [2022] extend upon their work by considering multiple Scan $B$-statistics in parallel and introduce online kernel CUSUM. Another method enabling the computation of MMD on data streams is NEWMA [Keriven et al., 2020], which is based on random Fourier features (RFFs; Rahimi and Recht 2007, Sriperumbudur and Szabó 2015). Harchaoui and Cappé [2007] apply kernel-based tests for offline change point detection on audio and brain-computer-interface data.

Non-kernel-based change detection algorithms include the classic ADWIN [Bifet and Gavaldà, 2007], which is limited to univariate data and only detects changes in the mean. ADWINK [Faithfull et al., 2019] alleviates the former by running one instance of ADWIN per dimension and issues a change if a predefined number of the instances agree that a change occurred. Hence, the approach can only detect changes in the means of the marginal distributions and changes in higher moments or the covariance structure can not be detected. WATCH [Faber et al., 2021] is a recent approach that uses a two-sample test based on the Wasserstein distance. The method by Dasu et al. [2009] is conceptually similar to our method, as it also uses two-sample tests and is non-parametric, but it relies on the Kullback-Leibler divergence. A conceptually different approach to find changes is using classifiers. D3 [Gözüaçik et al., 2019] maintains two consecutive sliding windows and trains a classifier to distinguish their elements. It reports a change if the classifier performance, measured by AUC, drops below a threshold. Another recent algorithm is IBDD [de Souza et al., 2021], which scales well with the number of features.

We summarize the main properties of related approaches in Table 6.1.[1] We consider the dimensionality $d$ as constant for the complexities where its influence is dominated by other terms and for approaches not restricted to Euclidean domains.

In this chapter, we introduce Maximum Mean Discrepancy on Exponential Windows (MMDEW), a change detection algorithm for data streams that solves the quadratic-time bottleneck of MMD. Specifically, our **contributions** include the following.

- Our main contribution is MMDEW, a change detector based on an efficient online approximation of MMD. When considering the entire history of $t$ observations, the proposed method has a memory

[1] [a]We refer to their used implementation of the Wasserstein distance computation and the discussion therein [Mérigot, 2011, Ch. 6]. [b]$m$ is the number of random Fourier features and $m \ll d$. [c]The complexity results from the matrix inversion of the logistic regression model, which has cubic runtime cost in practice. [d]Size of the constructed $q \times p$ image.

requirement of $O(\log t)$ and a runtime complexity of $O(\log^2 t)$ for each new observation. Otherwise, the algorithm has constant runtime and memory requirements.

- To achieve these complexities, we introduce a new data structure, which allows to approximate the quadratic-time MMD in an online setting. We accomplish the speedup by introducing windows that store summaries of the observations seen so far, and by storing a sample of logarithmic size of the observations per window.
- Our experiments on standard benchmark data sets show that MMDEW performs better than state-of-the-art change detectors on four out of the five tested data sets using the $F_1$-score. For the more challenging setting of short detection delays, the proposed algorithm is better on three out of five data sets.

The remainder of this chapter is structured as follows. We state the problem in Section 6.2 and detail our proposed solution in Section 6.3. Our experiments are in Section 6.4. We include illustrative proofs in the main text and defer technical proofs to Section 6.5.

## 6.2. Problem definition

The problem setting is as follows. We consider a data stream, that is, a possibly infinite sequence of observations, $x_1, x_2, \ldots, x_t, \ldots$ for $t = 1, 2, \ldots$, and $x_t \in \mathcal{X}$. Each $x_t$ is generated independently following some distribution $D_t \in \mathcal{M}_1^+(\mathcal{X})$. If there exists $t^*$ such that for $i < t^*$ and $j \geq t^*$ we have $D_i \neq D_j$, then $t^*$ is a change point, and our task is to detect it; in practice, a $D_t$ typically generates a range of i.i.d. observations. We note that these definitions place few assumptions on the type of data, that is, we only require the data to reside in a topological space.

## 6.3. Proposed algorithm

We introduce MMDEW in three steps. We first extend the threshold for the MMD two-sample test, (2.14), to samples of unequal sizes (Section 6.3.1). We then introduce our data structure that enables the efficient computation of MMD on data streams (Section 6.3.2). Last, we describe the complete algorithm in Section 6.3.3.

### 6.3.1. Threshold for the hypothesis test

Given a sequence of observations $\{x_1, \ldots, x_t\}$ up until time $t$ our goal is to test the null hypothesis $\mathbb{P} = \mathbb{Q}$ for any two neighboring windows $X \cdot Y = \{x_1, \ldots, x_i\} \cdot \{x_{i+1}, \ldots, x_t\}$, with $i = 1, \ldots, t-1$. Our following proposition extends Gretton et al. [2012, Theorem 8] (recalled in (2.14)), which considers the case $m = n$, giving the distribution-free acceptance region for $m \neq n$ (corresponding to the setting that one generally encounters in change detection).

**Proposition 6.3.1.** *Let $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_1^+(\mathcal{X})$, $\hat{\mathbb{P}}_m = \{x_1, \ldots, x_m\} \overset{i.i.d.}{\sim} \mathbb{P}$, $\hat{\mathbb{Q}}_n = \{y_1, \ldots, y_n\} \overset{i.i.d.}{\sim} \mathbb{Q}$. Assume that $0 \leq k(x, y) \leq K$ for all $x, y \in \mathcal{X}$ and $t > 0$. Then a hypothesis test of level at most $\alpha > 0$ for $\mathbb{P} = \mathbb{Q}$ has the acceptance region*

$$\mathrm{MMD}_k\left(\hat{\mathbb{P}}_m, \hat{\mathbb{Q}}_n\right) < \sqrt{\frac{K}{m} + \frac{K}{n}}\left(1 + \sqrt{2\log\alpha^{-1}}\right) =: \epsilon_\alpha.$$

Note that, when considering multiple possible change points, one needs to account for multiple testing in order to achieve an overall level of size $\alpha$. For example, one may adjust $\epsilon_\alpha$ through Bonferroni correction ($\epsilon'_\alpha = \epsilon_\alpha/(t-1)$) by dividing by the total number of tests.

To perform change detection with MMD, it is natural to consider the stopping time

$$T = \inf\left\{t : \max_{n=1,\ldots,t-1}\left[\mathrm{MMD}_k\left(\hat{\mathbb{P}}_m, \hat{\mathbb{Q}}_n\right) \geq \epsilon_\alpha\right] = 1\right\}, \tag{6.1}$$

for $m = t - n$, the empirical measures $\hat{\mathbb{P}}_m = \{x_1, \ldots, x_m\}$, $\hat{\mathbb{Q}}_n = \{x_{m+1}, \ldots, x_t\}$, and the brackets equal to one if their argument is true and zero otherwise [Graham et al., 1994]; we note that $\epsilon_\alpha := \epsilon_\alpha(m, n)$ depends on the respective sizes of the subsamples considered. In other words, a change is indicated by the first time any MMD estimated across all splits exceeds its threshold. However, due to the quadratic runtime requirements of MMD, the computation of (6.1) costs $O\left(t^3\right)$ for each new observation.

We now introduce our novel data structure that allows considering multiple possible change points efficiently.

### 6.3.2. Data structure

One common method to obtain a good runtime complexity in change detection algorithms is to slice the data into windows of exponentially increasing sizes [Bifet and Gavaldà, 2007]. Recent observations are collected in smaller windows, and older observations are grouped into larger windows. This leads to a fine-grained change detection in the recent past and more coarse-grained change detection in the distant past.

Our new data structure adopts this concept and, at the same time, facilitates the computation of MMD. In what follows, we first describe the properties of the proposed data structure. Then, we show how to update the data structure and explain its use for change detection.

#### 6.3.2.1. Properties

We use 2 as the basis for the exponential slicing. Then, after observing $t$ elements, the number of windows stored in the data structure corresponds to the number of ones in the binary representation of $t$. We may thus index the windows as $B_l, \ldots, B_0$ (in decreasing order), with the largest position being $l = \lfloor \log_2 t \rfloor$. A window does not exist if the binary representation of $t$ at this position is zero.

If it exists, a window $B_s = (\mathrm{X}_s, \mathrm{XX}_s, \mathrm{XY}_s)$ at position $s = 0, \ldots, l$ stores $2^s$ observations

$$\mathrm{X}_s = \left\{x_1^{(s)}, \ldots, x_{2^s}^{(s)}\right\}, \tag{6.2}$$

together with the summaries

$$\mathrm{XX}_s = \sum_{i,j=1}^{2^s} k\left(x_i^{(s)}, x_j^{(s)}\right), \tag{6.3}$$

$$\mathrm{XY}_s = \Bigg\{\underbrace{\sum_{i=1}^{2^s}\sum_{j=1}^{2^{s+1}} k\left(x_i^{(s)}, x_j^{(s+1)}\right)}_{=:\mathrm{XY}_s^{s+1}}, \ldots, \underbrace{\sum_{i=1}^{2^s}\sum_{j=1}^{2^l} k\left(x_i^{(s)}, x_j^{(l)}\right)}_{=:\mathrm{XY}_s^{l}}\Bigg\}, \tag{6.4}$$

where $\mathrm{XX}_s \in \mathbb{R}$ is the sum of the kernel $k$ evaluated on all pairs of the window's own observations, and $\mathrm{XY}_s$ stores a list of sums of the kernel evaluated on the window's own observations and the observations in windows coming before it.[2] Storing a list enables the efficient merging of windows, elaborated in Lemma 6.3.2. The length of the list $\mathrm{XY}_s$ equals the number of windows having observations older than window $B_s$ and is at most $\lfloor \log_2 t \rfloor$. We use $\mathrm{XY}_i^j$ to represent the entry in $\mathrm{XY}_i$ that refers to the window $B_j$. Specifically, in (6.4), $\mathrm{XY}_s^{s+1}$ stores the interaction of $B_s$ with $B_{s+1}$; similarly, $\mathrm{XY}_s^l$ stores its interaction with $B_l$.

**Remark 6.3.1.** *Given a stream of data $x_1, x_2, \ldots, x_t$, (6.2) corresponds to the mapping $x_i^{(s)} = x_\ell$, with $\ell = \sum_{j=s+1}^{\lfloor \log_2 t \rfloor} 2^j [B_j \text{ exists}] + i$, where the bracket is one if the argument is true and zero otherwise (using Iverson's convention; Graham et al. 1994). A bucket $B_j$ exists if the $j$-th right-most digit in the binary expansion of $t$ is* 1.

We summarize two of the main properties of the data structure as lemmas. Lemma 6.3.1 establishes that one can compute the value of MMD between two windows with constant complexity. The proof follows from comparing (6.3) and (6.4) with (2.13). Lemma 6.3.2 shows that windows can be merged with logarithmic runtime complexity. These results provide our first steps towards efficiently computing MMD in a data stream.

**Lemma 6.3.1.** *Let $B_{s+1}$ and $B_s$ be any two neighboring windows with elements $\mathrm{X}_{s+1} = \left\{x_1^{(s+1)}, \ldots, x_{2^{s+1}}^{(s+1)}\right\}$ and $\mathrm{X}_s = \left\{x_1^{(s)}, \ldots, x_{2^s}^{(s)}\right\}$, and sums as defined by* (6.3) *and* (6.4)*, respectively. Then*

$$\mathrm{MMD}_k^2\left(\mathrm{X}_{s+1}, \mathrm{X}_s\right) = \frac{1}{(2^{s+1})^2}\mathrm{XX}_{s+1} + \frac{1}{(2^s)^2}\mathrm{XX}_s - \frac{2}{(2^{s+1})(2^s)}\mathrm{XY}_s^{s+1},$$

*with a computational complexity of $O(1)$.*

**Lemma 6.3.2.** *Merging two windows $B_{s+1}$ and $B_s$ into a new window $B'$, such that $B'$ stores* (6.2)*,* (6.3)*, and* (6.4) *costs $O(\log t)$.*

Besides showing the result, the proof of Lemma 6.3.2 illustrates the steps that allow merging windows efficiently.

*Proof.* For computing $\mathrm{XX}'$, we use the symmetry of $k$ to obtain

$$\begin{aligned}
\mathrm{XX}' &= \sum_{i,j=1}^{2^{s+1}} k\left(x_i^{(s+1)}, x_j^{(s+1)}\right) + \sum_{i,j=1}^{2^s} k\left(x_i^{(s)}, x_j^{(s)}\right) + \sum_{i=1}^{2^{s+1}}\sum_{j=1}^{2^s} k\left(x_i^{(s+1)}, x_j^{(s)}\right) + \sum_{i=1}^{2^s}\sum_{j=1}^{2^{s+1}} k\left(x_i^{(s)}, x_j^{(s+1)}\right) \\
&= \sum_{i,j=1}^{2^{s+1}} k\left(x_i^{(s+1)}, x_j^{(s+1)}\right) + \sum_{i,j=1}^{2^s} k\left(x_i^{(s)}, x_j^{(s)}\right) + 2\sum_{i=1}^{2^{s+1}}\sum_{j=1}^{2^s} k\left(x_i^{(s+1)}, x_j^{(s)}\right) \\
&= \mathrm{XX}_{s+1} + \mathrm{XX}_s + 2\mathrm{XY}_s^{s+1}, \qquad (6.5)
\end{aligned}$$

which has a runtime complexity in $O(1)$.

To compute $\mathrm{XY}'$, we note that $B_{s+1}$ stores the list $\mathrm{XY}_{s+1}$ of kernel evaluations corresponding to all windows coming before it. The same holds for $B_s$, for which the list has one more element, $\mathrm{XY}_s^{s+1}$,

[2] Note that the superscript $(s)$ of the $x_i^{(s)}$-s indicates the corresponding window $B_s$.

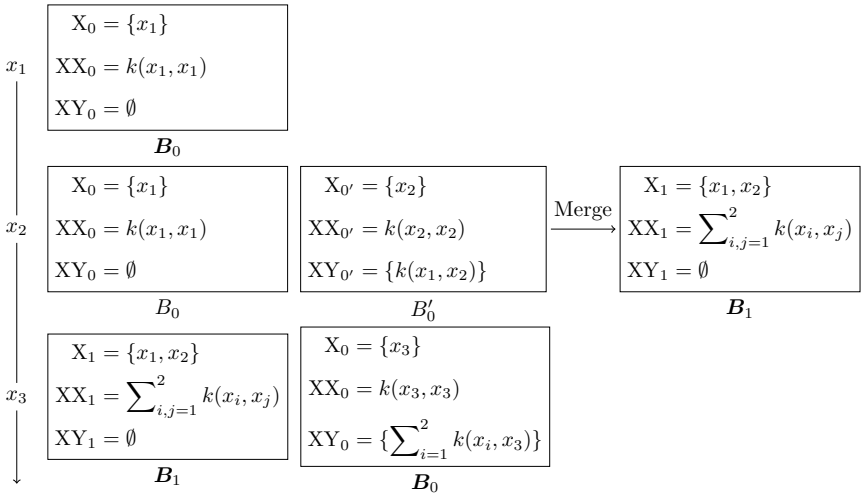


**Figure 6.1.:** Schematic representation of Example 6.3.1. For a given step, the proposed scheme stores the windows in bold face.

which was used in (6.5). All the elements in $\mathrm{XY}_s$ and $\mathrm{XY}_{s+1}$ are sums and thus additive; it suffices to merge both lists by adding their values element-wise, omitting $\mathrm{XY}_s^{s+1}$, and storing the result in $\mathrm{XY}'$. As each list has at most $\log t$ elements, merging them is in $O(\log t)$. □

Specifically, the scheme facilitates the merging of windows of equal size, enabling us to establish the exponential structure outlined in the next section.

#### 6.3.2.2. Insertion of observations

The structure is set up recursively. For each new observation, we create a new window $B_0$, with $\mathrm{XX}_0$ as defined by (6.3) and $\mathrm{XY}_0$ computed w.r.t. the already existing windows. If two windows have the same size, we merge them by Lemma 6.3.2, which costs $O(\log t)$. This yields $\lfloor \log t \rfloor$ windows of exponentially increasing sizes.

We illustrate the scheme in the following Example 6.3.1 and the corresponding Figure 6.1.

**Example 6.3.1.** *To set up the structure, we start with the first observation $x_1$ and create the first window $B_0$, with $\mathrm{XX}_0$ as defined by (6.3) and $\mathrm{XY}_0 = \emptyset$. When observing $x_2$, we similarly create a new window $B'_0$, now also computing $\mathrm{XY}_{0'}^0 = \{\mathrm{XY}_{0'}^0\}$. As $B_0$ and $B'_0$ have the same size, we merge them into $B_1$, computing $\mathrm{XX}_1$ with (6.5). No previous window exists so that $\mathrm{XY}_1 = \emptyset$. We repeat this for all new observations, for example, for $x_3$, one creates (a new) $B_0$, computing $\mathrm{XX}_0$ and $\mathrm{XY}_0 = \{\mathrm{XY}_0^1\}$, which results in two windows, $B_1$ and $B_0$.*

#### 6.3.2.3. MMD computation and change detection

We now show that we can compute the MMD statistic (2.13) at positions between windows with a runtime complexity of $O(\log t)$.

**Proposition 6.3.2.** *Let $B_l, \ldots, B_{s+1}, B_s, \ldots, B_0$ be a given list of windows with corresponding elements $\mathrm{X}_i$, $i = 0, \ldots, l$, as defined in* (6.2)*. For any split $s \in \{1, \ldots, l-1\}$, the computation of*

$$\mathrm{MMD}_k^2 \left( \bigcup_{i=s+1}^{l} \mathrm{X}_i, \bigcup_{i=0}^{s} \mathrm{X}_i \right), \tag{6.6}$$

*that is, the computation of MMD between the elements in windows coming before window $B_s$ and the elements in windows coming after (and including) $B_s$, has a runtime complexity of $O(\log t)$ for $0 < s < l$, with $s, l \in \mathbb{N}$.*

*Proof.* To obtain (6.6), one recursively merges $B_s, \ldots, B_0$ to $B'_s$ using Lemma 6.3.2, starting from the right, and similarly $B_l, \ldots, B_{s+1}$ to $B'_l$. One then obtains the statistic with Lemma 6.3.1, and by setting $\mathrm{XY}'^{l'}_{s'} = \sum_{i=1}^{l-s} \mathrm{XY}^i_{s'}$, that is, by summing all elements in the $\mathrm{XY}'_s$-list of $B'_s$. This concludes the proof as the logarithmic complexity was already established. □

The application of the presented data structure for change detection is as follows. For each new observation, we estimate MMD at any position between windows and compare it to the threshold $\epsilon'_\alpha = \frac{\epsilon_\alpha}{l}$ (with Bonferroni correction) from Proposition 6.3.1. We report a change when the value of MMD exceeds the threshold. As there are at most $\log t$ windows, we have at most $\log t - 1$ positions. Computing MMD for a position is in $O(\log t)$ by Proposition 6.3.2, and so the procedure has a total runtime complexity of $O(\log^2 t + t)$ per insert operation, where the term linear in $t$ results from computing $\mathrm{XY}_0$ when inserting a new observation. We may equivalently consider the proposed method as providing a more coarse-grained estimate of (6.1), taking the form

$$T' = \inf \left\{ t : \max_{s=1,\ldots,l} \left[ \mathrm{MMD}_k \left( \bigcup_{i=s+1}^{l} \mathrm{X}_i, \bigcup_{i=0}^{s} \mathrm{X}_i \right) \geq \epsilon_\alpha \right] = 1 \right\}, \tag{6.7}$$

with the $X_i$-s defined as in Proposition 6.3.2.[3]

While the data structure in its current form allows to obtain the precise values of (2.13) in an incremental fashion, its runtime and memory complexity are $O(t)$ for each new observation; these complexities are unsuitable for deploying the algorithm in the streaming setting. We reduce the runtime by subsampling within the windows, which we present together with the complete algorithm in the following section.

### 6.3.3. MMDEW algorithm

Our algorithm builds upon the data structure discussed previously. But, we suggest that each window of size $2^s$, $s = 0, \ldots, l$, samples $s$ observations (of the total $2^s$), that is, a logarithmic amount, while keeping everything else as before.

In this section, we first analyze such subsampling and discuss its benefits. Afterwards, we present the complete algorithm.

**Proposition 6.3.3.** *With subsampling, the number of terms in the sum $\mathrm{XX}_l$ for a window at position $l$, $1 \leq l$ ($l \in \mathbb{N}$) is*

$$n_{\mathrm{XX}_l} = 2^{l-1} \left( l^2 - l + 4 \right) = \frac{t}{2} \left( \log_2^2 t - \log_2 t + 4 \right),$$

[3] Recall that Remark 6.3.1 gives the precise value of $l$. Further, some split positions $s$ may not exist.

**Input:** Data stream $x_1, x_2, \ldots$, level $\alpha \in (0, 1)$
**Output:** Change points in $x_1, x_2, \ldots$; detection times

1: $windows \leftarrow \emptyset$ ▷ List of windows
2: **for** each $x_i \in \{x_1, x_2, \ldots\}$ **do**
3: $X_0 \leftarrow x_i$ ▷ Initialize $B_0$
4: $\mathrm{XX}_0 \leftarrow k(x_i, x_i)$
5: **for** each $B_j \in windows$ **do**
6: $\mathrm{XY}_0^j \leftarrow \sum_{x_k^{(j)} \in B_j} k(x_i, x_k^{(j)})$
7: $B_0 = (\mathrm{X}_0, \mathrm{XX}_0, \mathrm{XY}_0)$
8: $windows \leftarrow windows \cup B_0$
9: **for** each split $s$ in $windows = \{B_l, \ldots, B_{s+1}, B_s, \ldots, B_0\}$) **do** ▷ Detect changes
10: **if** $\mathrm{MMD}_k \left( \bigcup_{j=s+1}^{l} \mathrm{X}_j, \bigcup_{j=0}^{s} \mathrm{X}_j \right) \geq \epsilon'_\alpha$ **then**
11: **print** "Change at $s$ detected at time $i$"
12: $windows \leftarrow B_s, \ldots, B_0$ ▷ Drop windows
13: **while** two windows have the same size $2^l$ **do** ▷ Maintain exponential structure
14: Merge windows following Lemma 6.3.2 into $B_{l+1}$
15: Store a uniform sample of size $l + 1$ in $\mathrm{X}_{l+1}$ of $B_{l+1}$

**Algorithm 6.1.:** Proposed MMDEW change detection algorithm.

*with $t = 2^l$ the number of observations of $B_l$. The number of terms of $\mathrm{XY}_l^l$ for windows of the same size, which occur prior to merging, is*

$$n_{\mathrm{XY}_l^l} = 2^l l = t \log_2 t.$$

**Remark 6.3.2.** *The number of terms in the sums of* (2.13) *acts as a proxy for the quality of the estimate. It is optimal when no subsampling takes place; this number is $O(t^2)$. When subsampling a logarithmic number of observations per window with our data structure (as we propose), one achieves polylogarithmic runtime and logarithmic memory complexity. At the same time, one achieves a better approximation quality than naively sampling a logarithmic number of observations without the summary data structure. While such sampling would also yield a memory complexity of $O(\log t)$ when using the naive approach for change detection—that is, splitting the sample into two neighboring windows and computing $\mathrm{MMD}^2$—the number of terms in* (2.13) *would be $O(\log^2 t)$. Proposition 6.3.3 shows that the summary data structure improves upon this by a factor of approximately $t/2$ for $n_{\mathrm{XX}_l}$ and a factor of $t/\log_2 t$ for $n_{\mathrm{XY}_l^l}$ (we neglect logarithmic and constant terms in the former due to their small contribution).*

Algorithm 6.1 now summarizes the complete algorithm, with MMD in Line 10 referring to the computation of MMD as in Proposition 6.3.2. MMDEW stores only a uniform sample of size $l + 1$, that is, of size logarithmic in the number of observations, while keeping the respective $\mathrm{XX}_s$ and $\mathrm{XY}_s$, $s = 0, \ldots, l$, computed before. With this approach, the number of samples in a window increases by one each time the window is merged, and the memory complexity is logarithmic in the number of observations. Note that one recovers the previous algorithm (Section 6.3.2.3) and therefore the precise value of (6.6) if one omits Line 15. Further, changes in Line 15 allow to adjust the subsampling, for example, the user may defer the sampling until windows contain a minimum number of observations, or choose a different function to control the sample size.

The following example illustrates the procedure. Figure 6.2 expands upon Example 6.3.2 and shows the evolution of the data structure upon observing $x_1, \ldots, x_6$ and when merging windows.

**Example 6.3.2.** *We assume that there is a stream of i.i.d. observations $x_1, x_2, \ldots$. Note that the i.i.d. assumption implies that there are no changes. MMDEW receives the first observation, $x_1$ and creates a window $B_0$ storing $x_1$, $\mathrm{XX}_0 = k(x_1, x_1)$, and $\mathrm{XY}_0 = \emptyset$. For the next observation, $x_2$, it creates a new window $B_{0'}$, storing $x_2$, $\mathrm{XX}_{0'} = k(x_2, x_2)$, and $\mathrm{XY}_{0'} = \{k(x_1, x_2)\}$ and detects no change. As $B_0$ and $B_{0'}$ have the same size, MMDEW merges them into window $B_1$, storing a sample of size $\log_2 2 = 1$, say, it stores $x_1$ and discards $x_2$, and computes $\mathrm{XX}_1 = k(x_1, x_1) + k(x_2, x_2) + 2k(x_1, x_2)$, following (6.3). As no previous window exists, the computation of $\mathrm{XY}_1$ is not required. We see that the number of terms in $\mathrm{XX}_1$ equals four, while $B_1$ stores only one observation (established in Proposition 6.3.3). Next, the algorithm observes $x_3$ and creates a new window, $B_0$, storing $x_3$, $\mathrm{XX}_0 = k(x_3, x_3)$, and computing $\mathrm{XY}_0$ to the window coming before, that is, $B_1$, so that $\mathrm{XY}_0 = \{\mathrm{XY}_0^1\}$. In the next step, MMDEW receives $x_4$, again creating a new window $B_{0'}$. The algorithm now recursively merges the windows, that is, $B_0$ and $B_{0'}$ become $B_{1'}$, and $B_1$ and $B_{1'}$ then become $B_2$. Upon receiving $x_5$, the algorithm creates a new window $B_0$, storing $x_5$, the kernel evaluation $k(x_5, x_5)$, and the interaction of $x_5$ with $\{x_1, x_4\}$ from $B_3$. We conclude the example with $x_6$, which leads to the creation of a new window $B_0'$. As in steps 2 and 4, $B_0$ and $B_0'$ will now be merged to obtain $B_1$.*

Algorithm 6.1 has a runtime cost of $O(\log^2 t)$ per insert operation and a total memory complexity of $O(\log t)$. This allows it to scale to very large data streams. Nevertheless, if one strictly requires constant time and memory, one can simply limit the number of windows at the expense of detecting changes only up to a certain time in the past. In the latter configuration, MMDEW fulfills the requirements for streaming algorithms laid out by Domingos and Hulten [2003].

## 6.4. Experiments

This section compares our approach on synthetic data with the quadratic-time estimator in Section 6.4.1, where we use the McDiarmid-based bound (Proposition 6.3.1). We evaluate the runtime of MMDEW and its contenders in Section 6.4.2. A comparison to other kernel-based approaches with an optimally chosen threshold is in Section 6.4.3, and a comparison to univariate approaches is in Section 6.4.4. We showcase MMDEW on streams derived from real-world classification tasks in Section 6.4.5. We ran all experiments on a server running Ubuntu 20.04 with 124GB RAM, and 32 cores with 2GHz each.

### 6.4.1. Comparison with the quadratic-time MMD estimator

To evaluate the average run length (ARL) and the mean time to detection (MTD) in a controlled environment, we first conduct experiments on synthetic data, comparing MMDEW to the MMD estimate (2.13) as baseline.[4]

The ARL quantifies the expected number of observations processed before a change detector flags a change, assuming $H_0$ holds. In the static setting, this corresponds to the type I error. Formally, for $\tilde{T}$ corresponding to the stopping time captured by Algorithm 6.1, that is, (6.7) with subsampling applied, we are interested in $\mathbb{E}_{H_0}\tilde{T}$.

The error under the alternative ($H_1$ holds) is captured by the expected detection delay (EDD), also called "mean time to detection (MTD)". Specifically, a change detector processes a stream that contains a change at a known observation $\kappa$ ($\kappa = 1, \ldots, t$) and we want to know the delay until the change is

[4] For computational reasons, we compute MMD as described in the discussion following Proposition 6.3.1. For a fair comparison, we use the distribution-free bound of Proposition 6.3.1 for both algorithms.

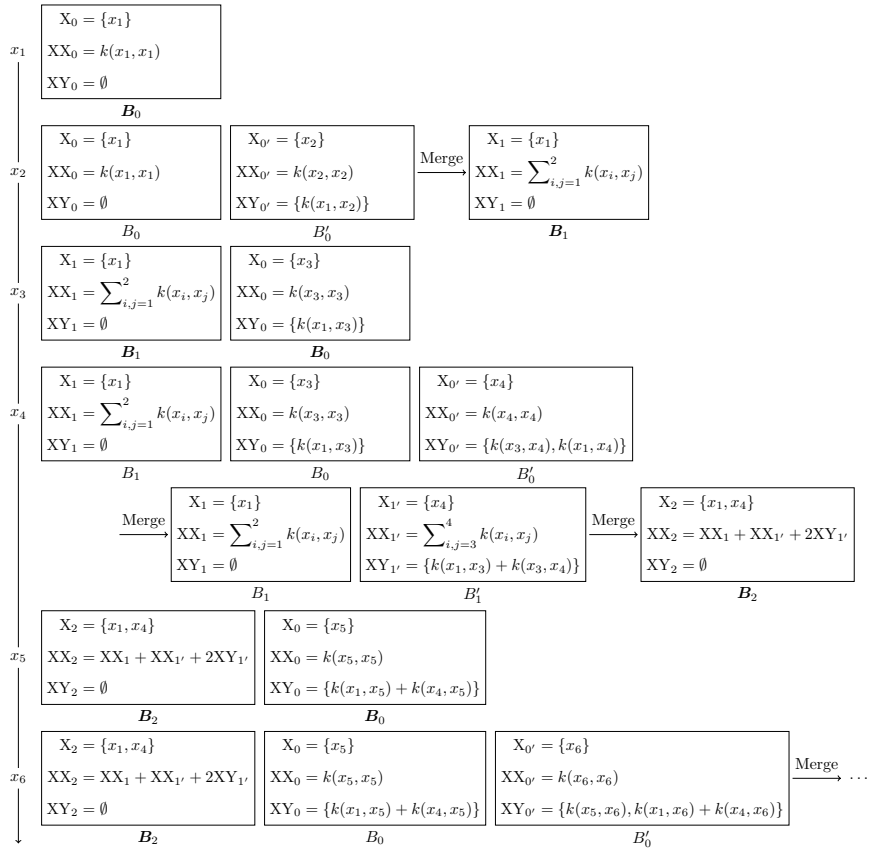


**Figure 6.2.:** Set up of data structure with subsampling upon inserting $x_1, \ldots, x_6$. MMDEW stores the windows in bold face at the end of the merge operations. Observations $x_2$ and $x_3$ are not stored explicitly due to the sampling applied. $x_4$ is split into two lines for readability. See Example 6.3.2 for a detailed discussion.

reported, that is, $\mathbb{E}_{x_1,\ldots,x_{\kappa-1}\sim\mathbb{P},x_\kappa,\ldots,x_t\sim\mathbb{Q}}\tilde{T}$, with $\mathbb{P}$ the pre-change distribution and $\mathbb{Q} \neq \mathbb{P}$ the post-change distribution. In other words, the EDD quantifies how many samples of $\mathbb{Q}$ must be processed to flag a change after having observed $\kappa - 1$ samples from $\mathbb{P}$. Note that $\kappa$ must be chosen large enough to allow MMD to capture the difference in $\mathbb{P}$ and $\mathbb{Q}$, and we assume that the statistic does not exceed the threshold on the first $\kappa$ samples. In a static setting, the EDD is comparable with the type II error. Note that existing literature [Xie and Siegmund, 2013, Wei and Xie, 2022] usually considers a fixed threshold $b$ for the stopping time for both ARL and EDD, while our threshold $\epsilon_\alpha$ depends on the position of the split considered. Experiments for a fixed value of $b$ are in Section 6.4.3.

To approximate the ARL and the EDD in different scenarios, we simulate 5-dimensional data distributed according to the multivariate normal $\mathcal{N}(\mathbf{0}, \mathbf{I}_5)$, the uniform $\mathrm{Unif}(-\mathbf{1}_5, \mathbf{1}_5)$, the $\mathrm{Laplace}(0, \sigma\mathbf{I}_5)$, and a mixed distribution, respectively. The mixed distribution is taken to be $\mathcal{N}(\mathbf{0}, \mathbf{I}_5)$ with probability 0.3 and $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_5)$ with probability 0.7. We set $\sigma = 3$.

To compute ARL, we consider 10,000 observations distributed according to either the uniform, the Laplace, or the mixed distribution. Hence, the data does not contain any changes. For MTD, we first run both algorithms on 512 ($= 2^9$) and 1024 ($= 2^{10}$) observations, respectively, leading to MMDEW summarizing the data in one window in both cases. These observations are distributed according to $\mathcal{N}(\mathbf{0}, \mathbf{I}_5)$ and then followed by either the uniform, Laplace, or mixed distribution. That is, we induce a change point at $\kappa = 513$ (resp. $\kappa = 1025$), and then count the number of observations processed from the new distribution until the algorithms report a change.

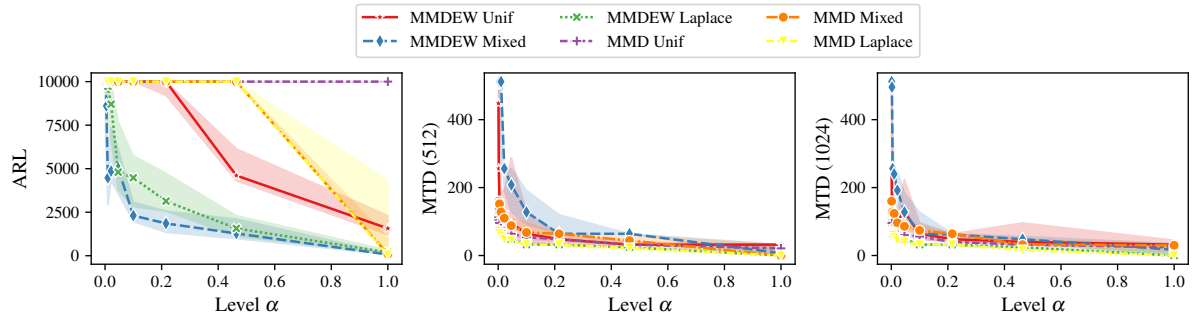


**Figure 6.3.:** Average run length (ARL) and expected detection delay/mean time to detection (EDD/MTD) of MMDEW and the quadratic-time estimator on synthetically generated data.

Figure 6.3 collects the median results over 20 repetitions. The left plot shows that an increase in the level $\alpha \in (0, 1)$ leads to a decrease in ARL. This is expected as the test becomes more sensitive, leading to more false positives. The baseline achieves a higher ARL but at the cost of an increased runtime. The MTD plots (center and r.h.s.) mirror the ARL observation: The MTD decreases with increasing $\alpha$. We further observe that the detection delay depends on the post-change distribution. The delay is comparably large when changing from the multivariate standard normal to the mixed distribution. This matches our intuition: the mixed distribution is relatively similar to the pre-change distribution, rendering it difficult to detect a change between them. For larger values of $\alpha$, that is, $\alpha \geq 0.2$, MMDEW performs similarly to the baseline in all cases. Comparing the MTD when the change happens after 512 observations to the MTD when the change happens after 1024 observations, the results show that more pre-change samples render the algorithms more sensitive to detecting changes, due to more samples improving the approximation of the mean embedding. Overall, the results on these synthetic streams indicate that MMDEW is (i) robust to the choice of $\alpha$ and (ii) that $\alpha$ has the expected influence on the behavior of the algorithm.

### 6.4.2. Runtime evaluation

We now compare the runtime of MMDEW to that of its contenders and additionally validate the runtime guarantees that we derived analytically in Section 6.3.3.

To this end, we generate a constant stream of $10^6$ one-dimensional observations, that is, the observed stream contains no change. Note that, while the dimensionality of the data affects the runtime depending on the used kernel, its influence is the same across all kernel-based algorithms, hence we limit our considerations to the univariate case.

Figure 6.4 shows the average results over 10 runs. The left plot reveals that the fixed cost per insert operation of MMDEW is relatively large, as processing a small number of observations requires comparably much time. However, the runtime does not increase by much with the number of observations. The figure also shows that the proposed algorithm's runtime is better than that of an alternate kernel-based method, Scan $B$-statistics, where we use a window size of $\omega = 100$ in the runtime experiments. For $t > 0.05 \cdot 10^6$, MMDEW also outperforms IBDD. Still, the other algorithms run faster than MMDEW but achieve a lower $F_1$ score in our later experiments.

The right plot of Figure 6.4 verifies the analytically derived runtime of $O(\log^2 t)$ by fitting the corresponding curve ($t \mapsto c \log^2 t$) to the measured data with the least squares method. The resulting mean squared error is approximately $10^{-6}$, which confirms the preceding asymptotic runtime analysis.

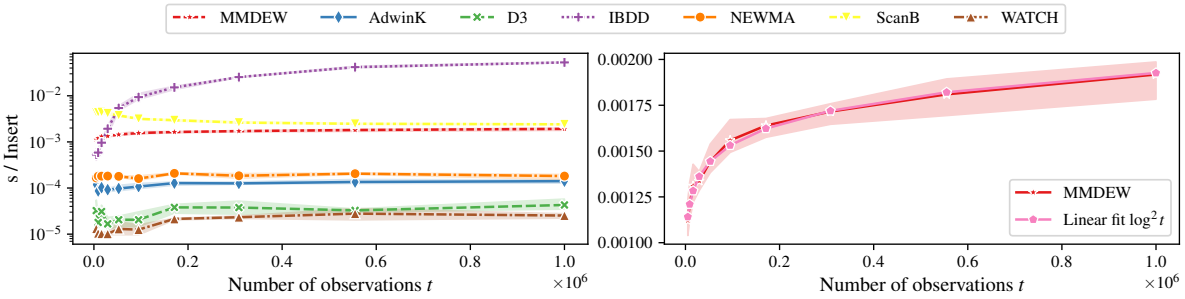


**Figure 6.4.:** Comparison of runtimes per insert operation (l.h.s.) and least squares fit validating the theoretical runtime complexity of MMDEW w.r.t. the runtime observed in practice (r.h.s.).

### 6.4.3. Comparison with kernel-based approaches

The following experiments compare the EDD of the kernel-based online change detection approaches for a fixed ARL on toy data, extending the experiments of Wei and Xie [2022, Figure 4]. We note that all approaches considered in this section compute, for each new observation, a test statistic and compare the statistic to a threshold. If the statistic exceeds the threshold, a change is flagged. In the case of the proposed algorithm, multiple test statistics (one for each possible split) are computed. To allow for a comparison, we select the maximum MMD value across all splits, that is, for MMD, we consider the stopping rule

$$T'' = \inf\left\{t : \max_{s=1,\dots,l} \mathrm{MMD}_k\left(\bigcup_{i=s+1}^{l} \mathrm{X}_i, \bigcup_{i=0}^{s} \mathrm{X}_i\right) \geq b\right\}, \tag{6.8}$$

with a fixed $b > 0$, instead of (6.7). Similarly, for MMDEW, we consider (6.8) but with subsampling applied to the $X_i$-s. The experimental setup is as follows.

To achieve a fixed target ARL $\mathbb{E}_{H_0}T$ for a given stopping time $T$, we run 25 Monte Carlo simulations on 150,000 samples from $\mathbb{P} = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ with $d = 20$ and select $b$ as the $1 - 1/(\text{target ARL})$-quantile of the collected test statistics as threshold. For online kernel CUSUM, we set its parameters $B_{\max} = 50$ and $N = 15$, matching the settings of Wei and Xie [2022, Figure 4]. Similarly, for Scan $B$-statistics and NEWMA, we set $B_0 = 50$; the remaining parameters of NEWMA then follow from the heuristics detailed by the authors [Keriven et al., 2020].

For approximating the EDD of MMDEW and NEWMA for a threshold $b$, we draw 64 and 400 samples from $\mathbb{P}$, respectively, before sampling from $\mathbb{Q}$. Online kernel CUSUM and Scan $B$-statistics each receive 1,000 samples from $\mathbb{P}$ upfront, for computing the variance estimate they require and to use as a reference sample. All approaches use the Gaussian kernel with the bandwidth set by the median heuristic [Garreau et al., 2018].[5]

Figure 6.5 collects our results, with each subfigure corresponding to a different post-change distribution, of which we sample and process 500 elements to find the first time the test statistics exceeds the threshold. We consider the mean result over 100 repetitions. The results show that OKCUSUM and NEWMA perform similarly across all experiments, with Scan $B$-statistics performing generally worse in three of the cases. MMDEW achieves the lowest EDD throughout. Specifically, on the mixed distribution $\mathbb{Q} = \gamma\mathcal{N}(0, \mathbf{I}_d) + (1-\gamma)\mathcal{N}(0, \sigma^2\mathbf{I}_d)$ ($\gamma = 0.3$), the EDD of the proposed method is between 1.62 and

[5] Note that NEWMA uses random Fourier features [Rahimi and Recht, 2007] to approximate the kernel.

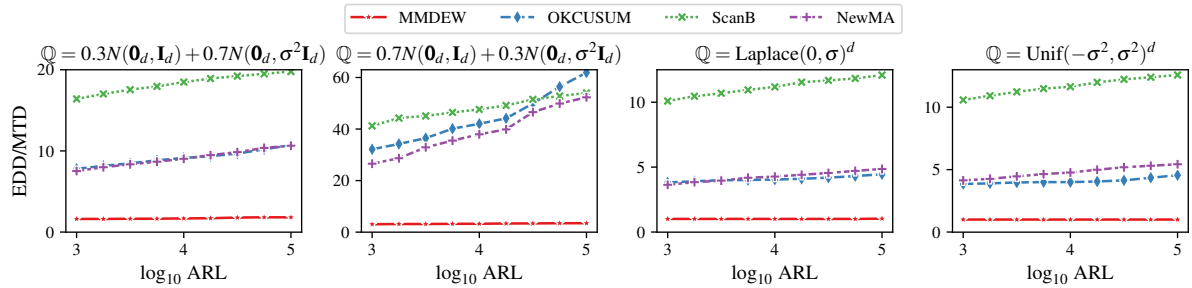


**Figure 6.5.:** EDD/MTD of kernel-based change detectors with a pre-change distribution of $\mathbb{P} = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, $d = 20$, and the indicated post-change distribution ($\sigma = 2$).

1.82. In the more challenging setting of $\gamma = 0.7$, the EDD of MMDEW is between 3.06 and 3.46. Here, OKCUSUM performs second-best, with an EDD of 32.15 for a target ARL of 1,000 and 61.92 for a target ARL of 100,000. On the Laplace and Uniform distributions, the proposed method improves upon the results of OKCUSUM and NEWMA as well, albeit by a smaller margin.

While these experiments show that the proposed method improves upon the state-of-the-art, we note that the experiments require obtaining samples from $\mathbb{P}$, which is rarely feasible in practice. In this case, we recommend setting the threshold of MMDEW by the McDiarmid-based bound (Proposition 6.3.1), as we do in Section 6.4.1.

### 6.4.4. Comparison with univariate approaches

In this section, we show that our MMD-based approach detects changes in the covariance structure of multivariate data, which aggregated univariate approaches cannot detect reliably.

In particular, we compare the test statistics of the proposed MMDEW, MMD, the Cramer-von-Mises change point model (CvM CPM; Ross and Adams 2012, Ross 2015), and the recent non-parametric Focus [Romano et al., 2024] approach on 20-dimensional multivariate normal data with a mean and correlation shift, respectively. CvM CPM and Focus handle univariate data only. To run each on multivariate data, we run one instance per dimension and consider the means of their test statistics. For MMD and MMDEW, our settings are the same as detailed in Section 6.4.3. We set the pre-change distribution to $\mathbb{P} = \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$; the respective post-change distributions $\mathbb{Q}$ are indicated in Figure 6.6. For CPM, which updates all previous test statistics upon observing a new sample, we report the test statistics computed after processing 500 samples from the pre-change distribution and 100 samples from the respective post-change distribution; for all other approaches, we report the test statistic computed upon observing each sample.

Our results are in Figure 6.6. A change in either the distribution mean (l.h.s.) or the correlation (r.h.s.) lead to an increase of the test statistic of MMDEW and MMD, respectively. Hence, these approaches allow detecting such changes. CPM and Focus correctly identify the change in mean, which is reflected in the univariate marginals they consider. CPM correctly identifies the change point, that is, the maximum value of the test statistic is at 500. When regarding the change in the correlation (the marginals the univariate approaches consider do not change), Focus' test statistic does not reflect the change point. Surprisingly, for CPM, the change in the correlation structure leads to a change in the test-statistic—but the change is identified incorrectly, with the maximum of the test statistic occurring after approximately 530 samples. We conclude that using MMD or the proposed MMDEW is preferable

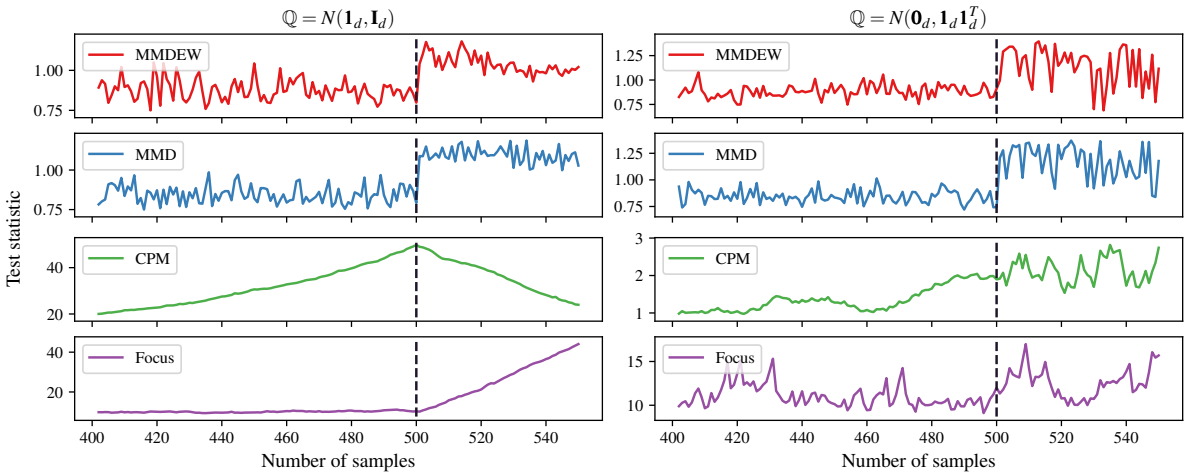


**Figure 6.6.:** Maximum values of the respective test-statistics (20 repetitions, $d = 20$). A change (indicated by a dashed line) occurs after 500 samples, from $\mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ to the distribution indicated on top of the columns, respectively. For the univariate approaches (CPM, Focus), we run one instance per dimension and consider the mean.

**Table 6.2.:** Overview of data sets.

| Data set | $n$ | $d$ | #CPs |
|---|---|---|---|
| CIFAR10 [Krizhevsky et al., 2009] | 60,000 | 1,024 | 9 |
| FashionMNIST [Xiao et al., 2017] | 70,000 | 784 | 9 |
| Gas [Vergara et al., 2012] | 13,910 | 128 | 5 |
| HAR [Anguita et al., 2013] | 10,299 | 561 | 5 |
| MNIST [Deng, 2012] | 70,000 | 784 | 9 |

to aggregating univariate change detectors when processing multivariate data, when the changes are not reflected in the marginals.

### 6.4.5. Streams from real-world classification data

To obtain our change detection quality estimates, we use well-known classification data sets and interpret them as streaming data.[6] This is common in the literature, for example, Faithfull et al. [2019], Faber et al. [2021], as only few high-dimensional annotated change detection data sets are publicly available.

For each data set, we first order the observations by their classes; a change occurs if the class changes. To introduce variation into the order of change points, we randomly permute the order of the classes before each run but use the same permutation across all algorithms. For preprocessing, we apply min-max scaling to all data sets. Table 6.2 summarizes the data sets, where $n$ is the number of observations, $d$ is the data dimensionality, and #CP is the number of change points.

We run a grid parameter optimization per data set and algorithm and report the best result w.r.t. the $F_1$-score. We note that such an optimization is difficult to perform in practice—here one typically prefers approaches with fewer or easy-to-set parameters—but allows a fair comparison. Table 6.3 lists

[6] While MMDEW is not limited to Euclidean data, Euclidean data is the type of data most frequently encountered in practice, and our experiments target this setting.

**Table 6.3.:** Values for the parameter optimization.

| Algorithm | Parameters | Parameter values |
|---|---|---|
| MMDEW | $\alpha$ | $\alpha \in \{0.001, 0.01, 0.1, \ldots, 0.9, 0.99, 0.999\}$ |
| ADWINK | $\delta, k$ | $\delta \in \{0.05, 0.1, 0.2, 0.9, 0.99\}$, $k \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$ |
| D3 | $\omega, \rho, \tau$, d | $\omega \in \{100, 200, 500\}$, $\rho \in \{0.1, 0.3, 0.5\}$, $\tau \in \{0.7, 0.8, 0.9\}$, $d = 1$ |
| IBDD | $m, w$ | $m \in \{10, 20, 50, 100\}$, $w \in \{20, 100, 200, 300\}$ |
| NEWMA | $\omega, \alpha$ | $\omega \in \{20, 50, 100\}$, $\alpha \in \{0.01, 0.02, 0.05, 0.1\}$ |
| Scan $B$ | $B, \omega, \alpha$ | $B \in \{2, 3\}$, $\omega \in \{100, 200, 300\}$, $\alpha \in \{0.01, 0.05\}$ |
| WATCH | $\epsilon, \kappa, \mu, \omega$ | $\epsilon \in \{1, 2, 3\}$, $\kappa \in \{25, 50, 100\}$, $\mu \in \{10, 20, 50, 100, 1000, 2000\}$, $\omega \in \{100, 250, 500, 1000\}$ |

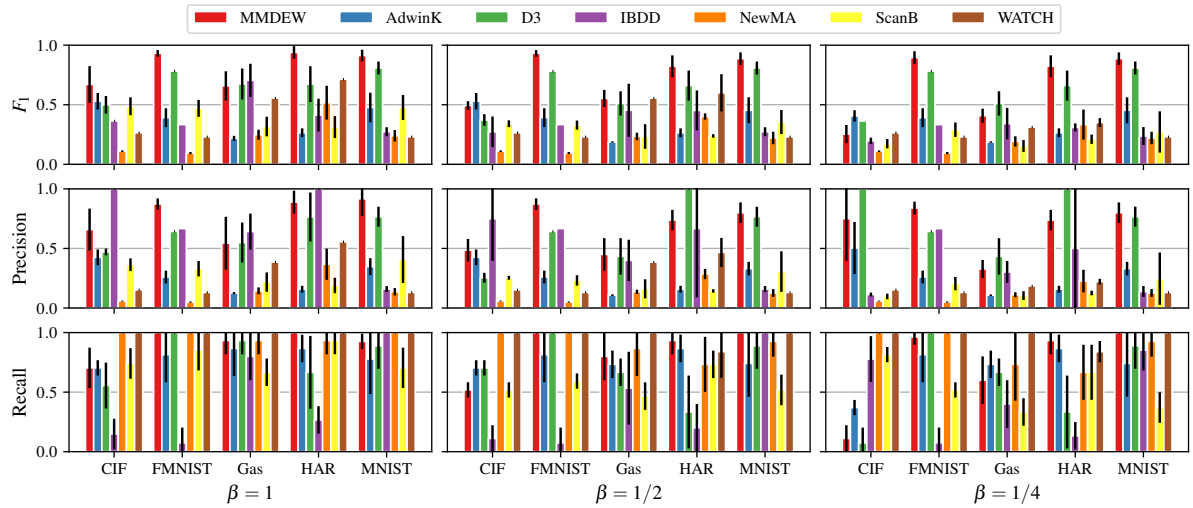


**Figure 6.7.:** Average $F_1$-score, precision and recall. The bars show the standard deviation over 10 permutations of the data.

all the parameters we tested. We note that the grid parameter optimization allowed us to obtain better $F_1$-scores than the heuristics proposed in Keriven et al. [2020] for NEWMA and Scan $B$-statistics. We exclude the squared time estimator of MMD due to its prohibitive runtime. For kernel-based algorithms (MMDEW, NEWMA, and Scan $B$-statistics) we use the Gaussian kernel $k(x, y) = \exp\left(-\gamma \|x - y\|^2\right)$ ($\gamma > 0$) and set $\gamma$ using the median heuristic [Garreau et al., 2018] on the first 100 observations. We also supply the first 100 observations to competitors requiring data to estimate further parameters (IBDD, WATCH) upfront.

**$F_1$-score, precision, and recall.** We compute the precision, the recall, and the $F_1$-score, which are common to evaluate change detection algorithms [Keriven et al., 2020, van den Burg and Williams, 2020, Faber et al., 2021]. Specifically, for a fixed $\Delta_T \in \mathbb{N}_{>0}$, we proceed as follows. If a change is detected, and there is an actual change point within the $\Delta_T$ previous time steps, we consider it a true positive (tp). If a change is detected, and there is no change point within the $\Delta_T$ previous steps, we consider it a false positive (fp). If no change is detected within $\Delta_T$ steps of a change point, we consider it a false negative (fn). We count at most one true positive for each actual change point. With these definitions, the precision is $\text{Prec} = \text{tp}/(\text{tp} + \text{fp})$, the recall is $\text{Rec} = \text{tp}/(\text{tp} + \text{fn})$, and the $F_1$-score is their harmonic mean $F_1 = 2 \cdot (\text{Prec} \cdot \text{Rec}) / (\text{Prec} + \text{Rec})$. Note that, while some algorithms allow to infer where in the data a change happens, including the proposed MMDEW, we only evaluate the time at which they report a change, as all tested approaches allow reporting this value.

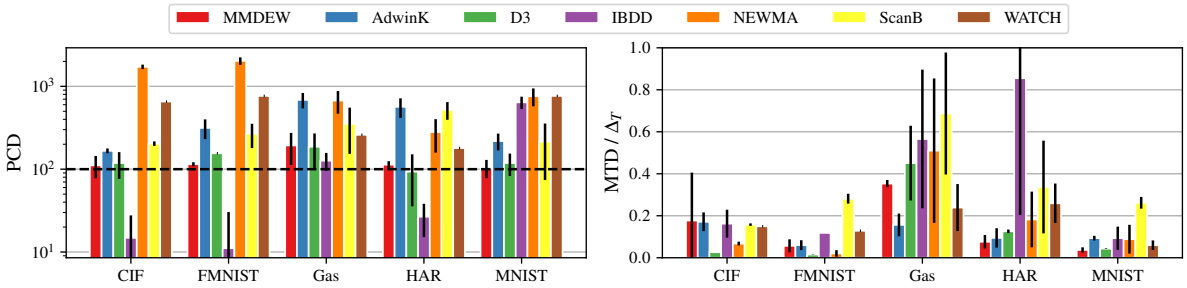


**Figure 6.8.:** Average of percentage of changes detected (PCD) and of mean time to detection (MTD). The dashed line indicates the optimum for PCD. For MTD lower values are better.

Figure 6.7 shows our results. As $\Delta_T$ is an evaluation-specific parameter, we vary it relative to the average distance between change points by a factor $\beta > 0$: Given a data set of length $N$ with $n$ changes, we set $\Delta_T = \beta \cdot N/(n+1)$. For $\beta = 1$ ($\Delta_T$ is equal to the average number of steps between change points per respective data set), MMDEW achieves a higher $F_1$-score than all competitors on all data sets except for Gas, where it still obtains a competitive result. Throughout, the proposed algorithm obtains a good balance between precision and recall. Other approaches either have very low precision (for example, less than 20%), or an inferior recall and precision, down to a few exceptions. With a reduced $\beta$, that is, we allow only a shorter detection delay, the performance of all algorithms decreases on average. For $\beta = 1/2$, MMDEW achieves the best $F_1$ score also on four data sets, and, for $\beta = 1/4$ (the most challenging setting) on three of the tested data sets.

We conclude that the proposed method achieves very good results across all these experiments—especially when taking into account the fewer hyperparameters compared to the other approaches that we tested.

**Percentage of changes detected and detection delay.** To obtain a complete picture of the performance of MMDEW, we also report the "percentage of changes detected" (PCD), that is, the ratio of the number of reported changes and the number of actual change points, and its MTD on the data streams derived from real-world data. In our context, MTD coincides with the expected detection delay.

Figure 6.8 collects our results. For PCD, results closer to 100% are better. Here, MMDEW is on par with the closest competitors and consistently, that is, across all data sets, detects an approximately correct number of change points. D3, NEWMA, Scan $B$-statistics, and WATCH detect too many change points in all cases. This behavior is also reflected in their comparably large recall in Figure 6.7.

For MTD, lower values are better. Here, the classification-based D3 performs best in most of the cases. MMDEW performs a bit worse than D3 but better than the other algorithms on most data sets, with the Gas data set the major exception. As the experiments in Figure 6.7 show, a lower $\Delta_T$ tends to lead to a lower $F_1$-score of MMDEW. In other words, MMDEW tends to detect changes with some delay, but it detects them consistently.

## 6.5. Proofs

This section contains additional proofs. The proof of Proposition 6.3.1 is in Section 6.5.1. Proposition 6.3.3 is proved in Section 6.5.2.

### 6.5.1. Proof of Proposition 6.3.1

Proposition 6.3.1 follows from the more general result that we state below. The statement and proof are similar to Gretton et al. [2012, Theorem 8] but do not assume $m = n$. Note that we recover Gretton et al. [2012, Theorem 8] in the case that $m = n$. We prove Proposition 6.3.1 afterwards.

**Proposition 6.5.1.** *Let $\mathbb{P}$, $\mathbb{Q}$, $\hat{\mathbb{P}}_m$, $\hat{\mathbb{Q}}_n$ be defined as in the main text, assume $0 \leq k(x, y) \leq K$ for all $x, y \in \mathcal{X}$, $\mathbb{P} = \mathbb{Q}$, and $t > 0$. Then*

$$P\left(\mathrm{MMD}_k\left(\hat{\mathbb{P}}_m, \hat{\mathbb{Q}}_n\right) - \left(\frac{K}{m} + \frac{K}{n}\right)^{\frac{1}{2}} \geq t\right) \leq e^{-\frac{t^2 mn}{2K(m+n)}}.$$

*Proof.* First, we bound the difference of $\mathrm{MMD}_k\left(\hat{\mathbb{P}}_m, \hat{\mathbb{Q}}_n\right)$ to its expected value. Changing a single one of either $x_i$ or $y_j$ in this function results in changes of at most $2\sqrt{K}/m$ and $2\sqrt{K}/n$, giving

$$\sum_{i=1}^{n+m} c_i^2 = 4K\frac{n+m}{nm}.$$

We now apply the bounded differences inequality (recalled in Theorem A.4.1) to obtain

$$P\left(\mathrm{MMD}_k\left(\hat{\mathbb{P}}_m, \hat{\mathbb{Q}}_n\right) - \mathbb{E}\,\mathrm{MMD}_k\left(\hat{\mathbb{P}}_m, \hat{\mathbb{Q}}_n\right) \geq t\right) \leq e^{-\frac{t^2 mn}{2K(m+n)}}. \tag{6.9}$$

The last step is to bound the expectation, which yields

$$\begin{aligned}
\mathbb{E}\,\mathrm{MMD}_k\left(\hat{\mathbb{P}}_m, \hat{\mathbb{Q}}_n\right) &= \mathbb{E}\left(\frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i, x_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i, y_j) - \frac{1}{mn}\sum_{i,j=1}^{m,n} k(x_i, y_j) - \frac{1}{mn}\sum_{j,i=1}^{n,m} k(y_j, x_i)\right)^{\frac{1}{2}} \\
&\leq \left(\frac{1}{m}\mathbb{E}k(X,X) + \frac{1}{n}\mathbb{E}k(Y,Y) + \frac{1}{m}(m-1)\mathbb{E}k(X,Y) + \frac{1}{n}(n-1)\mathbb{E}k(Y,X) - 2\mathbb{E}k(X,Y)\right)^{\frac{1}{2}} \\
&= \left(\frac{1}{m}\mathbb{E}k(X,X) + \frac{1}{n}\mathbb{E}k(Y,Y) - \frac{1}{m}\mathbb{E}k(X,Y) - \frac{1}{n}\mathbb{E}k(X,Y)\right)^{\frac{1}{2}} \\
&= \left(\frac{1}{m}\mathbb{E}\left[k(X,X) - k(X,Y)\right] + \frac{1}{n}\mathbb{E}\left[k(X,X) - k(X,Y)\right]\right)^{\frac{1}{2}} \leq \left(\frac{K}{m} + \frac{K}{n}\right)^{\frac{1}{2}},
\end{aligned} \tag{6.10}$$

where we used Jensen's inequality and the linearity of the expectation for the first inequality; the second inequality follows from the boundedness assumption. By combining (6.10) with (6.9), we obtain the stated result. □

Proposition 6.3.1 is now a corollary of Proposition 6.5.1, which follows by setting $\alpha = e^{-\frac{t^2 mn}{2K(m+n)}}$ and solving for $t$ to obtain a test of level $\alpha$.

### 6.5.2. Proof of Proposition 6.3.3

To find $n_{\mathrm{XY}_l^l}$, we use our implementation of MMDEW and the On-Line Encyclopedia of Integer Sequences (OEIS) to discover that $n_{\mathrm{XY}_l^l}$ follows the sequence $1, 2, 8, 24, 64, 160, \ldots$ for $l = 0, 1, 2, \ldots$. Thus

$$n_{\mathrm{XY}_l^l} = 2^l l, \quad \text{for } l > 0 \tag{6.11}$$

and $n_{\mathrm{XY}_0^0} = 1$ [Sloane, 1999b].

To find $n_{\mathrm{XX}_l}$, notice that $n_{\mathrm{XX}_l}$ only changes when one merges two windows, which happens for windows of the same size $n_{\mathrm{XX}_{l-1}}$. The algorithm adds to this $2 \cdot n_{\mathrm{XY}_{l-1}^{l-1}}$ terms, see (6.5), and, for $l = 0, 1, 2, \ldots$, we obtain the recurrence relation

$$n_{\mathrm{XX}_l} = \begin{cases} 1 & \text{if } l = 0, \\ 4 & \text{if } l = 1, \\ 2 \cdot n_{\mathrm{XX}_{l-1}} + 2 \cdot n_{\mathrm{XY}_{l-1}^{l-1}} & \text{if } l > 1, \end{cases}$$

with $n_{\mathrm{XX}_{-1}} := 0$. Now write

$$n_{\mathrm{XX}_l} = 2 \cdot n_{\mathrm{XX}_{l-1}} + l \cdot 2^l - 2^l + 2 \cdot [l = 0] + 2 \cdot [l = 1], \tag{6.12}$$

where the brackets are equal to one if their argument is true and zero otherwise (using Iverson's convention; Graham et al. 1994). To find a closed-form expression for (6.12), we define the ordinary generating function $A(z) = \sum_l a_l z^l$. Now, we multiply (6.12) by $z_l$ and sum on $l$, to obtain

$$A(z) = \frac{-8z^3 + 2z - 1}{(2z-1)^3}$$

after some algebra, so that

$$n_{\mathrm{XX}_l} = [z^l] \frac{-8z^3 + 2z - 1}{(2z-1)^3},$$

where $[z^l]$ is the coefficient of $z^l$ in the series expansion of the generating function $A(z)$. To extract coefficients, we first decompose $A(z)$ as

$$A(z) = \frac{3}{1-2z} - \frac{2}{(1-2z)^2} + \frac{1}{(1-2z)^3} - 1,$$

which allows us to then find the coefficients as

$$[z^l] \frac{3}{1-2z} \overset{(a)}{=} 3 \cdot 2^l, \qquad [z^l] - \frac{2}{(2z-1)^2} \overset{(b)}{=} -(l+1)2^{l+1}, \qquad [z^l] \frac{1}{(1-2z)^3} \overset{(c)}{=} (l+1)(l+2)2^{l-1},$$

where Graham et al. [1994, Table 335] implies (a), (b) is (6.11) shifted, and (c) is Sloane [1999a] shifted. We omit the last term as it corresponds to $[z^0]$, which we do not need. Now, adding all terms gives

$$3 \cdot 2^l - (l+1)2^{l+1} + (l+1)(l+2)2^{l-1} = 2^{l-1}(l^2 - l + 4),$$

concluding the proof.

# 7. Conclusions and Future Work

Kernel-based information theoretical measures are powerful and enjoy broad applications, but their "classic" estimators suffer from a quadratic runtime requirement w.r.t. the sample size. This thesis proposed accelerated estimators in the offline- and online setting, studied their computational-statistical trade-off, and showed the state-of-the-art performance of all new algorithms on diverse benchmarks. We also established the minimax lower bound of HSIC estimation, which implies the minimax optimality of our accelerated HSIC estimator.

Our contributions open the door to exciting future research, which we detail in the following.

**Nyström $M$-HSIC [Kalinke and Szabó, 2023].** Among other settings, our accelerated HSIC estimator (Chapter 3) allows tackling challenging causal discovery tasks, as shown in Section 3.4, but, in this configuration, requires enumerating all possible directed acyclic graphs (DAGs). However, many causal discovery algorithms, for example, the famous PC-algorithm [Spirtes and Glymour, 1991], rely on conditional independence tests, which limits the number of DAGs that one has to consider, reducing the overall runtime requirement. Kernel-based interaction- [Sejdinovic et al., 2013a] and conditional independence measures [Fukumizu et al., 2007, Zhang et al., 2011, Klebanov et al., 2020, Huang et al., 2022, Sheng and Sriperumbudur, 2023] exist, but these also suffer from quadratic or even cubic runtime requirements, limiting their practical applicability. Hence, the interest is in accelerating these estimators and understanding their statistical properties.

**Minimax rate of HSIC estimation [Kalinke and Szabó, 2024].** Our minimax result in Chapter 4 fully settles the lower bound of HSIC estimation on Euclidean domains for translation-invariant kernels. It is still open if faster rates on non-Euclidean domains are possible. Further, we studied the problem in the classical minimax setting, without any constraint on the estimators in the infimum. In the context of sparse PCA [Berthet and Rigollet, 2013, Wang et al., 2016] or canonical correlation analysis [Gao et al., 2017], computational constraints have been taken into account, and such considerations could shed light onto the computational lower-bound of HSIC estimation corresponding to our statistical lower bound. Such an analysis has also not been carried out for the related MMD.

**Nyström KSD [Kalinke et al., 2025b].** In Chapter 5, together with our accelerated KSD estimator, we proposed an accelerated wild bootstrap for goodness-of-fit testing. Our guarantees show that the KSD estimator preserves the statistical accuracy of the quadratic-time estimator. One interesting line of future research is analyzing the impact of the Nyström method on the wild bootstrap-based test.[1] Another worthwhile avenue is exploring the sub-Gaussian requirement to handle the unbounded Stein feature map. Our variance proxy assumption on the inner product (Assumption 5.2.1) enables us to use tight concentration results for controlling the concentration of the covariance operator (Theorem A.3.1). Results with weaker assumptions exist [Zhivotovskiy, 2024, Nakakita et al., 2024], but did not allow us to obtain the $\sqrt{n}$-rate. A more refined analysis might be possible and permit weakening the assumption. Independently, similar to the HSIC lower-bound, obtaining a minimax result for KSD estimation would yield insights into the quality of existing estimators.

[1] A related analysis has recently been accomplished for a permutation-based two-sample test [Chatalic et al., 2025].

**MMD for change detection [Kalinke et al., 2025a].** We introduced a novel change detection algorithm with a new data structure, MMDEW, that builds upon two-sample testing with MMD. While the algorithm achieves state-of-the art results on benchmark tasks, the statistical analysis of the subsampling scheme is challenging. The logical next step is combining our proposed method with random Fourier features, which allow storing the feature maps explicitly, to obtain an efficient change detection algorithm with stronger theoretical foundations while featuring a similar asymptotic runtime. Another promising direction is investigating the connection of recent results on sequential two-sample tests by betting [Shekhar and Ramdas, 2024] and change detection [Shin et al., 2023].

Overall, this dissertation advanced the state of the art of estimating kernel-based information theoretical measures and answered the computational-statistical trade-off of Nyström-based HSIC and KSD. With their solid statistical foundations, all our methods will yield great benefits in practice, and we hope to see them find numerous practical applications.

# A. External Results

The structure of the appendix mirrors that of the main part of this thesis and collects the external results that we use. The appendix to Chapter 3 is in Section A.1, that to Chapter 4 is in Section A.2, that to Chapter 5 is in Section A.3, and that to Chapter 6 is in Section A.4.

## A.1. Appendix to Nyström $M$-Hilbert-Schmidt Independence Criterion

In this section two theorems and two lemmas used in Chapter 3 are recalled for self-completeness. Theorem A.1.1 is about bounding the error of Nyström mean embeddings [Chatalic et al., 2022, Theorem 4.1]. Theorem A.1.2 is a well-known result [Serfling, 1980, Section 5.6, Theorem A] for bounding the deviation of U-statistics. Lemma A.1.1 is about the connection between U- and V-statistics. Lemma A.1.2 recalls Markov's inequality.

**Theorem A.1.1** (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 $\phi_k$. Assume that there exists a constant $K \in (0, \infty)$ such that $\sup_{x \in \mathcal{X}} \sqrt{k(x,x)} \leq K$. Let $C_{\mathbb{P},k} = \mathbb{E}_{\mathbb{P}}\left[\phi_k(X) \otimes \phi_k(X)\right]$. 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 $n' \leq n$ 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*

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

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

$$n' \geq \max\left(67, 12K^2 \left\|C_{\mathbb{P},k}\right\|_{\mathrm{op}}^{-1}\right) \log\left(\frac{n'}{\delta}\right),$$

*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 $X_1, \ldots, X_n \sim \mathbb{P}$ $(n \geq m)$ with form

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

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.1.2** (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 $a \leq h(x_1, \ldots, x_m) \leq b$. Then, for any $u > 0$ and $n \geq m$,*

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

Similar to (A.1) one can consider an alternative (slightly biased) estimator of $\theta$, which is called V-statistic:

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

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.1.1** (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* (A.1) *and* (A.2)*, respectively. Then it holds that*

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

**Lemma A.1.2** (Markov inequality)**.** *For a real-valued random variable $X$ with probability distribution $\mathbb{P}$ and $a > 0$, it holds that*

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

## A.2. Appendix to The Minimax Rate of HSIC Estimation for Translation-Invariant Kernels

For self-completeness, we include the external statements that we use in Chapter 4. The well-known result by Bochner, stated in Theorem A.2.1, completely characterizes continuous bounded translation-invariant kernels. Theorem A.2.2 allows expressing MMD with continuous bounded translation-invariant kernels in terms of characteristic functions and Theorem A.2.3 gives an equivalent condition for a continuous bounded translation-invariant kernel to be characteristic. Theorem A.2.4 connects characteristic kernels to characteristic product kernels and to $\mathcal{I}$-characteristic product kernels on $\mathbb{R}^d$ (we include only the part relevant to this thesis for brevity). We recall Le Cam's method in Theorem A.2.5 and collect results on the Kullback-Leibler divergence in Lemma A.2.1 and Lemma A.2.2.

**Theorem A.2.1** (Bochner; Theorem 6.6; Wendland [2005])**.** *A continuous function $\kappa : \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if it is the Fourier transform of a finite nonnegative Borel measure $\Lambda$ on $\mathbb{R}^d$, that is,*

$$\kappa(\mathbf{x}) = \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}, \boldsymbol{\omega}\rangle} \mathrm{d}\Lambda(\boldsymbol{\omega}), \quad \text{for all } \mathbf{x} \in \mathbb{R}^d.$$

**Theorem A.2.2** (Corollary 4(i); Sriperumbudur et al. [2010])**.** *Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a continuous bounded translation-invariant kernel. Then, for any $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_1^+\left(\mathbb{R}^d\right)$,*

$$\mathrm{MMD}_k^2(\mathbb{P}, \mathbb{Q}) = \left\|\psi_{\mathbb{P}} - \psi_{\mathbb{Q}}\right\|^2_{L^2\left(\mathbb{R}^d, \Lambda_k\right)},$$

*with $\psi_{\mathbb{P}}$ and $\psi_{\mathbb{Q}}$ being the characteristic functions of $\mathbb{P}$ and $\mathbb{Q}$, respectively, and $\Lambda_k$ defined in* (2.4).

**Theorem A.2.3** (Theorem 9; Sriperumbudur et al. [2010]). *Suppose $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a continuous bounded translation-invariant kernel. Then $k$ is characteristic if and only if* $\operatorname{supp}(\Lambda_k) = \mathbb{R}^d$, *with $\Lambda_k$ defined as in* (2.4).

**Theorem A.2.4** (Theorem 4; Szabó and Sriperumbudur [2018]). *Suppose $k_m : \mathbb{R}^{d_m} \times \mathbb{R}^{d_m} \to \mathbb{R}$ is a continuous bounded and translation-invariant kernel for all $m \in [M]$. Then the following statements are equivalent:*

*(i)* $(k_m)_{m=1}^M$*-s are characteristic;*

*(ii)* $\otimes_{m=1}^M k_m$ *is characteristic;*

*(iii)* $\otimes_{m=1}^M k_m$ *is $\mathcal{I}$-characteristic.*

The next statement follows directly from Tsybakov [2009, Eq. (2.9)] and Tsybakov [2009, Theorem 2.2].

**Theorem A.2.5** (Theorem 2.2; Tsybakov [2009]). *Let $\mathcal{X}$ be a measurable space, $(\Theta, d)$ a semi-metric space, and $\mathcal{P}_\Theta = \{\mathbb{P}_\theta : \theta \in \Theta\}$ a class of probability measures on $\mathcal{X}$ indexed by $\Theta$. We observe data $D \sim \mathbb{P}_\theta \in \mathcal{P}_\Theta$ with some unknown parameter $\theta$. The goal is to estimate $\theta$. Let $\hat{\theta} = \hat{\theta}(D)$ be an estimator of $\theta$ based on $D$. Assume that there exist $\theta_0, \theta_1 \in \Theta$ such that $d(\theta_0, \theta_1) \geq 2s > 0$ and $\mathrm{KL}(\mathbb{P}_{\theta_1} || \mathbb{P}_{\theta_0}) \leq \alpha < \infty$ for $\alpha > 0$. Then*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{P}_\theta\left(d\left(\hat{\theta}, \theta\right) \geq s\right) \geq \max\left(\frac{e^{-\alpha}}{4}, \frac{1 - \sqrt{\alpha/2}}{2}\right).$$

We have the following property of the Kullback-Leibler divergence for product measures [Tsybakov, 2009, p. 85].

**Lemma A.2.1** (KL divergence of product measures). *Let $\mathbb{P} = \otimes_{i=1}^n \mathbb{P}_i$ and $\mathbb{Q} = \otimes_{i=1}^n \mathbb{Q}_i$. Then*

$$\mathrm{KL}(\mathbb{P} || \mathbb{Q}) = \sum_{i \in [n]} \mathrm{KL}(\mathbb{P}_i || \mathbb{Q}_i).$$

The following lemma [Duchi, 2007, p. 13] shows that the Kullback-Leibler divergence of multivariate Gaussians can be computed in closed form.

**Lemma A.2.2** (KL divergence of Gaussians). *The KL divergence of two normal distributions $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ and $\mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$ on $\mathbb{R}^d$ is*

$$\mathrm{KL}(\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1) || \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)) = \frac{\operatorname{tr}(\Sigma_0^{-1}\Sigma_1) + (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\mathsf{T} \Sigma_0^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) - d + \ln\left(\frac{|\Sigma_0|}{|\Sigma_1|}\right)}{2}.$$

## A.3. Appendix to Nyström Kernel Stein Discrepancy

This section collects the external statements that we use in Chapter 5. Lemma A.3.1 gives equivalent norms for $f \otimes f$. We collect properties of Orlicz norms in Lemma A.3.2. Theorem A.3.1 is about the concentration of the empirical covariance, and Theorem A.3.2 recalls Bernstein's inequality for separable Hilbert spaces. Theorem A.3.3 is a concentration result for bounded random variables in a separable Hilbert space.

**Lemma A.3.1** (Lemma B.8; Sriperumbudur and Sterge [2022]). *Define $B = f \otimes f$, where $f \in \mathcal{H}$ and $\mathcal{H}$ is a separable Hilbert space. Then $\|B\|_{\mathrm{op}} = \|B\|_{\mathcal{H}\otimes\mathcal{H}} = \operatorname{tr} B = \|f\|_{\mathcal{H}}^2$.*

We refer to the following sources for the items in Lemma A.3.2. Item 1 is Vershynin [2018, Lemma 2.6.8], Item 2 is Vershynin [2018, Exercise 2.7.10], Item 3 recalls van der Vaart and Wellner [1996, p. 95], and Item 4 is Vershynin [2018, Lemma 2.7.6].

**Lemma A.3.2** (Collection of Orlicz properties). *Let $X$ be a real-valued random variable.*

1. *If $X$ is sub-Gaussian, then $X - \mathbb{E}X$ is also sub-Gaussian, and*
$$\|X - \mathbb{E}X\|_{\psi_2} \leq \|X\|_{\psi_2} + \|\mathbb{E}X\|_{\psi_2} \lesssim \|X\|_{\psi_2}.$$
2. *If $X$ is sub-exponential, then $X - \mathbb{E}X$ is also sub-exponential, and satisfies*
$$\|X - \mathbb{E}X\|_{\psi_1} \leq \|X\|_{\psi_1} + \|\mathbb{E}X\|_{\psi_1} \lesssim \|X\|_{\psi_1}.$$
3. *If $X$ is sub-Gaussian, it is sub-exponential. Specifically, it holds that $\|X\|_{\psi_1} \leq \sqrt{\log 2}\,\|X\|_{\psi_2}$.*
4. *$X$ is sub-Gaussian if and only if $X^2$ is sub-exponential. Moreover,*
$$\left\|X^2\right\|_{\psi_1} = \|X\|_{\psi_2}^2.$$

**Theorem A.3.1** (Theorem 9; Koltchinskii and Lounici [2017]). *Let $X, X_1, \ldots, X_n$ be i.i.d. square integrable centered random vectors in a Hilbert space $\mathcal{H}$ with covariance operator $C$. Let the empirical covariance operator be $\hat{C}_n = \frac{1}{n}\sum_{i=1}^n X_i \otimes X_i$. If $X$ is sub-Gaussian, then there exists a constant $c > 0$ such that, for all $\delta \in (0,1)$, with probability at least $1-\delta$,*
$$\left\|\hat{C}_n - C\right\|_{\mathrm{op}} \leq c\,\|C\|_{\mathrm{op}} \max\left(\sqrt{\frac{r(C)}{n}},\ \sqrt{\frac{\log(1/\delta)}{n}}\right),$$
*provided that $n \geq \max\{r(C), \log(1/\delta)\}$, where $r(C) := \frac{\operatorname{tr}(C)}{\|C\|_{\mathrm{op}}}$.*

The following theorem by Yurinsky [1995] is quoted from Sriperumbudur and Sterge [2022].

**Theorem A.3.2** (Bernstein bound for separable Hilbert spaces; Theorem 3.3.4; Yurinsky [1995]). *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{H}$ a separable Hilbert space, $B > 0$, $\sigma > 0$, and $\eta_1, \ldots, \eta_n : \Omega \to \mathcal{H}$ centered i.i.d. random variables that satisfy*
$$\mathbb{E}\,\|\eta_1\|_{\mathcal{H}}^p \leq \frac{1}{2} p!\,\sigma^2 B^{p-2}$$
*for all $p \geq 2$. Then, for any $\delta \in (0,1)$ it holds with probability at least $1-\delta$ that*
$$\left\|\frac{1}{n}\sum_{i=1}^n \eta_i\right\|_{\mathcal{H}} \leq \frac{2B\log(2/\delta)}{n} + \sqrt{\frac{2\sigma^2\log(2/\delta)}{n}}.$$

**Theorem A.3.3** (Concentration in separable Hilbert spaces; Lemma E.1; Chatalic et al. [2022])**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables with zero mean in a separable Hilbert space $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ such that $\max_{i\in[n]} \|X_i\|_{\mathcal{H}} \leq b$ almost surely, for some $b > 0$. Then for any $\delta \in (0, 1)$, it holds with probability at least $1 - \delta$ that*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \right\|_{\mathcal{H}} \leq b \frac{\sqrt{2 \log(2/\delta)}}{\sqrt{n}}.$$

## A.4. Appendix to Maximum Mean Discrepancy on Exponential Windows for Online Change Detection

To prove Proposition 6.3.1, we recall McDiarmid's concentration inequality [McDiarmid, 1989] from Vershynin [2018].

**Theorem A.4.1** (Bounded differences inequality)**.** *Let $X = (X_1, \ldots, X_n)$ be a random vector with independent components. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a measurable function. Assume that the value of $f(\mathbf{x})$ can change by at most $c_i > 0$ under an arbitrary change of a single coordinate of $\mathbf{x} = (c_1, \ldots, c_n) \in \mathbb{R}^n$. Then, for any $t > 0$, we have*

$$P\{f(X) - \mathbb{E}f(X) \geq t\} \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right).$$

# B. Bibliography


S. Abbasi and A. Haq. Optimal CUSUM and adaptive CUSUM charts with auxiliary information for process mean. *Journal of Statistical Computation and Simulation*, 89(2):337–361, 2019.

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.

S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 28:131–142, 1966.

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 computational statistics: A review of some recent developments. *Statistical Science*, 38(1):120–139, 2023.

A. Anaya-Isaza and L. Mera-Jiménez. Data augmentation and transfer learning for brain tumor detection in magnetic resonance imaging. *IEEE Access*, 10:23217–23233, 2022.

D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *European Symposium on Artificial Neural Networks (ESANN)*, 2013.

R. G. Antonini. Subgaussian random variables in Hilbert spaces. *Rendiconti del Seminario Matematico della Università di Padova*, 98:89–99, 1997.

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

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

C. R. Baker. Mutual information for Gaussian processes. *SIAM Journal on Applied Mathematics*, 19: 451–458, 1970.

C. R. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

K. Balasubramanian, T. Li, and M. Yuan. On the optimality of kernel-embedding based goodness-of-fit tests. *Journal of Machine Learning Research*, 22(1):1–45, 2021.

R. Bardenet, A. Doucet, and C. C. Holmes. Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach. In *International Conference on Machine Learning (ICML)*, pages 405–413, 2014.

L. Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88:190–206, 2004.

L. Baringhaus and N. Henze. A consistent test for multivariate normality based on the empirical characteristic function. *Metrika. International Journal for Theoretical and Applied Statistics*, 35(6): 339–348, 1988.

J. Baum, H. Kanagawa, and A. Gretton. A kernel Stein test of goodness of fit for sequential models. In *International Conference on Machine Learning (ICML)*, pages 1936–1953, 2023.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory (COLT)*, pages 1046–1066, 2013.

T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 591–596, 2011.

R. Bhatia. *Positive Definite Matrices*. Princeton University Press, 2007.

A. Bifet and R. Gavaldà. Learning from time-changing data with adaptive windowing. In *SIAM International Conference on Data Mining (SDM)*, pages 443–448, 2007.

M. Bilodeau and A. G. Nangue. Tests of mutual or serial independence of random vectors with applications. *Journal of Machine Learning Research*, 18:1–40, 2017.

M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations (ICLR)*, 2018.

P. Bonnier, H. Oberhauser, and Z. Szabó. Kernelized cumulants: Beyond kernel mean embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11049–11074, 2023.

K. Borgwardt, E. Ghisu, F. Llinares-López, L. O'Bray, and B. Riec. Graph kernels: State-of-the-art and future challenges. *Foundations and Trends in Machine Learning*, 13(5-6):531–712, 2020.

D. Bouche, R. Flamary, F. d'Alché Buc, R. Plougonven, M. Clausel, J. Badosa, and P. Drobinski. Wind power predictions from nowcasts to 4-hour forecasts: a learning approach with variable selection. *Renewable Energy*, 211:938–947, 2023.

L. L. Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1:38–53, 1973.

G. Camps-Valls, J. M. Mooij, and B. Schölkopf. Remote sensing feature selection by kernel dependence measures. *IEEE Geoscience and Remote Sensing Letters*, 7(3):587–591, 2010.

C. Canonne. Tail bounds for maximum of sub-Gaussian random variables. Mathematics Stack Exchange, 2021. `https://math.stackexchange.com/q/4002713` (version: 2023-12-21).

A. Caponnetto and E. De Vito. Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

G. Casella and R. L. Berger. *Statistical inference*. Wadsworth & Brooks/Cole, 1990.

S. Chakraborty and X. Zhang. Distance metrics for measuring joint dependence with application to causal inference. *Journal of the American Statistical Association*, 114(528):1638–1650, 2019.

A. Chatalic, N. Schreuder, A. Rudi, and L. Rosasco. Nyström kernel mean embeddings. In *International Conference on Machine Learning (ICML)*, pages 3006–3024, 2022.

A. Chatalic, M. Letizia, N. Schreuder, and L. Rosasco. An efficient permutation-based kernel two-sample test. Technical report, 2025. `https://arxiv.org/abs/2502.13570`.

L. H. Y. Chen. Stein's method of normal approximation: Some recollections and reflections. *The Annals of Statistics*, 49(4):1850–1863, 2021.

W. Y. Chen, L. Mackey, J. Gorham, F.-X. Briol, and C. J. Oates. Stein points. In *International Conference on Machine Learning (ICML)*, pages 844–853, 2018.

W. Y. Chen, A. Barp, F.-X. Briol, J. Gorham, M. Girolami, L. Mackey, and C. Oates. Stein point Markov chain Monte Carlo. In *International Conference on Machine Learning (ICML)*, pages 1011–1021, 2019.

K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1972–1980, 2015.

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

H. Climente-González, C.-A. Azencott, S. Kaski, and M. Yamada. Block HSIC Lasso: model-free biomarker detection for ultra-high dimensional data. *Bioinformatics*, 35(14):i427–i435, 2019.

D. L. Cohn. *Measure Theory*. Birkhäuser/Springer, second edition, 2013.

J. Coullon, L. F. South, and C. Nemeth. Efficient and generalizable tuning strategies for stochastic gradient MCMC. *Statistics and Computing*, 33(3):66, 2023.

I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica. Combinatorics, Geometry and Topology (CoGeTo)*, 2: 299–318, 1967.

M. Cuturi. Fast global alignment kernels. In *International Conference on Machine Learning (ICML)*, pages 929–936, 2011.

M. Cuturi and J.-P. Vert. The context-tree kernel for strings. *Neural Networks*, 18(8):1111–1123, 2005.

M. Cuturi, J.-P. Vert, O. Birkenes, and T. Matsui. A kernel for time series based on global alignments. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 413–416, 2007.

T. Dasu, S. Krishnan, D. Lin, S. Venkatasubramanian, and K. Yi. Change (detection) you can believe in: Finding distributional shifts in data streams. In *International Symposium on Intelligent Data Analysis (IDA)*, pages 21–34, 2009.

V. M. A. de Souza, A. R. S. Parmezan, F. A. Chowdhury, and A. Mueen. Efficient unsupervised drift detector for fast and high-dimensional data streams. *Knowledge and Information Systems*, 63(6): 1497–1527, 2021.

A. Della Vecchia, J. Mourtada, E. De Vito, and L. Rosasco. Regularized ERM on random subspaces. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 4006–4014, 2021.

L. Deng. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, pages 141–142, 2012.

J. Diestel and J. J. Uhl. *Vector Measures*. American Mathematical Society, 1977.

P. Domingos and G. Hulten. A general framework for mining massive data streams. *Journal of Computational and Graphical Statistics*, 12(4):945–949, 2003.

J. Duchi. Derivations for linear algebra and optimization. Technical report, 2007. `https://web.stanford.edu/~jduchi/projects/general_notes.pdf`.

R. M. Dudley. *Real analysis and probability*. Cambridge University Press, 2002.

G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 258–267, 2015.

K. Faber, R. Corizzo, B. Sniezynski, M. Baron, and N. Japkowicz. WATCH: Wasserstein change point detection for high-dimensional time series data. In *IEEE International Conference on Big Data*, pages 4450–4459, 2021.

W. J. Faithfull, J. J. R. Diez, and L. I. Kuncheva. Combining univariate approaches for ensemble change detection in multivariate data. *Information Fusion*, pages 202–214, 2019.

A. Fedorov, E. Geenjaar, L. Wu, T. Sylvain, T. P. DeRamus, M. Luck, M. Misiura, G. Mittapalle, R. D. Hjelm, S. M. Plis, et al. Self-supervised multimodal learning for group inferences from MRI data: Discovering disorder-relevant brain regions and multimodal links. *NeuroImage*, 285:120485, 2024.

N. Fellmann, C. Blanchet-Scalliet, C. Helbert, A. Spagnol, and D. Sinoquet. Kernel-based sensitivity analysis for (excursion) sets. *Technometrics*, 66(4):575–587, 2024.

T. Fernandez, N. Rivera, W. Xu, and A. Gretton. Kernelized Stein discrepancy tests of goodness-of-fit for time-to-event data. In *International Conference on Machine Learning (ICML)*, pages 3112–3122, 2020.

M. Freitas Gustavo, M. Hellström, and T. Verstraelen. Sensitivity analysis for ReaxFF reparametrization using the Hilbert–Schmidt independence criterion. *Journal of Chemical Theory and Computation*, 19 (9):2557–2573, 2023.

R. Fukuda. Exponential integrability of sub-Gaussian vectors. *Probability Theory and Related Fields*, 85 (4):505–521, 1990.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 489–496, 2007.

F. Futami, Z. Cui, I. Sato, and M. Sugiyama. Bayesian posterior approximation via greedy particle optimization. In *AAAI Conference on Artificial Intelligence*, pages 3606–3613, 2019.

J. Gama. *Knowledge discovery from data streams.* CRC Press, 2010.

C. Gao, Z. Ma, and H. H. Zhou. Sparse CCA: adaptive estimation and computational barriers. *The Annals of Statistics*, 45(5):2074–2101, 2017.

D. Garreau, W. Jitkrittum, and M. Kanagawa. Large sample analysis of the median heuristic. Technical report, 2018. `https://arxiv.org/abs/1707.07269`.

T. Gärtner, P. Flach, A. Kowalczyk, and A. Smola. Multi-instance kernels. In *International Conference on Machine Learning (ICML)*, pages 179–186, 2002.

T. Gärtner, P. Flach, and S. Wrobel. On graph kernels: Hardness results and efficient alternatives. *Learning Theory and Kernel Machines*, pages 129–143, 2003.

G. Giorgobiani, V. Kvaratskhelia, and V. Tarieladze. Notes on sub-Gaussian random elements. In *Applications of Mathematics and Informatics in Natural Sciences and Engineering (AMINSE)*, pages 197–203, 2020.

T. Górecki, M. Krzyśko, and W. Wolyński. Independence test and canonical correlation analysis based on the alignment between kernel matrices for multivariate functional data. *Artificial Intelligence Review*, pages 1–25, 2018.

J. Gorham and L. Mackey. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 226–234, 2015.

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

J. Gorham, A. Raj, and L. Mackey. Stochastic Stein discrepancies. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 17931–17942, 2020.

Ö. Gözüaçik, A. Büyükçakir, H. R. Bonab, and F. Can. Unsupervised concept drift detection with a discriminative classifier. In *International Conference on Information and Knowledge Management (CIKM)*, pages 2365–2368, 2019.

R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics*. Addison-Wesley, 1994.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory (ALT)*, pages 63–78, 2005.

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

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

J. Guevara, R. Hirata, and S. Canu. Cross product kernels for fuzzy set similarity. In *International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6, 2017.

O. Hagrass, B. Sriperumbudur, and B. Li. Spectral regularized kernel two-sample tests. *The Annals of Statistics*, 52(3):1076–1101, 2024a.

O. Hagrass, B. K. Sriperumbudur, and B. Li. Spectral regularized kernel goodness-of-fit tests. *Journal of Machine Learning Research*, 25(309):1–52, 2024b.

O. Hagrass, B. Sriperumbudur, and K. Balasubramanian. Minimax optimal goodness-of-fit testing with kernel Stein discrepancy. *Bernoulli*, 2025. (accepted; preprint: `https://arxiv.org/abs/2404.08278`).

Z. Harchaoui and O. Cappé. Retrospective multiple change-point estimation with kernels. In *IEEE/SP Workshop on Statistical Signal Processing*, pages 768–772, 2007.

D. Haussler. Convolution kernels on discrete structures. Technical report, University of California at Santa Cruz, 1999. `http://cbse.soe.ucsc.edu/sites/default/files/convolutions.pdf`.

M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability measures. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 136–143, 2005.

S. Herrando-Pérez and F. Saltré. Estimating extinction time using radiocarbon dates. *Quaternary Geochronology*, 79:101489, 2024.

W. Hoeffding. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19:293–325, 1948.

Z. Huang, N. Deb, and B. Sen. Kernel partial correlation coefficient – a measure of conditional dependence. *Journal of Machine Learning Research*, 23(216):1–58, 2022.

J. H. Huggins and L. Mackey. Random feature Stein discrepancies. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1899–1909, 2018.

Y. I. Ingster and I. A. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*. Springer, 2003.

M. Izzatullah, R. Baptista, L. Mackey, Y. Marzouk, and D. Peter. Bayesian seismic inversion: Measuring Langevin MCMC sample quality with kernels. In *SEG International Exposition and Annual Meeting*, 2020.

Y. Jiao and J.-P. Vert. The Kendall and Mallows kernels for permutations. In *International Conference on Machine Learning (ICML)*, pages 2982–2990, 2016.

W. Jitkrittum, Z. Szabó, K. Chwialkowski, and A. Gretton. Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 181–189, 2016.

W. Jitkrittum, Z. Szabó, and A. Gretton. An adaptive test of independence with analytic kernel embeddings. In *International Conference on Machine Learning (ICML)*, pages 1742–1751, 2017a.

W. Jitkrittum, W. Xu, Z. Szabó, K. Fukumizu, and A. Gretton. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 262–271, 2017b.

W. Jitkrittum, H. Kanagawa, and B. Schölkopf. Testing goodness of fit of conditional density models with kernels. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 221–230, 2020.

F. Kalinke and Z. Szabó. Nyström M-Hilbert-Schmidt independence criterion. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 1005–1015, 2023.

F. Kalinke and Z. Szabó. The minimax rate of HSIC estimation for translation-invariant kernels. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 108468–108489, 2024.

F. Kalinke, M. Heyden, G. Gntuni, E. Fouché, and K. Böhm. Maximum mean discrepancy on exponential windows for online change detection. *Transactions on Machine Learning Research*, 2025a.

F. Kalinke, Z. Szabó, and B. K. Sriperumbudur. Nyström kernel Stein discrepancy. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 388–396, 2025b.

N. Keriven, D. Garreau, and I. Poli. NEWMA: A new method for scalable model-free online change-point detection. *IEEE Transactions on Signal Processing*, 68:3515–3528, 2020.

F. J. Király and H. Oberhauser. Kernels for sequentially ordered data. *Journal of Machine Learning Research*, 20:1–45, 2019.

I. Klebanov, I. Schuster, and T. J. Sullivan. A rigorous theory of conditional mean embeddings. *SIAM Journal on Mathematics of Data Science*, 2(3):583–606, 2020.

D. Koller and N. Friedman. *Probabilistic graphical models*. MIT Press, 2009.

A. N. Kolmogorov. Sulla determinazione empirica delle leggi di probabilita. *Giornale dell'Istituto Italiano degli Attuari*, 4(1), 1933.

V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017.

A. Korattikara, Y. Chen, and M. Welling. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *International Conference on Machine Learning (ICML)*, pages 181–189, 2014.

A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. Technical report, 2009. `https://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf`.

S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22: 79–86, 1951.

H. C. L. Law, D. J. Sutherland, D. Sejdinovic, and S. R. Flaxman. Bayesian approaches to distribution regression. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1167–1176, 2018.

E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer, 2021.

S. Li, Y. Xie, H. Dai, and L. Song. Scan *B*-statistic for kernel change-point detection. *Sequential Analysis*, 38(4):503–544, 2019.

Y. Li, K. Swersky, and R. S. Zemel. Generative moment matching networks. In *International Conference on Machine Learning (ICML)*, pages 1718–1727, 2015.

Z. Li, J.-F. Ton, D. Oglic, and D. Sejdinovic. Towards a unified analysis of random Fourier features. *Journal of Machine Learning Research*, 22(108):1–51, 2021.

J. N. Lim, M. Yamada, B. Schölkopf, and W. Jitkrittum. Kernel Stein tests for multiple model comparison. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2243–2253, 2019.

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 (ICML)*, pages 6316–6326, 2020.

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

H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, 2002.

G. Lorden. On excess over the boundary. *Annals of Mathematical Statistics*, 41:520–527, 1970.

G. Lorden and M. Pollak. Nonanticipating estimation applied to sequential analysis and changepoint detection. *The Annals of Statistics*, 33(3):1422–1454, 2005.

R. Lyons. Distance covariance in metric spaces. *The Annals of Probability*, 41:3284–3305, 2013.

D. Martinez-Taboada and E. Kennedy. Counterfactual density estimation using kernel Stein discrepancies. In *International Conference on Learning Representations (ICLR)*, 2024.

C. McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

Q. Mérigot. A multiscale approach to optimal transport. *Computer Graphics Forum*, 30(5):1583–1592, 2011.

C. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006.

J. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17:1–102, 2016.

K. Muandet, K. Fukumizu, F. Dinuzzo, and B. Schölkopf. Learning from distributions via support measure machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10–18, 2011.

K. Muandet, K. Fukumizu, B. K. Sriperumbudur, and B. Schölkopf. 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*, 29:429–443, 1997.

S. Nakakita, P. Alquier, and M. Imaizumi. Dimension-free bounds for sums of dependent matrices and operators with heavy-tailed distributions. *Electronic Journal of Statistics*, 18(1):1130–1159, 2024.

E. T. Nalisnick, A. Matsukawa, Y. W. Teh, and B. Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using a test for typicality. In *Workshop on Bayesian Deep Learning. Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

G. Nikolentzos and M. Vazirgiannis. Graph alignment kernels using Weisfeiler and Leman hierarchies. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2019–2034, 2023.

Y. Nishiyama and K. Fukumizu. Characteristic kernels and infinitely divisible distributions. *Journal of Machine Learning Research*, 17:1–28, 2016.

E. J. Nyström. Über die Praktische Auflösung von Integralgleichungen mit Anwendungen auf Randwertaufgaben. *Acta Mathematica*, 54(1):185–204, 1930.

C. J. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):695–718, 2017.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41:100–115, 1954.

V. I. Paulsen and M. Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press, 2016.

N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):5–31, 2018.

A. Podkopaev, P. Blöbaum, S. Kasiviswanathan, and A. Ramdas. Sequential kernelized independence testing. In *International Conference on Machine Learning (ICML)*, pages 27957–27993, 2023.

N. Quadrianto, L. Song, and A. Smola. Kernelized sorting. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1289–1296, 2009.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1177–1184, 2007.

G. Romano, I. A. Eckley, and P. Fearnhead. A log-linear non-parametric online changepoint detection algorithm based on functional pruning. *IEEE Transactions on Signal Processing*, 72:594–606, 2024.

G. J. Ross. Parametric and nonparametric sequential change detection in R: The cpm package. *Journal of Statistical Software*, 66:1–20, 2015.

G. J. Ross and N. M. Adams. Two nonparametric control charts for detecting arbitrary distribution changes. *Journal of Quality Technology*, 44(2):102–116, 2012.

P. K. Rubenstein, O. Bousquet, J. Djolonga, C. Riquelme, and I. O. Tolstikhin. Practical and consistent estimation of $f$-divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4072–4082, 2019.

A. Rudi, R. Camoriano, and L. Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1657–1665, 2015.

S. Saitoh and Y. Sawano. *Theory of Reproducing Kernels and Applications*. Springer, 2016.

R. Salakhutdinov. Learning deep generative models. *Annual Review of Statistics and Its Application*, 2: 361–385, 2015.

M. Sarvmaili, H. Sajjad, and G. Wu. Data-centric prediction explanation via kernelized Stein discrepancy. In *International Conference on Learning Representations (ICLR)*, 2025.

B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

A. Schrab, B. Guedj, and A. Gretton. KSD aggregated goodness-of-fit test. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 32624–32638, 2022a.

A. Schrab, I. Kim, B. Guedj, and A. Gretton. Efficient aggregated kernel tests using incomplete U-statistics. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 18793–18807, 2022b.

A. Schrab, I. Kim, B. Guedj, and A. Gretton. Efficient aggregated kernel tests using incomplete U-statistics. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 18793–18807, 2022c.

D. Sejdinovic, A. Gretton, and W. Bergsma. A kernel test for three-variable interactions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1124–1132, 2013a.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, 41:2263–2291, 2013b.

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

S. Shekhar and A. Ramdas. Nonparametric two-sample testing by betting. *IEEE Transactions on Information Theory*, 70(2):1178–1203, 2024.

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

T. Sheng and B. K. Sriperumbudur. On distance and kernel measures of conditional independence. *Journal of Machine Learning Research*, 24(7):1–16, 2023.

W. A. Shewhart. The application of statistics as an aid in maintaining quality of a manufactured product. *Journal of the American Statistical Association*, 20(152):546–548, 1925.

J. Shin, A. Ramdas, and A. Rinaldo. E-detectors: a nonparametric framework for sequential change detection. Technical report, 2023. `https://arxiv.org/abs/2203.03532`.

D. Siegmund and E. S. Venkatraman. Using the generalized likelihood ratio statistic for sequential detection of a change-point. *The Annals of Statistics*, 23(1):255–271, 1995.

N. J. A. Sloane. Entry A001788 in The On-Line Encyclopedia of Integer Sequences, 1999a. `https://oeis.org/A001788`.

N. J. A. Sloane. Entry A036289 in The On-Line Encyclopedia of Integer Sequences, 1999b. `https://oeis.org/A036289`.

N. V. Smirnov. Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, 19:279–281, 1948.

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory (ALT)*, pages 13–31, 2007.

L. Song, A. J. Smola, A. Gretton, and K. M. Borgwardt. A dependence maximization view of clustering. In *International Conference on Machine Learning (ICML)*, pages 815–822, 2007.

L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(1):1393–1434, 2012.

R. S. Sparks. CUSUM charts for signalling varying location shifts. *Journal of Quality Technology*, 32(2): 157–171, 2000.

P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72, 1991.

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

B. K. Sriperumbudur and N. Sterge. Approximate kernel PCA: computational versus statistical trade-off. *The Annals of Statistics*, 50(5):2713–2736, 2022.

B. K. Sriperumbudur and Z. Szabó. Optimal rates for random Fourier features. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1144–1152, 2015.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Berkeley Symposium on Mathematical Statistics and Probability*, pages 583–602, 1972.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

J. Stenger, F. Gamboa, M. Keller, and B. Iooss. Optimal uncertainty quantification of a risk measurement from a thermal-hydraulic code using canonical moments. *International Journal for Uncertainty Quantification*, 10(1), 2020.

N. Sterge and B. K. Sriperumbudur. Statistical optimality and computational efficiency of Nyström kernel PCA. *Journal of Machine Learning Research*, 23(337):1–32, 2022.

Z. Szabó and B. K. Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18(233):1–29, 2018.

Z. Szabó, B. K. Sriperumbudur, B. Póczos, and A. Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40, 2016.

G. Székely and M. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5:1249–1272, 2004.

G. Székely and M. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93: 58–80, 2005.

G. J. Székely and M. L. Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 3:1236–1265, 2009.

G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35:2769–2794, 2007.

M. Talagrand. Regularity of Gaussian processes. *Acta Mathematica*, 159(1-2):99–149, 1987.

I. Tolstikhin, B. Sriperumbudur, and B. Schölkopf. Minimax estimation of maximal mean discrepancy with radial kernels. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1930–1938, 2016.

I. Tolstikhin, B. Sriperumbudur, and K. Muandet. Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18:1–47, 2017.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

R. v. Mises. On the asymptotic distribution of differentiable statistical functions. *Annals of Mathematical Statistics*, 18:309–348, 1947.

G. J. J. van den Burg and C. K. I. Williams. An evaluation of change point detection algorithms. Technical report, 2020. `https://arxiv.org/abs/2003.06222`.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.

S. D. Veiga. Global sensitivity analysis with dependence measures. *Journal of Statistical Computation and Simulation*, 85(7):1283–1305, 2015.

A. Vergara, S. Vembu, T. Ayhan, M. A. Ryan, M. L. Homer, and R. Huerta. Chemical gas sensor drift compensation using classifier ensembles. *Sensors and Actuators B: Chemical*, pages 320–329, 2012.

R. Vershynin. *High-dimensional probability*. Cambridge University Press, 2018.

S. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt. Graph kernels. *Journal of Machine Learning Research*, 11(40):1201–1242, 2010.

A. Wang, J. Du, X. Zhang, and J. Shi. Ranking features to promote diversity: An approach based on sparse distance correlation. *Technometrics*, 64(3):384–395, 2022.

H. Wang and Y. Xie. Sequential change-point detection: Computation versus statistical performance. *Wiley Interdisciplinary Reviews: Computational Statistics*, 16(1):e1628, 2024.

T. Wang, Q. Berthet, and R. J. Samworth. Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics*, 44(5):1896–1930, 2016.

C. Watkins. Dynamic alignment kernels. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 39–50, 1999.

L. Wehbe and A. Ramdas. Nonparametric independence testing for small sample sizes. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3777–3783, 2015.

S. Wei and Y. Xie. Online kernel CUSUM for change-point detection. Technical report, 2022. `https://arxiv.org/abs/2211.15070`.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning (ICML)*, pages 681–688, 2011.

H. Wendland. *Scattered data approximation*. Cambridge University Press, 2005.

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

C. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 682–688, 2001.

S. J. Wright. Modified Cholesky factorizations in interior-point algorithms for linear programming. *SIAM Journal on Optimization*, 9(4):1159–1191, 1999.

G. Wynne, M. Kasprzak, and A. B. Duncan. A spectral representation of kernel Stein discrepancy with application to goodness-of-fit tests for measures on infinite dimensional Hilbert spaces. *Bernoulli*, 2024. (accepted; preprint: `https://arxiv.org/abs/2206.04552`).

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. Technical report, 2017. https://arxiv.org/abs/1708.07747.

L. Xie, G. V. Moustakides, and Y. Xie. Window-limited CUSUM for sequential change detection. *IEEE Transactions on Information Theory*, 69(9):5990–6005, 2023.

Y. Xie and D. Siegmund. Sequential multi-sensor change-point detection. *The Annals of Statistics*, 41(2): 670–692, 2013.

W. Xu and G. Reinert. A Stein goodness-of-test for exponential random graph models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 415–423, 2021.

M. Yamada, W. Jitkrittum, L. Sigal, E. P. Xing, and M. Sugiyama. High-dimensional feature selection by feature-wise kernelized Lasso. *Neural Computation*, 26(1):185–207, 2014.

J. Yang, Q. Liu, V. Rao, and J. Neville. Goodness-of-fit testing for discrete distributions via Stein discrepancy. In *International Conference on Machine Learning (ICML)*, pages 5561–5570, 2018.

J. Yang, V. A. Rao, and J. Neville. A Stein-Papangelou goodness-of-fit test for point processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 226–235, 2019.

V. Yurinsky. *Sums and Gaussian vectors*. Springer, 1995.

W. Zaremba, A. Gretton, and M. B. Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In *Neural Information Processing Systems (NeurIPS)*, pages 755–763, 2013.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 804–813, 2011.

Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):1–18, 2018.

T. Zhang. Effective dimension and generalization of kernel learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 454–461, 2002.

N. Zhivotovskiy. Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *Electronic Journal of Probability*, 29, 2024.

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

Y. Zhou, D.-R. Chen, and W. Huang. A class of optimal estimators for the covariance operator in reproducing kernel Hilbert spaces. *Journal of Multivariate Analysis*, 169:166–178, 2019.

V. Zolotarev. Probability metrics. *Theory of Probability and its Applications*, 28:278–302, 1983.