Journal of Machine Learning Research 23 (2022) 1-32 Submitted 7/21; Revised 8/22; Published 10/22

Statistical Optimality and Computational Efficiency of Nyström Kernel PCA

Nicholas Sterge NSTERGE@US.FLOWTRADERS.COM
Department of Statistics
Pennsylvania State University
University Park, PA 16802, USA

Bharath K. Sriperumbudur BKS18@PSU.EDU
Department of Statistics
Pennsylvania State University
University Park, PA 16802, USA

Editor: Sanjiv Kumar

Abstract

Kernel methods provide an elegant framework for developing nonlinear learning algorithms from simple linear methods. Though these methods have superior empirical performance in several real data applications, their usefulness is inhibited by the significant computational burden incurred in large sample situations. Various approximation schemes have been proposed in the literature to alleviate these computational issues, and the approximate kernel machines are shown to retain the empirical performance. However, the theoretical properties of these approximate kernel machines are less well understood. In this work, we theoretically study the trade-off between computational complexity and statistical accuracy in Nyström approximate kernel principal component analysis (KPCA), wherein we show that the Nyström approximate KPCA matches the statistical performance of (non-approximate) KPCA while remaining computationally beneficial. Additionally, we show that Nyström approximate KPCA outperforms the statistical behavior of another popular approximation scheme, the random feature approximation, when applied to KPCA.

Keywords: Principal component analysis, kernel PCA, Nyström approximation, reproducing kernel Hilbert space, covariance operator, U-statistics

1. Introduction

Principal component analysis (PCA) (Jolliffe, 1986) is an unsupervised learning technique in which a random variable X taking values in $\mathbb{R}^d$ is projected onto the direction $\mathbf{a} \in \mathbb{R}^d$ such that $\mathrm{Var}[\mathbf{a}^\top X]$ is maximized. Further, for some $\ell < d$, PCA may be used to find an ℓ-dimensional subspace retaining the maximum possible variance of X, making PCA a popular methodology for dimension reduction and feature extraction. This low-dimensional subspace is the ℓ-eigenspace, i.e., the span of the eigenvectors associated with the top ℓ eigenvalues of the covariance matrix $\mathbb{E}XX^\top - \mathbb{E}X\mathbb{E}X^\top$. The respective eigenvectors are referred to as the principal components of the data, and a lower-dimensional representation of the input data may be computed by projecting onto the principal components.

The principal components outputted by PCA are linearly related to the original coordinates; however, in many cases a non-linear component provides a better description

of the data. Kernel PCA (KPCA) (Schölkopf et al., 1998) is a non-linear extension of PCA which maps the original data into a reproducing kernel Hilbert space (RKHS) (Aronszajn, 1950) where PCA is performed, resulting in principal components which are non-linearly related to the original data. Specifically, for an RKHS $\mathcal{H}$ with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, KPCA solves $\sup\{\mathrm{Var}[f(X)] : \|f\|_{\mathcal{H}} = 1\}$. Analogous to linear PCA, the principal components in KPCA are the eigenfunctions of the covariance operator $\Sigma = \mathbb{E}[\Phi(X) \otimes_{\mathcal{H}} \Phi(X)] - \mathbb{E}\Phi(X) \otimes_{\mathcal{H}} \mathbb{E}\Phi(X)$, where $\Phi(X) = k(\cdot, X)$ is the *feature map* and $\otimes_{\mathcal{H}}$ denotes the tensor product on $\mathcal{H}$, defined as $(f_1 \otimes_{\mathcal{H}} f_2)g = \langle f_2, g\rangle_{\mathcal{H}} f_1$ for any $f_1, f_2, , g \in \mathcal{H}$. Similarly, an ℓ-dimensional representation of X is obtained by projecting onto the ℓ-eigenspace of Σ. Kernel PCA has been employed successfully in tasks such as image denoising (Mika et al., 1999; Jade et al., 2003; Teixeira et al., 2008; Phophalia and Mitra, 2017), image/systems modeling (Kim et al., 2005; Li et al., 2015), novelty/fault detection (Hoffmann, 2007; Samuel and Cao, 2016; de Moura and de Seixas, 2017), feature extraction (Chang and Wu, 2015), and computer vision (Lampert, 2009; Peter et al., 2019).

Empirically, given $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{P}$, the eigenfunctions of Σ are estimated by those of the empirical covariance operator

$$\widehat{\Sigma} = \frac{1}{2n(n-1)} \sum_{i \neq j}^{n} (\Phi(X_i) - \Phi(X_j)) \otimes_{\mathcal{H}} (\Phi(X_i) - \Phi(X_j)), \tag{1}$$

yielding empirical KPCA (EKPCA). Though this may require solving an infinite dimensional system, it can be shown (see Proposition 1) that the eigenfunctions of $\widehat{\Sigma}$ can be obtained by solving an n-dimensional eigenvalue problem, which has a computational requirement of $O(n^3)$ and a memory requirement of $O(n^2)$. This means, EKPCA scales quite poorly with large sample sizes, a behavior shared by many kernel methods. This has lead to much research activity in constructing approximation methods which relieve the computational burden. Nyström method (Reinhardt, 1985; Williams, 1998) is a popular approximation scheme, which uses a subsample of the original data to construct a low-rank approximation to the Gram matrix $\mathbf{K} = [k(X_i, X_j)]_{i,j}$, which in turn is closely related to $\widehat{\Sigma}$. More precisely, the Gram matrix $\mathbf{K}$ is approximated by $\tilde{\mathbf{K}} = \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{nm}^{\top}$, where $\mathbf{K}_{nm}$ is the matrix formed by randomly selecting m columns of $\mathbf{K}$ and $\mathbf{K}_{mm}$ is the intersection of those m columns and m rows of $\mathbf{K}$. The approximate Gram matrix can be used in subsequent learning tasks, resulting in computational saving for $m < n$. In the case of KPCA, the computational complexity is reduced from $O(n^3)$ to $O(nm^2)$. Of course, the question of interest is whether this computational saving comes at the cost of statistical accuracy. In this work we establish the consistency of Nyström KPCA (NY-EKPCA), and study the relationship between statistical behavior and computational complexity.

1.1 Contributions

The contributions of the paper are as follows:

(i) In Section 3.2, we propose Nyström empirical KPCA (NY-EKPCA), and demonstrate its computational complexity. In Section 4, we compare the performance of empirical KPCA (EKPCA) with that of NY-EKPCA in terms of the reconstruction error of the ℓ-eigenspaces in $\mathcal{H}$. We show that NY-EKPCA matches the statistical performance of EKPCA with

less computational complexity, provided the number of subsamples, m, is large enough, and the number of eigenfunctions used in the reconstruction, i.e., ℓ, is not too large. We note that a similar analysis has been performed working with the uncentered covariance operator (Sterge et al., 2020), that is, $\widehat{\Sigma}$ is defined as a V-statistic estimator where the mean element is assumed to be zero ($\mathbb{E}_{X\sim\mathbb{P}}k(\cdot,X)=0$), i.e., $\widehat{\Sigma}=\frac{1}{n}\sum_{i=1}^{n}k(\cdot,X_i)\otimes_{\mathcal{H}}k(\cdot,X_i)$. This assumption of $\mathbb{E}_{X\sim\mathbb{P}}k(\cdot,X)=0$ is highly restrictive, as it is not satisfied by virtually all common kernels, e.g., Gaussian, Matérn, inverse multiquadric, that induce an infinite dimensional RKHS. However, if this assumption is relaxed, the resulting V-statistic estimator, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n}k(\cdot,X_i)\otimes_{\mathcal{H}}k(\cdot,X_i)-\left(\frac{1}{n}\sum_{i=1}^{n}k(\cdot,X_i)\right)\otimes_{\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^{n}k(\cdot,X_i)\right)$$

is no longer unbiased. Since unbiasedness is crucial for a tighter analysis, we consider a U-statistic estimator of Σ as shown in (1) and develop the analysis based on a Bernstein-type inequality for operator valued U-statistics, which we proposed in our earlier work (Sriperumbudur and Sterge, 2022). Thus, the current work provides a non-trivial extension to our previous results (note that the previous results in Sterge et al., 2020 deal with sums of independent $\mathcal{H}$-valued random variables unlike U-statistics that appear in this work) by relaxing a significant assumption of $\mathbb{E}_{X\sim\mathbb{P}}k(\cdot,X)=0$.

(ii) In Section 4.2, to foster a comparison with random Fourier features (RFF) approximation (Rahimi and Recht, 2008a; Sriperumbudur and Sterge, 2022), we study the performance of NY-EKPCA in terms of the reconstruction error of the ℓ-eigenspace in $L^2(\mathbb{P})$. Comparing NY-EKPCA with EKPCA and random feature approximate EKPCA (RF-EKPCA), we show that NY-EKPCA again recovers the statistical performance of EKPCA with less computational complexity. However, unlike in $\mathcal{H}$ where the number of eigenfunctions, ℓ used in the reconstruction cannot be too large, the result in $L^2(\mathbb{P})$ holds regardless of the number of eigenfunctions, ℓ, used in the reconstruction. Additionally, we show the Nyström approximation to be superior to that of the random features approximation by showing that NY-EKPCA outperforms RF-EKPCA in terms of the reconstruction error while enjoying better computational complexity—similar observation was already made in the context of kernel ridge regression (Rudi et al., 2015; Rudi and Rosasco, 2017).

1.2 Related Work

The statistical behavior of EKPCA has been well studied. The statistical consistency of EKPCA is established by Shawe-Taylor et al. (2005), where the reconstruction error of the empirical ℓ-eigenspace is shown to converge at the rate $\sqrt{\ell/n}$. Blanchard et al. (2007) and Rudi et al. (2013) obtain improved rates for uncentered KPCA by considering the decay rate of the eigenvalues of the covariance operator. Additionally, Blanchard et al. (2007) shows $\sqrt{\ell/n}$-rate for the excess reconstruction error of EKPCA with empirical recentering. In Corollary 5, we show that EKPCA with empirical recentering based on U-statistics has the same statistical convergence behavior to that of uncentered EKPCA as considered in Rudi et al. (2013). On the other hand, the setting of our paper is different from that of Blanchard et al. (2007), wherein ours deals with growing ℓ while the latter deals with fixed ℓ as $n\to\infty$.

Outside of the Nyström method, several other approximation strategies have been proposed in the kernel methods literature. These include incomplete Cholesky (Fine and Scheinberg, 2001; Bach and Jordan, 2005), random features (Rahimi and Recht, 2008a), sketching (Yang et al., 2017), and sparse greedy approximation (Smola and Schölkopf, 2000). These methods, including Nyström, offer significant reductions in computational complexity, and, empirically, have been shown to provide performance competitive to their more expensive counterparts without approximation (Rahimi and Recht, 2008b; Kumar et al., 2009; Yang et al., 2012). The theoretical analysis of the Nyström method has primarily concerned the distance between the Gram matrix, $\mathbf{K}$, and its low-rank approximation (Drineas and Mahoney, 2005; Gittens and Mahoney, 2013; Jin et al., 2013). Recent research has focused on the impact of Nyström approximation in specific learning tasks, allowing one to observe the trade-off between statistical accuracy and computational complexity. The supervised setting has been studied heavily (Bach, 2013; Alaoui and Mahoney, 2015; Rudi et al., 2015), where it has been shown that Nyström approximation can achieve best possible statistical performance with better computational complexity. Significantly less is known in the unsupervised setting; however, approximate KPCA has been studied (Lopez-Paz et al., 2014; Ullah et al., 2018; Sriperumbudur and Sterge, 2022) using the random feature approximation (Rahimi and Recht, 2008a).

Lopez-Paz et al. (2014) approximated the Gram matrix by a random feature based Gram matrix and showed the operator norm of their difference to converge to zero at the rate $n\sqrt{(\log n)/m}$ as $m, n \to \infty$, with m being the number of random features. This rate is too loose as m has to grow faster than n to achieve convergence to zero, which defeats the purpose of approximation. Ullah et al. (2018) compared the ℓ-eigenspaces (with ℓ fixed) of EKPCA and its random feature version (call it RF-EKPCA) by comparing certain inner product of the uncentered covariance operator with the difference between the projection operators associated with ℓ-eigenspaces of KPCA and EKPCA (*resp.* RF-EKPCA), after embedding them all as Hilbert-Schmidt operators on $L^2(\mathbb{P})$. Through upper bounds on these differences of inner products, they argued that $m = \sqrt{n}$ random features are sufficient for RF-EKPCA to have similar statistical behavior to that of EKPCA, thereby guaranteeing better computational complexity at no statistical loss. However, the work lacks on two fronts: (i) The comparison is made using only upper bounds on the performance criterion (i.e., difference of inner products) and no matching lower bounds are provided to establish their sharpness, which means the sufficiency of $\sqrt{n}$ random features is inconclusive, and (ii) the criterion used for comparison has no clear interpretation. Using the reconstruction error as a criterion, recently, Sriperumbudur and Sterge (2022) developed a comprehensive analysis and showed RF-KPCA to have better computational complexity than EKPCA with no loss in statistical performance. Similar type of results are shown by Sterge et al. (2020) for Nyström approximate KPCA, that it can yield computational benefit without statistical loss; however, as mentioned in Section 1.1, this result hinges on the highly-restrictive assumption that $\mathbb{E}_{X\sim\mathbb{P}}k(\cdot, X) = 0$. As is the case in the present paper, Hallgren (2021) also generalized NY-EKPCA of (Sterge et al., 2020) by relaxing the assumption of $\mathbb{E}_{X\sim\mathbb{P}}k(\cdot, X) = 0$, and provided excess-error bounds in the style of Blanchard et al. (2007) in the fixed ℓ setting with a convergence rate of $n^{-1/2}$. However, (Hallgren, 2021) does not provide any details about the requirement on the number of Nyström subsamples to achieve this convergence rate and no discussion is provided on the optimality of this rate as well.

2. Definitions and Notation

For $\mathbf{a} := (a_1, \ldots, a_d) \in \mathbb{R}^d$ and $\mathbf{b} := (b_1, \ldots, b_d) \in \mathbb{R}^d$, define $\|\mathbf{a}\|_2 := \sqrt{\sum_{i=1}^d a_i^2}$ and $\langle \mathbf{a}, \mathbf{b} \rangle_2 := \sum_{i=1}^d a_i b_i$. $\mathbf{a} \otimes_2 \mathbf{b} := \mathbf{a}\mathbf{b}^\top$ denotes the tensor product of $\mathbf{a}$ and $\mathbf{b}$. $\mathbf{I}_n$ denotes an $n \times n$ identity matrix and $\mathbf{1}_n = (1, \overset{n}{\ldots}, 1)^\top$. We define $\mathbf{C}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ and $\mathbf{H}_n = n\mathbf{C}_n$. For a matrix $A \in \mathbb{R}^{n \times m}$, $A^+ \in \mathbb{R}^{m \times n}$ denotes the *Moore-Penrose generalized inverse* of A. $a \wedge b := \min(a, b)$ and $a \vee b := \max(a, b)$. Define $[n] := \{1, \ldots, n\}$ for $n \in \mathbb{N}$. For sequences a_n and b_n indexed by $n \in \mathbb{N}$, $a_n \lesssim b_n$ (*resp.* $a_n \gtrsim b_n$) denotes that there exists a positive constant c (*resp.* c') independent of n such that $a_n \leq c b_n$ (*resp.* $a_n \geq c' b_n$). For a random variable A with law P and a constant b, $A \lesssim_p b$ denotes that for any $\delta > 0$, there exists a positive constant $c_\delta < \infty$ such that $P(A \leq c_\delta b) \geq \delta$.

For $x, y \in H$, a Hilbert space, $x \otimes_H y$ is an element of the tensor product space $H \otimes H$, and can also be seen as an operator from H to H since $(x \otimes_H y)z = x\langle y, z \rangle_H$ for any $z \in H$. $\alpha \in \mathbb{R}$ is called an *eigenvalue* of a bounded self-adjoint operator S if there exists an $x \neq 0$ such that $Sx = \alpha x$ and such an x is called the *eigenvector/eigenfunction* of S and α. An eigenvalue is said to be *simple* if it has multiplicity one. For an operator $S : H \to H$, $\|S\|_{\mathcal{L}^1(H)}$, $\|S\|_{\mathcal{L}^2(H)}$ and $\|S\|_{\mathcal{L}^\infty(H)}$ denote the trace, Hilbert-Schmidt and operator norms of S, respectively. $\mathcal{R}(S)$ and $\mathcal{N}(S)$ denote the range space and null space of an operator S. $L^r(\mathcal{X}, \mu)$ denotes the Banach space of r-power ($r \geq 1$) μ-integrable functions. For $f \in L^r(\mathcal{X}, \mu)$, $\|f\|_{L^r(\mu)} := \left(\int_{\mathcal{X}} |f|^r \, d\mu\right)^{1/r}$ denotes the L^r-norm of f for $1 \leq r < \infty$.

3. Kernel PCA and its Variations

We first start with an assumption that will be used throughout the paper.

Assumption 1 *$(\mathcal{X}, \mathcal{B})$ is a completely separable space endowed with Borel σ-algebra $\mathcal{B}$. $\mathcal{H}$ is an RKHS of $\mathbb{R}$-valued functions on $\mathcal{X}$ with bounded continuous strictly positive definite kernel satisfying $\sup_{x \in \mathcal{X}} k(x, x) =: \kappa < \infty$.*

The assumption that $\mathcal{X}$ is completely separable, also called second countable, ensures that $L^r(\mathcal{X}, \mu)$ is separable for any σ-finite measure μ on $\mathcal{B}$ and $r \in [1, \infty)$ (Cohn, 2013, Proposition 3.4.5). Also, the second-countability of $\mathcal{X}$ along with the continuity of k guarantees that $\mathcal{H}$ is separable (Steinwart and Christmann, 2008, Lemma 4.33), which in turn combined with the bounded of k ensures $k(\cdot, x) : \mathcal{X} \to \mathcal{H}$ is Bochner-measurable for all $x \in \mathcal{X}$ (Dinculeanu, 2000, Theorem 8 on p.5).

3.1 Kernel PCA in the population and sample

Kernel PCA (Schölkopf et al., 1998) is an unsupervised learning method in which classical PCA is performed on data which has been mapped to a reproducing kernel Hilbert space. That is, kernel PCA (KPCA) finds $f \in \mathcal{H}$ with unit norm such that $\mathrm{Var}[f(X)] = \mathbb{E}[f(X) - \mathbb{E}[f(X)]]^2$ is maximized. Using the reproducing property, we have $\mathrm{Var}[f(X)] = \mathbb{E}[\langle f, k(\cdot, X) \rangle_{\mathcal{H}} - \langle f, m_{\mathbb{P}} \rangle_{\mathcal{H}}]^2$ with $m_{\mathbb{P}} \in \mathcal{H}$ being the unique *mean element* of $\mathbb{P}$ in $\mathcal{H}$, defined for all $f \in \mathcal{H}$ by

$$\langle f, m_{\mathbb{P}} \rangle_{\mathcal{H}} = \mathbb{E}[f(X)] = \mathbb{E}[\langle f, k(\cdot, X) \rangle_{\mathcal{H}}] = \left\langle f, \int_{\mathcal{X}} k(\cdot, x) d\mathbb{P}(x) \right\rangle_{\mathcal{H}}, \tag{2}$$

where the last equality of (2) holds via Riesz representation theorem (Reed and Simon, 1980) and the boundedness of k, which jointly ensure Bochner $\mathbb{P}$-integrability (Diestel and Uhl, 1977) of $k(\cdot, X)$. Therefore, we may write $\mathrm{Var}[f(X)] = \langle f, \Sigma f \rangle_{\mathcal{H}}$, where

$$\Sigma = \int_{\mathcal{X}} k(\cdot, x) \otimes_{\mathcal{H}} k(\cdot, x) d\mathbb{P}(x) - m_{\mathbb{P}} \otimes_{\mathcal{H}} m_{\mathbb{P}} \tag{3}$$

is the covariance operator on $\mathcal{H}$ associated with $\mathbb{P}$. Thus, the KPCA problem may be expressed as

$$\sup\{\langle f, \Sigma f \rangle_{\mathcal{H}} : f \in \mathcal{H},\ \|f\|_{\mathcal{H}} = 1\}, \tag{4}$$

bearing a strong resemblance to classical PCA. In fact, KPCA can be seen as a generalization of classical linear PCA, as taking $\mathcal{H} = \mathbb{R}^d$ with $k(x, y) = x^\top y$ yields classical PCA with covariance matrix $\Sigma = \mathbb{E}[XX^\top] - \mathbb{E}[X]\mathbb{E}[X]^\top$. The boundedness of k in Assumption 1 ensures that Σ is trace class and thus compact. Since Σ is also positive and self-adjoint, the spectral theorem (Reed and Simon, 1980) gives

$$\Sigma = \sum_{i \in I} \lambda_i \phi_i \otimes_{\mathcal{H}} \phi_i,$$

where $(\lambda_i)_{i \in I} \subset \mathbb{R}^+$ and $(\phi_i)_{i \in I}$ are the eigenvalues and eigenfunctions, respectively, of Σ. $(\phi_i)_{i \in I}$ form an orthonormal system spanning $\overline{\mathcal{R}(\Sigma)}$, where the index set I is either finite or countable, in which case $\lambda_i \to 0$ as $i \to \infty$. The solution to (4) is simply the eigenfunction of Σ corresponding to its largest eigenvalue. We make the following simplifying assumption for ease of presentation.

Assumption 2 *The eigenvalues $(\lambda_i)_{i \in I}$ of Σ are simple, positive, and w.l.o.g. satisfy a decreasing rearrangement, i.e., $\lambda_1 > \lambda_2 > \ldots$*

Based on Assumption 2, the orthogonal projection operator onto the ℓ-eigenspace of Σ, i.e., $\mathrm{span}\{(\phi_i)_{i=1}^{\ell}\}$ can be written as

$$P^{\ell}(\Sigma) = \sum_{i=1}^{\ell} \phi_i \otimes_{\mathcal{H}} \phi_i. \tag{5}$$

The discussion so far corresponds to the population version of KPCA, i.e., when the data-generating distribution $\mathbb{P}$ is known. In practice, the knowledge of $\mathbb{P}$ is available only through a sample $\{X_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$. Therefore, performing KPCA in practice requires one to replace Σ in (3) with an estimate. In the vast majority of the literature (e.g., Schölkopf et al. 1998; Shawe-Taylor et al. 2005), the assumption $m_{\mathbb{P}} = 0$ is made and the corresponding V-statistic estimator of Σ, i.e., $\frac{1}{n}\sum_{i=1}^n k(\cdot, X_i) \otimes_{\mathcal{H}} k(\cdot, X_i)$ is used. However, the assumption $m_{\mathbb{P}} = 0$ is quite restrictive as many popular kernels, such as the Gaussian, do not satisfy this condition. Therefore, to make the setting and results more general, we make no such assumption on $m_{\mathbb{P}}$; however, this relaxation causes the resultant V-statistic estimator to be biased. To mediate the technical difficulties arising from a biased estimator, we consider the following U-statistic estimator,

$$\widehat{\Sigma} = \frac{1}{2n(n-1)} \sum_{i \neq j}^{n} (k(\cdot, X_i) - k(\cdot, X_j)) \otimes_{\mathcal{H}} (k(\cdot, X_i) - k(\cdot, X_j)),$$

conceived from the following alternate representation of Σ:

$$\Sigma = \frac{1}{2}\int_{\mathcal{X}\times\mathcal{X}} \left(k(\cdot,x)-k(\cdot,y)\right)\otimes_{\mathcal{H}}\left(k(\cdot,x)-k(\cdot,y)\right)\, d\mathbb{P}(x)\, d\mathbb{P}(y).$$

Using the reproducing property, it is easy to verify that

$$\widehat{\mathrm{Var}}[f(X)] := \frac{1}{2n(n-1)}\sum_{i\neq j}^{n}\left(f(X_i)-f(X_j)\right)^2 = \left\langle \widehat{\Sigma} f, f\right\rangle_{\mathcal{H}}.$$

Therefore, substituting for Σ in (4) yields the objective of empirical KPCA (EKPCA):

$$\sup\left\{\langle f, \widehat{\Sigma} f\rangle_{\mathcal{H}} : f\in\mathcal{H},\, \|f\|_{\mathcal{H}}=1\right\}. \tag{6}$$

Of course $\widehat{\Sigma}$ is self-adjoint and positive. Also, the assumption of strict positive definiteness of k (see Assumption 1) yields that $\mathrm{rank}(\widehat{\Sigma}) = n-1$ $\mathbb{P}$-a.s. (therefore compact). Thus the spectral theorem (Reed and Simon, 1980) yields

$$\widehat{\Sigma} = \sum_{i=1}^{n-1}\widehat{\lambda}_i\widehat{\phi}_i\otimes_{\mathcal{H}}\widehat{\phi}_i, \tag{7}$$

where $(\widehat{\lambda}_i)_{i=1}^{n-1}\subset\mathbb{R}^+$ and $(\widehat{\phi}_i)_{i=1}^{n-1}\subset\mathcal{H}$ are the eigenvalues and eigenfunctions of $\widehat{\Sigma}$. We make the following simplifying assumption regarding the spectrum of $\widehat{\Sigma}$.

Assumption 3 *The eigenvalues of $(\widehat{\lambda}_i)_{i=1}^{n-1}$ of $\widehat{\Sigma}$ are simple $\mathbb{P}$-a.s., positive and w.l.o.g. satisfy a decreasing rearrangement, i.e., $\widehat{\lambda}_1 > \widehat{\lambda}_2 > \ldots$ $\mathbb{P}$-a.s.*

We would like to mention that the simplicity of the eigenvalues of $\widehat{\Sigma}$ is not really required for the results of this paper to hold. However, this assumption simplifies the notation and proofs, and therefore for the sake of simplicity and clarity, we resort to the above assumption. For any $\ell \leq n-1$, since $(\widehat{\phi}_i)_{i=1}^{\ell}$ forms an orthonormal coordinate system in $\mathcal{H}$, it yields the following low-dimensional Euclidean representation of $k(\cdot,x)$:

$$\left(\left\langle k(\cdot,x),\widehat{\phi}_1\right\rangle_{\mathcal{H}},\cdots,\left\langle k(\cdot,x),\widehat{\phi}_\ell\right\rangle_{\mathcal{H}}\right)^\top = \left(\widehat{\phi}_1(x),\cdots,\widehat{\phi}_\ell(x)\right)^\top,$$

for any $x\in\mathcal{X}$. Moreover, following Assumption 3, the orthogonal projector onto $\mathrm{span}\{\widehat{\phi}_i : i=1,\ldots,\ell\}$ is given by

$$P^{\ell}(\widehat{\Sigma}) = \sum_{i=1}^{\ell}\widehat{\phi}_i\otimes_{\mathcal{H}}\widehat{\phi}_i. \tag{8}$$

Though $\widehat{\Sigma}$ is finite rank, its eigenfunctions are solution to a possibly infinite dimensional linear system. The following result, quoted from Sriperumbudur and Sterge (2022), shows that the eigenvalues of $\widehat{\Sigma}$ can computed by solving an n-dimensional system.

Proposition 1 (Sriperumbudur and Sterge 2022, Proposition 1) *Let $(\widehat{\lambda}_i, \widehat{\phi}_i)_i$ be the eigensystem of $\widehat{\Sigma}$. Define $\mathbf{K} = [k(X_i, X_j)]_{i,j\in[n]}$. Then*

$$\widehat{\phi}_i = \frac{1}{\widehat{\lambda}_i} \sum_{j=1}^{n} \gamma_{i,j} k(\cdot, X_j),$$

where $\boldsymbol{\gamma}_i = (\gamma_{i,1}, \cdots, \gamma_{i,n}) = \frac{1}{n(n-1)} \mathbf{H}_n \widehat{\boldsymbol{\alpha}}_i$ with $\widehat{\boldsymbol{\alpha}}_i \notin \mathcal{N}(\mathbf{H}_n)$, and $(\widehat{\lambda}_i, \widehat{\boldsymbol{\alpha}}_i)_i$ are the eigenvalues and eigenvectors of $\frac{1}{n(n-1)} \mathbf{K}\mathbf{H}_n$.

Since the eigensystem of $\widehat{\Sigma}$ can be obtained by solving an $n \times n$ system, finding $(\widehat{\lambda}_i, \widehat{\phi}_i)_{i=1}^{\ell}$ for $\ell < n$ has a space complexity of $O(n^2)$ and a time complexity of $O(n^2\ell)$–e.g., by partial SVD methods such as Krylov subspace method (see Halko et al. 2011, Sections 3.3.2 & 3.3.3).

3.2 Approximate kernel PCA using the Nyström method

For large sample sizes, performing EKPCA amounts to a significant computational burden, motivating many approximation schemes. We explore the popular Nyström approximation (Reinhardt, 1985; Williams and Seeger, 2001; Drineas and Mahoney, 2005) to speed up EKPCA and study the trade-offs between computational gains and statistical accuracy. The general idea in Nyström method is to obtain a low-rank approximation to the Gram matrix $\mathbf{K}$, and replace $\mathbf{K}$ by this approximation in kernel algorithms, resulting in computational speedup. Since the eigenspace of $\mathbf{K}\mathbf{H}_n$ is related to that of $\widehat{\Sigma}$ (as noted in Proposition 1), Nyström method also yields a low rank approximation to $\widehat{\Sigma}$, which is what we exploit in developing Nyström approximate KPCA. It follows from Proposition 1 that the eigenfunctions of $\widehat{\Sigma}$ lie in

$$\bar{\mathcal{H}}_n := \left\{ f \in \mathcal{H} \middle| f = \sum_{i=1}^{n} \left(n\alpha_i - \sum_{j=1}^{n} \alpha_j \right) k(\cdot, X_i) : \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R} \right\}.$$

Thus, we could instead express the objective in (6) as an optimization over $\bar{\mathcal{H}}_n$, or equivalently, over $\boldsymbol{\alpha} \in \mathbb{R}^n$. Define a random subspace

$$\bar{\mathcal{H}}_m := \left\{ f \in \mathcal{H} \middle| f = \sum_{j=1}^{m} \left(m\beta_j - \sum_{l=1}^{m} \beta_l \right) k(\cdot, X_{r_j}) : \boldsymbol{\beta} = (\beta_1, \ldots, \beta_m) \subset \mathbb{R} \right\}$$

of $\mathcal{H}$, where $m < n$ indices $\{r_1, \ldots, r_m\}$ are sampled uniformly without replacement from $[n]$, yielding the subsample $\{X_{r_j}\}_{j=1}^{m}$. Nyström KPCA (NY-EKPCA) optimizes the EKPCA objective in (6) over $\bar{\mathcal{H}}_m$, that is, NY-EKPCA is the solution to the following problem:

$$\sup \left\{ \left\langle f, \widehat{\Sigma} f \right\rangle_{\mathcal{H}} : f \in \bar{\mathcal{H}}_m, \, \|f\|_{\mathcal{H}} = 1 \right\}. \tag{9}$$

The following result, whose proof is presented in Section 6.1, shows that the solution to (9) is obtained by solving a finite dimensional linear system, which has better computational

complexity than that of EKPCA, provided $m < n$. To this end, we first introduce some notation before stating the result,

$$\mathbf{K}_{mm} = [k(X_{r_j}, X_{r_l})]_{j,l\in[m]} \in \mathbb{R}^{m\times m},\ \mathbf{K}_{nm} = [k(X_i, X_{r_j})]_{i\in[n],j\in[m]} \in \mathbb{R}^{n\times m},$$

and $\mathbf{K}_{mn} = \mathbf{K}_{nm}^\top$.

Proposition 2 *Define the $m \times m$ matrix $\mathbf{M} = \mathbf{K}_{mm}^{-1/2}\mathbf{K}_{mn}\mathbf{H}_n\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1/2}$. The solution to (9) is given by*

$$\tilde{\phi}_1 = \tilde{S}_m^* \mathbf{K}_{mm}^{-1/2}\mathbf{u}_1,$$

where $\mathbf{u}_1 = (u_{1,1}, \ldots, u_{1,m})$ is the unit eigenvector of $\frac{1}{n(n-1)}\mathbf{M}$ corresponding to its largest eigenvalue, denoted $\tilde{\lambda}_1$ and $\tilde{S}_m^* : \mathbb{R}^m \to \mathcal{H}$, $\boldsymbol{\beta} \mapsto \sum_{j=1}^m \beta_j k(\cdot, X_{r_j})$.

The complexity of computing $\mathbf{M}$ and its eigendecomposition is $O(m^2\ell + nm^2 + m^3)$—e.g., by partial SVD methods. Therefore, for $m < \sqrt{n\ell}$ (i.e., $m = o(\sqrt{n\ell})$ as $n, \ell \to \infty$), the complexity of solving (9) scales as $O(nm^2)$, which is a reduction from the $O(n^2\ell)$ complexity of solving EKPCA. It is worth noting the connection between our Nyström approximation to KPCA and the traditional Nyström approximation of Williams and Seeger (2001) to the Gram matrix, given by

$$\tilde{\mathbf{K}} = \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}.$$

Observing

$$\begin{aligned} &\mathbf{M}\mathbf{u} = \mathbf{K}_{mm}^{-1/2}\mathbf{K}_{mn}\mathbf{H}_n\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1/2}\mathbf{u} = n(n-1)\tilde{\lambda}\mathbf{u} \\ \implies &\tilde{\mathbf{K}}\mathbf{H}_n\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1/2}\mathbf{u} = n(n-1)\tilde{\lambda}\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1/2}\mathbf{u}, \end{aligned}$$

it is clear that $\tilde{\mathbf{K}}\mathbf{H}_n$ will have the same eigenvalues as $\mathbf{M}$. All eigenvalues of $\tilde{\mathbf{K}}\mathbf{H}_n$ and $\mathbf{M}$ will be positive as

$$\begin{aligned} \mathbf{u}^\top\mathbf{M}\mathbf{u} &= \left\langle \mathbf{K}_{mm}^{-1/2}\mathbf{K}_{mn}\mathbf{H}_n\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1/2}\mathbf{u}, \mathbf{u}\right\rangle_2 \\ &= n(n-1)\left\langle \widehat{\Sigma}\tilde{S}_m^*\mathbf{K}_{mm}^{-1/2}\mathbf{u}, \tilde{S}_m^*\mathbf{K}_{mm}^{-1/2}\mathbf{u}\right\rangle_{\mathcal{H}} \geq 0. \end{aligned}$$

We will make the following simplifying assumption on $\tilde{\mathbf{K}}\mathbf{H}_n$ and its eigenvalues:

Assumption 4 *The eigenvalues $(\tilde{\lambda}_i)_{i=1}^{m-1}$ of $\frac{\tilde{\mathbf{K}}\mathbf{H}_n}{n(n-1)}$ are simple $\mathbb{P}$-a.s., positive, and w.l.o.g. satisfy a decreasing rearrangement, i.e., $\tilde{\lambda}_1 > \tilde{\lambda}_2 > \ldots$ $\mathbb{P}$-a.s.*

As shown in the proof of Proposition 2 (see Section 6.1), $(\tilde{\phi}_i)_i$ form an orthonormal system. Thus, for some $\ell < m$, the orthogonal projector onto $\operatorname{span}\{(\tilde{\phi}_i)_{i=1}^\ell\}$ is given by

$$P_{nys}^\ell(\widehat{\Sigma}) = \sum_{i=1}^\ell \tilde{\phi}_i \otimes_{\mathcal{H}} \tilde{\phi}_i. \tag{10}$$

One may ask if $\tilde{\phi}_1$ is the eigenfunction of some operator. Denoting $\bar{P}_m$ as the orthogonal projector onto $\bar{\mathcal{H}}_m$, since $\tilde{S}_m : \mathcal{H} \to \mathbb{R}^m$, $f \mapsto (f(X_{r_1}), f(X_{r_2}), \ldots, f(X_{r_m}))^\top$, it can be shown (see Section 6.2) that

$$\bar{P}_m = \tilde{S}_m^*\mathbf{H}_m(\mathbf{H}_m\mathbf{K}_{mm}\mathbf{H}_m)^+\mathbf{H}_m\tilde{S}_m, \tag{11}$$

and that $(\tilde{\phi}_i)_i$ are the orthonormal eigenfunctions of $\bar{P}_m \widehat{\Sigma} \bar{P}_m$ with corresponding eigenvalues $(\tilde{\lambda}_i)_i$, that is,

$$\bar{P}_m \widehat{\Sigma} \bar{P}_m \tilde{\phi}_i = \tilde{\lambda}_i \tilde{\phi}_i. \tag{12}$$

Therefore, we may think of $\bar{P}_m \widehat{\Sigma} \bar{P}_m$ as a low-rank approximation to $\widehat{\Sigma}$.

4. Computational vs. Statistical Trade-off

In this section, we explore whether the computational saving achieved by NY-EKPCA is obtained at the expense of statistical performance. We measure the statistical performance of KPCA, EKPCA, and NY-EKPCA in terms of the *reconstruction error*, which is detailed below.

4.1 Reconstruction error in $\mathcal{H}$-norm

In linear PCA, the reconstruction error defined as

$$\mathbb{E}_{X\sim\mathbb{P}} \left\| (X-\mu) - \sum_{i=1}^{\ell} \langle X-\mu, \phi_i \rangle_2 \, \phi_i \right\|_2^2 = \mathbb{E}_{X\sim\mathbb{P}} \left\| (X-\mu) - P^{\ell}(\Sigma)(X-\mu) \right\|_2^2,$$

is the error involved in reconstructing a centered random variable X by projecting it onto the ℓ-eigenspace (i.e., span of the top-ℓ eigenvectors) associated with its covariance matrix, $\Sigma = \mathbb{E}[XX^\top] - \mathbb{E}[X]\mathbb{E}[X]^\top$ through the orthogonal projection operator $P^{\ell}(\Sigma) := \sum_{i=1}^{\ell} \phi_i \otimes_2 \phi_i$. Clearly, the error is zero when $\ell = d$. The analogs of the reconstruction error in KPCA, EKPCA and NY-EKPCA, can be similarly stated in terms of their projection operators, (5), (8), and (10) as follows:

$$R^{\ell}(\Sigma) = \mathbb{E}_{X\sim\mathbb{P}} \left\| (k(\cdot,X) - m_{\mathbb{P}}) - P^{\ell}(\Sigma)(k(\cdot,X) - m_{\mathbb{P}}) \right\|_{\mathcal{H}}^2,$$

$$R^{\ell}(\widehat{\Sigma}) = \mathbb{E}_{X\sim\mathbb{P}} \left\| (k(\cdot,X) - m_{\mathbb{P}}) - P^{\ell}(\widehat{\Sigma})(k(\cdot,X) - \widehat{m}_{\mathbb{P}}) \right\|_{\mathcal{H}}^2, \tag{13}$$

$$R^{\ell}_{nys}(\widehat{\Sigma}) = \mathbb{E}_{X\sim\mathbb{P}} \left\| (k(\cdot,X) - m_{\mathbb{P}}) - P^{\ell}_{nys}(\widehat{\Sigma})(k(\cdot,X) - \widehat{m}_{\mathbb{P}}) \right\|_{\mathcal{H}}^2. \tag{14}$$

In (13) and (14), we center with $\widehat{m}_{\mathbb{P}} := \frac{1}{n}\sum_{i=1}^{n} k(\cdot, X_i)$ to ensure that the low-dimensional representation of the reconstructed random variable is computable in practice, as the mean element $m_{\mathbb{P}}$ is likely unknown. Throughout the rest of the analysis we will drop the $X \sim \mathbb{P}$ subscript, and assume expectations are with respect to $X \sim \mathbb{P}$ unless otherwise noted. The following theorem, proved in Section 6.3, provides a finite-sample bound on the reconstruction error associated with NY-EKPCA, under uniform sampling, as well as a new result for centered EKPCA, from which convergence rates may be obtained.

Theorem 3 *Suppose Assumptions 1-4 hold. For any $t > 0$, define $\mathcal{N}_{\Sigma}(t) = \mathrm{tr}(\Sigma(\Sigma + tI)^{-1})$ and $\mathcal{N}_{C,\infty}(t) = \sup_{x\in\mathcal{X}} \left\langle k(\cdot,x), (C+tI)^{-1} k(\cdot,x) \right\rangle_{\mathcal{H}}$, where $C = \int_{\mathcal{X}} k(\cdot,x) \otimes_{\mathcal{H}} k(\cdot,x) d\mathbb{P}(x)$ is the uncentered covariance operator. Then the following hold:*

(i)

$$R^{\ell}(\Sigma) = \sum_{i>\ell} \lambda_i.$$

(ii) For any $0 \le \delta \le \frac{1}{2}$, $n \ge 2\log\frac{2}{\delta}$ *and* $\frac{140\kappa}{n}\log\frac{16\kappa n}{\delta} \le t \le \|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})}$,

$$\mathbb{P}^n\left\{\sum_{i>\ell}\lambda_i \le R^\ell(\widehat{\Sigma}) \le 3\mathcal{N}_\Sigma(t)(\lambda_{\ell+1}+t) + \frac{32\kappa\log\frac{2}{\delta}}{n}\right\} \ge 1-5\delta.$$

(iii) For any $0 \le \delta \le \frac{1}{2}$,

$$\mathbb{P}^n\left\{\sum_{i>\ell}\lambda_i \le R^\ell_{nys}(\widehat{\Sigma}) \le 6\mathcal{N}_\Sigma(t)(\lambda_{\ell+1}+9t) + \frac{32\kappa\log\frac{2}{\delta}}{n}\right\} \ge 1-11\delta,$$

provided the following conditions are satisfied:

1. $\left(\frac{140\kappa}{n}\log\frac{16\kappa n}{\delta} \vee \frac{9\kappa}{n}\log\frac{n}{\delta}\right) \le t \le \|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})} \wedge \|C\|_{\mathcal{L}^\infty(\mathcal{H})}$,
2. $m \ge (67 \vee 5\mathcal{N}_{C,\infty}(t))\log\frac{4\kappa}{t\delta} \vee \frac{140\kappa}{t}\log\frac{8}{t\delta}$,
3. $n \ge 2\log\frac{2}{\delta}$.

Remark 4 *Since* Σ *is trace-class and* $\lambda_\ell \to 0$ *as* $\ell \to \infty$, *it follows that* $R^\ell(\Sigma) \to 0$ *as* $\ell \to \infty$. *The rate of this convergence may be analyzed after making assumptions on the decay rate of* $(\lambda_i)_i$, *which will be presented in the upcoming corollaries. The behavior of the empirical versions (of KPCA) depends significantly on* t *and* $\mathcal{N}_\Sigma(t)$. $\mathcal{N}_\Sigma(t)$ *is referred to as the effective dimension or degrees of freedom (Caponnetto and Vito, 2007), which measures the capacity of* $\mathcal{H}$. *Upon making assumptions regarding the decay rate of* $(\lambda_i)_i$, *the size of* $\mathcal{N}_\Sigma(t)$ *can be quantified and convergence rates for* $R^\ell(\widehat{\Sigma})$ *and* $R^\ell_{nys}(\widehat{\Sigma})$ *can be obtained. The upper bounds for* $R^\ell(\widehat{\Sigma})$ *and* $R^\ell_{nys}(\widehat{\Sigma})$ *are equivalent up to constants; however, the conditions imposed on* m *and* t *in (iii) will dictate whether this behavior of* $R^\ell_{nys}(\widehat{\Sigma})$ *may be achieved with a reduced computational complexity (*$m < n$*). We also would like to highlight that the results presented in Theorem 3, which are obtained for the U-statistic estimator* $\widehat{\Sigma}$ *of the centered covariance operator,* Σ, *matches up to constants, the results in Theorem 2 of Sterge et al. (2020), which were derived for the uncentered covariance operator,* C.

The following corollaries derive convergence rates from the bounds in Theorem 3 under the polynomial and exponential decay assumption on the eigenvalues of Σ.

Corollary 5 *Suppose* $\underline{A}i^{-\alpha} \le \lambda_i \le \bar{A}i^{-\alpha}$ *for some* $\alpha > 1$ *and* $0 < \underline{A} < \bar{A} < \infty$. *Let* $\ell = n^{\frac{\theta}{\alpha}}$, *where* $0 < \theta \le \alpha$. *Then*

(i)

$$n^{-\theta(1-\frac{1}{\alpha})} \lesssim R^\ell(\Sigma) \lesssim n^{-\theta(1-\frac{1}{\alpha})}.$$

There exists an $N \in \mathbb{N}$ *such that for all* $n > N$, *the following hold:*

(ii)

$$n^{-\theta(1-\frac{1}{\alpha})} \lesssim R^\ell(\widehat{\Sigma}) \lesssim_{\mathbb{P}^n} \begin{cases} n^{-\theta(1-\frac{1}{\alpha})}, & \theta < 1 \\ \left(\frac{\log n}{n}\right)^{1-\frac{1}{\alpha}}, & \theta \ge 1 \end{cases};$$

(iii)

$$n^{-\theta(1-\frac{1}{\alpha})} \lesssim R^{\ell}_{nys}(\widehat{\Sigma}) \lesssim_{\mathbb{P}^n} \begin{cases} n^{-\theta(1-\frac{1}{\alpha})}, & \theta < 1,\, m \gtrsim n^{\theta}\log n \\ \left(\frac{\log n}{n}\right)^{1-\frac{1}{\alpha}}, & \theta \geq 1\,, m \gtrsim \frac{n}{\log n}\log\frac{n}{\log n} \end{cases}.$$

Remark 6 *(i) Of course, $\alpha > 1$ is required to ensure that Σ is trace class. Observing (ii) and (iii) of Corollary 5, we see that the convergence rates of $R^{\ell}(\widehat{\Sigma})$ and $R^{\ell}_{nys}(\widehat{\Sigma})$ rely heavily on the growth of ℓ through θ. Comparing $R^{\ell}(\widehat{\Sigma})$ to $R^{\ell}(\Sigma)$, EKPCA will match the convergence rate of KPCA provided ℓ does not grow faster than $n^{1/\alpha}$. We note that $0 < \theta < 1$ is the only sensible regime both computationally and statistically, as $\theta \geq 1$ increases the computational complexity while the rate plateaus at $(\log n/n)^{1-\frac{1}{\alpha}}$.*

(ii) When $\theta < 1$, the convergence rate of $R^{\ell}_{nys}(\widehat{\Sigma})$ is equal to that of $R^{\ell}(\Sigma)$ and $R^{\ell}(\widehat{\Sigma})$, provided $m \gtrsim n^{\theta}\log n$, i.e., if ℓ grows to infinity not faster than $n^{1/\alpha}$ and the number of subsamples m grows sufficiently fast, then NY-EKPCA and EKPCA enjoy the same statistical behavior. From a computational perspective, the computational complexity of EKPCA is $O(n^{2+\frac{\theta}{\alpha}})$, while the complexity of NY-EKPCA is $O(nm^2 + m^2\ell) = O(nm^2)$. Thus, NY-EKPCA will offer a computational advantage with no loss in statistical performance, if $\theta < \frac{1}{2} + \frac{\theta}{2\alpha}$, i.e., $\theta < \frac{\alpha}{2\alpha-1}$. This means NY-EKPCA has better computational complexity and same statistical rates for $\theta < \frac{\alpha}{2\alpha-1}$ while it loses the computational edge with no loss in the statistical behavior when $\frac{\alpha}{2\alpha-1} \leq \theta < 1$. Note that the first few principal components are often the greatest interest in practice; thus, the case $\theta < 1$ may be more relevant in applications.

Corollary 7 (Exponential decay of eigenvalues) *Suppose $\underline{B}e^{-\tau i} \leq \lambda_i \leq \bar{B}e^{-\tau i}$ for $\tau > 0$ and $\underline{B}, \bar{B} \in (0,\infty)$. Let $\ell = \frac{1}{\tau}\log n^{\theta}$ for $\theta > 0$. Then*

(i)

$$n^{-\theta} \lesssim R^{\ell}(\Sigma) \lesssim n^{-\theta}.$$

There exists an $N \in \mathbb{N}$ such that for all $n > N$, the following hold:

(ii)

$$n^{-\theta} \lesssim R^{\ell}(\widehat{\Sigma}) \lesssim_{\mathbb{P}^n} \begin{cases} n^{-\theta}\log n, & \theta < 1 \\ n^{-1}(\log n)^2, & \theta \geq 1 \end{cases};$$

(iii)

$$n^{-\theta} \lesssim R^{\ell}_{nys}(\widehat{\Sigma}) \lesssim_{\mathbb{P}^n} \begin{cases} n^{-\theta}\log n, & \theta < 1,\, m \gtrsim n^{\theta}\log n \\ n^{-1}(\log n)^2, & \theta \geq 1,\, m \gtrsim \frac{n}{\log n}\log\frac{n}{\log n} \end{cases}.$$

Corollary 7 shares similar behavior to that Corollary 5 as discussed in Remark 6 but yields faster rates than in Corollary 5 because of the faster decay rate of eigenvalues. We do not include the proof of Corollary 7 in Section 6 as it is very similar to that (Sterge et al., 2020, Corollary 4.3).

4.2 Reconstruction error in $L^2(\mathbb{P})$-norm

While defining the reconstruction error of NY-KPCA in $\mathcal{H}$ is the most natural construction, we additionally study the reconstruction error of NY-KPCA in $L^2(\mathbb{P})$ to allow for comparison with another popular approximation based on random features. Random feature approximation (Rahimi and Recht, 2008a) computes a random low-dimensional approximation of the kernel, which is then used in place of k in learning methodologies to achieve reduced computational complexity. To elaborate, suppose

$$k(x,y) = \int_{\Theta} \varphi(x,\theta)\varphi(y,\theta)d\Lambda(\theta),$$

where $\varphi(x,\cdot) \in L^2(\Theta,\Lambda)$ for all $x \in \mathcal{X}$ and Λ is a probability measure (w.l.o.g.) on a measurable space, Θ. For a random sample $(\theta_i)_{i=1}^m \overset{i.i.d.}{\sim} \Lambda$, the random feature approximation to k is constructed as

$$k_m(x,y) = \frac{1}{m}\sum_{i=1}^{m} \varphi(x,\theta_i)\varphi(y,\theta_i) = \langle \Phi_m(x), \Phi_m(y)\rangle_2 ,$$

where $\Phi_m(x) = \frac{1}{\sqrt{m}}\left(\varphi(x,\theta_1),\ldots,\varphi(x,\theta_m)\right))^\top$ is the *approximate feature map*, i.e., an approximation to the *canonical feature map*, $\Phi(x) := k(\cdot,x)$. It can be shown (Sriperumbudur and Sterge, 2022, Section 3.3) that k_m is the reproducing kernel of an m-dimensional RKHS, denoted $\mathcal{H}_m$, which is isometrically isomorphic to $\mathbb{R}^m$. Random feature KPCA (RF-KPCA) involves solving

$$\arg\sup \left\{ \langle f, \Sigma_m f\rangle_{\mathcal{H}_m} : f \in \mathcal{H}_m, \|f\|_{\mathcal{H}_m} = 1 \right\},$$

where

$$\Sigma_m = \int_{\mathcal{X}} k_m(\cdot,x) \otimes_{\mathcal{H}_m} k_m(\cdot,x) d\mathbb{P}(x) - \left(\int_{\mathcal{X}} k_m(\cdot,x)d\mathbb{P}(x)\right) \otimes_{\mathcal{H}_m} \left(\int_{\mathcal{X}} k_m(\cdot,x)d\mathbb{P}(x)\right),$$

is the *approximate* covariance operator on $\mathcal{H}_m$ induced by k_m. Note that RF-KPCA is exactly KPCA but with Σ and $\mathcal{H}$ being replaced by their approximate counterparts, i.e., Σ_m and $\mathcal{H}_m$, respectively. This means, the solution to RF-KPCA is the eigenfunction that corresponds to the maximum eigenvalue of Σ_m. The empirical version of RF-KPCA, referred to as RF-EKPCA, involves solving RF-KPCA with Σ replaced by its U-statistic estimator

$$\widehat{\Sigma}_m = \frac{1}{2n(n-1)}\sum_{i\neq j}^{n} \left(k_m(\cdot,X_i) - k_m(\cdot,X_j)\right) \otimes_{\mathcal{H}_m} \left(k_m(\cdot,X_i) - k_m(\cdot,X_j)\right).$$

Note that the computation of top-ℓ eigenfunctions, $(\widehat{\phi}_{m,i})_{i=1}^{\ell} \subset \mathcal{H}_m$, of $\widehat{\Sigma}_m$ by RF-EKPCA has a computational complexity $O(m^2\ell + m^2 n)$, an improvement upon EKPCA provided $m < \sqrt{n\ell}$, i.e., $m = o(\sqrt{n\ell})$ as $n,\ell \to \infty$. The orthogonal projection operator

$$P_{rf}^{\ell}(\widehat{\Sigma}_m) = \sum_{i=1}^{\ell} \widehat{\phi}_{m,i} \otimes_{\mathcal{H}_m} \widehat{\phi}_{m,i}$$

may then be used to compute a low-dimensional representation of $X_i \in \mathcal{X}$ (see Sriperumbudur and Sterge, 2022, Section 3 for details). We emphasize that in Nyström approximation, m is the number of subsampled indices, while in random features, m is the number of random features sampled.

Unlike NY-EKPCA, where the eigenfunctions reside in $\mathcal{H}$, the eigenfunctions $(\widehat{\phi}_{m,i})_i$ lie in $\mathcal{H}_m$. Therefore, the reconstruction error of RF-EKPCA in $\mathcal{H}$-norm is ill-defined. To remedy this issue, Sriperumbudur and Sterge (2022) proposed a reconstruction error named *Reconstruct and Embed* (R-E), in which elements in $\mathcal{H}$ and $\mathcal{H}_m$ are mapped to a common space, $L^2(\mathbb{P})$ through the inclusion and approximation operators,

$$\mathfrak{I} : \mathcal{H} \to L^2(\mathbb{P}),\ f \mapsto f - f_{\mathbb{P}},$$

and

$$\mathfrak{A} : \mathcal{H}_m \to L^2(\mathbb{P}), \quad f = \sum_{i=1}^m \beta_i \varphi_i \mapsto \sum_{i=1}^m \beta_i(\varphi_i - \varphi_{i,\mathbb{P}}) = f - f_{\mathbb{P}},$$

where $\varphi_{i,\mathbb{P}} := \int_{\mathcal{X}} \varphi_i(x)\, d\mathbb{P}(x)$, $\varphi_i := \varphi(\cdot, \theta_i)$ and $f_{\mathbb{P}} = \int_{\mathcal{X}} f(x) d\mathbb{P}(x)$. As the name suggests, the functions are first reconstructed in $\mathcal{H}$ based on principal components and then embedded into $L^2(\mathbb{P})$. Formally, the reconstruction error for KPCA in $L^2(\mathbb{P})$ is defined as

$$T^\ell(\Sigma) = \mathbb{E}\left\| \mathfrak{I}(k(\cdot, X) - m_{\mathbb{P}}) - \mathfrak{I}P^\ell(\Sigma)(k(\cdot, X) - m_{\mathbb{P}}) \right\|^2_{L^2(\mathbb{P})}.$$

We refer the reader to (Sriperumbudur and Sterge, 2022, Sections 4.1 and 4.2) for details about the properties of R-E. In the following, we analyze the behavior of NY-EKPCA w.r.t. R-E and compare it to that of RF-EKPCA. To this end, the R-E reconstruction errors for EKPCA, RF-EKPCA and NY-EKPCA are defined as

$$\begin{aligned}
T^\ell(\widehat{\Sigma}) &= \mathbb{E}\left\| \mathfrak{I}(k(\cdot, X) - m_{\mathbb{P}}) - \mathfrak{I}P^\ell(\widehat{\Sigma})(k(\cdot, X) - \widehat{m}_{\mathbb{P}}) \right\|^2_{L^2(\mathbb{P})}, \\
T^\ell_{rf}(\widehat{\Sigma}_m) &= \mathbb{E}\left\| \mathfrak{I}(k(\cdot, X) - m_{\mathbb{P}}) - \mathfrak{A}P^\ell_{rf}(\widehat{\Sigma}_m)(k_m(\cdot, X) - \widehat{m}_{\mathbb{P},m}) \right\|^2_{L^2(\mathbb{P})}, \\
T^\ell_{nys}(\widehat{\Sigma}) &= \mathbb{E}\left\| \mathfrak{I}(k(\cdot, X) - m_{\mathbb{P}}) - \mathfrak{I}P^\ell_{nys}(\widehat{\Sigma})(k(\cdot, X) - \widehat{m}_{\mathbb{P}}) \right\|^2_{L^2(\mathbb{P})},
\end{aligned}$$

respectively, where $\widehat{m}_{\mathbb{P},m} = \frac{1}{n}\sum_{i=1}^n k_m(\cdot, X_i)$. Of course, $\|\cdot\|_{L^2(\mathbb{P})}$ is weaker than $\|\cdot\|_{\mathcal{H}}$, so naturally we can expect better error behavior of $T^\ell_{nys}(\widehat{\Sigma})$ when compared with $R^\ell_{nys}(\widehat{\Sigma})$. However, a more interesting comparison is between $T^\ell_{nys}(\widehat{\Sigma})$ and $T^\ell_{rf}(\widehat{\Sigma})$. To facilitate this, we first present the following result, which gives finite sample bounds on $T^\ell(\Sigma)$, $T^\ell(\widehat{\Sigma})$ and $T^\ell_{rf}(\widehat{\Sigma}_m)$.

Theorem 8 (Sriperumbudur and Sterge 2022, Theorem 6) *Suppose Assumptions 1-4 hold. For any $t > 0$, define $\mathcal{N}_\Sigma(t) = \operatorname{tr}(\Sigma(\Sigma + tI)^{-1})$. Then the following hold:*

(i)

$$T^\ell(\Sigma) = \sum_{i>\ell} \lambda_i^2.$$

(ii) For any $\delta > 0$ *with* $n \geq 2\log\frac{2}{\delta}$ *and* $\frac{140\kappa}{n}\log\frac{16\kappa n}{\delta} \leq t \leq \|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})}$,

$$\mathbb{P}^n\left\{\sum_{i\geq\ell}\lambda_i^2 \leq T^\ell(\widehat{\Sigma}) \leq 9\mathcal{N}_\Sigma(t)(\lambda_{\ell+1}+t)^2 + \frac{64\kappa^2\log\frac{2}{\delta}}{n}\right\} \geq 1-3\delta.$$

(iii) Suppose m *random features are sampled i. i. d. from a probability measure* Λ. *For any* $\delta > 0$, $n \geq 2\log\frac{2}{\delta}$, $m \geq \left(2 \vee \frac{1024\kappa^2}{\sum_{i>\ell}\lambda_i^2}\right)\log\frac{2}{\delta}$ *and* $\frac{140\kappa}{n}\log\frac{16\kappa n}{\delta} \vee \frac{86\kappa}{m}\log\frac{16\kappa m}{\delta} \leq t \leq \frac{\|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})}}{3}$, *with probability at least* $1-12\delta$ *over the joint measure* $\mathbb{P}^n \times \Lambda^m$:

$$\frac{1}{4}\sum_{i>\ell}\lambda_i^2 \leq T_{rf}^\ell(\widehat{\Sigma}_m) \leq 162\mathcal{A}_1(t)(\lambda_{\ell+1}+t)^2 + \frac{640\kappa^2\log\frac{2}{\delta}}{3n} + \frac{256\kappa^2\log\frac{2}{\delta}}{m},$$

where $\mathcal{A}_1(t) := \mathcal{N}_\Sigma(t) + \frac{16\kappa\log\frac{2}{\delta}}{tm} + \sqrt{\frac{8\kappa\mathcal{N}_\Sigma(t)\log\frac{2}{\delta}}{tm}}$.

We must note that unlike *(i)-(ii)*, the probability statement in *(iii)* is with respect to a joint measure, as it must consider the randomness introduced by the random features. In light of the previously presented result, we now present an analogous result for NY-EKPCA in $L^2(\mathbb{P})$-norm.

Theorem 9 *Under the same assumptions as in Theorem 3,*

$$\mathbb{P}^n\left\{\sum_{i>\ell}\lambda_i^2 \leq T_{nys}^\ell(\widehat{\Sigma}) \leq 36\mathcal{N}_\Sigma(t)(\lambda_{\ell+1}+9t)^2 + \frac{32\kappa^2\log\frac{2}{\delta}}{n}\right\} \geq 1-11\delta,$$

provided the following conditions are satisfied:

1. $\left(\frac{140\kappa}{n}\log\frac{16\kappa n}{\delta} \vee \frac{9\kappa}{n}\log\frac{n}{\delta}\right) \leq t \leq \|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})} \wedge \|C\|_{\mathcal{L}^\infty(\mathcal{H})}$,
2. $m \geq (67 \vee 5\mathcal{N}_{C,\infty}(t))\log\frac{4\kappa}{t\delta} \vee \frac{140\kappa}{t}\log\frac{8}{t\delta}$,
3. $n \geq 2\log\frac{2}{\delta}$.

Remark 10 *(i) Comparing Theorem 9 with Theorem 8(ii), we note that the bounds for EKPCA and NY-EKPCA are identical up to constants, similar to the case in Theorem 3. Compared to their counterparts in Theorem 3,* $T^\ell(\widehat{\Sigma})$ *and* $T_{nys}^\ell(\widehat{\Sigma})$ *have similar dependence on the effective dimension, but a squared dependence on* $\lambda_{\ell+1}$ *and* t*—in contrast to a linear dependence in Theorem 3. This is a byproduct of working with the* $L^2(\mathbb{P})$*-norm, which is weaker than the RKHS norm, and therefore will result in faster convergence rates, as will be evident in Corollary 11. Additionally,* $T^\ell(\Sigma)$ *will decay as* $\ell \to \infty$ *more rapidly than* $R^\ell(\Sigma)$, *as it depends on the sum of squared eigenvalues. The error in estimating the mean element is not improved in the move to* $L^2(\mathbb{P})$*-norm; it is bounded as* n^{-1} *in all of the empirical varieties regardless of norm.*

(ii) An immediate difference between NY-EKPCA and RF-EKPCA is the dependence on m. *This difference can be seen in both the upper bounds of* $T_{rf}^\ell(\widehat{\Sigma}_m)$ *and* $T_{nys}^\ell(\widehat{\Sigma})$, *as well as the*

size requirements on m. This is primarily due to the approximation error incurred by RF-EKPCA, which approximates $\mathcal{H}$ with an m-dimensional RKHS. Of course, this dependence on m is crucial in analyzing the computational vs. statistical trade-off between the two methods.

The following corollaries specialize Theorems 8 and 9 under the polynomial and exponential decay assumption on the eigenvalues of Σ.

Corollary 11 *Suppose $\underline{A} i^{-\alpha} \leq \lambda_i \leq \bar{A} i^{-\alpha}$ for some $\alpha > 1$ and $0 < \underline{A} < \bar{A} < \infty$. Let $\ell = n^{\frac{\theta}{\alpha}}$, where $0 < \theta \leq \alpha$. Then*

(i)

$$n^{-2\theta(1-\frac{1}{2\alpha})} \lesssim T^{\ell}(\Sigma) \lesssim n^{-2\theta(1-\frac{1}{2\alpha})}.$$

There exists an $N \in \mathbb{N}$ such that for all $n > N$, the following hold:

(ii)

$$n^{-2\theta(1-\frac{1}{2\alpha})} \lesssim T^{\ell}(\widehat{\Sigma}) \lesssim_{\mathbb{P}^n} \begin{cases} n^{-2\theta(1-\frac{1}{2\alpha})}, & \theta < \frac{\alpha}{2\alpha-1} \\ \frac{1}{n}, & \theta \geq \frac{\alpha}{2\alpha-1} \end{cases};$$

(iii)

$$n^{-2\theta(1-\frac{1}{2\alpha})} \lesssim T^{\ell}_{nys}(\widehat{\Sigma}) \lesssim_{\mathbb{P}^n} \begin{cases} n^{-2\theta(1-\frac{1}{2\alpha})}, & \theta < \frac{\alpha}{2\alpha-1}, m \gtrsim n^{\theta} \log n \\ \frac{1}{n}, & \theta \geq \frac{\alpha}{2\alpha-1}, m \gtrsim n^{\frac{\alpha}{2\alpha-1}} \log n \end{cases};$$

(iv)

$$n^{-2\theta(1-\frac{1}{2\alpha})} \mathbb{1}_{\{\gamma \geq \theta(2-\frac{1}{\alpha})\}} \lesssim T^{\ell}_{rf}(\widehat{\Sigma}_m) \lesssim_{\mathbb{P}^n \times \Lambda^m} \begin{cases} n^{-2\theta(1-\frac{1}{2\alpha})}, & \gamma \geq \theta(2-\frac{1}{\alpha}), \theta < \frac{\alpha}{2\alpha-1} \\ n^{-\gamma}, & \gamma < 1 \wedge \theta(2-\frac{1}{\alpha}) \end{cases},$$

where $m = n^{\gamma}$ for $0 < \gamma \leq 1$.

Remark 12 *(i), (ii), and (iv) of Corollary 11 are quoted from Sriperumbudur and Sterge (2022, Corollary 7).*
(i) We first highlight the difference between the $L^2(\mathbb{P})$ and $\mathcal{H}$ norms through the comparison of Corollaries 5 and 11. $T^{\ell}(\Sigma)$ decays at a rate of $n^{-2\theta(1-\frac{1}{2\alpha})}$, which is faster than that of its analogue in $\mathcal{H}$-norm, i.e., $R^{\ell}(\Sigma)$. While $R^{\ell}(\widehat{\Sigma})$ and $R^{\ell}_{nys}(\widehat{\Sigma})$ recover the optimal convergence rate (compared to KPCA) in the range $\theta < 1$, $T^{\ell}(\widehat{\Sigma})$ and $T^{\ell}_{nys}(\widehat{\Sigma})$ are only able to recover the optimal rate for $\theta < \frac{\alpha}{2\alpha-1}$. The $\frac{1}{n}$ term, which arises due to the empirical recentering, is never dominant in $R^{\ell}(\widehat{\Sigma})$ and $R^{\ell}_{nys}(\widehat{\Sigma})$; however, it can dominate in $T^{\ell}(\widehat{\Sigma})$ and $T^{\ell}_{nys}(\widehat{\Sigma})$ depending on the range of θ.
(ii) Observing $T^{\ell}(\widehat{\Sigma})$ and $T^{\ell}_{nys}(\widehat{\Sigma})$ we see that, as in Corollary 5, NY-EKPCA and EKPCA have similar convergence behavior, provided m is large enough. Therefore, it follows from Remark 6, that in both $\mathcal{H}$ and $L^2(\mathbb{P})$, NY-EKPCA will provide less computational cost with no loss in statistical performance compared to EKPCA. However in $L^2(\mathbb{P})$, unlike in $\mathcal{H}$,

NY-EKPCA is computationally advantageous than EKPCA regardless of the size of θ.
(iii) When $\theta < \frac{\alpha}{2\alpha-1}$, both RF-EKPCA and NY-EKPCA achieve the optimal convergence rate of $n^{-2\theta(1-\frac{1}{2\alpha})}$, but with NY-EKPCA being more computationally efficient than RF-EKPCA. This is because, in this regime, RF-EKPCA scales as $O(n^{1+2\gamma})$ and NY-EKPCA as $O(n^{1+2\theta})$ with $\gamma \geq \theta(2-\frac{1}{\alpha}) > \theta$. In the range $\theta \geq \frac{\alpha}{2\alpha-1}$, NY-EKPCA achieves faster convergence rate of $\frac{1}{n}$, while RF-EKPCA converges as $n^{-\gamma}$ with $\gamma < 1$. Further, in this range of θ, NY-EKPCA will offer less computational cost for $\gamma \geq \frac{\alpha}{2\alpha-1}$, as RF-EKPCA scales as $O(n^{1+2\gamma})$ and NY-EKPCA as $O(n^{1+\frac{2\alpha}{2\alpha-1}})$. However, we would like to warn that these comparisons are made based on only lower bounds on m and it is not clear whether these bounds are sharp.

Corollary 13 *Suppose $\underline{B}e^{-\tau i} \leq \lambda_i \leq \bar{B}e^{-\tau i}$ for $\tau > 0$ and $\underline{B}, \bar{B} \in (0,\infty)$. Let $\ell = \frac{1}{\tau}\log n^{\theta}$ for $\theta > 0$. Then*

(i)

$$n^{-2\theta} \lesssim T^{\ell}(\Sigma) \lesssim n^{-2\theta}.$$

There exists $\tilde{n} \in \mathbb{N}$ such that for all $n > \tilde{n}$, the following hold:

(ii)

$$n^{-2\theta} \lesssim T^{\ell}(\widehat{\Sigma}) \lesssim_{\mathbb{P}^n} \begin{cases} n^{-2\theta}\log n, & \theta \leq \frac{1}{2} \\ \frac{1}{n}, & \theta > \frac{1}{2} \end{cases} ;$$

(iii)

$$n^{-2\theta} \lesssim T^{\ell}_{nys}(\widehat{\Sigma}) \lesssim_{\mathbb{P}^n} \begin{cases} n^{-2\theta}\log n, & \theta \leq \frac{1}{2},\ m \gtrsim n^{\theta}\log n \\ \frac{\log n}{n}, & \theta > \frac{1}{2},\ m \gtrsim \sqrt{n}\log n \end{cases} ;$$

(iv) For $0 < \gamma \leq 1$ and $m = n^{\gamma}$,

$$n^{-2\theta}\mathbb{1}_{\{\gamma \geq 2\theta\}} \lesssim_{\Lambda^m} T^{\ell}_{rf}(\widehat{\Sigma}_m) \lesssim_{\mathbb{P}^n \times \Lambda^m} \begin{cases} n^{-2\theta}\log n, & \gamma \geq 2\theta,\ \theta \leq \frac{1}{2} \\ n^{-\gamma}, & \gamma < 2\theta,\ \gamma \leq 1 \end{cases} .$$

5. Discussion

We have studied the trade-off between statistical accuracy and computational complexity in Nyström-based approximate kernel PCA. While it is clear that Nyström kernel PCA is computationally advantageous to EKPCA for Nyström subsamples $m < n$, we have showed the error in reconstructing $k(\cdot, X)$ in $\mathcal{H}$ using ℓ-eigenfunctions in Nyström kernel PCA to be statistically optimal, provided m is large enough, and ℓ small enough. Additionally, the size of m depends on the number of eigenfunctions ℓ; larger ℓ requires more subsamples to achieve the best possible statistical behavior. Unlike several existing theoretical works on kernel PCA, we derived these results by not assuming the mean element of k to be zero. Further, we adapted our notion of reconstruction error to the $L^2(\mathbb{P})$ setting in order to compare Nyström kernel PCA with random feature-based kernel PCA. In $L^2(\mathbb{P})$, we showed Nyström kernel PCA to achieve the best possible statistical behavior, while maintaining its

computational edge regardless of the number of eigenfunctions, ℓ. In comparison to random features, we showed Nyström KPCA to require fewer Nyström samples than the number of random features for RF-EKPCA in order to achieve optimal statistical rates. However, we express caution that this observation is based only on lower bounds on m and the sharpness of these bounds have to be established. Further comparison to random features could be explored by considering alternative definitions of reconstruction error in $L^2(\mathbb{P})$ as proposed in Sriperumbudur and Sterge (2022).

While this work considers only plain Nyström, that is, choosing subsamples uniformly, it is possible to choose these subsamples with probabilities proportional to their individual leverage scores, defined as the diagonal entries of the matrix $(\mathbf{K} + nt\mathbf{I}_n)^{-1}\mathbf{K}$. Consideration of Nyström subsampling according to the leverage scores has yielded success in kernel ridge regression (Rudi et al., 2015; Alaoui and Mahoney, 2015), where it has led to relaxed requirements on the size of m necessary to achieve best possible statistical behavior. Thus, the computational benefit provided by Nyström is more pronounced when points are sampled according to the leverage score distribution. Though leverage score sampling has been studied successfully in uncentered Nyström KPCA (Sterge et al., 2020), the U-statistic estimator considered in this work introduces significant technical challenges, and so is relegated to future work.

6. Proofs

In this section, we present the proofs of the results of the paper.

6.1 Proof of Proposition 2

Note that any $f \in \bar{\mathcal{H}}_m$ can be written as $\tilde{S}_m^* \mathbf{H}_m \boldsymbol{\alpha}$ for some $\boldsymbol{\alpha} \in \mathbb{R}^m$. Then the objective in (9) can be written as

$$\begin{aligned}\left\langle f, \widehat{\Sigma} f \right\rangle_{\mathcal{H}} &= \left\langle \tilde{S}_m^* \mathbf{H}_m \boldsymbol{\alpha}, \widehat{\Sigma} \tilde{S}_m^* \mathbf{H}_m \boldsymbol{\alpha} \right\rangle_{\mathcal{H}} = \frac{1}{n(n-1)} \left\langle \tilde{S}_m^* \mathbf{H}_m \boldsymbol{\alpha}, S_n^* \mathbf{H}_n S_n \tilde{S}_m^* \mathbf{H}_m \boldsymbol{\alpha} \right\rangle_{\mathcal{H}} \\ &= \frac{1}{n(n-1)} \left\langle S_n \tilde{S}_m^* \mathbf{H}_m \boldsymbol{\alpha}, \mathbf{H}_n S_n \tilde{S}_m^* \mathbf{H}_m \boldsymbol{\alpha} \right\rangle_2 = \frac{1}{n(n-1)} \left\langle \mathbf{K}_{nm} \mathbf{H}_m \boldsymbol{\alpha}, \mathbf{H}_n \mathbf{K}_{nm} \mathbf{H}_m \boldsymbol{\alpha} \right\rangle_2 \\ &= \frac{1}{n(n-1)} \left\langle \mathbf{K}_{mn} \mathbf{H}_n \mathbf{K}_{nm} \mathbf{H}_m \boldsymbol{\alpha}, \mathbf{H}_m \boldsymbol{\alpha} \right\rangle_2,\end{aligned}$$

where we have used Lemma A.7 in several steps. Similarly, the constraint in (9) can be written as

$$\begin{aligned}\|f\|_{\mathcal{H}}^2 &= \left\| \tilde{S}_m^* \mathbf{H}_m \boldsymbol{\alpha} \right\|_{\mathcal{H}}^2 = \left\langle \tilde{S}_m^* \mathbf{H}_m \boldsymbol{\alpha}, \tilde{S}_m^* \mathbf{H}_m \boldsymbol{\alpha} \right\rangle_{\mathcal{H}} = \left\langle \tilde{S}_m \tilde{S}_m^* \mathbf{H}_m \boldsymbol{\alpha}, \mathbf{H}_m \boldsymbol{\alpha} \right\rangle_2 \\ &= \left\langle \mathbf{K}_{mm} \mathbf{H}_m \boldsymbol{\alpha}, \mathbf{H}_m \boldsymbol{\alpha} \right\rangle_2 = \left\| \mathbf{K}_{mm}^{1/2} \mathbf{H}_m \boldsymbol{\alpha} \right\|_2^2,\end{aligned}$$

which also follows from Lemma A.7. Thus, we may express (9) as

$$\sup_{\substack{\boldsymbol{\alpha} \in \mathbb{R}^m \\ \left\| \mathbf{K}_{mm}^{1/2} \mathbf{H}_m \boldsymbol{\alpha} \right\|_2 = 1}} \frac{1}{n(n-1)} \left\langle \mathbf{K}_{mn} \mathbf{H}_n \mathbf{K}_{nm} \mathbf{H}_m \boldsymbol{\alpha}, \mathbf{H}_m \boldsymbol{\alpha} \right\rangle_2.$$

Let $\mathbf{u} = \mathbf{K}_{mm}^{1/2}\mathbf{H}_m\boldsymbol{\alpha}$, and the above problem may be written as

$$\sup_{\substack{\mathbf{u}\in\mathcal{R}(\mathbf{K}_{mm}^{1/2}\mathbf{H}_m)\\ \mathbf{u}^\top\mathbf{u}=1}} \frac{1}{n(n-1)}\left\langle \mathbf{K}_{mn}\mathbf{H}_n\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1/2}\mathbf{u}, \mathbf{K}_{mm}^{-1/2}\mathbf{u}\right\rangle_2 = \sup_{\substack{\mathbf{u}\in\mathcal{R}(\mathbf{K}_{mm}^{1/2}\mathbf{H}_m)\\ \mathbf{u}^\top\mathbf{u}=1}} \frac{1}{n(n-1)}\mathbf{u}^\top\mathbf{M}\mathbf{u},$$

where $\mathbf{M} := \mathbf{K}_{mm}^{-1/2}\mathbf{K}_{mn}\mathbf{H}_n\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1/2}$. The solution to the above problem is the unit eigenvector of $\frac{1}{n(n-1)}\mathbf{M}$ corresponding to its largest eigenvalue; denote this eigenvector as $\mathbf{u}_1$ with eigenvalue $\tilde{\lambda}_1$. We then have $\mathbf{H}_m\boldsymbol{\alpha}_1 = \mathbf{K}_{mm}^{-1/2}\mathbf{u}_1$ yielding the function in $\bar{\mathcal{H}}_m$,

$$\tilde{\phi}_1 = \tilde{S}_m^*\mathbf{K}_{mm}^{-1/2}\mathbf{u}_1,$$

solving (9). Subsequent eigenfunctions $\tilde{\phi}_i$ may be computed in a similar manner from the eigenvectors of $\frac{1}{n(n-1)}\mathbf{M}$, and the orthonormality of the $\{\tilde{\phi}_i\}_i$ follows from the orthonormality of the $\{\mathbf{u}_i\}_i$, i.e.,

$$\left\langle \tilde{\phi}_i, \tilde{\phi}_j\right\rangle_{\mathcal{H}} = \left\langle \tilde{S}_m^*\mathbf{K}_{mm}^{-1/2}\mathbf{u}_i, \tilde{S}_m^*\mathbf{K}_{mm}^{-1/2}\mathbf{u}_j\right\rangle_2 = \left\langle \mathbf{K}_{mm}^{-1/2}\mathbf{K}_{mm}\mathbf{K}_{mm}^{-1/2}\mathbf{u}_i, \mathbf{u}_j\right\rangle_2 = \delta_{ij}.$$

6.2 Proofs of (11) and (12)

(11) is immediate because $\bar{\mathcal{H}}_m = \mathcal{R}(\tilde{S}_m^*\mathbf{H}_m)$ and

$$\mathbf{H}_m\tilde{S}_m\tilde{S}_m^*\mathbf{H}_m = \mathbf{H}_m\mathbf{K}_{mm}\mathbf{H}_m.$$

To verify (12), because $\tilde{\phi}_i \in \bar{\mathcal{H}}_m$ for all $i \in [m]$, we have

$$\begin{aligned}
\bar{P}_m\widehat{\Sigma}\bar{P}_m\tilde{\phi}_i = \bar{P}_m\widehat{\Sigma}\tilde{\phi}_i &= \frac{1}{n(n-1)}\bar{P}_m S_n^*\mathbf{H}_n S_n\tilde{S}_m^*\mathbf{K}_{mm}^{-1/2}\mathbf{u}_i \\
&= \frac{1}{n(n-1)}\tilde{S}_m^*\mathbf{H}_m(\mathbf{H}_m\mathbf{K}_{mm}\mathbf{H}_m)^+\mathbf{H}_m\tilde{S}_m S_n^*\mathbf{H}_n S_n\tilde{S}_m^*\mathbf{K}_{mm}^{-1/2}\mathbf{u}_i \\
&= \frac{1}{n(n-1)}\tilde{S}_m^*\mathbf{H}_m(\mathbf{H}_m\mathbf{K}_{mm}\mathbf{H}_m)^+\mathbf{H}_m\mathbf{K}_{mn}\mathbf{H}_n\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1/2}\mathbf{u}_i \\
&= \frac{1}{n(n-1)}\tilde{S}_m^*\mathbf{H}_m(\mathbf{H}_m\mathbf{K}_{mm}\mathbf{H}_m)^+\mathbf{H}_m\mathbf{K}_{mm}^{1/2}\mathbf{M}\mathbf{u}_i \\
&= \tilde{\lambda}_i\tilde{S}_m^*\mathbf{H}_m(\mathbf{H}_m\mathbf{K}_{mm}\mathbf{H}_m)^+\mathbf{H}_m\mathbf{K}_{mm}\mathbf{H}_m\boldsymbol{\alpha}_i = \tilde{\lambda}_i\bar{P}_m\tilde{S}_m^*\mathbf{H}_m\boldsymbol{\alpha}_i = \tilde{\lambda}_i\tilde{\phi}_i,
\end{aligned}$$

which follows from Lemma A.7, thereby completing the proof.

6.3 Proof of Theorem 3

For notational convenience, we define $\bar{k}(\cdot, X) := k(\cdot, X) - m_{\mathbb{P}}$.

(i) From Lemma A.6, we have

$$R^\ell(\Sigma) = \left\|(I - P^\ell(\Sigma))\Sigma^{1/2}\right\|_{\mathcal{L}^2(\mathcal{H})}^2 = \operatorname{tr}\left((I - P^\ell(\Sigma))\Sigma(I - P^\ell(\Sigma))\right) = \sum_{i>\ell}\lambda_i.$$

(*ii*) *Upper Bound:* We write

$$\begin{aligned} R^\ell(\widehat{\Sigma}) &= \mathbb{E}\left\|(k(\cdot,X)-m_{\mathbb{P}}) - P^\ell(\widehat{\Sigma})(k(\cdot,X)-\widehat{m}_{\mathbb{P}})\right\|_{\mathcal{H}}^2 \\ &= \mathbb{E}\left\|(I-P^\ell(\widehat{\Sigma}))\bar{k}(\cdot,X)\right\|_{\mathcal{H}}^2 + \left\|P^\ell(\widehat{\Sigma})(m_{\mathbb{P}}-\widehat{m}_{\mathbb{P}})\right\|_{\mathcal{H}}^2 \\ &\qquad -2\mathbb{E}\left\langle (I-P^\ell(\widehat{\Sigma}))\bar{k}(\cdot,X), P^\ell(\widehat{\Sigma})(m_{\mathbb{P}}-\widehat{m}_{\mathbb{P}})\right\rangle_{\mathcal{H}} \\ &= \underbrace{\left\|(I-P^\ell(\widehat{\Sigma}))\Sigma^{1/2}\right\|_{\mathcal{L}^2(\mathcal{H})}^2}_{\text{(A)}} + \underbrace{\left\|P^\ell(\widehat{\Sigma})(m_{\mathbb{P}}-\widehat{m}_{\mathbb{P}})\right\|_{\mathcal{H}}^2}_{\text{(B)}}, \end{aligned} \tag{15}$$

where the last equality holds because $\mathbb{E}[\bar{k}(\cdot,X)] = 0$ and we have employed Lemma A.6. For any $t > 0$ we have

$$\begin{aligned} \text{(A)} &= \left\|(I-P^\ell(\widehat{\Sigma}))(\widehat{\Sigma}+tI)^{1/2}(\widehat{\Sigma}+tI)^{-1/2}(\Sigma+tI)^{1/2}(\Sigma+tI)^{-1/2}\Sigma^{1/2}\right\|_{\mathcal{L}^2(\mathcal{H})}^2 \\ &\le \left\|(\Sigma+tI)^{-1/2}\Sigma^{1/2}\right\|_{\mathcal{L}^2(\mathcal{H})}^2 \left\|(\widehat{\Sigma}+tI)^{-1/2}(\Sigma+tI)^{1/2}\right\|_{\mathcal{L}^\infty(\mathcal{H})}^2 \\ &\qquad \times \left\|(I-P^\ell(\widehat{\Sigma}))(\widehat{\Sigma}+tI)^{1/2}\right\|_{\mathcal{L}^\infty(\mathcal{H})}^2 \\ &\overset{(\dagger)}{\le} \mathcal{N}_\Sigma(t)\left\|(\widehat{\Sigma}+tI)^{-1/2}(\Sigma+tI)^{1/2}\right\|_{\mathcal{L}^\infty(\mathcal{H})}^2 (\widehat{\lambda}_{\ell+1}+t), \end{aligned} \tag{16}$$

where we have used

$$\left\|(\Sigma+tI)^{-1/2}\Sigma^{1/2}\right\|_{\mathcal{L}^2(\mathcal{H})}^2 = \operatorname{tr}\left(\Sigma^{1/2}(\Sigma+tI)^{-1}\Sigma^{1/2}\right) = \mathcal{N}_\Sigma(t),$$

which holds via invariance of trace under cyclic permutations, in (†). The result follows from applying Lemma A.1 to (16) and Lemma A.3 to (B), noticing that

$$\text{(B)} \le \left\|P^\ell(\widehat{\Sigma})\right\|_{\mathcal{L}^\infty(\mathcal{H})}^2 \|m_{\mathbb{P}}-\widehat{m}_{\mathbb{P}}\|_{\mathcal{H}}^2 \le \|m_{\mathbb{P}}-\widehat{m}_{\mathbb{P}}\|_{\mathcal{H}}^2 .$$

Lower Bound: It is clear from (15) that $R^\ell(\widehat{\Sigma}) \ge$ (A). We will show that

$$R^\ell(\Sigma) = \inf_{\{\psi_i\}_i \in Q} \left\|(I-P_{\psi,\ell})\,\Sigma^{1/2}\right\|_{\mathcal{L}^2(\mathcal{H})}^2$$

where $P_{\psi,\ell} = \sum_{i=1}^\ell \psi_i \otimes_{\mathcal{H}} \psi_i$ and

$$Q = \left\{\{\psi_i\}_{i=1}^\ell \subset \mathcal{H} : \langle \psi_i, \psi_j\rangle_{\mathcal{H}} = \delta_{ij}, \forall i,j \in [\ell]\right\},$$

which in turn implies that (A) $\ge R^\ell(\Sigma)$. We have

$$\begin{aligned} \left\|(I-P_{\psi,\ell})\,\Sigma^{1/2}\right\|_{\mathcal{L}^2(\mathcal{H})}^2 &= \operatorname{tr}\left[\Sigma^{1/2}(I-P_{\psi,\ell})(I-P_{\psi,\ell})\Sigma^{1/2}\right] = \operatorname{tr}\left[(I-P_{\psi,\ell})\Sigma\right] \\ &= \sum_{i\ge 1}\lambda_i - \langle P_{\psi,\ell}, \Sigma\rangle_{\mathcal{L}^2(\mathcal{H})} . \end{aligned} \tag{17}$$

Clearly the l.h.s. of (17) is minimized if and only if $\langle P_{\psi,\ell}, \Sigma\rangle_{\mathcal{L}^2(\mathcal{H})}$ is maximized over Q, which occurs only when $\psi_i = \phi_i$, yielding $\langle P_{\psi,\ell}, \Sigma\rangle_{\mathcal{L}^2(\mathcal{H})} = \sum_{i=1}^{\ell} \lambda_i$.

(*iii*) *Upper Bound:* We first establish notation for the uncentered covariance operator,

$$C := \int_{\mathcal{X}} k(\cdot, x) \otimes_{\mathcal{H}} k(\cdot, x) d\mathbb{P}(x),$$

and its empirical estimate

$$C_n = \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i) \otimes_{\mathcal{H}} k(\cdot, X_i),$$

which will be necessary for our upcoming analysis. We decompose the reconstruction error as

$$\begin{aligned} R^{\ell}_{nys}(\widehat{\Sigma}) &= \mathbb{E} \left\| (k(\cdot, X) - m_{\mathbb{P}}) - P^{\ell}_{nys}(\widehat{\Sigma})(k(\cdot, X) - \widehat{m}_{\mathbb{P}}) \right\|^2_{\mathcal{H}} \\ &= \underbrace{\mathbb{E} \left\| (I - P^{\ell}_{nys}(\widehat{\Sigma}))\bar{k}(\cdot, X) \right\|^2_{\mathcal{H}}}_{\text{(C)}} + \underbrace{\left\| P^{\ell}_{nys}(\widehat{\Sigma})(m_{\mathbb{P}} - \widehat{m}_{\mathbb{P}}) \right\|^2_{\mathcal{H}}}_{\text{(D)}} \\ &\quad -2\mathbb{E} \left\langle (I - P^{\ell}_{nys}(\widehat{\Sigma}))\bar{k}(\cdot, X), P^{\ell}_{nys}(\widehat{\Sigma})(m_{\mathbb{P}} - \widehat{m}_{\mathbb{P}}) \right\rangle_{\mathcal{H}}. \end{aligned} \tag{18}$$

Now $\mathbb{E}[\bar{k}(\cdot, X)] = 0$ implies the third term in (18) is 0. (D) can be bound by writing

$$\text{(D)} \leq \left\| P^{\ell}_{nys}(\widehat{\Sigma}) \right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \|m_{\mathbb{P}} - \widehat{m}_{\mathbb{P}}\|^2_{\mathcal{H}} \leq \|m_{\mathbb{P}} - \widehat{m}_{\mathbb{P}}\|^2_{\mathcal{H}},$$

and applying Lemma A.3, yields

$$\mathbb{P}^n \left\{ \text{(D)} \leq \frac{32\kappa \log \frac{2}{\delta}}{n} \right\} \geq 1 - \delta. \tag{19}$$

Regarding (C), for any $t > 0$, we have

$$\begin{aligned} \text{(C)} &\overset{(\star)}{=} \left\| (I - P^{\ell}_{nys}(\widehat{\Sigma}))\Sigma^{1/2} \right\|^2_{\mathcal{L}^2(\mathcal{H})} \\ &= \left\| (I - P^{\ell}_{nys}(\widehat{\Sigma}))(\widehat{\Sigma} + tI)^{1/2}(\widehat{\Sigma} + tI)^{-1/2}\Sigma^{1/2} \right\|^2_{\mathcal{L}^2(\mathcal{H})} \\ &\leq \left\| (I - P^{\ell}_{nys}(\widehat{\Sigma}))(\widehat{\Sigma} + tI)^{1/2} \right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \left\| (\widehat{\Sigma} + tI)^{-1/2}\Sigma^{1/2} \right\|^2_{\mathcal{L}^2(\mathcal{H})} \\ &\leq \left\| (I - P^{\ell}_{nys}(\widehat{\Sigma}))(\widehat{\Sigma} + tI)^{1/2} \right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \left\| (\widehat{\Sigma} + tI)^{-1/2}(\Sigma + tI)^{1/2} \right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \\ &\qquad \times \left\| (\Sigma + tI)^{-1/2}\Sigma^{1/2} \right\|^2_{\mathcal{L}^2(\mathcal{H})} \\ &\overset{(\dagger)}{\leq} 2\mathcal{N}_{\Sigma}(t) \left\| (I - P^{\ell}_{nys}(\widehat{\Sigma}))(\widehat{\Sigma} + tI)^{1/2} \right\|^2_{\mathcal{L}^\infty(\mathcal{H})}, \end{aligned} \tag{20}$$

where we have used Lemma A.6 in ($\star$) and Lemma A.1 in (†). For convenience, we now let $\widehat{\Sigma}_t = \widehat{\Sigma} + tI$. Observing the last term, we have

$$
\begin{aligned}
&\left\|(I - P^{\ell}_{nys}(\widehat{\Sigma}))(\widehat{\Sigma} + tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \\
&\leq \left(\left\|(I-\bar{P}_m)\widehat{\Sigma}_t^{1/2}\right\|_{\mathcal{L}^\infty(\mathcal{H})} + \left\|(\bar{P}_m - P^{\ell}_{nys}(\widehat{\Sigma}))\widehat{\Sigma}_t^{1/2}\right\|_{\mathcal{L}^\infty(\mathcal{H})}\right)^2 \\
&\leq 2\left\|\left(I-\bar{P}_m\right)\widehat{\Sigma}_t^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} + 2\left\|\left(\bar{P}_m - P^{\ell}_{nys}(\widehat{\Sigma})\right)\widehat{\Sigma}_t^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \\
&\leq \underbrace{2\left\|(I-\bar{P}_m)\Sigma_t^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})}\left\|\Sigma_t^{-1/2}\widehat{\Sigma}_t^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})}}_{\text{(C1)}} + 2\left\|(I - P^{\ell}_{nys}(\widehat{\Sigma}))\bar{P}_m\widehat{\Sigma}_t^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \qquad (21) \\
&= 2\,\text{(C1)} + 2\left\|(I - P^{\ell}_{nys}(\widehat{\Sigma}))\bar{P}_m\widehat{\Sigma}_t\bar{P}_m(I - P^{\ell}_{nys}(\widehat{\Sigma}))\right\|_{\mathcal{L}^\infty(\mathcal{H})} \\
&\leq 2\,\text{(C1)} + 2\left\|(I - P^{\ell}_{nys}(\widehat{\Sigma}))\bar{P}_m\widehat{\Sigma}\bar{P}_m(I - P^{\ell}_{nys}(\widehat{\Sigma}))\right\|_{\mathcal{L}^\infty(\mathcal{H})} \\
&\qquad + 2t\left\|(I - P^{\ell}_{nys}(\widehat{\Sigma}))\bar{P}_m(I - P^{\ell}_{nys}(\widehat{\Sigma}))\right\|_{\mathcal{L}^\infty(\mathcal{H})} \\
&\leq 2\,\text{(C1)} + 2\left(\tilde{\lambda}_{\ell+1} + t\right), \qquad (22)
\end{aligned}
$$

where we have used $\mathcal{R}(P^{\ell}_{nys}(\widehat{\Sigma})) \subset \mathcal{R}(\bar{P}_m)$ in (21), and (22) holds because $P^{\ell}_{nys}(\widehat{\Sigma})$ projects onto the ℓ-eigenspace of $\bar{P}_m\widehat{\Sigma}\bar{P}_m$. Lemmas A.5 and A.1*(iii)* give

$$
\mathbb{P}^n\left\{\text{(C1)} \leq 3t\right\} \geq 1 - 4\delta. \qquad (23)
$$

Continuing, we have

$$
\tilde{\lambda}_{\ell+1} + t \leq |\tilde{\lambda}_{\ell+1} - \widehat{\lambda}_{\ell+1}| + \widehat{\lambda}_{\ell+1} + t \overset{(\dagger)}{\leq} \frac{1}{n(n-1)}\left\|\left(\tilde{\mathbf{K}} - \mathbf{K}\right)\mathbf{H}_n\right\|_{\mathcal{L}^\infty(\mathbb{R}^n)} + \widehat{\lambda}_{\ell+1} + t, \qquad (24)
$$

where (†) uses the Hoffman-Wielandt inequality (Bhatia and Elsner, 1994) because $\tilde{\lambda}_{\ell+1}$ (*resp.* $\widehat{\lambda}_{\ell+1}$) is an eigenvalue of $\tilde{\mathbf{K}}\mathbf{H}_n$ (*resp.* $\mathbf{K}\mathbf{H}_n$). Letting P_m to be the orthogonal projector onto $\text{span}\{k(\cdot, X_{r_j})|j \in [m]\}$, it is easy to verify that $P_m = \tilde{S}_m^*\mathbf{K}_{mm}^{-1}\tilde{S}_m$ (Rudi et al., 2015, Lemma 1). Using $S_n\tilde{S}_m^* = \mathbf{K}_{nm}$, which follows from Lemma A.7*(ii)*, we have the expression

$$
\tilde{\mathbf{K}} = \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn} = S_n\tilde{S}_m^*\mathbf{K}_{mm}^{-1}\tilde{S}_mS_n^* = S_nP_mS_n^*.
$$

Thus, we may write,

$$
\begin{aligned}
\left\|\left(\tilde{\mathbf{K}} - \mathbf{K}\right)\mathbf{H}_n\right\|_{\mathcal{L}^\infty(\mathbb{R}^n)} &\leq \left\|\tilde{\mathbf{K}} - \mathbf{K}\right\|_{\mathcal{L}^\infty(\mathbb{R}^n)}\|\mathbf{H}_n\|_{\mathcal{L}^\infty(\mathbb{R}^n)} = n\,\|S_n(I-P_m)S_n^*\|_{\mathcal{L}^\infty(\mathbb{R}^n)} \\
&= n\,\|(I-P_m)S_n^*S_n(I-P_m)\|_{\mathcal{L}^\infty(\mathcal{H})} \\
&= n^2\,\|(I-P_m)C_n(I-P_m)\|_{\mathcal{L}^\infty(\mathcal{H})}\,, \qquad (25)
\end{aligned}
$$

where we have used Lemma A.7*(iv)* in (25). Proceeding,

$$\begin{aligned}(25) &= n^2 \left\| C_n^{1/2}(I-P_m)^2 C_n^{1/2} \right\|_{\mathcal{L}^\infty(\mathcal{H})} \\ &\leq n^2 \left\| C_n^{1/2}(C+tI)^{-1/2} \right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \left\| (C+tI)^{1/2}(I-P_m) \right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \\ &\leq n^2 \left\| C_n^{1/2}(C_n+tI)^{-1/2} \right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \left\| (C_n+tI)^{1/2}(C+tI)^{-1/2} \right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \\ &\qquad \times \left\| (C+tI)^{1/2}(I-P_m) \right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \\ &\leq n^2 \left\| (C_n+tI)^{1/2}(C+tI)^{-1/2} \right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \left\| (C+tI)^{1/2}(I-P_m) \right\|^2_{\mathcal{L}^\infty(\mathcal{H})}. \end{aligned} \tag{26}$$

Applying Lemmas A.2 and A.4 to (26) and Lemma A.1*(iv)* in (24) gives

$$\mathbb{P}^n \left\{ \tilde{\lambda}_{\ell+1} + t \leq \frac{9n^2 t}{2n(n-1)} + \frac{3}{2}(\lambda_{\ell+1}+t) \right\} \geq 1 - 4\delta. \tag{27}$$

Combining (27) with (23) in (22) gives

$$\mathbb{P}^n \left\{ \left\| (I - P^\ell_{nys}(\widehat{\Sigma}))(\widehat{\Sigma}+tI)^{1/2} \right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \leq 27t + 3\lambda_{\ell+1} \right\} \geq 1 - 8\delta, \tag{28}$$

where we note that $\frac{1}{n-1} \leq \frac{2}{n}$ for $n \geq 2$. The result follows by combining (28), (20), and (19) in (18).

Lower Bound: Using (18) we have

$$R^\ell_{nys}(\widehat{\Sigma}) = \textcircled{C} + \textcircled{D} \geq \left\| \left(I - P^\ell_{nys}(\widehat{\Sigma}) \right) \Sigma^{1/2} \right\|^2_{\mathcal{L}^2(\mathcal{H})}.$$

Since we have shown in *(ii)* that

$$R^\ell(\Sigma) = \inf_{\{\psi_i\}_i \in Q} \left\| (I - P_{\psi,\ell})\, \Sigma^{1/2} \right\|^2_{\mathcal{L}^2(\mathcal{H})},$$

the lower bound follows immediately.

6.4 Proof of Corollary 5

(i) From Theorem 3*(i)* we have

$$R^\ell(\Sigma) = \sum_{i>\ell} \lambda_i \lesssim \sum_{i>\ell} i^{-\alpha} \lesssim \int_\ell^\infty x^{-\alpha} dx \lesssim \ell^{1-\alpha} = n^{-\theta(1-\frac{1}{\alpha})}.$$

Similarly,

$$R^\ell(\Sigma) = \sum_{i>\ell} \lambda_i \gtrsim \sum_{i>\ell} i^{-\alpha} \gtrsim \int_{\ell+1}^\infty x^{-\alpha} dx \gtrsim (\ell+1)^{1-\alpha} \gtrsim n^{-\theta(1-\frac{1}{\alpha})}.$$

(ii) Theorem 3(*ii*) and (Sriperumbudur and Sterge, 2022, Lemma B.9) yield

$$R^{\ell}(\widehat{\Sigma}) \lesssim_{\mathbb{P}^n} t^{-1/\alpha}(n^{-\theta}+t)+\frac{1}{n},$$

for $\frac{\log n}{n} \lesssim t \lesssim 1$ where we have used $\lambda_\ell \lesssim \ell^{-\alpha} = n^{-\theta}$. Now if $\theta < 1$

$$\inf\left\{t^{-1/\alpha}(n^{-\theta}+t)+\frac{1}{n} : \frac{\log n}{n} \lesssim t \lesssim 1\right\} \lesssim n^{-\theta(1-\frac{1}{\alpha})}+\frac{1}{n},$$

and for $\theta \geq 1$

$$\inf\left\{t^{-1/\alpha}(n^{-\theta}+t)+\frac{1}{n} : \frac{\log n}{n} \lesssim t \lesssim 1\right\} \lesssim \left(\frac{\log n}{n}\right)^{\frac{\alpha-1}{\alpha}}+\frac{1}{n},$$

yielding the result.

(*iii*) Theorem 3(*iii*) and (Sriperumbudur and Sterge, 2022, Lemma B.9) yield

$$R_{\widehat{\Sigma},\ell} \lesssim_{\mathbb{P}^n} t^{-1/\alpha}(n^{-\theta}+t)+\frac{1}{n},$$

with $\frac{\log n}{n} \lesssim t \lesssim 1$ and $m \gtrsim \left(\frac{1}{t} \vee \mathcal{N}_{C,\infty}(t)\right)\log\frac{1}{t}$. Since $\mathcal{N}_{C,\infty}(t) \lesssim \frac{1}{t}$, we have $m \gtrsim \frac{1}{t}\log\frac{1}{t}$, and the result follows as in (*ii*).

6.5 Proof of Theorem 9

Upper Bound: We have

$$\begin{aligned}
T^{\ell}_{nys}(\widehat{\Sigma}) &= \mathbb{E}\left\|\mathfrak{I}\bar{k}(\cdot,X)-\mathfrak{I}P^{\ell}_{nys}(\widehat{\Sigma})(k(\cdot,X)-\widehat{m}_{\mathbb{P}})\right\|^2_{L^2(\mathbb{P})} \\
&= \underbrace{\mathbb{E}\left\|\mathfrak{I}(I-P^{\ell}_{nys}(\widehat{\Sigma}))\bar{k}(\cdot,X)\right\|^2_{L^2(\mathbb{P})}}_{\text{(E)}} + \underbrace{\left\|\mathfrak{I}P^{\ell}_{nys}(\widehat{\Sigma})(m_{\mathbb{P}}-\widehat{m}_{\mathbb{P}})\right\|^2_{L^2(\mathbb{P})}}_{\text{(F)}} \\
&\quad -2\mathbb{E}\left\langle\mathfrak{I}(I-P^{\ell}_{nys}(\widehat{\Sigma}))\bar{k}(\cdot,X),\mathfrak{I}P^{\ell}_{nys}(\widehat{\Sigma})(m_{\mathbb{P}}-\widehat{m}_{\mathbb{P}})\right\rangle_{L^2(\mathbb{P})}. \qquad (29)
\end{aligned}$$

The third term of (29) is 0, because $\mathbb{E}[\bar{k}(\cdot,X)]=0$. Using $\Sigma = \mathfrak{I}^*\mathfrak{I}$ (Sriperumbudur and Sterge, 2022, Proposition C.2 (*iii*)) we write

$$\begin{aligned}
\text{(E)} &= \mathbb{E}\left\|\mathfrak{I}(I-P^{\ell}_{nys}(\widehat{\Sigma}))\bar{k}(\cdot,X)\right\|^2_{L^2(\mathbb{P})} = \left\langle\Sigma(I-P^{\ell}_{nys}(\widehat{\Sigma}))\bar{k}(\cdot,X),(I-P^{\ell}_{nys}(\widehat{\Sigma}))\bar{k}(\cdot,X)\right\rangle_{\mathcal{H}} \\
&\overset{(\dagger)}{=} \left\|\Sigma^{1/2}(I-P^{\ell}_{nys}(\widehat{\Sigma}))\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})}, \qquad (30)
\end{aligned}$$

where (†) follows from Lemma A.6. Now (30) is similar to (C) in the proof of Theorem 3(*iii*), and the proof will proceed similarly. Using a similar argument to that in (20), and the idempotency of $I-P^{\ell}_{nys}(\widehat{\Sigma})$, we have

$$\begin{aligned}
&\left\|\Sigma^{1/2}(I-P^{\ell}_{nys}(\widehat{\Sigma}))\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} \\
&\leq \mathcal{N}_{\Sigma}(t)\left\|\Sigma_t^{-1/2}\Sigma^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})}\left\|\Sigma_t^{1/2}(I-P^{\ell}_{nys}(\widehat{\Sigma}))\right\|^4_{\mathcal{L}^\infty(\mathcal{H})} \\
&\leq \mathcal{N}_{\Sigma}(t)\left\|\widehat{\Sigma}_t^{-1/2}\Sigma_t^{1/2}\right\|^4_{\mathcal{L}^\infty(\mathcal{H})}\left\|(I-P^{\ell}_{nys}(\widehat{\Sigma}))\widehat{\Sigma}_t^{1/2}\right\|^4_{\mathcal{L}^\infty(\mathcal{H})}. \qquad (31)
\end{aligned}$$

The last term in (31) is simply the square of the last term of (20); therefore, we simply apply the result from (28), yielding

$$\mathbb{P}^n \left\{ \left\| \widehat{\Sigma}_t^{1/2} (I - P_{nys}^{\ell}(\widehat{\Sigma})) \right\|_{\mathcal{L}^\infty(\mathcal{H})}^4 \leq 9(9t + \lambda_{\ell+1})^2 \right\} \geq 1 - 8\delta. \tag{32}$$

Continuing,

$$\begin{aligned} \text{(F)} &= \left\| \Sigma^{1/2} P_{nys}^{\ell}(\widehat{\Sigma})(m_{\mathbb{P}} - \widehat{m}_{\mathbb{P}}) \right\|_{\mathcal{H}}^2 \leq \|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})}^2 \left\| P_{nys}^{\ell}(\widehat{\Sigma}) \right\|_{\mathcal{L}^\infty(\mathcal{H})}^2 \|m_{\mathbb{P}} - \widehat{m}_{\mathbb{P}}\|_{\mathcal{H}}^2 \\ &\leq \kappa \|m_{\mathbb{P}} - \widehat{m}_{\mathbb{P}}\|_{\mathcal{H}}^2 \leq \frac{32\kappa^2 \log \frac{2}{\delta}}{n}, \end{aligned} \tag{33}$$

where last inequality holds with probability at least $1 - \delta$ from Lemma A.3. The result follows by applying Lemma A.1(ii) to the middle term in (31) and combining with (32) and (33).

Lower Bound: The proof of (Sriperumbudur and Sterge, 2022, Theorem 6(ii)) gives

$$\sum_{i>\ell} \lambda_i^2 = \inf_{\{\psi_i\}_i \in Q} \left\| \Sigma^{1/2} \left(I - P_{\psi,\ell} \right) \Sigma^{1/2} \right\|_{\mathcal{L}^2(\mathcal{H})}^2.$$

The result therefore follows by noticing that

$$T_{nys}^{\ell}(\widehat{\Sigma}) \geq \text{(E)} = \left\| \Sigma^{1/2} \left(I - P_{nys}^{\ell}(\widehat{\Sigma}) \right) \Sigma^{1/2} \right\|_{\mathcal{L}^2(\mathcal{H})}^2.$$

6.6 Proof of Corollary 11

(i), (ii), (iv) are provided in (Sriperumbudur and Sterge, 2022, Corollary 7).
(iii) The lower bound follows immediately from previous results. For the upper bound, Theorem 9 and the proof of Corollary 5 *(ii)* yield

$$T_{nys}^{\ell}(\widehat{\Sigma}) \lesssim_{\mathbb{P}^n} t^{-1/\alpha}(t + n^{-\theta})^2 + \frac{1}{n}, \tag{34}$$

for $\frac{\log n}{n} \lesssim t \lesssim 1$ and $m \gtrsim \left(\frac{1}{t} \vee \mathcal{N}_{C,\infty}(t) \right) \log \frac{1}{t}$. Now

$$\mathcal{N}_{C,\infty}(t) = \sup_{x \in \mathcal{X}} \left\langle k(\cdot, x), (C + tI)^{-1} k(\cdot, x) \right\rangle_{\mathcal{H}} \lesssim \frac{1}{t};$$

thus, $m \gtrsim \frac{1}{t} \log \frac{1}{t}$. Larger values of t correspond to smaller requirement on m; thus, to optimize the performance of NY-EKPCA we select the largest value of t such that the behavior of (34) matches that of $T^{\ell}(\widehat{\Sigma})$. Setting $t = n^{-\theta}$ when $\theta < \frac{\alpha}{2\alpha - 1}$ and $t = n^{-\frac{\alpha}{2\alpha-1}}$ when $\theta \geq \frac{\alpha}{2\alpha-1}$ yields the result.

References

A. Alaoui and M. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 775–783. Curran Associates, Inc., 2015.

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

F. Bach. Sharp analysis of low-rank kernel matrix approximations. In S. Shalev-Shwartz and I. Steinwart, editors, *Proc. of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 185–209. PMLR, 2013.

F. Bach and M. I. Jordan. Predictive low-rank decomposition for kernel methods. In L. D. Raedt and S. Wrobel, editors, *Proc. of the 22nd International Conference on Machine Learning*, pages 33–40, 2005.

R. Bhatia and L. Elsner. The Hoffman-Wielandt inequality in infinite dimensions. In *Proceedings of the Indian Academy of Sciences - Mathematical Sciences*, volume 104, pages 483–494, 1994.

G. Blanchard, O. Bousquet, and L. Zwald. Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2):259–294, 2007.

A. Caponnetto and E. De Vito. Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

P. C. Chang and J. L. Wu. A critical feature extraction by kernel PCA in stock trading model. *Soft Computing*, 19:1393–1408, 2015.

D. L. Cohn. *Measure Theory*. Birkhäuser, 2013.

N. N. de Moura and J. M. de Seixas. Novelty detection in passive sonar systems using a kernel approach. In *13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, pages 116–122, 2017.

J. Diestel and J. J. Uhl. *Vector Measures*. American Mathematical Society, Providence, 1977.

N. Dinculeanu. *Vector Integration and Stochastic Integration in Banach Spaces*. John-Wiley & Sons, Inc., 2000.

P. Drineas and M. W. Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2153–2175, December 2005.

S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264, 2001.

A. Gittens and M. Mahoney. Revisiting the Nyström method for improved large-scale machine learning. In S. Dasgupta and D. McAllester, editors, *Proc. of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 567–575. PMLR, 2013.

N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

F. Hallgren. Kernel PCA with the Nyström method. 2021. http://arxiv.org/abs/2109.05578v2.

H. Hoffmann. Kernel PCA for novelty detection. *Pattern Recognition*, 40:863–874, 2007.

A. M. Jade, B. Srikanth, V. K. Jayaraman, B. D. Kulkarni, J. P. Jog, and L. Priya. Feature extraction and denoising using kernel PCA. *Chemical Engineering Science*, 58(19):4441–4448, 2003.

R. Jin, T. Yang, M. Mahdavi, Y-F. Li, and Z-H. Zhou. Improved bounds for the Nyström method with application to kernel classification. *IEEE Transactions on Information Theory*, 59(10):6939–6949, 2013.

I. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, USA, 1986.

K. I. Kim, M. O. Franz, and B. Schölkopf. Iterative kernel principal component analysis for image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(9):1351–1366, 2005.

S. Kumar, M. Mohri, and A. Talwalkar. Ensemble Nyström method. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1060–1068. Curran Associates, Inc., 2009.

C. Lampert. Kernel methods in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 4(3):193–285, 2009.

Z. Li, U. Kruger, L. Xie, A. Almansoori, and H. Su. Adaptive KPCA modeling of nonlinear systems. *IEEE Transactions on Signal Processing*, 63(9):2364–2376, 2015.

D. Lopez-Paz, S. Sra, A. Smola, Z. Ghahramani, and B. Schölkopf. Randomized nonlinear component analysis. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1359–1367. PMLR, 2014.

S. Mika, B. Schölkopf, A. J. Smola, K-R. Müller, M. Scholz, and G. Rätsch. Kernel PCA and de-noising in feature spaces. In M. J. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 536–542. MIT Press, 1999.

M. Peter, J-L. Minoi, and I. H. M. Hipiny. 3d face recognition using kernel-based PCA approach. In R. Alfred, Y. Lim, A.A.A. Ibrahim, and P. Anthony, editors, *Computational Science and Technology*, pages 77–86, Singapore, 2019. Springer Singapore.

A. Phophalia and S. K. Mitra. 3d MR image denoising using rough set and kernel PCA method. *Magnetic Resonance Imaging*, 36:135–145, 2017.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc., 2008a.

A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *NIPS*, pages 1313–1320, 2008b.

M. Reed and B. Simon. *Methods of Modern Mathematical Physics: Functional Analysis I.* Academic Press, New York, 1980.

H.-J. Reinhardt. *Analysis of Approximation Methods for Differential and Integral Equations.* Springer-Verlag, New York, 1985.

A. Rudi and L. Rosasco. Generalization properties of learning with random features. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3215–3225. Curran Associates, Inc., 2017.

A. Rudi, G. D. Canas, and L. Rosasco. On the sample complexity of subspace learning. In *Advances in Neural Information Processing Systems 26*, pages 2067–2075, 2013.

A. Rudi, R. Camoriano, and L. Rosasco. Less is more: Nyström computational regularization. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1657–1665. Curran Associates, Inc., 2015.

R. T. Samuel and Y. Cao. Nonlinear process fault detection and identification using kernel PCA and kernel density estimation. *Systems Science & Control Engineering*, 4(1):165–174, 2016.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

J. Shawe-Taylor, C. Williams, N. Christianini, and J. Kandola. On the eigenspectrum of the Gram matrix and the generalisation error of kernel PCA. *EEE Transactions on Information Theory*, 51(7):2510–2522, 2005.

A. J. Smola and B. Schölkopf. Sparse greedy matrix approximation for machine learning. In *Proc. 17th International Conference on Machine Learning*, pages 911–918. Morgan Kaufmann, San Francisco, CA, 2000.

B. K. Sriperumbudur and N. Sterge. Approximate kernel PCA using random features: Computational vs. statistical trade-off. *Annals of Statistics*, 2022. To appear, http://arxiv.org/abs/1706.06296v4.

I. Steinwart and A. Christmann. *Support Vector Machines.* Springer, New York, 2008.

N. Sterge, B. Sriperumbudur, L. Rosasco, and A. Rudi. Gain with no pain: Efficiency of kernel-PCA by Nyström sampling. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3642–3652. PMLR, 26–28 Aug 2020.

A. R. Teixeira, A.M. Tomé, K. Stadlthanner, and E.W. Lang. KPCA denoising and the pre-image problem revisited. *Digital Signal Processing*, 18(4):568–580, 2008.

E. Ullah, P. Mianjy, T. V. Marinov, and R. Arora. Streaming kernel PCA with $\tilde{O}(\sqrt{n})$ random features. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7322–7332. Curran Associates, Inc., 2018.

C. K. I. Williams. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In M. Jordan, editor, *Learning in Graphical Models*, Cambridge, MA, 1998. MIT Press.

C. K. I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In V. Tresp T. K. Leen, T. G. Diettrich, editor, *Advances in Neural Information Processing Systems 13*, pages 682–688, Cambridge, MA, 2001. MIT Press.

T. Yang, Y. Li, M. Mahdavi, R. Jin, and Z-H. Zhou. Nyström method vs random Fourier features: A theoretical and empirical comparison. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 476–484. Curran Associates, Inc., 2012.

Y. Yang, M. Pilanci, and M. J. Wainwright. Randomized sketches for kernels: Fast and optimal non-parametric regression. *Annals of Statistics*, 45(3):991–1023, 2017.

Appendix A. Technical Results

The following are a collection of technical results that are needed to prove the main results of this paper.

Lemma A.1 (Sriperumbudur and Sterge, 2022, Lemma B.2) *Let H be a separable Hilbert space and X a separable topological space. Define*

$$\mathfrak{C} = \frac{1}{2}\int_{X\times X} (\nu(x)-\nu(y)) \otimes_H (\nu(x)-\nu(y))\, dP(x,y)$$

where $\nu : X \to H$ is Bochner measurable with $\sup_{x\in X}\|\nu(x)\| = \kappa$. For $\{X_i\}_{i=1}^r \overset{i.i.d.}{\sim} P$ with $r \geq 2$, define

$$\widehat{\mathfrak{C}} = \frac{1}{2r(r-1)}\sum_{i\neq j}^{r} (\nu(X_i)-\nu(X_j)) \otimes_H (\nu(X_i)-\nu(X_j)).$$

Then the following hold for any $0 \leq \delta \leq \frac{1}{2}$ and $\frac{140\kappa}{r}\log\frac{16\kappa r}{\delta} \leq t \leq \|\mathfrak{C}\|_{\mathcal{L}^\infty(H)}$:

(i) $P^r\left\{\left\|(\mathfrak{C}+tI)^{-1/2}(\widehat{\mathfrak{C}}-\mathfrak{C})(\mathfrak{C}+tI)^{-1/2}\right\|_{\mathcal{L}^\infty(H)} \leq \frac{1}{2}\right\} \geq 1-2\delta;$

(ii) $P^r\left\{\sqrt{\frac{2}{3}} \leq \left\|(\mathfrak{C}+tI)^{1/2}(\widehat{\mathfrak{C}}+tI)^{-1/2}\right\|_{\mathcal{L}^\infty(H)} \leq \sqrt{2}\right\} \geq 1-2\delta;$

(iii) $P^r\left\{\left\|(\mathfrak{C}+tI)^{-1/2}(\widehat{\mathfrak{C}}+tI)^{1/2}\right\|_{\mathcal{L}^\infty(H)} \leq \sqrt{\frac{3}{2}}\right\} \geq 1-2\delta;$

(iv) $P^r\left\{\lambda_k(\widehat{\mathfrak{C}}) + t \le \frac{3}{2}(\lambda_k(\mathfrak{C}) + t)\right\} \ge 1 - 2\delta$ *for all* $k \ge 1$*;*

(v) $P^r\left\{\lambda_k(\mathfrak{C}) + t \le 2(\lambda_k(\widehat{\mathfrak{C}}) + t)\right\} \ge 1 - 2\delta$ *for all* $k \ge 1$.

Lemma A.2 (Sterge et al., 2020, Lemma A.1) *Suppose Assumption 1 holds and*

$$\frac{9\kappa}{n}\log\frac{n}{\delta} \le t \le \|C\|_{\mathcal{L}^\infty(\mathcal{H})}$$

for any $0 < \delta < 1$*. Then*

$$\mathbb{P}^n\left\{\left\|(C_n + tI)^{1/2}(C + tI)^{-1/2}\right\|_{\mathcal{L}^\infty(\mathcal{H})} \le \sqrt{\frac{3}{2}}\right\} \ge 1 - \delta,$$

where $C = \int k(\cdot, x) \otimes_{\mathcal{H}} k(\cdot, x)\, d\mathbb{P}(x)$*,* $C_n = \frac{1}{n}\sum_{i=1}^n k(\cdot, X_i) \otimes_{\mathcal{H}} k(\cdot, X_i)$ *and* $(X_i)_{i\in[n]} \overset{i.i.d.}{\sim} \mathbb{P}$.

Lemma A.3 (Sriperumbudur and Sterge, 2022, Lemma B.5*(i)*) *Suppose Assumption 1 holds and* $n \ge 2\log\frac{2}{\delta}$ *for any* $0 < \delta < 1$*. Then*

$$\mathbb{P}^n\left\{\|m_{\mathbb{P}} - \widehat{m}_{\mathbb{P}}\|_{\mathcal{H}}^2 \le \frac{32\kappa\log\frac{2}{\delta}}{n}\right\} \ge 1 - \delta,$$

where $m_{\mathbb{P}} = \int k(\cdot, x)\, d\mathbb{P}(x)$ *and* $\widehat{m}_{\mathbb{P}} = \frac{1}{n}\sum_{i=1}^n k(\cdot, X_i)$.

Lemma A.4 (Rudi et al. 2015, Lemma 6) *Suppose Assumption 1 holds, and for some* $m < n$*, the set of indices* $\{i_1, \cdots, i_m\}$ *is drawn uniformly without replacement from* $[n]$*. For some* $t > 0$*, define* $\mathcal{N}_{C,\infty}(t) = \sup_{x\in\mathcal{X}}\left\langle k(\cdot, x), (C + tI)^{-1}k(\cdot, x)\right\rangle_{\mathcal{H}}$*, where* $C = \int_{\mathcal{X}} k(\cdot, x) \otimes_{\mathcal{H}} k(\cdot, x)\, d\mathbb{P}(x)$ *is the uncentered covariance operator. Then, for any* $\delta > 0$ *and* $m \ge (67 \vee 5\mathcal{N}_{C,\infty}(t))\log\frac{4\kappa}{t\delta}$*, we have*

$$\mathbb{P}^n\left\{\left\|(I - P_m)(C + tI)^{1/2}\right\|_{\mathcal{L}^\infty(\mathcal{H})}^2 \le 3t\right\} \ge 1 - \delta,$$

where P_m *is the orthogonal projector onto* $\operatorname{span}\{k(\cdot, X_{i_j}) | j \in [m]\}$.

The following is an adaption of Rudi et al. (2015, Lemma 6) for U-statistics.

Lemma A.5 *Suppose Assumption 1 holds, and for some* $m < n$*, the set of indices* $\{i_j\}_{j=1}^m$ *is drawn uniformly from the set of all partitions of size* m *of* $\{1, 2, ..., n\}$*. Then, for* $0 \le \delta \le \frac{1}{2}$*,* $0 < t \le \|\Sigma\|_{\mathcal{L}^\infty(H)}$ *and* $m \ge \frac{140\kappa}{t}\log\frac{8}{t\delta}$*, we have*

$$\mathbb{P}^n\left\{\left\|(I - \bar{P}_m)(\Sigma + tI)^{1/2}\right\|_{\mathcal{L}^\infty(\mathcal{H})}^2 \le 2t\right\} \ge 1 - 2\delta,$$

where $\bar{P}_m$ *is the orthogonal projector onto* $\bar{\mathcal{H}}_m$ *as defined in Section 3.2.*

Proof Define

$$\widehat{\Sigma}_m := \frac{1}{2m(m-1)} \sum_{j \neq l}^{m} (k(\cdot, X_{i_j}) - k(\cdot, X_{i_l})) \otimes_{\mathcal{H}} (k(\cdot, X_{i_j}) - k(\cdot, X_{i_l})),$$

which means

$$\widehat{\Sigma}_m = \frac{1}{2m(m-1)} \tilde{S}_m^* \mathbf{H}_m \tilde{S}_m = \frac{1}{2(m-1)} \tilde{S}_m^* \mathbf{C}_m^2 \tilde{S}_m = Z^* Z,$$

where $Z^* = \frac{1}{\sqrt{2(m-1)}} \tilde{S}_m^* \mathbf{C}_m$. Note that Z^* has range $\bar{\mathcal{H}}_m$, and so $\mathcal{R}(\bar{P}_m) = \mathcal{R}(Z^*)$. Therefore, by Proposition 3 of Rudi et al. (2015), we have

$$\left\| (I - \bar{P}_m)(\Sigma + tI)^{1/2} \right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \leq t \left\| (\widehat{\Sigma}_m + tI)^{-1/2} (\Sigma + t)^{1/2} \right\|^2_{\mathcal{L}^\infty(\mathcal{H})}.$$

The proof is completed by applying Lemma A.1. ■

Lemma A.6 *For any orthogonal projector $P : \mathcal{H} \to \mathcal{H}$, the following holds:*

$$\mathbb{E}_{X \sim \mathbb{P}} \left\| (I - P)(k(\cdot, X) - m_{\mathbb{P}}) \right\|^2_{\mathcal{H}} = \left\| (I - P)\Sigma^{1/2} \right\|^2_{\mathcal{L}^2(\mathcal{H})}.$$

Proof By denoting $\bar{k}(\cdot, X) := k(\cdot, X) - m_{\mathbb{P}}$, we have

$$\begin{aligned} \mathbb{E} \left\| (I - P)\bar{k}(\cdot, X) \right\|^2_{\mathcal{H}} &= \mathbb{E} \left\langle (I - P)\bar{k}(\cdot, X), (I - P)\bar{k}(\cdot, X) \right\rangle_{\mathcal{H}} \\ &= \mathbb{E} \left\langle (I - P)\bar{k}(\cdot, X), \bar{k}(\cdot, X) \right\rangle_{\mathcal{H}} = \mathbb{E} \left\langle I - P, \bar{k}(\cdot, X) \otimes_{\mathcal{H}} \bar{k}(\cdot, X) \right\rangle_{\mathcal{L}^2(\mathcal{H})}, \end{aligned}$$

where we used $(I - P)^2 = (I - P)$. Since k is bounded, and thus Bochner integrable, it follows that

$$\begin{aligned} \mathbb{E} \left\langle I - P, \bar{k}(\cdot, X) \otimes_{\mathcal{H}} \bar{k}(\cdot, X) \right\rangle_{\mathcal{L}^2(\mathcal{H})} &= \left\langle I - P, \mathbb{E}[\bar{k}(\cdot, X) \otimes_{\mathcal{H}} \bar{k}(\cdot, X)] \right\rangle_{\mathcal{L}^2(\mathcal{H})} = \langle I - P, \Sigma \rangle_{\mathcal{L}^2(\mathcal{H})} \\ &= \operatorname{tr}\left((I - P)\Sigma\right) = \operatorname{tr}\left(\Sigma^{1/2}(I - P)^2\Sigma^{1/2}\right). \end{aligned} \tag{35}$$

The proof is completed by using $\left\| (I - P)\Sigma^{1/2} \right\|^2_{\mathcal{L}^2(\mathcal{H})} = \operatorname{tr}\left(\Sigma^{1/2}(I - P)^2\Sigma^{1/2}\right)$ in (35). ■

The following is a collection of results regarding the sampling operator S_n and its approximate version, $\tilde{S}_m$, which are quoted from Sriperumbudur and Sterge (2022) and Rudi et al. (2015).

Lemma A.7 *The sampling operator,*

$$S_n : \mathcal{H} \to \mathbb{R}^n,\ f \mapsto (f(X_1), f(X_2), \ldots, f(X_n))^\top,$$

and approximate sampling operator,

$$\tilde{S}_m : \mathcal{H} \to \mathbb{R}^m,\ f \mapsto (f(X_{r_1}), f(X_{r_2}), \ldots, f(X_{r_m}))^\top,$$

have the following properties:

(i) $S_n^* : \mathbb{R}^n \to \mathcal{H},\ \alpha \mapsto \sum_{i=1}^n \alpha_i k(\cdot, X_i)$;

(ii) $\frac{1}{n(n-1)} S_n^* \mathbf{H}_n S_n = \widehat{\Sigma}$;

(iii) $S_n S_n^* = \mathbf{K}$ *and* $\tilde{S}_m \tilde{S}_m^* = \mathbf{K}_{mm}$;

(iv) $\frac{1}{n} S_n^* S_n = \frac{1}{n} \sum_{i=1}^n k(\cdot, X_i) \otimes_{\mathcal{H}} k(\cdot, X_i) = C_n$;

(v) $\mathbf{K}_{nm} = S_n \tilde{S}_m^*$.

Proof We refer the reader to Sriperumbudur and Sterge (2022, Proposition C.1) for the proofs of *(i)*, *(ii)*, and *(iii)*. The proofs of *(iv)* and *(v)* are provided in Rudi et al. (2015, Section B). ■