Nyström Kernel Stein Discrepancy

Florian Kalinke[1], Zoltán Szabó[2], Bharath K. Sriperumbudur[3]
[1]Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]London School of Economics, London, UK
[3]The Pennsylvania State University, University Park, PA 16802, USA
florian.kalinke@kit.edu z.szabo@lse.ac.uk bks18@psu.edu

Abstract

Kernel methods underpin many of the most successful approaches in data science and statistics, and they allow representing probability measures as elements of a reproducing kernel Hilbert space without loss of information. Recently, the kernel Stein discrepancy (KSD), which combines Stein's method with the flexibility of kernel techniques, gained considerable attention. Through the Stein operator, KSD allows the construction of powerful goodness-of-fit tests where it is sufficient to know the target distribution up to a multiplicative constant. However, the typical U- and V-statistic-based KSD estimators suffer from a quadratic runtime complexity, which hinders their application in large-scale settings. In this work, we propose a Nyström-based KSD acceleration—with runtime $\mathcal{O}(mn + m^3)$ for n samples and $m \ll n$ Nyström points—, show its $\sqrt{n}$-consistency with a classical sub-Gaussian assumption, and demonstrate its applicability for goodness-of-fit testing on a suite of benchmarks.

1 INTRODUCTION

The kernel mean embedding, which involves mapping probability distributions into a reproducing kernel Hilbert space (RKHS; Aronszajn 1950) has found various far-reaching applications in the last 20 years. For example, it allows to measure the discrepancy between probability distributions through maximum mean discrepancy (MMD; Smola et al. 2007, Gretton et al. 2012), defined as the distance between the corresponding mean embeddings, which underpins powerful two-sample tests. MMD is also known as energy distance [Székely and Rizzo, 2004, 2005, Baringhaus and Franz, 2004] in the statistics literature; see Sejdinovic et al. [2013] for the equivalence. We refer to [Muandet et al., 2017] for a recent overview of kernel mean embeddings.

In addition to two-sample 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 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 work. 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., 2024]. 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].

Estimators for Langevin Stein operator-based KSD exist. But, the classical U-statistic- [Liu et al., 2016] and V-statistic-based [Chwialkowski et al., 2016] estimators 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. [2017] 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] 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.

Our main **contributions** are the following.

1. We introduce a Nyström-based acceleration of kernel Stein discrepancy. The proposed estimator runs in $\mathcal{O}\left(mn + m^3\right)$ time, with n samples and $m \ll n$ Nyström points.
2. 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.
3. We perform an extensive suite of experiments to demonstrate the applicability of the proposed method. Our proposed approach achieves competitive results throughout all experiments.

The paper is structured as follows. We introduce the notations used throughout the article (Section 2) followed by recalling the classical quadratic-time KSD estimators (Section 3). In Section 4.1, we detail our proposed Nyström-based estimator, alongside with its adaptation to a modified wild bootstrap goodness-of-fit test (Section 4.2), and our theoretical guarantees (Section 4.3). Experiments demonstrating the efficiency of our Nyström-KSD estimator are provided in Section 5. Proofs and additional experiments are deferred to the appendices.

2 NOTATIONS

In this section, we introduce our notations. Let $[N] := \{1, \ldots, N\}$ for a positive integer N. For $a_1, a_2 \geq 0$, $a_1 \lesssim a_2$ (resp. $a_1 \gtrsim a_2$) means that $a_1 \leq c a_2$ (resp. $a_1 \geq c' a_2$) 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 $\mathbb{1}_{(\cdot)}$ for the indicator function and $\{\{\cdot\}\}$ for a multiset. The n-dimensional vector of ones is denoted by $\mathbf{1}_n = (1, \ldots, 1)^\mathsf{T} \in \mathbb{R}^n$, that of n zeros by $\mathbf{0}_n = (0, \ldots, 0)^\mathsf{T} \in \mathbb{R}^n$. The identity matrix is $\mathbf{I}_n \in \mathbb{R}^{n \times n}$. For a matrix $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}$, $\mathbf{A}^- \in \mathbb{R}^{d_2 \times d_1}$ denotes its (Moore-Penrose) pseudo-inverse, and $\mathbf{A}^\mathsf{T} \in \mathbb{R}^{d_2 \times d_1}$ stands for the transpose of $\mathbf{A}$. We write $\mathbf{A}^{-1} \in \mathbb{R}^{d \times d}$ for the inverse of a non-singular matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$. For a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, let $\nabla_{\mathbf{x}} f(\mathbf{x}) = \left(\frac{\partial f(\mathbf{x})}{\partial x_i}\right)_{i=1}^d \in \mathbb{R}^d$.

Let $(\mathcal{X}, \tau_\mathcal{X})$ be a topological space and $\mathcal{B}(\tau_\mathcal{X})$ the corresponding Borel σ-algebra. Probability measures considered in this article are meant w.r.t. the measurable space $(\mathcal{X}, \mathcal{B}(\tau_\mathcal{X}))$ and are written as $\mathcal{M}_1^+(\mathcal{X})$; for instance, the set of Borel probability measures on $\mathbb{R}^d$ is $\mathcal{M}_1^+\left(\mathbb{R}^d\right)$. 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 $(\mathcal{X}_1, \tau_{\mathcal{X}_1})$ and $(\mathcal{X}_2, \tau_{\mathcal{X}_2})$ are topological spaces. For a sequence of i.i.d. real-valued random variables $X_n \sim \mathbb{P} \in \mathcal{M}_1^+(\mathbb{R})$ and a sequence of positive r_n-s, $X_n = \mathcal{O}_\mathbb{P}(r_n)$ means that $\frac{X_n}{r_n}$ is bounded in probability. The unit ball in a Hilbert space $\mathcal{H}$ is denoted by $B(\mathcal{H}) = \{f \in \mathcal{H} \,|\, \|f\|_\mathcal{H} \leq 1\}$. The reproducing kernel Hilbert space with $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as the reproducing kernel is denoted by $\mathcal{H}_k$. Throughout the paper, k is assumed to be measurable and $\mathcal{H}_k$ to be separable.[1] Given a closed linear subspace $U \subseteq \mathcal{H}_k$, the (orthogonal) projection of $h \in \mathcal{H}_k$ 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 $u \in U$, $\|h - P_U h\|_{\mathcal{H}_k} \leq \|h - u\|_{\mathcal{H}_k}$, that is, $P_U h$ is the closest element in U to h. A linear operator $A : \mathcal{H}_k \to \mathcal{H}_k$ is called bounded if $\|A\|_{\text{op}} := \sup_{\|h\|_{\mathcal{H}_k}=1} \|Ah\|_{\mathcal{H}_k} < \infty$; the set of $\mathcal{H}_k \to \mathcal{H}_k$ bounded linear operators is denoted by $\mathcal{L}(\mathcal{H}_k)$. An $A \in \mathcal{L}(\mathcal{H}_k)$ is called positive (shortly $A \geq 0$) if it is self-adjoint ($A^* = A$, where $A^* \in \mathcal{L}(\mathcal{H}_k)$ is defined by $\langle Af, g\rangle_{\mathcal{H}_k} = \langle f, A^* g\rangle_{\mathcal{H}_k}$ for all $f, g \in \mathcal{H}_k$), and $\langle Ah, h\rangle_{\mathcal{H}_k} \geq 0$ for all $h \in \mathcal{H}_k$. If $A \geq 0$,

[1] For instance, a continuous kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ implies both properties; see Steinwart and Christmann [2008, Lemma 4.33] for separability.

then there exists a unique $B \geq 0$ such that $B^2 = A$; we write $B = A^{\frac{1}{2}}$ and call B the square root of A. An $A \in \mathcal{L}(\mathcal{H}_k)$ is called trace-class if $\sum_{i\in I}\langle (A^*A)^{\frac{1}{2}} e_i, e_i\rangle_{\mathcal{H}_k} < \infty$ for some countable orthonormal basis (ONB) $(e_i)_{i\in I}$ of $\mathcal{H}_k$, and in this case $\mathrm{tr}(A) := \sum_{i\in I}\langle Ae_i, e_i\rangle_{\mathcal{H}_k} < \infty$.[2] 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 $A \in \mathcal{L}(\mathcal{H}_k)$ is called compact if $\overline{\{Ah \,|\, h \in B(\mathcal{H}_k)\}}$ is compact, where $\bar{\cdot}$ denotes the closure. A trace class operator is compact, and a compact positive operator A has largest eigenvalue $\|A\|_{\mathrm{op}}$. For any $A \in \mathcal{L}(\mathcal{H}_k)$, it holds that $\|A^*A\|_{\mathrm{op}} = \|A\|^2_{\mathrm{op}}$ (which is called the C^* property).

The mean embedding of a probability measure $\mathbb{P} \in \mathcal{M}_1^+(\mathbb{R}^d)$ into the RKHS associated to kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is $\mu_k(\mathbb{P}) = \int_{\mathbb{R}^d} k(\cdot, \mathbf{x})\, \mathrm{d}\mathbb{P}(\mathbf{x}) \in \mathcal{H}_k$, where the integral is meant in Bochner's sense [Diestel and Uhl, 1977, Chapter II.2]. The mean element $\mu_k(\mathbb{P})$ exists iff. $\int_{\mathbb{R}^d} \|k(\cdot, \mathbf{x})\|_{\mathcal{H}_k} \mathrm{d}\mathbb{P}(\mathbf{x}) < \infty$ [Diestel and Uhl, 1977, p. 45; Theorem 2].

Let $f, g \in \mathcal{H}_k$. Their tensor product is written as $f \otimes g \in \mathcal{H}_k \otimes \mathcal{H}_k$, where $\mathcal{H}_k \otimes \mathcal{H}_k$ is the tensor product Hilbert space; further, $f \otimes g : \mathcal{H}_k \to \mathcal{H}_k$ defines a rank-one operator by $h \mapsto f\langle g, h\rangle_{\mathcal{H}_k}$. It is known that $\mathcal{H}_k \otimes \mathcal{H}_k$ is also an RKHS [Berlinet and Thomas-Agnan, 2004, Theorem 13]. Given a probability measure $\mathbb{P} \in \mathcal{M}_1^+\left(\mathbb{R}^d\right)$ and a kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, the uncentered covariance operator

$$C_{\mathbb{P},k} = \int_{\mathbb{R}^d} k(\cdot, \mathbf{x}) \otimes k(\cdot, \mathbf{x})\, \mathrm{d}\mathbb{P}(\mathbf{x}) \in \mathcal{H}_k \otimes \mathcal{H}_k$$

exists if $\int_{\mathcal{X}} \|k(\cdot, \mathbf{x})\|^2_{\mathcal{H}_k} \mathrm{d}\mathbb{P}(\mathbf{x}) < \infty$; $C_{\mathbb{P},k}$ is a positive trace-class operator. We define $C_{\mathbb{P},k,\lambda} = C_{\mathbb{P},k} + \lambda I$, where I denotes the identity operator and $\lambda > 0$. The effective dimension of $\mathbb{P} \in \mathcal{M}_1^+\left(\mathbb{R}^d\right)$ is defined as $\mathcal{N}_{\mathbb{P},k}(\lambda) := \mathrm{tr}\left(C_{\mathbb{P},k} C^{-1}_{\mathbb{P},k,\lambda}\right) \leq \frac{\mathrm{tr}(C_{\mathbb{P},k})}{\lambda}$.[3] With $r \geq 1$ and a real-valued random variable $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\tau_{\mathbb{R}}))$, where $\mathcal{B}(\tau_{\mathbb{R}})$ denotes the Borel σ-field on $\mathbb{R}$, let $\|X\|_{L_r(\mathbb{P})} = \left[\int_\Omega |X(\omega)|^r \mathrm{d}\mathbb{P}(\omega)\right]^{\frac{1}{r}}$. 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 $\|X\|_{\psi_2} < \infty$. In the following, we specialize Definition 2 by Koltchinskii and Lounici [2017] stated for Banach spaces to (reproducing kernel) Hilbert spaces by using the Riesz representation theorem. A centered $\mathcal{H}_k$-valued random variable $X \sim \mathbb{Q} \in \mathcal{M}_1^+(\mathcal{H}_k)$ is called sub-Gaussian iff. there exists a universal constant $C > 0$ such that for all $u \in \mathcal{H}_k$:

$$\left\|\langle X, u\rangle_{\mathcal{H}_k}\right\|_{\psi_2} \leq C \left\|\langle X, u\rangle_{\mathcal{H}_k}\right\|_{L_2(\mathbb{Q})} < \infty. \tag{1}$$

3 PROBLEM FORMULATION

We now introduce our quantity of interest, the kernel Stein discrepancy. 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 $\mathbf{f} = (f_i)_{i=1}^d, \mathbf{g} = (g_i)_{i=1}^d \in \mathcal{H}_k^d$. Let $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_1^+\left(\mathbb{R}^d\right)$ be fixed; we refer to $\mathbb{P}$ as the target distribution and to $\mathbb{Q}$ as the sampling distribution. Assume that $\mathbb{P}$ is absolutely continuous w.r.t. the Lebesgue measure and let p be the corresponding density (w.r.t. Lebesgue measure). We assume that p is continuously differentiable with support $\mathbb{R}^d$, $p(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathbb{R}^d$, and $\lim_{\|\mathbf{x}\|\to\infty} f(\mathbf{x})p(\mathbf{x}) = 0$

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

[3] This 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 $(\lambda_i)_{i\in I}$ denote the eigenvalues of $C_{\mathbb{P},k}$.

for all $f \in \mathcal{H}_k$. The last property holds for instance if p is bounded and $\lim_{\|\mathbf{x}\|\to\infty} f(\mathbf{x}) = 0$ for all $f \in \mathcal{H}_k$. Further, we assume that k is continuously differentiable in both arguments. This condition will imply the measurability of h_p and the separability of $\mathcal{H}_{h_p}$, both quantities defined below. The Stein operator is defined as $(T_p\mathbf{f})(\mathbf{x}) = \langle \nabla_{\mathbf{x}}[\log p(\mathbf{x})], \mathbf{f}(\mathbf{x})\rangle + \langle \nabla_{\mathbf{x}}\mathbf{f}(\mathbf{x}), \mathbf{1}_d\rangle$ $\left(\mathbf{f} \in \mathcal{H}_k^d,\, \mathbf{x} \in \mathbb{R}^d\right)$. With this notation at hand, for all $\mathbf{f} \in \mathcal{H}_k^d$ and $\mathbf{x} \in \mathbb{R}^d$ one has

$$\begin{aligned}(T_p\mathbf{f})(\mathbf{x}) &= \langle \mathbf{f}, \boldsymbol{\xi}_p(\mathbf{x})\rangle_{\mathcal{H}_k^d}, \\ \boldsymbol{\xi}_p(\mathbf{x}) &= \left[\nabla_{\mathbf{x}}\left(\log p(\mathbf{x})\right) k(\cdot, \mathbf{x}) + \nabla_{\mathbf{x}} k(\cdot, \mathbf{x})\right] \in \mathcal{H}_k^d \end{aligned} \tag{2}$$

with the kernel (for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$)

$$h_p(\mathbf{x}, \mathbf{y}) = \langle \boldsymbol{\xi}_p(\mathbf{x}), \boldsymbol{\xi}_p(\mathbf{y})\rangle_{\mathcal{H}_k^d} = \langle h_p(\cdot, \mathbf{x}), h_p(\cdot, \mathbf{y})\rangle_{\mathcal{H}_{h_p}}; \tag{3}$$

notice that $\boldsymbol{\xi}_p(\mathbf{x})$ 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), takes the explicit form

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

The kernel Stein discrepancy (KSD; Chwialkowski et al. 2016, Liu et al. 2016 then is defined as an integral probability metric [Zolotarev, 1983, Müller, 1997]

$$\begin{aligned}S_p(\mathbb{Q}) &= \sup_{\mathbf{f}\in B\left(\mathcal{H}_k^d\right)} \underbrace{\mathbb{E}_{X\sim\mathbb{P}}\left[T_p\mathbf{f}(X)\right]}_{\overset{(a)}{=}0} - \mathbb{E}_{X\sim\mathbb{Q}}\left[T_p\mathbf{f}(X)\right] = \sup_{\mathbf{f}\in B\left(\mathcal{H}_k^d\right)} \langle \mathbf{f}, \mathbb{E}_{X\sim\mathbb{Q}}\boldsymbol{\xi}_p(X)\rangle_{\mathcal{H}_k^d} \\ &= \left\|\mathbb{E}_{X\sim\mathbb{Q}}\boldsymbol{\xi}_p(X)\right\|_{\mathcal{H}_k^d} \overset{(b)}{=} \left\|\mathbb{E}_{X\sim\mathbb{Q}} h_p(\cdot, X)\right\|_{\mathcal{H}_{h_p}},\end{aligned} \tag{4}$$

where (a) holds by the construction of KSD and (b) follows from (3).

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 $\hat{\mathbb{Q}}_n$; it takes the form

$$S_p^2\left(\hat{\mathbb{Q}}_n\right) = \frac{1}{n^2}\sum_{i,j=1}^{n} h_p(\mathbf{x}_i, \mathbf{x}_j), \tag{5}$$

and can be computed in $\mathcal{O}\left(n^2\right)$ time. The corresponding U-statistic-based estimator [Liu et al., 2016, (14)] omits the diagonal terms, that is, $S_{p,u}^2\left(\hat{\mathbb{Q}}_n\right) = \frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}^{n} h_p(\mathbf{x}_i, \mathbf{x}_j)$; it also has a runtime cost of $\mathcal{O}\left(n^2\right)$. For large-scale applications, the quadratic runtime is a significant bottleneck; this is the shortcoming we tackle in the following.

4 PROPOSED NYSTRÖM-KSD

To enable the efficient estimation of (4), we propose a Nyström technique-based estimator in Section 4.1 and an accelerated wild bootstrap test in Section 4.2. In Section 4.3, our consistency results are collected.

4.1 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 $\mathcal{H}_{h_p,m} := \operatorname{span}\{h_p(\cdot, \tilde{\mathbf{x}}_i) \mid i \in [m]\} \subset \mathcal{H}_{h_p}$, with feature map $h_p(\cdot, \tilde{\mathbf{x}}_i)$ and associated kernel h_p defined in (3). 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

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^m} \left\| \underbrace{\frac{1}{n}\sum_{i=1}^n h_p(\cdot, \mathbf{x}_i)}_{=\mathbb{E}_{X\sim\hat{\mathbb{Q}}_n} h_p(\cdot,X)} - \sum_{i=1}^m \alpha_i h_p(\cdot, \tilde{\mathbf{x}}_i) \right\|_{\mathcal{H}_{h_p}}, \tag{6}$$

which gives rise to the squared KSD estimator[4]

$$\tilde{S}_p^2\left(\hat{\mathbb{Q}}_n\right) := \left\| \sum_{i=1}^m \alpha_i h_p(\cdot, \tilde{\mathbf{x}}_i) \right\|_{\mathcal{H}_{h_p},m}^2 = \left\| P_{\mathcal{H}_{h_p,m}} \mathbb{E}_{X\sim\hat{\mathbb{Q}}_n} h_p(\cdot, X)) \right\|_{\mathcal{H}_{h_p,m}}^2. \tag{7}$$

Lemma 1 (Nyström-KSD Estimator)**.** *The squared KSD estimator* (7) *takes the form*

$$\tilde{S}_p^2\left(\hat{\mathbb{Q}}_n\right) = \boldsymbol{\beta}_p^{\mathsf{T}} \mathbf{K}_{h_p,m,m}^{-} \boldsymbol{\beta}_p, \tag{8}$$

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

Remark 1.

(a) ***Runtime complexity.*** *The computation of* (8) *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* $\mathcal{O}(mn)$ *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* $\mathcal{O}(m^3)$,[5] *dominating the cost of* $\mathcal{O}(m^2)$ *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* $\mathcal{O}\left(m^2\right)$*. Hence,* (8) *can be computed in* $\mathcal{O}\left(mn + m^3\right)$*, which means that for* $m = o\left(n^{2/3}\right)$*, our proposed Nyström-KSD estimator guarantees an asymptotic speedup.*

(b) ***Comparison of*** (5) ***and*** (8)***.*** *The Nyström estimator* (8) *recovers the V-statistic-based estimator* (5) *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* (8) *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 4.3.*

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

[5] Although faster algorithms for (pseudo) matrix inversion exist, we consider the runtime that one typically encounters in practice.

4.2 Nyström Bootstrap Testing

In this section, we discuss how one can use (8) 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 (8) 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

$$B_n = \frac{1}{n^2} \sum_{i,j=1}^{n} w_i w_j h_p \left(\mathbf{x}_i, \mathbf{x}_j\right), \tag{9}$$

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

$$w_i = \mathbb{1}_{(U_i > 0.5)} w_{i-1} - \mathbb{1}_{(U_i \leq 0.5)} w_{i-1}, \tag{10}$$

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

1. Calculate the test statistic (5).
2. Obtain D wild bootstrap samples $\{B_{n,i}\}_{i=1}^{D}$ with (9) and estimate the $1-\alpha$ empirical quantile of these samples.
3. Reject the null hypothesis if the test statistic (5) exceeds the quantile.

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

We propose the Nyström-based bootstrap

$$B_n^{\text{Nys}} = \frac{1}{n^2} \mathbf{w}^{\mathsf{T}} \mathbf{K}_{h_p,n,m} \mathbf{K}_{h_p,m,m}^{-} \mathbf{K}_{h_p,m,n} \mathbf{w}, \tag{11}$$

with $\mathbf{w} = (w_i)_{i=1}^{n} \in \mathbb{R}^n$ collecting the w_i-s ($i \in [n]$) defined in (10); $\mathbf{K}_{h_p,n,m}$ ($= \mathbf{K}_{h_p,m,n}^{\mathsf{T}}$) and $\mathbf{K}_{h_p,m,m}$ are defined as in Lemma 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 (11) approximates (9) but reduces the cost from $\mathcal{O}\left(n^2\right)$ to $\mathcal{O}\left(nm + m^3\right)$ if one computes from left to right (also refer to Remark 1(a)); in the case of $m = o\left(n^{2/3}\right)$ this guarantees an asymptotic speedup. We obtain a total cost of $\mathcal{O}\left(D\left(nm + m^3\right)\right)$ for obtaining the wild bootstrap samples. This acceleration allows KSD-based goodness-of-fit tests to be applied on large data sets.

4.3 Guarantees

This section is dedicated to the statistical behavior of the proposed Nyström-KSD estimator (8).

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 $\sup_{\mathbf{x} \in \mathbb{R}^d} \left\| h_p\left(\cdot, \mathbf{x}\right) \right\|_{\mathcal{H}_{h_p}} < \infty$, the consistency of (8) 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 $(i_j)_{j=1}^{m} \overset{\text{i.i.d.}}{\sim} \text{Unif}([n]) =: \Lambda$, which means that $\tilde{\mathbf{x}}_j = \mathbf{x}_{i_j}$ with $j \in [m]$.

Theorem 1 (Bounded case)**.** *Assume the setting of Lemma 1, $C_{\mathbb{Q},h_p} \neq 0$, $m \geq 4$ Nyström samples, and a bounded Stein feature map ($\sup_{\mathbf{x}\in\mathbb{R}^d} \|h_p(\cdot,\mathbf{x})\|_{\mathcal{H}_{h_p}} =: K < \infty$). 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 1 is not applicable, as it is illustrated in the following example with the frequently-used Gaussian kernel.

Example 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 (3), direct calculation yields (i) $\|\xi_p(\cdot,x)\|_{\mathcal{H}_k}^2 = x^2 + 2\gamma \overset{x\to\infty}{\to} \infty$ in the first, and (ii) $\|\xi_p(\cdot,x)\|_{\mathcal{H}_k}^2 = x^2c^{2\beta} - 2\beta c^{2(\beta-1)} \overset{x\to\infty}{\to} \infty$ in the second case.*

Remark 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., 2024, Remark 2], which clearly rules out Gaussian targets.*

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

Assumption 1. *The centered Stein feature map $\bar{h}_p(\cdot,X) = h_p(\cdot,X) - \mathbb{E}_{X\sim\mathbb{Q}}h_p(\cdot,X)$ with the sampling distribution $\mathbb{Q} \in \mathcal{M}_1^+\left(\mathbb{R}^d\right)$ is sub-Gaussian in the sense of (1), that is,*

$$\left\|\left\langle \bar{h}_p(\cdot,X), u\right\rangle_{\mathcal{H}_{h_p}}\right\|_{\psi_2} \lesssim \left\|\left\langle \bar{h}_p(\cdot,X), u\right\rangle_{\mathcal{H}_{h_p}}\right\|_{L_2(\mathbb{Q})} < \infty$$

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

We elaborate further on Assumption 1 in Remark 3(c), after we state our following main result.

Theorem 2 (Consistency of Nyström-KSD)**.** *Let Assumption 1 hold, $C_{\mathbb{Q},\bar{h}_p} \neq 0$, and assume the setting of Lemma 1. Then, for any $\delta \in (0,1)$ with $(\mathbb{Q}^n \otimes \Lambda^m)$-probability of at least $1-\delta$ it holds that*

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

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

To interpret the consistency guarantee of Theorem 2, we consider the three terms on the r.h.s. w.r.t. the magnitude of m. The first two terms converge with $\mathcal{O}\big(n^{-1/2}\big)$, independent of the choice of m. By using the universal upper bound $\mathcal{N}_{\mathbb{Q},\bar{h}_p}\left(\frac{c\,\mathrm{tr}\big(C_{\mathbb{Q},\bar{h}_p}\big)}{m}\right) \lesssim m$ on the effective dimension, the last term reveals that an overall rate of $\mathcal{O}\big(n^{-1/2}\big)$ 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 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 $\mathcal{O}\left(n^{-1/2}\right)$ can be achieved, depending on the properties of the covariance operator.

Corollary 1. *In the setting of Theorem 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*

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

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) $m \gtrsim \sqrt{n}\left(\log\left(1+\frac{c_1 n}{c\,\mathrm{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}\right)\log(12n/\delta)\log(12/\delta)\right)^{1/2}$ in the second case.

To interpret these rates—see Remark 3(d)—, we obtain the (matching) $\sqrt{n}$-consistency of the quadratic time estimator (5) in our following result.

Theorem 3 (Consistency of KSD)**.** *Assume that $\left\|\|h_p(\cdot, X)\|_{\mathcal{H}_{h_p}}\right\|_{\psi_2} < \infty$ and denote by $\hat{\mathbb{Q}}_n = \{X_1, \ldots, X_n\}$, where $\{X_i\}_{i\in[n]} \overset{i.i.d.}{\sim} \mathbb{Q}$. Then it holds that*

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

A few remarks are in order.

Remark 3.

(a) ***Runtime benefit.*** *Recall that — see Remark 1(a) —, our proposed Nyström estimator (8) 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 (5) is guaranteed for $\gamma < 1/2$; in the case of exponential decay, large enough n always suffices.*

(b) ***Comparison of Theorem 1 and Theorem 2.*** *Recall that both theorems target precisely the same estimators, Chatalic et al. [2022, (5)] and (8), 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.*

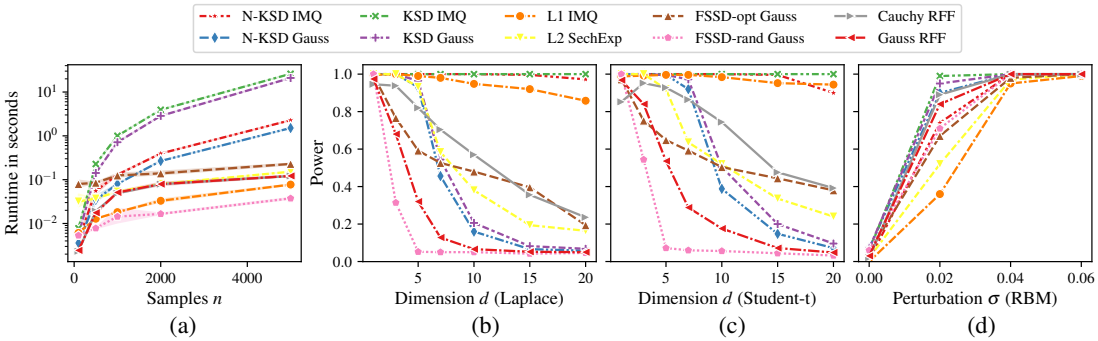

Figure 1: Comparison of goodness-of-fit tests w.r.t. their runtime and their power.

(c) ***Sub-Gaussian assumption.*** *Key to the proof of Theorem 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 selected Koltchinskii and Lounici [2017, Def. 2].*[6] *Specifically, this assumption allows us to derive our key Lemma B.1 and Lemma B.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.*[7] *The latter, Lemma B.3, intuitively states that norms of sub-Gaussian vectors whitened by $C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2}$ inherit the sub-Gaussian property. Together, these lemmas open the door to proving Theorem 2.*

(d) ***Comparison of Theorem 2 and Theorem 3.*** *With the weaker sub-Gaussian condition $\left\| \|h_p(\cdot, X)\|_{\mathcal{H}_{h_p}} \right\|_{\psi_2} < \infty$ (implied by Assumption 1, see Lemma B.3), Theorem 3 shows that the quadratic time estimator (5) converges with rate $\mathcal{O}\left(n^{-1/2}\right)$. Our Nyström result, Theorem 2 with Corollary 1, shows that a matching rate can be achieved (given an appropriate decay of the effective dimension) with $m = \tilde{\Theta}(\sqrt{n})$; 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., 2024] but we present them on $\mathbb{R}^d$, which one arguably most frequently encounters in practice, to simplify exposition.*

5 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,

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

[7] 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.

Chwialkowski et al., 2016], the linear-time goodness-of-fit test finite set Stein discrepancy (FSSD; Jitkrittum et al. 2017), 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).[8] 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 and implementations provided by the respective authors. We use the well-known Gaussian kernel $k(\mathbf{x}, \mathbf{y}) = \exp\left(-\gamma \left\|\mathbf{x} - \mathbf{y}\right\|_{\mathbb{R}^d}^2\right)$ $(\gamma > 0)$ with the median heuristic [Garreau et al., 2018], and the IMQ kernel $k(\mathbf{x}, \mathbf{y}) = \left(c^2 + \left\|\mathbf{x} - \mathbf{y}\right\|_{\mathbb{R}^d}^2\right)^{-\beta}$ [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 (11), setting $D = 500$. To allow an easy comparison, our experiments replicate goodness-of-fit testing experiments from Chwialkowski et al. [2016], Jitkrittum et al. [2017] and Huggins and Mackey [2018]. For additional results, we refer to Appendix D. We ran all experiments on a PC with Ubuntu 20.04, 124GB RAM, and 32 cores with 2GHz each.

Runtime. We set $m = 4\sqrt{n}$ for N-KSD to match the settings in our other experiments. As per recommendation, we fix the number of test locations $J = 10$ for L1 IMQ, L2 SechExp, 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 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 (5). 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 β for KSD IMQ and N-KSD IMQ as per the recommendation for L1 IMQ in the corresponding experiment by Huggins and Mackey [2018]. Figure 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 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 and Wang [2016], Jitkrittum et al. [2017], 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 σ^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. 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 1(d), using 100 repetitions to obtain the power. KSD with the IMQ and with the Gaussian kernel performs best. Our

[8] All the code replicating our experiments is available at `https://github.com/FlopsKa/nystroem-ksd`.

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.

Acknowledgements

This work was supported by the pilot program Core-Informatics of the Helmholtz Association (HGF). BKS is partially supported by the National Science Foundation (NSF) CAREER award DMS-1945396.

References

Andreas Anastasiou, Alessandro Barp, François-Xavier Briol, Bruno Ebner, Robert E Gaunt, Fatemeh Ghaderinezhad, Jackson Gorham, Arthur Gretton, Christophe Ley, Qiang Liu, et al. Stein's method meets computational statistics: A review of some recent developments. *Statistical Science*, 38(1):120–139, 2023.

Rita Giuliano Antonini. Subgaussian random variables in Hilbert spaces. *Rendiconti del Seminario Matematico della Università di Padova*, 98:89–99, 1997.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

Rémi Bardenet, Arnaud Doucet, and Christopher C. Holmes. Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach. In *International Conference on Machine Learning (ICML)*, pages 405–413, 2014.

Ludwig Baringhaus and Carsten Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88:190–206, 2004.

Ludwig Baringhaus and Norbert 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.

Jerome Baum, Heishiro Kanagawa, and Arthur Gretton. A kernel Stein test of goodness of fit for sequential models. In *International Conference on Machine Learning (ICML)*, pages 1936–1953, 2023.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

Rajendra Bhatia. *Positive Definite Matrices*. Princeton University Press, 2007.

Clément Canonne. Tail bounds for maximum of sub-Gaussian random variables. Mathematics Stack Exchange, 2021. `https://math.stackexchange.com/q/4002713` (version: 2023-12-21).

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

Louis H. Y. Chen. Stein's method of normal approximation: Some recollections and reflections. *The Annals of Statistics*, 49(4):1850–1863, 2021.

Wilson Ye Chen, Lester Mackey, Jackson Gorham, Francois-Xavier Briol, and Chris J. Oates. Stein points. In *International Conference on Machine Learning (ICML)*, pages 844–853, 2018.

Wilson Ye Chen, Alessandro Barp, Francois-Xavier Briol, Jackson Gorham, Mark Girolami, Lester Mackey, and Chris Oates. Stein point Markov chain Monte Carlo. In *International Conference on Machine Learning (ICML)*, pages 1011–1021, 2019.

Kacper Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1972–1980, 2015.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning (ICML)*, pages 2606–2615, 2016.

Jeremie Coullon, Leah F. South, and Christopher Nemeth. Efficient and generalizable tuning strategies for stochastic gradient MCMC. *Statistics and Computing*, 33(3):66, 2023.

Andrea Della Vecchia, Jaouad Mourtada, Ernesto De Vito, and Lorenzo Rosasco. Regularized ERM on random subspaces. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 4006–4014, 2021.

Joseph Diestel and John Jerry Uhl. *Vector Measures*. American Mathematical Society, 1977.

Tamara Fernandez, Nicolas Rivera, Wenkai Xu, and Arthur 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.

Ryoji Fukuda. Exponential integrability of sub-Gaussian vectors. *Probability Theory and Related Fields*, 85(4):505–521, 1990.

Futoshi Futami, Zhenghang Cui, Issei Sato, and Masashi Sugiyama. Bayesian posterior approximation via greedy particle optimization. In *AAAI Conference on Artificial Intelligence*, pages 3606–3613, 2019.

Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. Large sample analysis of the median heuristic. Technical report, 2018. `https://arxiv.org/abs/1707.07269`.

George Giorgobiani, Vakhtang Kvaratskhelia, and Vaja Tarieladze. Notes on sub-Gaussian random elements. In *Applications of Mathematics and Informatics in Natural Sciences and Engineering (AMINSE)*, pages 197–203, 2020.

Jackson Gorham and Lester Mackey. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 226–234, 2015.

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning (ICML)*, pages 1292–1301, 2017.

Jackson Gorham, Anant Raj, and Lester Mackey. Stochastic Stein discrepancies. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 17931–17942, 2020.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Omar Hagrass, Bharath Sriperumbudur, and Krishnakumar Balasubramanian. Minimax optimal goodness-of-fit testing with kernel Stein discrepancy. Technical report, 2024. https://arxiv.org/abs/2404.08278.

Jonathan H. Huggins and Lester Mackey. Random feature Stein discrepancies. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1899–1909, 2018.

Yuri I. Ingster and Irina A. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*. Springer, 2003.

Muhammad Izzatullah, Ricardo Baptista, Lester Mackey, Youssef Marzouk, and Daniel Peter. Bayesian seismic inversion: Measuring Langevin MCMC sample quality with kernels. In *SEG International Exposition and Annual Meeting*, 2020.

Wittawat Jitkrittum, Wenkai Xu, Zoltán Szabó, Kenji Fukumizu, and Arthur Gretton. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 262–271, 2017.

Wittawat Jitkrittum, Heishiro Kanagawa, and Bernhard Schölkopf. Testing goodness of fit of conditional density models with kernels. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 221–230, 2020.

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

Daphne Koller and Nir Friedman. *Probabilistic graphical models*. MIT Press, 2009.

Andrey N. Kolmogorov. Sulla determinazione empirica delle leggi di probabilita. *Giornale dell'Istituto Italiano degli Attuari*, 4(1), 1933.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017.

Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *International Conference on Machine Learning (ICML)*, pages 181–189, 2014.

Erich L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer, 2021.

Jen Ning Lim, Makoto Yamada, Bernhard Schölkopf, and Wittawat Jitkrittum. Kernel Stein tests for multiple model comparison. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2243–2253, 2019.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2378–2386, 2016.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning (ICML)*, pages 276–284, 2016.

Diego Martinez-Taboada and Edward Kennedy. Counterfactual density estimation using kernel Stein discrepancies. In *International Conference on Learning Representations (ICLR)*, 2024.

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

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29:429–443, 1997.

Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using a test for typicality. Technical report, 2019. `http://arxiv.org/abs/1906.02994`.

Chris J. Oates, Mark Girolami, and Nicolas Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B*, 79(3):695–718, 2017.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1177–1184, 2007.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1657–1665, 2015.

Ruslan Salakhutdinov. Learning deep generative models. *Annual Review of Statistics and Its Application*, 2:361–385, 2015.

Mahtab Sarvmaili, Hassan Sajjad, and Ga Wu. Data-centric prediction explanation via kernelized Stein discrepancy. Technical report, 2024. `https://arxiv.org/abs/2403.15576`.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, 41: 2263–2291, 2013.

Nikolai V. Smirnov. Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, 19:279–281, 1948.

Alexander Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory (ALT)*, pages 13–31, 2007.

Bharath K. Sriperumbudur and Nicholas Sterge. Approximate kernel PCA: computational versus statistical trade-off. *The Annals of Statistics*, 50(5):2713–2736, 2022.

Bharath K. Sriperumbudur and Zoltán Szabó. Optimal rates for random Fourier features. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1144–1152, 2015.

Charles 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.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.

Nicholas Sterge and Bharath K. Sriperumbudur. Statistical optimality and computational efficiency of Nyström kernel PCA. *Journal of Machine Learning Research*, 23(337):1–32, 2022.

Gábor Székely and Maria Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5: 1249–1272, 2004.

Gábor Székely and Maria Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93:58–80, 2005.

Michel Talagrand. Regularity of Gaussian processes. *Acta Mathematica*, 159(1-2):99–149, 1987.

Aad van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.

Roman Vershynin. *High-Dimensional Probability*. Cambridge University Press, 2018.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning (ICML)*, pages 681–688, 2011.

Christopher Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 682–688, 2001.

George Wynne, Mikołaj Kasprzak, and Andrew 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. (to appear).

Wenkai Xu and Gesine Reinert. A Stein goodness-of-test for exponential random graph models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 415–423, 2021.

Jiasen Yang, Qiang Liu, Vinayak Rao, and Jennifer Neville. Goodness-of-fit testing for discrete distributions via Stein discrepancy. In *International Conference on Machine Learning (ICML)*, pages 5561–5570, 2018.

Jiasen Yang, Vinayak A. Rao, and Jennifer Neville. A Stein-Papangelou goodness-of-fit test for point processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 226–235, 2019.

Vadim Yurinsky. *Sums and Gaussian vectors*. Springer, 1995.

V. Zolotarev. Probability metrics. *Theory of Probability and its Applications*, 28:278–302, 1983.

A PROOFS

This section is dedicated to the proofs of our results in the main text. Lemma 1 is proved in Section A.1. We prove our main result (Theorem 2) in Section A.2; Corollary 1 is shown in Section A.3. The proof of Theorem 3 is in Section A.4.

A.1 Proof of Lemma 1

By (4), KSD is the norm of the mean embedding of $\mathbb{Q}$ under $h_p(\cdot,\cdot)$, that is,

$$S_p(\mathbb{Q}) = \left\| \int_{\mathbb{R}^d} h_p(\cdot, \mathbf{x}) \, \mathrm{d}\mathbb{Q}(\mathbf{x}) \right\|_{\mathcal{H}_{h_p}} = \left\| \mu_{h_p}(\mathbb{Q}) \right\|_{\mathcal{H}_{h_p}}. \tag{A.1}$$

Hence, with Chatalic et al. [2022, (5)], the optimization problem (6) 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 (A.1), we have

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

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 $\mathbf{K}^\mathsf{T}_{h_p,m,n} = \mathbf{K}_{h_p,n,m}$, and that the Moore-Penrose inverse satisfies $\mathbf{A}^-\mathbf{A}\mathbf{A}^- = \mathbf{A}^-$ for any matrix $\mathbf{A}$.

A.2 Proof of Theorem 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 $\hat{\mu}^{\mathrm{Nys}}_{h_p} := P_{\mathcal{H}_{h_p,m}} \mu_{h_p}\left(\hat{\mathbb{Q}}_n\right)$.

First, we decompose the error as follows.

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

(a) is implied by (A.1) and (7); (b) follows from the reverse triangle inequality; $\pm\hat{\mu}_{h_p}$ and the triangle inequality yield (c); in (d), we use the distributive law 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 0; in (e) $I = C_{\mathbb{Q},\bar{h}_p,\lambda}^{1/2} C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2}$ 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 μ_{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 C.4) 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 1 and Lemma B.3, Y is sub-Gaussian; hence it is also sub-exponential (Lemma C.2(3)), and therefore (Lemma B.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 $B = \left\|Y\right\|_{\psi_1} \overset{(a)}{\lesssim} \left\|Y\right\|_{\psi_2} \overset{(b)}{\lesssim} \sqrt{\mathrm{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}$. (a) follows from Lemma C.2(3) and (b) is implied by Lemma B.3. As $\sigma \asymp B$, we also got that $\sigma \lesssim \sqrt{\mathrm{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}$. Having obtained a bound on the moments, we can apply Bernstein's inequality for separable Hilbert spaces (Yurinsky 1995; recalled in Theorem C.4) 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

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

  Note that (A.3) also holds with the measure $\mathbb{Q}^n \otimes \Lambda^m$, since the event considered in (A.3) has no randomness w.r.t. Λ^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 B.1 and obtain that for any $\delta \in (0,1)$ it holds that

$$(\mathbb{Q}^n \otimes \Lambda^m)\left(\left\|\left(I - P_{\mathcal{H}_{h_p,m}}\right) C_{\mathbb{Q},\bar{h}_p,\lambda}^{1/2}\right\|_{\mathrm{op}} \lesssim \sqrt{\lambda}\right) \geq 1 - \delta/3 \tag{A.4}$$

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

- **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.,

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

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

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

  where we used that $\pm\mathbb{E}_{X\sim\mathbb{Q}}h_p\left(\cdot,X\right) = 0$, the triangle inequality, and the homogeneity of the norm. An application of Theorem C.5 yields that, conditioned on the sample $(\mathbf{x}_i)_{i=1}^n$, it holds that

$$\Lambda^m\left((i_j)_{j=1}^m \,:\, t_3 \leq K\frac{\sqrt{2\log(12/\delta)}}{\sqrt{m}} \;\middle|\; (\mathbf{x}_i)_{i=1}^n\right) \geq 1 - \frac{\delta}{6}. \tag{A.5}$$

  Randomness in $\mathbf{x}_i$-s: Let $Z_i := \left\|C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2}\bar{h}_p\left(\cdot,\mathbf{x}_i\right)\right\|_{\mathcal{H}_{h_p}}$ $(i \in [n])$ with $(\mathbf{x}_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{Q}$. By Assumption 1 and Lemma B.3, the Z_i-s are sub-Gaussian random variables. Hence, by Lemma B.5, with probability at least $1 - \delta/6$, it holds that

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

By Lemma B.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

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

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 Λ^m, we obtain

$$\begin{aligned}
(\mathbb{Q}^n \otimes \Lambda^m)(A) &= \mathbb{E}_{\mathbb{Q}^n}[\Lambda^m(A \mid (\mathbf{x}_i)_{i=1}^n)] = \int_{(\mathbb{R}^d)^n} \Lambda^m(A \mid (\mathbf{x}_i)_{i=1}^n)\,\mathrm{d}\mathbb{Q}^n(\mathbf{x}_1,\ldots,\mathbf{x}_n) \\
&\geq \int_B \Lambda^m(A \mid (\mathbf{x}_i)_{i=1}^n)\,\mathrm{d}\mathbb{Q}^n(\mathbf{x}_1,\ldots,\mathbf{x}_n) \geq \int_B \Lambda^m(C \mid (\mathbf{x}_i)_{i=1}^n)\,\mathrm{d}\mathbb{Q}^n(\mathbf{x}_1,\ldots,\mathbf{x}_n) \\
&\overset{(a)}{\geq} \left(1 - \frac{\delta}{6}\right)\mathbb{Q}^n(B) \overset{(b)}{\geq} (1-\delta/6)^2 = 1 - \delta/3 + \delta^2/6^2 > 1 - \delta/3.
\end{aligned} \tag{A.7}$$

(a) is implied by the uniform lower bound established in (A.5). (b) was shown in (A.6).

Combination of t_1, t_2, and t_3. To conclude, we use decomposition (A.2), and union bound (A.3), (A.4), and (A.7). 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 λ, after rearranging, we get the stated claim.

A.3 Proof of Corollary 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 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 2 yields that

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

  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 2, the first two addends are $\mathcal{O}\left(n^{-1/2}\right)$. For the last addend, again noticing that the trace is constant, we have

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

  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.

A.4 Proof of Theorem 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 C.4) by gaining control on the moments of $\|\eta_i\|_{\mathcal{H}_{h_p}}$ with Lemma B.2.

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

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

We use the definition of η_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 C.2(1); finiteness is due to the imposed assumption.

Hence, $\|\eta_i\|_{\mathcal{H}_{h_p}}$ is sub-exponential (Lemma C.2(3)), and, by Lemma B.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 C.4 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.

B AUXILIARY RESULTS

This section collects our auxiliary results. Lemma B.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 B.2 states that a sub-exponential random variable satisfies Bernstein's conditions, and Lemma B.3 is about the sub-Gaussianity of norms of Hilbert space-valued random variables. In Lemma B.4, we show how tensor products interplay with linearly transformed vectors. Lemma B.5 is about the maximum of real-valued sub-Gaussian random variables; it is a slightly altered restatement of Canonne [2021]. In Lemma B.6 and Lemma B.7, we collect inequalities of positive operators and of norms of covariance operators, respectively.

Lemma B.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

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

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 $f \mapsto \frac{1}{\sqrt{m}}(f(\tilde{\mathbf{x}}_i))_{i=1}^m$. 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 $\mathcal{H}_{h_p,m} = \operatorname{span}\{h_p(\cdot,\tilde{\mathbf{x}}_i) \mid i\in[m]\}$ and notice that $\operatorname{range} P_{\mathcal{H}_{h_p,m}} = \overline{\operatorname{range} Z_m^*}$. We obtain

$$\begin{aligned}
\left\|\left(I - P_{\mathcal{H}_{h_p,m}}\right) C^{1/2}_{\mathbb{Q},\bar{h}_p,\lambda}\right\|^2_{\mathrm{op}} &\overset{(a)}{\leq} \lambda\left\|\left(Z_m^* Z_m + \lambda I\right)^{-1/2} C^{1/2}_{\mathbb{Q},\bar{h}_p,\lambda}\right\|^2_{\mathrm{op}} \overset{(b)}{=} \lambda\left\|C^{-1/2}_{\tilde{\mathbb{Q}}_m,h_p,\lambda} C^{1/2}_{\mathbb{Q},\bar{h}_p,\lambda}\right\|^2_{\mathrm{op}} \qquad \text{(B.8)}\\
&\overset{(c)}{\leq} \lambda\left\|C^{-1/2}_{\tilde{\mathbb{Q}}_m,\tilde{h}_p,\lambda} C^{1/2}_{\mathbb{Q},\bar{h}_p,\lambda}\right\|^2_{\mathrm{op}}
\end{aligned}$$

where (a) follows from Rudi et al. [2015, Proposition 3] with $X := C^{1/2}_{\mathbb{Q},\bar{h}_p,\lambda}$ therein. (b) is by Sterge and Sriperumbudur [2022, Lemma A.7(iv)]. Lemma B.6(5) with $C := C^{-1/2}_{\tilde{\mathbb{Q}}_m,h_p,\lambda}$, $D := C^{-1/2}_{\tilde{\mathbb{Q}}_m,\tilde{h}_p,\lambda}$, and $X := C^{1/2}_{\mathbb{Q},\bar{h}_p,\lambda}$ yields (c), as we obtain $C^*C = C^{-1}_{\tilde{\mathbb{Q}}_m,h_p,\lambda} \preccurlyeq C^{-1}_{\tilde{\mathbb{Q}}_m,\tilde{h}_p,\lambda} = D^*D$; the positive definite relationship holding by the following chain of inequalities

$$\begin{aligned}
C^{-1}_{\tilde{\mathbb{Q}}_m,h_p,\lambda} \preccurlyeq C^{-1}_{\tilde{\mathbb{Q}}_m,\tilde{h}_p,\lambda} &\overset{\text{Lemma } B.6(4)}{\iff} C_{\tilde{\mathbb{Q}}_m,h_p,\lambda} \succcurlyeq C_{\tilde{\mathbb{Q}}_m,\tilde{h}_p,\lambda} \overset{(d)}{\iff} C_{\tilde{\mathbb{Q}}_m,h_p} \succcurlyeq C_{\tilde{\mathbb{Q}}_m,\tilde{h}_p}\\
&\overset{(e)}{\iff} 0 \preccurlyeq \mu_{h_p}\left(\tilde{\mathbb{Q}}_m\right)\otimes\mu_{h_p}\left(\tilde{\mathbb{Q}}_m\right),
\end{aligned}$$

which is true as the r.h.s. is a positive operator. In (d), we subtract λI 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 (B.8) (we specialize $A := C_{\tilde{\mathbb{Q}}_m,\tilde{h}_p}$ and $B := C_{\mathbb{Q},\bar{h}_p}$ therein), we obtain

$$\lambda\left\|C^{-1/2}_{\tilde{\mathbb{Q}}_m,\tilde{h}_p,\lambda} C^{1/2}_{\mathbb{Q},\bar{h}_p,\lambda}\right\|^2_{\mathrm{op}} \leq \frac{\lambda}{1-\beta(\lambda)}, \qquad \text{(B.9)}$$

when $\beta(\lambda) < 1$. The combination of (B.8) and (B.9) 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

$$C_{\tilde{\mathbb{Q}}_m,\tilde{h}_p} = C_{\tilde{\mathbb{Q}}_m,\bar{h}_p} - \left[\tilde{\mu}_{h_p} - \mu_{h_p}\right]\otimes\left[\tilde{\mu}_{h_p} - \mu_{h_p}\right], \qquad \text{(B.10)}$$

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

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

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 (B.10) and (c) holds by the triangle inequality. Lemma B.4 and Lemma C.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 C.3). 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 η_{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 η_{i_j} in (g) proves their sub-Gaussianity by using that $(\dagger)<\infty$ according to Assumption 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 C.3 yields that, conditioned on the Nyström selection,

it holds with probability at least $1-\delta/2$ that

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

provided that $m \geq \max\left\{r\left(A\right),\log(2/\delta)\right\}$, with $r\left(A\right)=\frac{\mathrm{tr}(A)}{\|A\|_{\mathrm{op}}}$. Using Lemma B.6(2), $A \preccurlyeq I$, hence $\|A\|_{\mathrm{op}} \leq 1$. Moreover by Lemma B.7(3), $r(A) \leq \frac{2\,\mathrm{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}{\lambda}$, which implies that, with the same probability,

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

holds when $m \geq \max\left\{\frac{2\,\mathrm{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}{\lambda},\log(2/\delta)\right\}$. Therefore, one can take

$$m \gtrsim \max\left\{\frac{\mathrm{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}{\lambda},\log(2/\delta)\right\}$$

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 η_{i_j}-s are centered, and, by Lemma B.3, it holds for any $j\in[m]$ that

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

that is, the $\left\|\eta_{i_j}\right\|_{\mathcal{H}_{h_p}}$-s are sub-Gaussian. Hence, by Lemma D.2(3), they are sub-exponential, and, by Lemma B.2, they satisfy the Bernstein condition with $\sigma,B \lesssim \sqrt{\mathrm{tr}\left(C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1}C_{\mathbb{Q},\bar{h}_p}\right)}$. Therefore, application of Theorem D.4 yields that, conditioned on the Nyström choice, it holds with probability at least $1-\delta/2$ that

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

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 B.7(1) yields (b). The obtained bound means that choosing $m\gtrsim\max\left\{\frac{\mathrm{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}{\lambda},1\right\}\log(4/\delta)$ 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 $m \gtrsim \max\left\{\frac{\mathrm{tr}\left(C_{\mathbb{Q},\bar{h}_p}\right)}{\lambda}, 1\right\}\log(4/\delta)$, it holds with probability at least $1-\delta$ that $\beta(\lambda) < 1$. We lift the conditioning by integrating over all Nyström selections. □

Lemma B.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 $C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2}\bar{h}_p(\cdot, X)$ enjoy sub-Gaussian properties in terms of their respective $\mathcal{H}_{h_p}$ norms.

Lemma B.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 $\|\langle Y, u\rangle_{\mathcal{H}}\|_{\psi_2} \lesssim \|\langle Y, u\rangle_{\mathcal{H}}\|_{L_2(\mathbb{Q})}$ holds for all $u \in \mathcal{H}$. Then*

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

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

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

that is, both $\left\|\bar{h}_p(\cdot, X)\right\|_{\mathcal{H}_{h_p}}$ and $\left\|C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1/2}\bar{h}_p(\cdot, X)\right\|_{\mathcal{H}_{h_p}}$ 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 C.2(4), Parseval's identity yields (b), and the triangle inequality implies (c). (d) holds by Lemma C.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 B.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.

$A := I$ and $Y := \bar{h}_p(\cdot, X)$ yields $\left\|\left\|\bar{h}_p(\cdot, X)\right\|_{\mathcal{H}_{h_p}}\right\|_{\psi_2} \lesssim \operatorname{tr}\left(\mathbb{E}\left(\bar{h}_p(\cdot, X)\otimes\bar{h}_p(\cdot, X)\right)\right) = \operatorname{tr}(C_{\mathbb{Q},\bar{h}_p}) < \infty$, 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 B.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,

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

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 B.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

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

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 B.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 B.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 $r(\cdot) = \frac{\operatorname{tr}(\cdot)}{\|\cdot\|_{\mathrm{op}}}$ be defined on trace-class operators. Assume that $0 < \lambda \leq \|C_\mathbb{Q}\|_{\mathrm{op}}$. Then, the following hold.*

1. $\frac{1}{2} r(C_\mathbb{Q}) \leq \operatorname{tr}\left(C_{\mathbb{Q},\lambda}^{-1} C_\mathbb{Q}\right) \leq \frac{\operatorname{tr}(C_\mathbb{Q})}{\lambda}$,
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. $r\left(C_{\mathbb{Q},\lambda}^{-1/2} C_\mathbb{Q} C_{\mathbb{Q},\lambda}^{-1/2}\right) \leq \frac{2 \operatorname{tr}(C_\mathbb{Q})}{\lambda}$.

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 $\operatorname{tr}\left(C_{\mathbb{Q},\lambda}^{-1} C_\mathbb{Q}\right) = \sum_{i \in I} \frac{\lambda_i}{\lambda_i + \lambda} \geq \sum_{i \in I} \frac{\lambda_i}{2\|C_\mathbb{Q}\|_{\mathrm{op}}} = \frac{\operatorname{tr}(C_\mathbb{Q})}{2\|C_\mathbb{Q}\|_{\mathrm{op}}}$. The second one is footnote 3.
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 $(\dagger) \Leftrightarrow 2\lambda_1 \geq \lambda_1 + \lambda \Leftrightarrow (\|C_\mathbb{Q}\|_{\mathrm{op}} =)\lambda_1 \geq \lambda$, 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 (1.) after rewriting the term 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.). □

C EXTERNAL STATEMENTS

This section collects the external statements that we use. Lemma C.1 gives equivalent norms for $f \otimes f$. We collect properties of Orlicz norms in Lemma C.2. Theorem C.3 is about the concentration of the empirical covariance, and Theorem C.4 recalls Bernstein's inequality for separable Hilbert spaces. Theorem C.5 is a concentration result for bounded random variables in a separable Hilbert space.

Lemma C.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 C.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 C.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 C.3 (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 $r(C) := \frac{\mathrm{tr}(C)}{\|C\|_{\mathrm{op}}}$.

The following theorem by Yurinsky [1995] is quoted from Sriperumbudur and Sterge [2022].

Theorem C.4 (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 C.5 (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*

$$\left\|\frac{1}{n}\sum_{i=1}^n X_i\right\|_{\mathcal{H}} \leq b\frac{\sqrt{2\log(2/\delta)}}{\sqrt{n}}.$$

D ADDITIONAL EXPERIMENTS

In this section, we collect additional numerical results. Section D.1 discusses the trade-off between power and runtime of the tested approaches. Section D.2 shows the impact of the size of the Nyström sample.

D.1 Runtime vs. Power

Based on the experimental setup in Section 5, we performed an additional set of experiments to contrast runtime and power. We repeated each setup for 100 rounds to obtain the given power and average runtime. The quadratic-time approaches are considered as baseline.

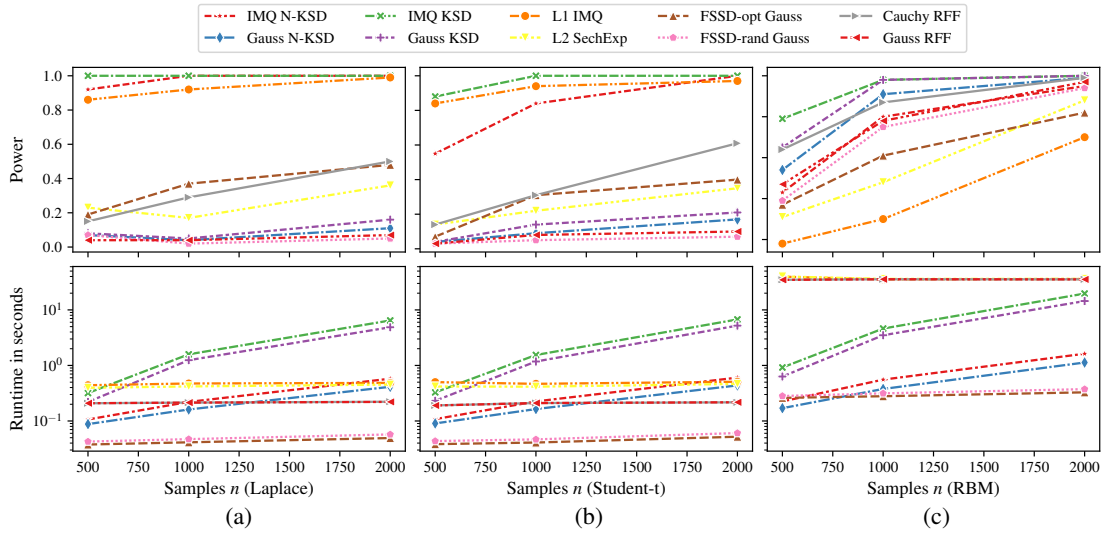

Figure 2: Runtime and power trade-off of the tested approximations.

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.

Figure 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.

Figure 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 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.

We summarize the results in Figure 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.

We conclude that the proposed approach features a good runtime/power trade-off.

D.2 Impact of the Size of the Nyström Sample

Figure 3(a–d) captures the impact of the choice of Nyström samples $m = c\sqrt{n}$ for $c \in \{1, 4, 8\}$; the $\sqrt{n}$ dependence follows from Corollary 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 the article in Section 5. 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.

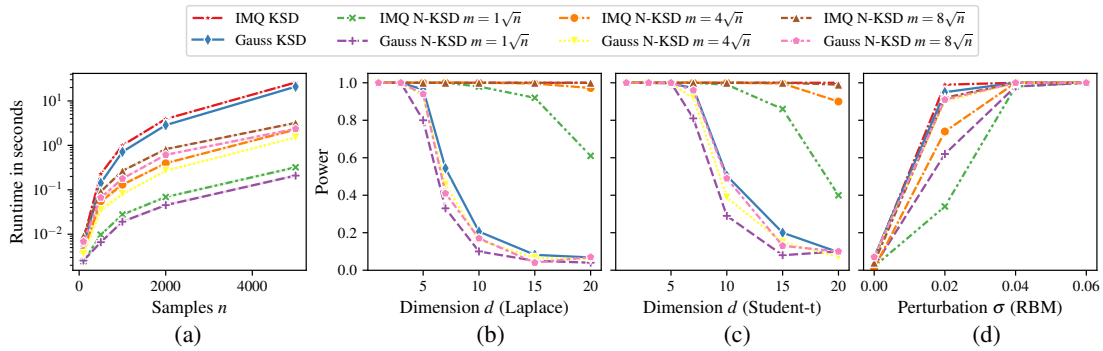

Figure 3: Impact of different choices of factor c for the number of Nyström samples $m = c\sqrt{n}$.