The Pennsylvania State University
The Graduate School

COMPUTATIONAL-STATISTICAL TRADEOFF IN APPOROXIMATE KERNEL PRINCIPAL COMPONENT ANALYSIS

A Dissertation in
Statistics
by
Nicholas J. Sterge

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2022

The dissertation of Nicholas J. Sterge was reviewed and approved by the following:

Bharath K. Sriperumbudur
Associate Professor of Statistics
Dissertation Advisor
Chair of Committee

Bing Li
Verne M. Willaman Professor of Statistics

John Harlim
Professor of Mathematics

Ethan X. Fang
Assistant Professor of Statistics

Ephraim Mont Hanks
Associate Professor of Statistics
Chair of Graduate Studies

Abstract

Kernel methods are powerful learning methodologies that provide a simple way to construct nonlinear algorithms from linear ones. Despite their popularity, they suffer from poor scalability in big data scenarios. Various approximation methods, including Nyström and random feature approximation have been proposed in the literature to alleviate the problem. Though these methods lead to significant computational saving, the statistical consistency of most of these approximate kernel methods are not well understood except for kernel ridge regression wherein these approximations have been shown to be not only computationally efficient but also statistically consistent with a minimax optimal rate of convergence. In this dissertation, we apply the Nyström and random features approximation to kernel principal component analysis (KPCA) and investigate the trade-off between computational and statistical behaviors of approximate KPCA. We show that, in each case, approximate KPCA is both computationally and statistically efficient compared to KPCA in terms of the error associated with reconstructing a kernel function based on its projection onto the corresponding eigenspaces.

With respect to previous works on KPCA, this work develops bounds which consider empirical recentering of the data, as opposed to previous works which make restrictive assumptions to avoid recentering of the data in the feature space. Thus, this work makes additional contributions to the existing literature by establishing tight bounds on the reconstruction error of empirical KPCA with recentering. The analysis hinges on Bernstein-type inequalities for the operator and Hilbert-Schmidt norms of a self-adjoint Hilbert-Schmidt operator-valued U-statistics, which is of independent interest.

Table of Contents

List of Figures

Index of Notation

Sets:

$\mathbb{N}$	set of positive integers
$\mathbb{Z}$	set of positive and negative integers, including 0
$\mathbb{R}$	set of real numbers
$\mathbb{C}$	set of complex numbers
$A \times B$	Cartesian product of A and B
$\overline{S}$	closure of a set S

Functions:

$\mathbb{1}_A$	indicator function on the set A
$a \wedge b$	minimum of a and b
$a \vee b$	maximum of a and b

Spaces:

$\mathcal{X}$	input space
$\mathcal{H}$	RKHS or general Hilbert space
H	general Hilbert space
$H \otimes H$	tensor product space
$f \otimes_H g$	tensor product of $f, g \in H$, i.e., $(f \otimes_H g)h = \langle g, h \rangle_H f$
$\mathcal{F}, \mathcal{G}$	arbitrary space of functions

$\mathcal{C}(\mathcal{X})$	space of continuous functions $f : \mathcal{X} \to \mathbb{R}$
$\mathcal{C}_b(\mathcal{X})$	space of bounded continuous functions $f : \mathcal{X} \to \mathbb{R}$
$\mathcal{C}_0(\mathcal{X})$	space of continuous functions $f : \mathcal{X} \to \mathbb{R}$ vanishing at infinity
$\mathcal{C}^r(\mathcal{X})$	space of $r-$differentiable functions $0 \leq r \leq \infty$
$L^p(\mathcal{X}, \mu)$	space of $p-$power ($p \geq 1$) $\mu-$integrable functions
$L^p(\mathcal{X})$	space of $p-$power ($p \geq 1$) Lebesgue integrable functions
$\ell_p(\mathcal{X}), \ell_p$	space of $p-$summable sequences
$\mathcal{L}^1(\mathcal{H}), \mathcal{L}^1(H)$	set of trace class operators
$\mathcal{L}^2(\mathcal{H}), \mathcal{L}^2(H)$	set of Hilbert-Schmidt operators
$\mathcal{L}^\infty(\mathcal{H}), \mathcal{L}^\infty(H)$	set of bounded operators

Norms:

$\|\cdot\|_2$	Euclidean norm
$\|\cdot\|_F$	Frobenius norm
$\|\cdot\|_p$	$p-$norm, $\|x\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$ for $x \in \mathbb{C}^d$
$\|\cdot\|_{L^p(\mathcal{X},\mu)}$	L^p norm
$\|\cdot\|_\infty$	sup $-$norm
$\|\cdot\|_\mathcal{H}, \|\cdot\|_H$	RKHS or general Hilbert space norm
$\langle\cdot,\cdot\rangle_2$	Euclidean inner product
$\langle\cdot,\cdot\rangle_S$	inner product w.r.t. an inner product space S
$\|\cdot\|_{\mathcal{L}^1(H)}$	trace norm, $\|A\|_{\mathcal{L}^1(H)} = \mathrm{tr}(\|A^*A\|^{1/2})$
$\|\cdot\|_{\mathcal{L}^\infty(H)}$	operator norm, $\|A\|_{\mathcal{L}^\infty(H)} = \sup_{f \in H} \frac{\|Af\|_H}{\|f\|_H}$
$\|\cdot\|_{\mathcal{L}^2(H)}$	Hilbert-Schmidt norm, $\|A\|_{\mathcal{L}^2(H)} = \mathrm{tr}(A^*A)$

Measures and Random Variables:

$(\mathcal{X}, \mathcal{A})$	measurable space with $\sigma-$algebra $\mathcal{A}$

μ	arbitrary measure, possibly a signed measure
$\operatorname{supp}\mu$	support of μ (also defined for functions)
δ_x	Dirac measure at $x \in \mathcal{X}$
$\mathbb{P}, \mathbb{Q}$	probability measures
$\mathbb{P}_n, \mathbb{Q}_m$	empirical estimates of $\mathbb{P}, \mathbb{Q}$
$\mathbb{E}$	expectation operator
$\mathbb{E}_{\mathbb{P}}$	expectation w.r.t. measure $\mathbb{P}$
$M_+^b(\mathcal{X}), M_+^b(\Theta)$	set of all finite Borel measures on $\mathcal{X}, \Theta$

Mean Element, Covariance and Projection Operators:

$m_{\mathbb{P}}$	mean element, $\int_{\mathcal{X}} k(\cdot, x) d\mathbb{P}(x)$
$\widehat{m}_{\mathbb{P}}$	empirical mean element, $\frac{1}{n}\sum_{i=1}^n k(\cdot, X_i)$
C	population uncentered covariance operator, $\int_{\mathcal{X}} k(\cdot, x) \otimes_{\mathcal{H}} k(\cdot, x) d\mathbb{P}(x)$
$\hat{C}$	empirical uncentered covariance operator, $\frac{1}{n}\sum_{i=1}^n k(\cdot, X_i) \otimes_{\mathcal{H}} k(\cdot, X_i)$
Σ	population centered covariance operator, $\int_{\mathcal{X}} k(\cdot, x) \otimes_{\mathcal{H}} k(\cdot, x) d\mathbb{P}(x) - m_{\mathbb{P}} \otimes_{\mathcal{H}} m_{\mathbb{P}}$
$\hat{\Sigma}$	empirical centered covariance operator, $\frac{1}{2n(n-1)}\sum_{i\neq j}(k(\cdot, X_i) - k(\cdot, X_j)) \otimes_{\mathcal{H}} (k(\cdot, X_i) - k(\cdot, X_j))$
k_m	random features approximate reproducing kernel
$m_{\mathbb{P},m}$	approximate mean element, $\int_{\mathcal{X}} k_m(\cdot, x) d\mathbb{P}(x)$
$\hat{m}_{\mathbb{P},m}$	empirical approximate mean element, $\frac{1}{n}\sum_{i=1}^n k_m(\cdot, X_i)$
Σ_m	approximate centered covariance operator, $\int_{\mathcal{X}} k_m(\cdot, x) \otimes_{\mathcal{H}_m} k_m(\cdot, x)\, d\mathbb{P}(x) - m_{\mathbb{P},m} \otimes_{\mathcal{H}_m} m_{\mathbb{P},m}$
$\widehat{\Sigma}_m$	approximate empirical centered covariance operator, $\frac{1}{2n(n-1)}\sum_{i\neq j}^n (k_m(\cdot, X_i) - k_m(\cdot, X_j)) \otimes_{\mathcal{H}_m} (k_m(\cdot, X_i) - k_m(\cdot, X_j))$
$P^\ell(A)$	orthogonal projector onto the top$-\ell$ eigenvectors/functions of A
$\lambda_\ell(A)$	ℓ-th top eigenvalue of A

Vectors, Matrices, and Operators:

$\mathrm{tr}(A)$	trace of a matrix/operator A
$\mathrm{ran}(A), \mathcal{R}(A)$	range of matrix/operator A
A^*	adjoint of operator A
$\mathbf{I}_n$	$n \times n$ identity matrix
$\mathbf{1}_n$	$n-$dimensional vector of 1's, $(1, \ldots, 1)^\top \in \mathbb{R}^n$
$\mathbf{C}_n$	centering matrix, $\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$
$\mathbf{H}_n$	scaled centering matrix, i.e., $\mathbf{H}_n = n\mathbf{C}_n$
$A \preceq B$	$B - A$ is positive definite

Miscellaneous:

$\forall$	for every/for all
$\exists$	there exists
$[n]$	$\{1, 2, \ldots, n\}$
$a := b$	a is defined by b
$a \lesssim b$	$\exists$ positive constant $c > 0$ such that $a \leq cb$
$X \lesssim_{\mathbb{Q}} b$	for random variable X and constant b, for any $\delta > 0$, there exists constant $c_\delta > 0$ such that $\mathbb{Q}(X \leq c_\delta b) \geq \delta$
RKHS	reproducing kernel Hilbert space
k	reproducing kernel
$\overline{k}(\cdot, X)$	$k(\cdot, X) - m_{\mathbb{P}}$
$\tilde{k}(\cdot, X)$	$k(\cdot, X) - \widehat{m}_{\mathbb{P}}$
$i.i.d.$	independent and identically distributed
r.h.s./l.h.s.	right hand side/left hand side
w.r.t.	with respect to
$\mathfrak{I}$	inclusion operator, $L^2(\mathbb{P}) \to \mathcal{H}$

$\mathfrak{A}$	approximation operator, $L^2(\mathbb{P}) \to \mathcal{H}_m$
S	sampling operator, $\mathcal{H} \to \mathbb{R}^n$
$\tilde{Z}_m$	approximate sampling operator, $\mathcal{H} \to \mathbb{R}^m$
R-E	reconstruct and embed
E-R	embed and reconstruct

Acknowledgments

Graduate studies are certainly not a solo endeavor. I would like to thank several individuals who have enhanced my graduate school experience. I must begin with my advisor, Bharath Sriperumbudur, who introduced me to the world of kernel methods and to the project that would ultimately become this dissertation. This dissertation would not have been possible without his mentorship. I have a great deal of admiration for his love of knowledge for its own sake and drive to deeply understand modern statistics and machine learning. I must give great thanks to my committee members: Bing Li, Ethan Fang, and John Harlim. I was introduced to Bing Li through his wonderful, and at times humorous, probability theory lectures. His passion for theoretical statistics made him an excellent choice for my committee. I admire Ethan's unique and diverse expertise, spanning across statistics, machine learning, and optimization, which compelled me to ask him to be a part of my committee. I was happy to have John Harlim bring a fresh perspective from outside of the statistics department due to his curiosity regarding kernel methods and their applications. Outside of my committee several faculty members enriched my experience at Penn State. Alexei Novikov provided me with an excellent introduction to functional analysis, a subject which is the basis of many of the proofs in this dissertation. Donald Richards' passion for probability and gambling puzzles made long nights of grading papers far more enjoyable than they ever should be. I need to additionally thank the wonderful staff in the department of statistics for their help keeping my academic life organized and on track.

Lorenzo Rosasco and Alessandro Rudi must be acknowledged for their brilliant contributions to the literature, as well their role as collaborators on my first published work. From my undergraduate institution, Wake Forest University, I owe thanks to Professors Edward Allen, Stephen Robinson, Ellen Kirkman, Kenneth Berenhaut, and Robert Erhardt; thank you all for an excellent college experience. Outside of the faculty and staff, my graduate life was enhanced by the fabulous graduate students at Penn State. I was blessed with a wonderful cohort to go through coursework and qualifying exams. I would like to especially thank my housemates Ben and Likun, as well as Christian and Kyongwon, for their friendship and engaging discussions on statistics, as well as sports and pop culture.

Lastly, I owe immense thanks to my family: Andy, Marylou, Jesse, and Regina. Without your love and support there is no way I could have gotten to this point. You provided a phenomenal environment to grow up in and I am incredibly fortunate. I would

especially like to thank my father for introducing me to many fascinating applications of mathematics from a young age.

Chapter 3 is based on joint work with Bharath Sriperumbudur, Lorenzo Rosasco, and Alessandro Rudi, appearing in Sterge et al. (2020), a longer version of Sterge et al. (2020) is currently under submission to the Journal of Machine Learning Research. Chapter 4 is based on joint work with Bharath Sriperumbudur, appearing in Sriperumbudur and Sterge (2020), currently under revision at the Annals of Statistics.

Chapter 1

Introduction

Given training data $D = \{(X_i, Y_i)\}_{i=1}^n$ with $X_i \in \mathbb{R}^d$ and $Y_i \in \{-1, 1\}$, let us consider the problem of constructing a binary classifier, wherein we attempt to learn a function $f : \mathbb{R}^d \to \mathbb{R}$ such that $Y_i = \text{sign}(f(X_i))$, for all $i \in 1, \ldots, n$, or at least a large fraction of D. In the case of linearly separable data (see Figure 1.1) there exists a hyperplane $\mathbf{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$ which correctly classifies the data, i.e., $Y_i(\langle \mathbf{w}, X_i \rangle_2 + b) = 1 \ \forall i \in [n]$. There exist many methods to learn linear classifiers, e.g., linear discriminant analysis, support vector machines, perceptron, etc.

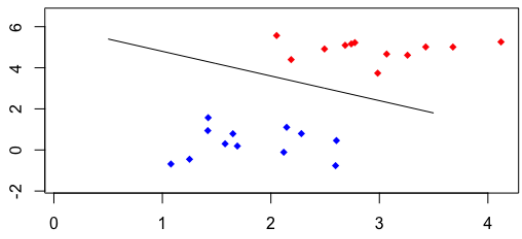

Figure 1.1. Linearly separable data in $\mathbb{R}^2$ with separating hyperplane.

In the vast majority of instances data is not linearly separable. However, there is often a way of mapping the input data to a higher dimensional space in which it is linearly separable. That is, given data in $\mathcal{X}$ for which no separating hyperplane exists, data may be mapped into a higher dimensional space, $\mathcal{H}$, and a linear classifier may be found in $\mathcal{H}$.

We refer to $\mathcal{X}$ as the *input space*, $\mathcal{H}$ as the *feature space*, and $\Phi : \mathcal{X} \to \mathcal{H}$ as the *feature map*. A simple instance of this procedure is demonstrated in Figure 1.2. Linear decision boundaries in the feature space $\mathcal{H}$ correspond to non-linear decision boundaries in the input space $\mathcal{X}$. This technique may be extended beyond binary classification, enabling the construction of nonlinear variants of regression, principal component analysis, and many other methods, and is commonly referred to in the literature as kernel methods (Schölkopf and Smola, 2002).

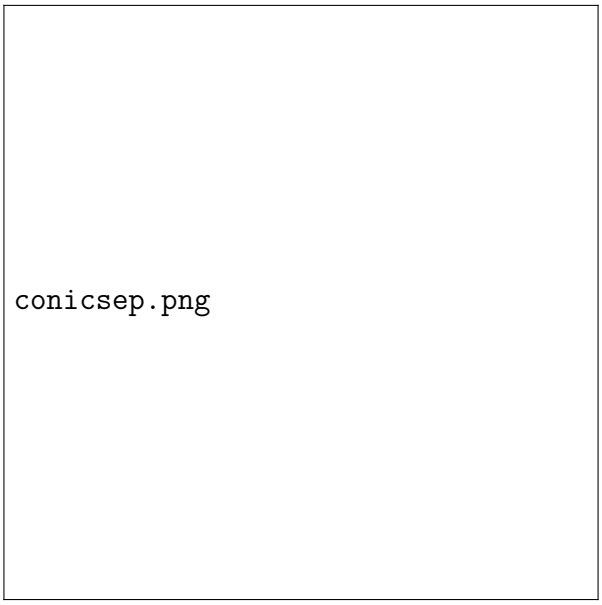

Figure 1.2. Data is not linearly separable in the input space, $\mathbb{R}^2$, though can be separated by an ellipse of the form, $ax_1^2 + bx_2^2 + cx_1x_2 + dx_1 + gx_2 + h = 0$. Thus, data are linearly separable when mapped to the feature space $\mathbb{R}^6$ through feature map $\Phi(x) = (x_1^2, x_2^2, x_1x_2, x_1, x_2, 1)^\top$.

1.1 Reproducing Kernel Hilbert Spaces

In kernel methods, input data are mapped to a canonical feature space, called as a reproducing kernel Hilbert space in which linear statistical methods are applied. This choice affords the user computational convenience and generates solutions that are non-linear in the input space. We present the following definitions:

Definition 1.1. *Let $\mathcal{H}$ be a vector space. A function $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is an inner product on $\mathcal{H}$ if it satisfies the following for all $\alpha, \beta \in \mathbb{R}$ and all $f, g, h \in \mathcal{H}$:*

1. $\langle \alpha f + \beta g, h \rangle_{\mathcal{H}} = \alpha \langle f, h \rangle_{\mathcal{H}} + \beta \langle g, h \rangle_{\mathcal{H}}$

2. $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$

3. $\langle f, f \rangle_{\mathcal{H}} \geq 0$ *and* $\langle f, f \rangle_{\mathcal{H}}$ *iff* $f = 0$.

The inner product induces the norm $\|f\|_{\mathcal{H}}^2 = \langle f, f \rangle_{\mathcal{H}}$. Of course, for $\mathcal{H} = \mathbb{R}^d$ the canonical inner product is given by $\langle \mathbf{a}, \mathbf{b} \rangle_2 = \mathbf{a}^\top \mathbf{b}$. A complete[1] vector space with an inner product is a *Hilbert space*.

Definition 1.2. *For a nonempty set $\mathcal{X}$, a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if there exists a Hilbert space $\mathcal{H}$ and a map $\Phi : \mathcal{X} \to \mathcal{H}$ such that*

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}, \quad \forall x, y \in \mathcal{X}.$$

In the preceding definition $\mathcal{X}$ is the input space, $\mathcal{H}$ is the feature space, and Φ is the feature map. Kernels satisfy the following properties, which may be proved using basic properties of inner products,

$$k(x, x) \geq 0 \quad \text{and} \quad |k(x, y)| \leq \sqrt{k(x, x)}\sqrt{k(y, y)} \quad \textit{(Cauchy-Schwarz).}$$

Let us now introduce the reproducing kernel and reproducing kernel Hilbert space, which are central to this work.

Definition 1.3. *Let $\mathcal{H}$ be a Hilbert space of $\mathbb{R}$-valued functions on non-empty set $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel of $\mathcal{H}$ and $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) if k satisfies the following:*

- $\forall x \in \mathcal{X}, \quad k(\cdot, x) \in \mathcal{H}$,
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \quad \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$.

From the definition it follows that

[1] A metric space X is complete if every Cauchy sequence in X converges to an element in X.

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}}, \quad \forall x, y \in \mathcal{X}.$$

Thus, from Definition 1.2 it follows that $\Phi(x) = k(\cdot, x)$ is a valid feature map from $\mathcal{X}$ to $\mathcal{H}$, and is referred to as the *canonical feature map*. Every Hilbert space with a reproducing kernel is an RKHS; conversely, every RKHS has a unique reproducing kernel (Aronszajn, 1950). An equivalent definition of reproducing kernel Hilbert space is a Hilbert space in which evaluation functionals are continuous for all points in the input space — and the equivalence follows from Riesz representation theorem (Reed and Simon, 1980); however, Definition 1.3 is more relevant in this work. For illustration, examples of kernel functions commonly seen in literature and application include:

- *Polynomial kernel:* $k(x, y) = (c + \langle x, y \rangle_2)^m$, for $x, y \in \mathbb{R}^d$, $c \geq 0$, and $m \in \mathbb{N}$. Setting $c = 0$ and $m = 1$ yields the linear kernel.
- *Gaussian kernel:* $k(x, y) = \exp\left(\frac{\|x-y\|_2^2}{\gamma^2}\right)$.

For a bivariate function k, it is not easy to verify that k is the reproducing kernel of some RKHS. Through the lens of positive definite functions, we can obtain a convenient characterization of reproducing kernels.

Definition 1.4. *A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if for all $n \geq 1$, $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$, and $(x_1, \ldots, x_n) \in \mathcal{X}^n$ the following holds:*

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0.$$

We say k is strictly positive definite if the strict inequality holds for all $(\alpha_1, \ldots, \alpha_n) \neq 0$.

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel if and only if it is symmetric and positive definite (Aronszajn, 1950). It is simple to verify that all kernels are symetric and positive definite; however, the proof that all symmetric and positive definite functions are kernels is non-trivial. All inner products are positive definite functions, and thus are associated with some RKHS. Functions in an RKHS will be linear combinations of the reproducing kernel evaluated at points in the input space, i.e.,

$$f = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i),$$

as well as limits of Cauchy sequences of functions of this type. Due to this, all functions in an RKHS $\mathcal{H}$ will inherit the properties of the reproducing kernel k, such as boundedness, measurability, differentiability, continuity, and integrability. We refer the reader to Schölkopf and Smola (2002) and Steinwart and Christmann (2008) for a detailed introduction.

1.2 Kernel Methods: Ridge Regression

The usefulness of kernel methods is easily illustrated through their application to regression. Suppose we are given training data $\{(X_i, Y_i)\}_{i=1}^n$ where $X_i \in \mathbb{R}^d$, $Y_i \in \mathbb{R}$. In ridge regression the goal is to find a linear function $f = \langle \mathbf{w}, \cdot \rangle_2$ such that $f(X_i)$ predicts Y_i well. That is, we solve

$$\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(\langle \mathbf{w}, X_i \rangle_2 - Y_i \right)^2 + \lambda \left\| \mathbf{w} \right\|_2^2, \tag{1.1}$$

where $\lambda > 0$ is a regularization parameter. Letting $\mathbf{X} = (X_1, \ldots, X_n) \in \mathbb{R}^{d \times n}$ and $\mathbf{Y} = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$, the solution to (1.1) is given by,

$$\mathbf{w} = \frac{1}{n} \left(\frac{1}{n} \mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}_d \right)^{-1} \mathbf{X}\mathbf{Y}. \tag{1.2}$$

Classical ridge regression restricts the regressor to be a linear function. If the linear methods do not suffice, then the user may map the input data to an RKHS and find linear regressors in $\mathcal{H}$ which are nonilinear functions of the data. Let $\mathcal{H}$ be an RKHS with reproducing kernel k, so the associated feature map is $\Phi(x) = k(\cdot, x) \in \mathcal{H}$. Kernel ridge regression solves the following problem:

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \left(\langle f, \Phi(X_i) \rangle_{\mathcal{H}} - Y_i \right)^2 + \lambda \left\| f \right\|_{\mathcal{H}}^2. \tag{1.3}$$

Letting $\Phi(\mathbf{X}) = (\Phi(X_1), \ldots, \Phi(X_n)) \in \mathbb{R}^{dim(\mathcal{H}) \times n}$, the solution to (1.3) is given by,

$$f = \frac{1}{n}\left(\frac{1}{n}\Phi(\mathbf{X})\Phi(\mathbf{X})^\top + \lambda \mathbf{I}_{dim(\mathcal{H})}\right)^{-1}\Phi(\mathbf{X})\mathbf{Y} \overset{(\dagger)}{=} \frac{1}{n}\Phi(\mathbf{X})\left(\frac{1}{n}\Phi(\mathbf{X})^\top\Phi(\mathbf{X}) + \lambda \mathbf{I}_n\right)^{-1}\mathbf{Y}, \tag{1.4}$$

where (†) follows from Woodbury's matrix identity. To make predictions on some $x_{test} \in \mathcal{X}$,

$$f(x_{test}) = \langle f, \Phi(x_{test})\rangle_{\mathcal{H}} = \frac{1}{n}\mathbf{Y}^\top\left(\frac{1}{n}\Phi(\mathbf{X})^\top\Phi(\mathbf{X}) + \lambda \mathbf{I}_n\right)^{-1}\Phi(\mathbf{X})^\top\Phi(x_{test}). \tag{1.5}$$

$\Phi(\mathbf{X})^\top\Phi(\mathbf{X})$ is often denoted $\mathbf{K}$ and is called the *kernel Gram matrix*. Further, the representer theorem (Kimeldorf and Wahba, 1970; Schölkopf and Smola, 2002) shows that the solution to (1.3) lies in $\text{span}\{(k(\cdot, X_i))_{i=1}^n\}$, and thus admits the representation

$$f = \sum_{i=1}^n \alpha_i k(\cdot, X_i) \quad \text{where} \quad \alpha = \left(\mathbf{K} + n\lambda \mathbf{I}_n\right)^{-1}\mathbf{Y}. \tag{1.6}$$

It is easy to verify, using the reproducing property, that (1.5) depends only on the kernel k, and does not require knowledge of the possibly infinite dimensional feature map (Schölkopf and Smola, 2002). This is often referred to as the 'kernel trick', as the solution depends on $\mathcal{H}$ only through inner products, i.e., kernel evaluations.

Though we may obtain a solution to (1.3) by solving a finite-dimensional system, the computational complexity scales as $O(n^3)$, as it requires the inversion of $\mathbf{K} + n\lambda\mathbf{I}_n$. Thus, kernel methods may be impractical in large sample scenarios due to their computational burden.

1.3 Kernel Principal Component Analysis

Let X be a zero-mean random variable with law $\mathbb{P}$ defined on $\mathcal{X}$. When $\mathcal{X} = \mathbb{R}^d$, classical PCA (Jolliffe, 1986) finds $\mathbf{a} \in \mathbb{R}^d$ such that $\text{Var}\left[\langle \mathbf{a}, X\rangle_2\right]$ is maximized, with the constraint $\|\mathbf{a}\|_2 = 1$. Defining $C := \mathbb{E}_{X\sim\mathbb{P}}[XX^\top]$, the solution is simply the unit eigenvector of C corresponding to its largest eigenvalue. More generally, it provides a low-dimensional representation of X that retains as much variance as possible, making it a popular statistical methodology for dimensionality reduction and feature extraction.

This low-dimensional representation of X is simply its projection onto the ℓ-eigenspace of C, i.e., the span of orthonormal eigenvectors associated with top ℓ eigenvalues of C where $\ell < d$. In practice, PCA is computed by replacing C with an empirical approximation $\hat{C} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top$ based on a centered random sample $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{P}$ where $\sum_{i=1}^{n} X_i = 0$.

The principal components outputted by PCA are linearly related to the original variables. In many cases, for example computer vision and facial recognition, it is more appropriate to obtain principal components which are non-linearly related to the original variables. By mapping input data to a reproducing kernel Hilbert space $\mathcal{H}$, Kernel PCA (KPCA) (Schölkopf et al., 1998) is an extension of classical PCA in which principal components are non-linear in the input space. For an RKHS $\mathcal{H}$ with reproducing kernel k defined on $\mathcal{X}$, KPCA finds $f \in \mathcal{H}$ with unit norm such that $\mathrm{Var}[f(X)] = \mathrm{Var}[\langle f, k(\cdot, X)\rangle_{\mathcal{H}}]$ is maximized. Assuming $\mathbb{E}[f(X)] = 0$ for all $f \in \mathcal{H}$, KPCA is solved in an analogous manner to classical PCA, by computing the eigenfunctions of covariance operator on $\mathcal{H}$ induced by $\mathbb{P}$, given by

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

As $\mathbb{P}$ is often unknown, C is replaced by the empirical approximation

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

where $(X_i)_{i=1}^{n} \overset{i.i.d.}{\sim} \mathbb{P}$, resulting in empirical KPCA (EKPCA). As the eigenfunctions of $\hat{C}$, denoted $(\phi_i(\hat{C}))_i$, live in a possibly infinite dimensional Hilbert space, it is not obvious how to compute them. However, the eigensystem $(\lambda_i(\hat{C}), \phi_i(\hat{C}))_{i=1}^{n}$ of $\hat{C}$ can be obtained by solving an n-dimensional eigenproblem involving the Gram matrix $\mathbf{K} = [k(X_i, X_j)]_{i,j\in[n]}$ (Schölkopf et al., 1998). In particular, the eigenvalues of $\mathbf{K}$ are related to those of $\hat{C}$ as $\lambda_i(\mathbf{K}) = n\lambda_i(\hat{C})$. Moreover, if $\mathbf{u}_i$ is an orthonormal eigenvector of $\mathbf{K}$ corresponding to

the eigenvalue $\lambda_i(\mathbf{K})$, then, under mild conditions on the kernel k, it holds for all $x \in \mathcal{X}$,

$$\phi_i(x) = \frac{1}{\sqrt{n\lambda_i(\hat{C})}} \sum_{j=1}^{n} k(x, x_j) u_{i,j}, \tag{1.7}$$

where the eigenvalues are assumed to satisfy a decreasing rearrangement. The above result proven in (Schölkopf et al., 2001) can be seen as a version of representer theorem for KPCA. Theoretical studies have established the consistency of EKPCA in terms of both the reconstruction error (Shawe-Taylor et al., 2005) and the excess error (Blanchard et al., 2007); the latter considers EKPCA both with and without empirical recentering.

Of course, as EKPCA requires diagonalizing the $n \times n$ Gram matrix $\mathbf{K}$, it suffers from the same computational burden as ridge regression. Various approximation schemes have been developed to alleviate these computational issues. It is known that these approximation methods offer significant improvements in computational complexity; however, whether these computational advantages come at the cost of statistical accuracy is in question. Early work has focused on the approximation quality of the gram matrix, $\mathbf{K}$, (Drineas et al., 2012; Drineas and Mahoney, 2005; Kumar et al., 2012) and the kernel k (Rahimi and Recht, 2008; Sriperumbudur and Szabó, 2015). However, the statistical consistency of approximation methods in downstream learning tasks has not received much attention until recently. Much of this attention has been directed toward kernel ridge regression, while unsupervised learning has not been well studied. This dissertation constructs approximate variations of kernel principal component analysis using Nyström method and random features approximation, and explores the compromise between computational complexity and statistical accuracy of said approximations.

1.4 Organization and Contributions

The central question this work addresses is whether kernel approximation methods may be utilized to develop a computationally efficient variant of kernel PCA, and what is the statistical performance of this approximate variant, especially when viewed in comparison

to empirical kernel PCA without approximation. In Chapter 2 we review methods for approximating kernels, detailing how specific approximation techniques may be used to construct learning algorithms. Specific attention is paid to the Nyström method and random features approximations. Section 3.1 describes how kernel PCA, without recentering, may be approximated using the Nyström method. A detailed theoretical analysis is given, including finite sample bounds and asymptotic convergence rates (Theorem 3.1, Corollaries 3.1 and 3.2). It is shown that Nyström method may be used to construct a computationally efficient approximation of kernel PCA which attains the best possible asymptotic behavior, that is, computational saving is achieved without loss in statistical accuracy. In Section 3.2 Nyström method is used to construct an approximate variation of kernel PCA with recentering, removing the restrictive assumption that the mean element of the RKHS is 0. To establish theoretical results a version of empirical KPCA using a U-statistic estimator is proposed and analyzed. Key in this analysis is a Bernstein type inequality for operator valued random variables in a separable Hilbert space (Theorem E.3), a result which is the first of its kind. After establishing the statistical behavior of empirical KPCA in the U-statistic case, Nyström approximate kernel PCA is analyzed, where it is shown that the approximate solution recovers the optimal convergence rate, while also being computationally beneficial (Theorem 3.2 and Corollary 3.3). Chapter 4 establishes approximate KPCA using random features. As the random features approximation computes principal components in $\mathbb{R}^m$, rather than $\mathcal{H}$, it is necessary to define a metric which captures the distance between the principal components in $\mathbb{R}^m$ and the true principal components in $\mathcal{H}$. To this end, two notions of reconstruction error for random features KPCA are proposed, *Reconstruct and Embed* (R-E) *Embed and Reconstruct* (E-R), where the crucial constructions are embedding operators into $L^2(\mathbb{P})$. Each of these notions of error are given thorough analysis, including establishing the statistical behavior of KPCA without approximation (Theorems 4.1 and 4.2). It is shown that random features may achieve the best possible convergence rate, while maintaining its computational edge (Corollaries 4.1, 4.2, 4.3, and 4.4). Additionally, in Section 4.4 we consider both a variation of E-R, obtained by projecting the random

variable onto a different system, and a generalization of R-E which is related to Schatten norms. Chapter 5 is dedicated to the comparison between Nyström and random features approximate PCA. Critical to this comparison, Theorem 5.1 establishes the consistency of Nyström KPCA in $L^2(\mathbb{P})$ under the R-E reconstruction error defined in Section 4. The Nyström approximate KPCA is shown to be superior to random features KPCA in terms of reconstruction error, while maintaining lesser computational complexity, similar to observations made in kernel ridge regression (Rudi and Rosasco, 2017). Chapter 6 highlights the contributions and directions for future work, including approximate two-sample testing and canonical correlation analysis. While proofs of the main results are contained in the body of the dissertation, necessary technical results are given in the appendix.

Chapter 2

Approximation Methods

The computational burden presented by kernel methods has been highlighted in Chapter 1 in the context of ridge regression and PCA; however, more broadly, these computational issues present themselves in all kernel learning tasks. This has lead to interest in approximation techniques aimed at alleviating these computational issues. Two popular approximation methods that are well studied in the literature are Nyström method and random features approximation along with few other methods.

2.1 Nyström Method

The Nyström method (Williams and Seeger, 2001) for approximating kernel matrices is inspired by numerical analysis methods for approximating solutions to eigenproblems. To overview (Kumar et al., 2012), suppose $\mathbf{K} \in \mathbb{R}^{n \times n}$ is a symmetric positive semidefinite Gram matrix with $\text{rank}(\mathbf{K}) = r \leq n$, with the singular value decomposition (SVD) of $\mathbf{K}$ being given by $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where $\mathbf{U}$ has orthogonal columns and $\mathbf{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_r\}$. The psuedo-inverse of $\mathbf{K}$ is $\mathbf{K}^+ = \sum_{i=1}^r \lambda_i^{-1} \mathbf{U}^{(i)} \mathbf{U}^{(i)^\top}$, where $\mathbf{U}^{(i)}$ is the i^{th} column of $\mathbf{U}$. Of course, if $\mathbf{K}$ is full rank, then $\mathbf{K}^+ = \mathbf{K}^{-1}$. The best rank-$q$ approximation to $\mathbf{K}$ is given by $\mathbf{K}_q = \sum_{i=1}^q \lambda_i \mathbf{U}^{(i)} \mathbf{U}^{(i)^\top}$, for some $q < r$, where the 'best' refers to distance in both spectral and Frobenius norms. Now, for a subsample of $m < n$ columns of $\mathbf{K}$ let $\mathbf{K}_{nm}$ be the $n \times m$ matrix consisting of these columns and $\mathbf{K}_{mm}$ the $m \times m$ matrix formed

by the intersection of these subsampled columns with the corresponding m rows. The rank-q Nyström approximation to the Gram matrix $\mathbf{K}$ is defined as

$$\tilde{\mathbf{K}}_q = \mathbf{K}_{nm}\mathbf{K}^{+}_{mm,q}\mathbf{K}^{\top}_{nm} \approx \mathbf{K}, \tag{2.1}$$

where $\mathbf{K}_{mm,q}$ is the best rank-q approximation of $\mathbf{K}_{mm}$. In this work, we assume that $\mathbf{K}^{-1}_{mm}$ exists and consider $q = m$, thus our approximation takes the form

$$\tilde{\mathbf{K}}_q = \mathbf{K}_{nm}\mathbf{K}^{-1}_{mm}\mathbf{K}^{\top}_{nm} \approx \mathbf{K},$$

though considering a lower rank Nyström approximation can lead to more significant computational saving. The approximation $\tilde{\mathbf{K}}_q$ can be substituted for $\mathbf{K}$ in kernel algorithms leading to reduction in computational complexity.

Early experimental and theoretical analysis of the Nyström method has focused on its quality of approximating the Gram matrix, rather than its efficacy in downstream learning tasks. Drineas and Mahoney (2005); Kumar et al. (2012) show that $\left\|\mathbf{K} - \tilde{\mathbf{K}}_q\right\| \lesssim \|\mathbf{K} - \mathbf{K}_q\| + O(\frac{1}{m^{1/4}}) \cdot \sum_{i=1}^{n} k(X_i, X_i)^2$, in both the spectral and Frobenius norms. A significant experimental treatment of the approximation quality is given in Gittens and Mahoney (2013). Much attention has also been given to improving the approximation quality through better subsampling methods (Cohen et al., 2015; Drineas et al., 2012; Kumar et al., 2012). Alaoui and Mahoney (2015) consider sampling according to leverage score weights where the leverage score weight of a column $\mathbf{K}^{(i)}$ is given by $\frac{\left\|\mathbf{K}^{(i)}\right\|_2^2}{\|\mathbf{K}\|_F^2}$. Kumar et al. (2012) propose adaptive sampling algorithms in which the distribution from which the columns are chosen is updated at each iteration. The aforementioned sampling techniques have been shown, both theoretically and experimentally, to provide the same degree of accuracy as uniform sampling; however, they require less columns to be subsampled from $\mathbf{K}$, enhancing the computational saving. In terms of the efficacy of Nyström in downstream learning tasks, the overwhelming majority of research has focused on approximate kernel ridge regression and classification (Alaoui and Mahoney, 2015; Cortes et al., 2010; Jin et al., 2013; Rudi et al., 2015; Yang et al., 2012). More

specifically, Rudi et al. (2015) study kernel ridge regression, i.e.,

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 . \tag{2.2}$$

It is easy to verify that the solution to (2.2) lies in the span of the $\mathcal{H}$-embedded training data, that is, $\{f \in \mathcal{H} | f = \sum_{i=1}^{n} \alpha_i k(\cdot, X_i), \alpha_1, \ldots, \alpha_n \in \mathbb{R}\}$ (Schölkopf and Smola, 2002). As shown in (1.4), computing the solution to (2.2) will scale as $O(n^3)$. To mediate this computational issue, Rudi et al. (2015) propose Nyström approximate kernel ridge regression, where the minimization of (2.2) is restricted to the set $\{f \in \mathcal{H} | f = \sum_{i=1}^{m} \alpha_{i_j} k(\cdot, X_{i_j}), \alpha \in \mathbb{R}^m\}$, where $m < n$ and $\{X_{i_1}, \ldots, X_{i_j}\}$ is a subsample of the training data. The solution to Nyström approximate kernel ridge regression is given by

$$\hat{f}_{\lambda,m} = \sum_{j=1}^{m} \tilde{\alpha}_{i_j} k(\cdot, X_{i_j}), \quad \text{where} \quad \tilde{\alpha} = \left(\mathbf{K}_{nm}^\top \mathbf{K}_{nm} + n\lambda \mathbf{K}_{mm} \right)^+ \mathbf{K}_{nm}^\top \mathbf{Y} \tag{2.3}$$

This approximation of kernel ridge regression incurs a computational complexity of $O(m^3 + m^2 n)$ and is shown (Rudi et al., 2015, Theorem 1) to be minimax optimal for $m \gtrsim n^\alpha \log n$ where $\alpha < 1$. Therefore, Nyström approximate kernel ridge regression achieves the best possible statistical behavior, while incurring less computational cost than standard kernel ridge regression.

In the unsupervised setting, kernel k-NN clustering has been studied (Wang et al., 2019), where relative error bounds are given; however, theoretical results in KPCA have not been previously established.

2.2 Random Features

An alternative to matrix approximation methods, the computational issue may be addressed by approximating the feature map Φ by a finite-dimensional map Φ_m, i.e., $\Phi_m(x) \in \mathbb{R}^m$. Clearly, this is equivalent to approximating the kernel, and thus approximating the RKHS $\mathcal{H}$ by an m-dimensional RKHS. An elegant approach to approximating the feature map is the random feature approximation introduced by Rahimi and Recht

(2008), which involves computing a finite dimensional feature map that approximates the kernel function. Suppose say k is a continuous translation invariant kernel on $\mathbb{R}^d$, i.e., $k(x,y) = \psi(x-y)$, $x, y \in \mathbb{R}^d$ where ψ is a continuous positive definite function on $\mathbb{R}^d$. Bochner's theorem (Wendland, 2005, Theorem 6.6) states that ψ is the Fourier transform of a finite non-negative Borel measure Λ on $\mathbb{R}^d$, i.e.,

$$k(x,y) = \int_{\mathbb{R}^d} e^{-\sqrt{-1}\langle x-y,\omega\rangle_2}\, d\Lambda(\omega) \overset{(\star)}{=} \int_{\mathbb{R}^d} \cos(\langle x-y,\omega\rangle_2)\, d\Lambda(\omega), \tag{2.4}$$

where $\langle\cdot,\cdot\rangle_2$ denotes the usual Euclidean inner product and $(\star)$ follows from the fact that ψ is real-valued and symmetric. Since $\Lambda(\mathbb{R}^d) = \psi(0)$, we can write (2.4) as $k(x,y) = \psi(0)\int_{\mathbb{R}^d}\cos(\langle x-y,\omega\rangle_2)\, d\frac{\Lambda}{\psi(0)}(\omega)$ where $\frac{\Lambda}{\psi(0)}$ is a probability measure on $\mathbb{R}^d$. Therefore, without loss of generality, we assume that Λ is a probability measure. Rahimi and Recht (2008) proposed a random approximation to k by replacing the integral with Monte Carlo sums constructed from $(\omega_i)_{i=1}^m \overset{i.i.d.}{\sim} \Lambda$, i.e., $k_m(x,y) = \psi_m(x-y) = \frac{1}{m}\sum_{i=1}^m \cos(\langle x-y,\omega_i\rangle_2) \overset{(\dagger)}{=} \langle\Phi_m(x),\Phi_m(y)\rangle_2$, where

$$\Phi_m(x) = \frac{1}{\sqrt{m}}(\cos\langle x,\omega_1\rangle_2, \ldots, \cos\langle x,\omega_m\rangle_2, \sin\langle x,\omega_1\rangle_2, \ldots, \sin\langle x,\omega_m\rangle_2)^\top$$

and $(\dagger)$ holds based on the trigonometric identity: $\cos(a-b) = \cos a\cos b + \sin a\sin b$. This kind of random approximation to k can be constructed for a more general class of kernels of the form

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

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

$$\Phi_m(x) = \frac{1}{\sqrt{m}}(\varphi(x,\theta_1), \ldots, \varphi(x,\theta_m))^\top,$$

$\varphi(x,\cdot) \in L^2(\Theta,\Lambda)$ for all $x \in \mathcal{X}$, $(\theta_i)_{i=1}^m \overset{i.i.d.}{\sim} \Lambda$, with $\mathcal{X}$ and Θ being measurable spaces.

To illustrate random features approximation in practice, consider a standard supervised learning problem in which i.i.d. data $\{y_i, X_i\}_{i=1}^n \subset \mathbb{R}\times\mathcal{X}$ is used to infer a function $\hat{f}: \mathcal{X} \to \mathbb{R}$ such that $\hat{f}(X_{test}) \approx y_{test}$. In kernel ridge regression, previously highlighted

in Section 1.2, this function has the form

$$\hat{f}_\lambda(x) = \sum_{i=1}^{n} \alpha_i k(x, X_i),$$

where k is a symmetric positive-definite kernel, $\lambda > 0$ a regularization parameter, and

$$\alpha = \frac{1}{n}\Phi(\mathbf{X}) \left(\frac{1}{n}\Phi(\mathbf{X})^\top \Phi(\mathbf{X}) + \lambda \mathbf{I}_n\right)^{-1} \mathbf{Y} = (\mathbf{K} + n\lambda \mathbf{I}_n)^{-1}\mathbf{y}, \tag{2.5}$$

is the solution (see (1.4)). Using the approximate feature map, $\Phi_m : \mathcal{X} \to \mathbb{R}^m$, in (2.5), an approximate variation to kernel ridge regression is obtained through,

$$\begin{aligned} \alpha &\approx \Phi_m(\mathbf{X}) \left(\Phi_m(\mathbf{X})^\top \Phi_m(\mathbf{X}) + n\lambda \mathbf{I}_n\right)^{-1} \mathbf{Y} \\ &\overset{(\clubsuit)}{=} \frac{1}{n} \left(\frac{1}{n}\Phi_m(\mathbf{X})\Phi_m(\mathbf{X})^\top + \lambda \mathbf{I}_m\right)^{-1} \Phi_m(\mathbf{X})\mathbf{Y}, \end{aligned} \tag{2.6}$$

where (♣) follows from Woodbury's matrix identity and

$$\Phi_m(\mathbf{X}) = (\Phi_m(X_1), \ldots, \Phi_m(X_n)) \in \mathbb{R}^{m \times n}.$$

Computation of (2.6) scales as $O(m^2 n)$, which, for $m \ll n$, is a significant improvement upon the $O(n^3)$ complexity of standard kernel ridge regression presented in (1.4). From a theoretical perspective, Rudi and Rosasco (2017) show that the excess risk of approximate kernel ridge regression with m random features, when compared to the Bayes' risk, achieves the minimax optimal rate, equivalent the case without approximation, provided $m \gtrsim n^\alpha$, where $\frac{1}{2} \leq \alpha < 1$. Thus, random features is computationally beneficial with no statistical loss.

2.3 Other Approximations

As the bottleneck stems from working with the $n \times n$ Gram matrix $\mathbf{K}$, several of low-rank matrix approximations have been applied in to kernels. Incomplete Cholesky factorization (Bach et al., 2005; Fine and Scheinberg, 2001) computes a sparse approximation to the

Cholesky factorization of $\mathbf{K}$. That is, $\mathbf{K} = \mathbf{L}\mathbf{L}^\top \approx \mathbf{Q}\mathbf{Q}^\top$, where $\mathbf{L}$ is lower triangular and $\mathbf{Q}$ is a sparse approximation of $\mathbf{L}$. This sparse lower triangular matrix can be used in kernel algorithms to speed up computations.

Yang et al. (2017) studies an approximation to kernel ridge regression based on random projections, or sketches, of the kernel matrix $\mathbf{K}$. Sketched kernel ridge regression limits the solution α (see (1.6)) to an $m-$dimensional subspace of $\mathbb{R}^n$, where $m < n$. This is accomplished by defining a so called *sketching matrix* $\mathbf{S} \in \mathbb{R}^{m\times n}$, such that the $m-$dimensional subspace is the row space of $\mathbf{S}$. The kernel ridge regression problem (1.3) can equivalently be solved via the quadratic program

$$\hat{\alpha} = \arg\min_{\alpha\in\mathbb{R}^n} \left\{ \frac{1}{2}\alpha\mathbf{K}^2\alpha - \frac{1}{\sqrt{n}}\alpha^\top\mathbf{K}\mathbf{Y} + \lambda\alpha\mathbf{K}\alpha \right\}.$$

From this, the sketched kernel ridge regression estimate (Yang et al., 2017) is computed by solving

$$\hat{\alpha} = \arg\min_{\alpha\in\mathbb{R}^m} \left\{ \frac{1}{2}\alpha^\top\mathbf{S}\mathbf{K}\mathbf{K}\mathbf{S}^\top\alpha - \frac{1}{n}\alpha^\top\mathbf{S}\mathbf{K}\mathbf{Y} + \lambda\alpha\mathbf{S}\mathbf{K}\mathbf{S}^\top\alpha \right\}, \tag{2.7}$$

resulting in the function

$$f = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{S}^\top\hat{\alpha})_i k(\cdot, X_i).$$

The sketched kernel ridge regression estimate may be computed with complexity $O(m^2n + m^3)$. Yang et al. (2017) considers both sub-Gaussian sketching matrices and randomized orthogonal system sketches such as discrete Fourier transform and random Hadamard matrices.

Differing from low-rank and projection based approximations, divide-and-conquer (Zhang et al., 2015) is a distributed algorithm which computes kernel estimators efficiently by partitioning the training data. Specifically, the training data $\{(X_i, Y_i)\}_{i=1}^n$ is partitioned randomly into m even subsets, the kernel ridge estimator (see (1.4)) is computed for each of the m subsets, and the divide-and-conquer estimate is computed via model averaging, i.e.,

$$\bar{f} = \frac{1}{m}\sum_{j=1}^{m}\hat{f}_j,$$

where $\hat{f}_j$ is the KRR estimate from the j^{th} subset of the partition. The computation and memory of divide-and-conquer kernel ridge regression scale as $O(n^3/m^2)$ and $O(n^2/m^2)$ respectively. Divide-and-conquer has also been applied to classification via support vector machines (Hsieh et al., 2014).

Chapter 3

Nyström Kernel PCA

We begin our study of approximate kernel PCA by analyzing the convergence behavior of kernel PCA both in the population (KPCA) and empirical (EKPCA) settings. We assume the following throughout the dissertation:

Assumption 3.1. *$(\mathcal{X}, \mathcal{B})$ is a second countable (i.e. completely separable) space endowed with $\sigma-$algebra $\mathcal{B}$. $(\mathcal{H}, k)$ a separable RKHS of real-valued functions on $\mathcal{X}$ endowed with bounded continuous positive definite kernel k satisfying $\sup_{x\in\mathcal{X}} k(x,x) = \kappa < \infty$.*

The second countability of $\mathcal{X}$ ensures that $\mathcal{B}$ is countably generated and therefore for any σ-finite measure μ defined on $\mathcal{B}$, $L^r(\mathcal{X}, \mu)$ is separable for any $r \in [1, \infty)$ (Cohn, 2013, Proposition 3.4.5). The assumption of $\mathcal{H}$ being separable and k being bounded continuous ensures that $k(\cdot, x) : \mathcal{X} \to \mathcal{H}$ is Bochner-measurable for all $x \in \mathcal{X}$ (Dinculeanu, 2000, Theorem 8 on p.5). The separability of $L^r(\mathcal{X}, \mu)$ and Bochner-measurability of $k(\cdot, x)$ will be crucial in our analysis.

3.1 Nyström Kernel PCA: Uncentered Covariance Operator

We recall the KPCA problem originally defined in Section 1.3. KPCA finds $f \in \mathcal{H}$ with unit norm such that $\mathrm{Var}[f(X)]$ is maximized. Since $\mathrm{Var}[f(X)] = \langle f, Cf\rangle_{\mathcal{H}}$ assuming

$\mathbb{E}[f(X)] = 0$ for all $f \in \mathcal{H}$, we have $f^* = \arg\sup\{\langle f, Cf\rangle_{\mathcal{H}} : \|f\|_{\mathcal{H}} = 1\}$ where C is the (uncentered) covariance operator on $\mathcal{H}$ defined as

$$C := \int_{\mathcal{X}} k(\cdot, x) \otimes_{\mathcal{H}} k(\cdot, x)\, d\mathbb{P}(x). \tag{3.1}$$

The boundedness of k in Assumption 3.1 ensures that C is trace class and thus compact. Since C is positive and self-adjoint, the spectral theorem (Reed and Simon, 1980) gives

$$C = \sum_{i \in I} \lambda_i \phi_i \otimes_{\mathcal{H}} \phi_i, \tag{3.2}$$

where $(\lambda_i)_{i\in I} \subset \mathbb{R}^+$ are the eigenvalues and $(\phi_i)_{i\in I}$ are the orthonormal system of eigenfunctions that span $\overline{\mathcal{R}(C)}$ with index set I either being finite or countable, in which case $\lambda_i \to 0$ as $i \to \infty$. The solution to the KPCA problem is thus the eigenfunction of C corresponding to its largest eigenvalue. We make the following simplifying assumption for ease of presentation.

Assumption 3.2. *The eigenvalues $(\lambda_i)_{i\in I}$ of C are simple, positive, and w.l.o.g. they satisfy a decreasing rearrangement, i.e., $\lambda_1 > \lambda_2, \ldots$*

Assumption 3.2 ensures that $(\phi_i)_{i\in I}$ form an orthonormal basis and the eigenspace corresponding to each λ_i is one-dimensional. This means the orthogonal projection operator onto the ℓ-eigenspace of C, i.e., $\text{span}\{(\phi_i)_{i=1}^{\ell}\}$, is given by

$$P^{\ell}(C) = \sum_{i=1}^{\ell} \phi_i \otimes_{\mathcal{H}} \phi_i. \tag{3.3}$$

The above construction corresponds to population version of KPCA when the data distribution $\mathbb{P}$ is known. If $\mathbb{P}$ is unknown and the knowledge of $\mathbb{P}$ is available only through the training set $(X_i)_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$, then KPCA cannot be carried out as C depends on $\mathbb{P}$. Therefore, an approximation to C is used to perform KPCA. Most commonly, this approximation is chosen to be the empirical estimator of C defined as

$$\hat{C} := \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i) \otimes_{\mathcal{H}} k(\cdot, X_i) \tag{3.4}$$

resulting in empirical kernel PCA (EKPCA). Note that $\hat{C}$ is a finite rank, positive, and self-adjoint operator. Thus the spectral theorem (Reed and Simon, 1980) yields

$$\hat{C} = \sum_{i=1}^{n} \hat{\lambda}_i \widehat{\phi}_i \otimes_{\mathcal{H}} \widehat{\phi}_i, \tag{3.5}$$

where $(\hat{\lambda}_i)_{i=1}^n \subset \mathbb{R}^+$ and $(\widehat{\phi}_i)_{i=1}^n \subset \mathcal{H}$ are the eigenvalues and eigenfunctions of $\hat{C}$. Similar to Assumption 3.2, we assume the following:

Assumption 3.3. *$\hat{C}$ has full rank, the eigenvalues $(\hat{\lambda}_i)_{i=1}^n$ of $\hat{C}$ are simple and w.l.o.g. they satisfy a decreasing rearrangement, i.e., $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots$.*

The eigensystem $(\hat{\lambda}_i, \widehat{\phi}_i)_{i=1}^n$ of $\hat{C}$ can be obtained by solving an n-dimensional system involving the eigendecomposition of the Gram matrix $\mathbf{K} = [k(X_i, X_j)]_{i,j\in[n]}$, which scales as $O(n^3)$ (Schölkopf et al., 1998). In particular, the eigenvalues of $\mathbf{K}$ are related to those of $\hat{C}$ as $\lambda_i(\mathbf{K}) = n\hat{\lambda}_i$. Moreover, if $\mathbf{u}_i$ is an orthonormal eigenvector of $\mathbf{K}$ corresponding to the eigenvalue $\lambda_i(\mathbf{K})$, then it holds for all $x \in \mathcal{X}$,

$$\phi_i(x) = \frac{1}{\sqrt{n\hat{\lambda}_i}} \sum_{j=1}^{n} k(x, x_j) \boldsymbol{u}_{i,j}. \tag{3.6}$$

The above result proven in (Schölkopf et al., 2001) can be seen as a representer theorem (Kimeldorf and Wahba, 1971) for KPCA. Finally, note that, for some $\ell \leq n$, the orthogonal projection operator onto $\text{span}\{(\widehat{\phi}_i)_{i=1}^{\ell}\}$ is given by

$$P^{\ell}(\hat{C}) = \sum_{i=1}^{\ell} \widehat{\phi}_i \otimes_{\mathcal{H}} \widehat{\phi}_i. \tag{3.7}$$

3.1.1 Approximate Kernel PCA using Nyström Method

The general idea in Nyström method is to obtain a low-rank approximation to the Gram matrix $\mathbf{K}$, and utilize this approximation in kernel algorithms, resulting in computational speedup. As the eigendecomposition of $\mathbf{K}$ is related to $\hat{C}$ (see (3.6)), Nyström method can be viewed as a low rank approximation to $\hat{C}$, which is what we exploit in obtaining a Nyström approximate KPCA. It follows from (3.6) that the eigenfunctions of $\hat{C}$ lie in

the space

$$\mathcal{H}_n = \left\{ f \in \mathcal{H} \,\middle|\, f = \sum_{i=1}^{n} \alpha_i k(\cdot, X_i), \alpha_1, ..., \alpha_n \in \mathbb{R} \right\}.$$

Therefore, it can be seen that EKPCA is a solution to the following problem

$$\arg\sup \left\{ \left\langle f, \hat{C} f \right\rangle_{\mathcal{H}} : f \in \mathcal{H}_n,\ \|f\|_{\mathcal{H}} = 1 \right\}.$$

Extending this representation, we propose Nyström KPCA (NY-KPCA) as a solution to the following problem:

$$\arg\sup \left\{ \left\langle f, \hat{C} f \right\rangle_{\mathcal{H}} : f \in \mathcal{H}_m,\ \|f\|_{\mathcal{H}} = 1 \right\}, \tag{3.8}$$

where

$$\mathcal{H}_m = \left\{ f \in \mathcal{H} \,\middle|\, f = \sum_{j=1}^{m} \alpha_j k(\cdot, X_{r_j}), \alpha_1, ..., \alpha_m \in \mathbb{R} \right\}$$

is a low-dimensional subspace of $\mathcal{H}_n$, where, for some $m < n$, indices $\{r_1, ..., r_m\}$ are sampled from $[n]$, yielding the subsample $\{X_{r_j}\}_{j=1}^{m}$. Basically, we are considering a plain Nyström approximation where the points $\{\tilde{X}_1, \ldots, \tilde{X}_m\}$ are sampled uniformly without replacement from $\{X_1, \ldots, X_n\}$; however, other subsampling methods are possible, see Section 3.1.2. The following result, proved in Section 3.3.1, shows that the solution to (3.8) is obtained by solving a finite dimensional linear system, which has better computational complexity than that of EKPCA.

Proposition 3.1. *Define the $m \times m$ matrix $\mathbf{M} = \mathbf{K}_{mm}^{-1/2} \mathbf{K}_{mn} \mathbf{K}_{nm} \mathbf{K}_{mm}^{-1/2}$. The solution to (3.8) is given by*

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

where $\mathbf{u}_{m,1}$ is the eigenvector of $\frac{1}{n}\mathbf{M}$ corresponding to its largest eigenvalue and $\tilde{Z}_m^* : \mathbb{R}^m \to \mathcal{H}$, $\alpha \mapsto \sum_{i=1}^{m} \alpha_i k(\cdot, \tilde{X}_i)$.

The cost of computing $\mathbf{M}$ is $O(nm^2 + m^3)$ and the cost of computing its eigendecomposition is $O(m^3)$. Thus, for $m < n$, the cost of NY-KPCA scales as $O(nm^2)$, faster than the $O(n^3)$ cost of EKPCA. It is easy to verify that $\mathbf{M}$ and $\tilde{\mathbf{K}}$ have same eigenvalues since $\mathbf{M} = \mathbf{K}_{mm}^{-1/2} \mathbf{K}_{mn} \left(\mathbf{K}_{mm}^{-1/2} \mathbf{K}_{mn} \right)^\top$ and $\tilde{\mathbf{K}} = \left(\mathbf{K}_{mm}^{-1/2} \mathbf{K}_{mn} \right)^\top \mathbf{K}_{mm}^{-1/2} \mathbf{K}_{mn}$,

and $\text{rank}(\mathbf{M}) = \text{rank}(\tilde{\mathbf{K}})$. Therefore we work with $\tilde{\mathbf{K}}$ and make the following assumption on its eigenvalues.

Assumption 3.4. $\text{rank}(\tilde{\mathbf{K}}) = m$. *The eigenvalues* $(\hat{\lambda}_{m,i})_{i=1}^m$ *of* $\frac{1}{n}\tilde{\mathbf{K}}$ *are simple and w.l.o.g. they satisfy a decreasing rearrangement, i.e.,* $\hat{\lambda}_{m,1} > \hat{\lambda}_{m,2} \ldots > \hat{\lambda}_{m,m}$.

The symmetry of $\mathbf{M}$ guarantees orthonormality of $(\mathbf{u}_{m,i})_i$, and the orthonormality of $(\hat{\phi}_{m,i})_i$ follows. For some $\ell \leq m$, the orthogonal projector onto $\text{span}\{\hat{\phi}_{m,i}\}_{i=1}^{\ell}$ is given by

$$P_m^{\ell}(\hat{C}) = \sum_{i=1}^{\ell} \hat{\phi}_{m,i} \otimes_{\mathcal{H}} \hat{\phi}_{m,i}. \tag{3.9}$$

One may ask if $\hat{\phi}_{m,i}$ are eigenfunctions of some operator on $\mathcal{H}$. Denote P_m as the orthogonal projector onto $\mathcal{H}_m$. It is simple to verify (Rudi et al., 2015, Theorem 2) that $P_m = \tilde{Z}_m^* \mathbf{K}_{mm}^{-1} \tilde{Z}_m$ and that $\left(\hat{\lambda}_{m,i}, \hat{\phi}_{m,i}\right)$ are the orthonormal eigenfunctions of $P_m \hat{C} P_m$, i.e.,

$$P_m \hat{C} P_m \hat{\phi}_{m,i} = \hat{\lambda}_{m,i} \hat{\phi}_{m,i} \text{ for all } i \in [m]. \tag{3.10}$$

Therefore, we may think of $P_m \hat{C} P_m$ as a low-rank approximation to $\hat{C}$.

3.1.2 Approximate Leverage Scores

In the above discussion on Nyström KPCA, $\{X_{r_j}\}_{j=1}^m$ is a subset of the training set sampled uniformly without replacement. As an alternative to uniform sampling, the subsamples $\{X_{r_j}\}_{j=1}^m$ can be generated according to the leverage score distribution (Alaoui and Mahoney, 2015; Cohen et al., 2015; Drineas et al., 2012). For any $s > 0$, the leverage scores associated with the training data $\mathbf{X}$ are defined as

$$(l_i(s))_{i=1}^n, \quad l_i(s) = [\mathbf{K}(\mathbf{K} + ns\mathbf{I}_n)^{-1}]_{ii}, i \in [n]$$

with the leverage score distribution being $p_i(s) = \frac{l_i(s)}{\sum_{i=1}^n l_i(s)}$ according to which we may sample independently with replacement to achieve a subsample of the original data. Since the leverage scores are computationally intensive to compute, they are often approximated in practice. To control the approximation error associated with approximating leverage

scores, we consider T-approximate leverage scores, guaranteeing a certain threshold of accuracy is met.

Definition 3.1. *(T-approximate leverage scores) For a given $s > 0$, let $(l_i(s))_{i=1}^n$ be the leverage scores associated with the training data $\{X_1, ..., X_n\}$. Let $\delta > 0$, $s_0 > 0$, and $T \geq 1$. $(\hat{l}_i(s))_{i=1}^n$ are T-approximate leverage scores, with confidence δ, if the following holds with probability at least $1 - \delta$:*

$$\frac{1}{T} l_i(s) \leq \hat{l}_i(s) \leq T l_i(s), \quad \forall i \in [n], \quad s > s_0.$$

Given T-approximate leverage scores for $s > s_0$, $\{X_{r_j}\}_{j=1}^m$ is obtained by sampling from original training data, $\mathbf{X}$, with replacement, according to the distribution $\hat{p}_i(s) = \hat{l}_i(s) / \sum_{i=1}^n \hat{l}_i(s)$. Having obtained $\{X_{r_j}\}_{j=1}^m$, (3.8) can be solved exactly as in Proposition 3.1. We refer to this method as approximate leverage score (ALS) Nyström subsampling.

3.1.3 Computational vs. Statistical Trade-Off: Main Results

Nyström kernel PCA approximates the solution to empirical kernel PCA with less computational expense. In this section, we explore whether this computational saving is obtained at the expense of statistical performance. We measure the statistical performance of KPCA, EKPCA, and NY-KPCA in terms of reconstruction error. In linear PCA, the reconstruction error, given by

$$\mathbb{E}_{X \sim \mathbb{P}} \left\| \left(I - P^\ell(C) \right) X \right\|_2^2, \tag{3.11}$$

is the error involved in reconstructing a random variable X by projecting it onto the ℓ-eigenspace (i.e., span of the top-ℓ eigenvectors) associated with its covariance matrix, $C = \mathbb{E}[XX^\top]$ through the orthogonal projection operator $P^\ell(C)$. Clearly, the error is zero when $\ell = d$. The analog of the reconstruction error in KPCA, as well as EKPCA and NY-KPCA, can be similarly stated in terms of their projection operators, (3.3), (3.7), and (3.9) as follows. For any orthogonal projection operator $P : \mathcal{H} \to \mathcal{H}$, define

the reconstruction error as

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

For the linear kernel this is exactly the reconstruction error of PCA. In the following, we often make use of the following identity, whose proof follows immediately.

Lemma 3.1.

$$R(P) = \|(I-P)C^{1/2}\|_{\mathcal{L}^2(\mathcal{H})}^2.$$

Proof. Note that

$$\begin{aligned} R(P) &= \mathbb{E}\, \|(I-P)k(\cdot,X)\|_{\mathcal{H}}^2 = \mathbb{E}\, \langle (I-P)k(\cdot,X), (I-P)k(\cdot,X)\rangle_{\mathcal{H}} \\ &= \mathbb{E}\, \langle (I-P)k(\cdot,X), k(\cdot,X)\rangle_{\mathcal{H}} \\ &= \mathbb{E}\, \langle (I-P), k(\cdot,X)\otimes_{\mathcal{H}} k(\cdot,X)\rangle_{\mathcal{L}^2(\mathcal{H})}\,, \end{aligned} \tag{3.12}$$

where we used $\langle Bf, g\rangle_{\mathcal{H}} = \langle B, f\otimes_{\mathcal{H}} g\rangle_{\mathcal{L}^2(\mathcal{H})}$ and $(I-P)^2 = (I-P)$ in (3.12). Since k is bounded, it follows that

$$\mathbb{E}\, \langle (I-P), k(\cdot,X)\otimes_{\mathcal{H}} k(\cdot,X)\rangle_{\mathcal{L}^2(\mathcal{H})} = \langle (I-P), \mathbb{E}[k(\cdot,X)\otimes_{\mathcal{H}} k(\cdot,X)]\rangle_{\mathcal{L}^2(\mathcal{H})}\,.$$

The result follows by using the above in (3.12) and noting that

$$\begin{aligned} \langle (I-P), C\rangle_{\mathcal{L}^2(\mathcal{H})} &= \operatorname{tr}\left((I-P)C\right) = \operatorname{tr}\left(C^{1/2}(I-P)(I-P)C^{1/2}\right) \\ &= \left\langle (I-P)C^{1/2}, (I-P)C^{1/2}\right\rangle_{\mathcal{L}^2(\mathcal{H})} = \left\|(I-P)C^{1/2}\right\|_{\mathcal{L}^2(\mathcal{H})}^2, \end{aligned}$$

where we have used the invariance of trace under cyclic permutations. □

Based on this definition, the reconstruction error in KPCA, EKPCA and NY-KPCA are given by

$$R_{C,\ell} := R(P^\ell(C)),\ R_{\hat{C},\ell} := R(P^\ell(\hat{C})),\ \text{and}\ R^{nys}_{\hat{C},\ell} := R(P^\ell_m(\hat{C})) \tag{3.13}$$

respectively. The following theorem, proved in Section 3.3.2, provides finite-sample bounds on the reconstruction error associated with NY-KPCA, under both uniform and approximate leverage score subsampling, from which convergence rates may be obtained.

Theorem 3.1. *Suppose Assumptions 3.1-3.4 hold. For any $t > 0$, define $\mathcal{N}_C(t) = \mathrm{tr}((C+tI)^{-1}C)$ and $\mathcal{N}_{C,\infty}(t) = \sup_{x\in\mathcal{X}} \langle k(\cdot,x), (C+tI)^{-1}k(\cdot,x)\rangle_{\mathcal{H}}$. Then the following hold:*

(i) Suppose $n > 3$, $0 < \delta < 1$, $\frac{9\kappa}{n}\log\frac{n}{\delta} \le t \le \lambda_1$, and $m \ge (67 \vee 5\mathcal{N}_{C,\infty}(t))\log\frac{4\kappa}{t\delta}$. Then, for plain Nyström subsampling:

$$\mathbb{P}^n\left\{(X_i)_{i=1}^n : R^{nys}_{\hat{C},\ell} \le \mathcal{N}_C(t)\,(6\lambda_\ell + 42t)\right\} \ge 1 - 2\delta. \tag{3.14}$$

(ii) For $0 < \delta < 1$, suppose there exists $T \ge 1$ such that $(\hat{l}_i(s))_{i=1}^n$ are $T-$approximate leverage scores with confidence δ for any $t \ge \frac{19\kappa}{n}\log\frac{2n}{\delta}$. Assume approximate leverage score Nyström subsampling is used with

$$t = \min\left\{\frac{19\kappa}{n}\log\frac{2n}{\delta} \le t \le \lambda_1 \,\middle|\, 78T^2\mathcal{N}_C(t)\log\frac{8n}{\delta} \le m\right\}.$$

If $n \ge 1655\kappa + 223\kappa\log\frac{2\kappa}{\delta}$ and $m \ge 334\log\frac{8n}{\delta}$, then

$$\mathbb{P}^n\left\{(X_i)_{i=1}^n : R^{nys}_{\hat{C},\ell} \le \mathcal{N}_C(t)\,(6\lambda_\ell + 42t)\right\} \ge 1 - 3\delta. \tag{3.15}$$

To understand the significance of Theorem 3.1, we have to compare it to the behavior of the reconstruction error associated with EKPCA, i.e., $R_{\hat{C},\ell}$. (Rudi et al., 2015, Theorem 3.1) showed that for $n > 3$, $0 < \delta < 1$ and $\frac{9\kappa}{n}\log\frac{n}{\delta} \le t \le \lambda_1$,

$$\mathbb{P}^n\left\{(X_i)_{i=1}^n : R_{\hat{C},\ell} \le 9\mathcal{N}_C(t)\,(\lambda_\ell + t)\right\} \ge 1 - \delta. \tag{3.16}$$

Comparing (3.14) and (3.15) to (3.16), it is clear that NY-KPCA has a statistical behavior similar to that EKPCA. However, it is not obvious whether such a behavior is achieved for $m < n$, i.e., the order of dependence of m on n is not clear. To clarify this, in

the following, we present two corollaries to Theorem 3.1, proved in Sections 3.3.3 and 3.3.4, which compare the asymptotic convergence rates of $R_{C,\ell}$, $R_{\hat{C},\ell}$ and $R^{nys}_{\hat{C},\ell}$ under an additional assumption on the decay rate of eigenvalues of C.

Corollary 3.1 (Polynomial decay of eigenvalues)**.** *Suppose* $\underline{A} i^{-\alpha} \leq \lambda_i \leq \bar{A} i^{-\alpha}$ *for* $\alpha > 1$ *and* $\underline{A}, \bar{A} \in (0, \infty)$*. Let* $\ell = n^{\frac{\theta}{\alpha}}$*,* $\theta > 0$*. Then the following hold:*

(i)

$$n^{-\theta(1-\frac{1}{\alpha})} \lesssim R_{C,\ell} \lesssim n^{-\theta(1-\frac{1}{\alpha})};$$

(ii)

$$R_{\hat{C},\ell} \lesssim_{\mathbb{P}^n} \begin{cases} n^{-\theta(1-\frac{1}{\alpha})}, & \theta < 1 \\ \left(\frac{\log n}{n}\right)^{1-\frac{1}{\alpha}}, & \theta \geq 1 \end{cases};$$

(iii) *For plain Nyström subsampling:*

$$R^{nys}_{\hat{C},\ell} \lesssim_{\mathbb{P}^n} \begin{cases} n^{-\theta(1-\frac{1}{\alpha})}, \ \theta < 1, \ m \gtrsim n^{\theta} \log n \\ \left(\frac{\log n}{n}\right)^{1-\frac{1}{\alpha}}, \ \theta \geq 1, \ m \gtrsim \frac{n}{\log n} \log \frac{n}{\log n} \end{cases};$$

(iv) *For approximate leverage score Nyström subsampling:*

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

Remark 3.1. (i) *The above result shows that the reconstruction errors associated with KPCA and EKPCA have similar asymptotic behavior as long as* ℓ *does not grow to infinity too fast, i.e.,* $\theta < 1$*. On the other hand, for* $\theta \geq 1$*, the reconstruction error of EKPCA has slower asymptotic convergence to zero than that of KPCA. If* ℓ *grows to infinity faster with the rate controlled by* θ*, then the variance term dominates the bias resulting in a slower convergence rate compared to that of KPCA.*

(ii) *Comparing* (ii) *and* (iii) *in the above result, we note that EKPCA and NY-KPCA*

have similar convergence behavior as long as m is large enough where the size of m is controlled by the growth of ℓ through θ. For the case of $\theta \geq 1$ in (iii), we require $m \gtrsim \frac{n}{\log n} \log \frac{n}{\log n}$ which means asymptotically m should be of the same order as n. On the other hand, the approximate leverage score Nyström subsampling gives same convergence rates as that of EKPCA but requiring far fewer samples than that for NY-KPCA with plain Nyström subsampling. These results show that for the interesting case of $\theta < 1$ where EKPCA performance matches with that of KPCA, NY-KPCA also achieves similar performance, albeit with lower computational requirement.

Corollary 3.2 (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 the following hold:*

(*i*)

$$n^{-\theta} \lesssim R_{C,\ell} \lesssim n^{-\theta};$$

(*ii*)

$$R_{\hat{C},\ell} \lesssim_{\mathbb{P}^n} \begin{cases} \frac{\log n}{n^\theta}, & \theta < 1 \\ \frac{(\log n)^2}{n}, & \theta \geq 1 \end{cases};$$

(*iii*) *For plain Nyström subsampling:*

$$R^{nys}_{\hat{C},\ell} \lesssim_{\mathbb{P}^n} \begin{cases} \frac{\log n}{n^\theta}, \ \theta < 1, \ m \gtrsim n^\theta \log n \\ \frac{(\log n)^2}{n}, \ \theta \geq 1, \ m \gtrsim \frac{n}{\log n} \log \frac{n}{\log n} \end{cases};$$

(*iv*) *For approximate leverage score Nyström subsampling:*

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

Corollary 3.2 shares similar behavior to that Corollary 3.1 as discussed in Remark 3.1 but just that it yields faster rates since the RKHS is smooth as determined by the rate of decay of eigenvalues. In addition, the approximate leverage score Nyström subsampling

based KPCA requires only $(\log n)^2$ subsamples to match the performance of EKPCA resulting in substantial computational savings without any loss in statistical accuracy.

3.2 Nyström Kernel PCA: Centered Covariance Operator

The results of Section 3.1.1 deal with the uncentered covariance operator, where the mean element is assumed to be zero, i.e., $\mathbb{E}_{X\sim\mathbb{P}}k(\cdot, X) = 0$. The overwhelming majority of theoretical results in KPCA make a similar assumption (Shawe-Taylor et al., 2005; Ullah et al., 2018). 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 and develop the analysis based on Bernstein-type inequality for operator valued U-statistics.

3.2.1 Empirical Kernel PCA with the Centered Covariance Operator

The uncentered method in Section 3.1.1 finds $f \in \mathcal{H}$ with unit norm such that $\mathbb{E}[f(X)^2]$ is maximized; effectively, it maximizes $\mathrm{Var}[f(X)]$ under the assumption $\mathbb{E}[f(X)] = 0$. Throughout the rest of this work, we will no longer make this assumption, instead, defining Kernel PCA (KPCA) as finding $f \in \mathcal{H}$ 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 $\mathrm{Var}[f(X)] = \mathbb{E}[\langle f, k(\cdot, X)\rangle_{\mathcal{H}} - \langle f, m_{\mathbb{P}}\rangle_{\mathcal{H}}]^2$ where $m_{\mathbb{P}} \in \mathcal{H}$ is the unique *mean element* of $\mathbb{P}$ in $\mathcal{H}$, defined

for all $f \in \mathcal{H}$ by

$$\langle f, m_{\mathbb{P}} \rangle_{\mathcal{H}} = \mathbb{E}[f(X)] = \mathbb{E}[\langle f, k(\cdot, X) \rangle_{\mathcal{H}}] = \left\langle f, \int_{\mathcal{X}} k(\cdot, x) d\mathbb{P}(x) \right\rangle_{\mathcal{H}}, \tag{3.17}$$

where the last equality of (3.17) holds via Riesz representation theorem (Reed and Simon, 1980) and the boundedness of k from Assumption 3.1 which ensures $k(\cdot, X)$ is Bochner integrable (Diestel and Uhl, 1977) with respect to $\mathbb{P}$. Therefore, we may write $\mathrm{Var}[f(X)] = \langle f, \Sigma f \rangle_{\mathcal{H}}$ where

$$\Sigma = \int_{\mathcal{X}} k(\cdot, x) \otimes_{\mathcal{H}} k(\cdot, x) d\mathbb{P}(x) - m_{\mathbb{P}} \otimes_{\mathcal{H}} m_{\mathbb{P}}, \tag{3.18}$$

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

$$\sup\{\langle f, \Sigma f \rangle_{\mathcal{H}} : f \in \mathcal{H},\ \|f\|_{\mathcal{H}} = 1\}, \tag{3.19}$$

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 3.1 ensures that Σ is trace class and thus compact. Since Σ is 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 (3.19) is simply the eigenfunction of Σ corresponding to its largest eigenvalue. We make the following simplifying assumption for ease of presentation.

Assumption 3.5. *The eigenvalues $(\lambda_i)_{i \in I}$ of Σ are simple, positive, and w.l.o.g. satisfy a decreasing rearrangement, i.e., $\lambda_1 > \lambda_2 > \ldots$*

Assumption 3.5 allows one to express the orthogonal projection operator onto the ℓ-

eigenspace of Σ, i.e. $\text{span}\{(\phi_i)_{i=1}^{\ell}\}$, as

$$P^{\ell}(\Sigma) = \sum_{i=1}^{\ell} \phi_i \otimes_{\mathcal{H}} \phi_i. \tag{3.20}$$

The above construction corresponds to population version when the data distribution $\mathbb{P}$ is known. In practice, the knowledge of $\mathbb{P}$ is available only through the sample $\{X_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} \mathbb{P}$. Therefore, performing KPCA in practice requires one to replace Σ in (3.18) with an estimate. We consider the following unbiased 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:

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

Using the reproducing property, it is easy to verify that

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

Therefore, substituting for Σ in (3.19) yields the objective of empirical KPCA (EKPCA):

$$\sup\{\langle f, \widehat{\Sigma} f \rangle_{\mathcal{H}} : f \in \mathcal{H}, \, \|f\|_{\mathcal{H}} = 1\}. \tag{3.21}$$

Of course $\widehat{\Sigma}$ is self-adjoint, positive and has rank at most $n-1$, thus is compact. Thus the spectral theorem (Reed and Simon, 1980) yields

$$\widehat{\Sigma} = \sum_{i=1}^{n-1} \widehat{\lambda}_i \widehat{\phi}_i \otimes_{\mathcal{H}} \widehat{\phi}_i, \tag{3.22}$$

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 $\widehat{\Sigma}$. Similar to Assumption 3.5, we make the following simplifying assumption regarding the spectrum of $\widehat{\Sigma}$.

Assumption 3.6. *The rank of $\widehat{\Sigma}$ is $n-1$, eigenvalues $(\widehat{\lambda}_i)_{i=1}^{n-1}$ of $\widehat{\Sigma}$ are simple, positive and w.l.o.g. satisfy a decreasing rearrangement, i.e., $\widehat{\lambda}_1 > \widehat{\lambda}_2 > \dots$*

For any $\ell \leq n-1$, since $\{\widehat{\phi}_i\}_{i=1}^{\ell}$ forms an orthogonal 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 $x \in X$. Moreover, following Assumption 3.6, the orthogonal projector onto $\operatorname{span}\{\widehat{\phi}_i : i = 1, \dots, \ell\}$ is given by

$$P^\ell(\widehat{\Sigma}) = \sum_{i=1}^{\ell} \widehat{\phi}_i \otimes_{\mathcal{H}} \widehat{\phi}_i. \tag{3.23}$$

Though $\widehat{\Sigma}$ is finite rank, its eigenfunctions are solution to a possibly infinite dimensional linear system. The following result, proved in Section 3.3.5, shows that the eigenvalues of $\widehat{\Sigma}$ can computed by solving an n-dimensional system.

Proposition 3.2. *Let $(\widehat{\lambda}_i, \widehat{\phi}_i)_i$ be the eigensystem of $\widehat{\Sigma}$ in (3.22). 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 $\widehat{\boldsymbol{\alpha}}_i \notin \operatorname{null}(\mathbf{H}_n)$, and $(\widehat{\lambda}_i, \widehat{\boldsymbol{\alpha}}_i)_i$ are the eigenvalues and eigenvectors of $\frac{1}{n(n-1)} \mathbf{K}\mathbf{H}_n$.

As computation of the eigensystem of $\widehat{\Sigma}$ is obtained by solving an $n \times n$ system, computation of $(\widehat{\lambda}_i, \widehat{\phi}_i)_{i=1}^{\ell}$ for $\ell < n$ has a space complexity $O(n^2)$ and a time complexity of $O(n^2 \ell)$ via the Lanczos method.

3.2.2 Approximate Kernel PCA using the Nyström Method: Centered Covariance Operator

Similar to the uncentered variant described in Section 3.1.1, we obtain an approximation of KPCA with respect to the centered covariance operator Σ by restricting the eigenfunctions to live in a lower dimensional space. To this end, it follows from Proposition 3.2 that the eigenfunctions of $\widehat{\Sigma}$ lie in the following space,

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

Thus, we could instead express the objective in (3.21) as an optimization over $\bar{\mathcal{H}}_n$, or equivalently, over $\boldsymbol{\alpha} \in \mathbb{R}^n$. From this, suppose for some $m < n$ indices $\{r_1, ..., r_m\}$ are sampled uniformly without replacement from $[n]$, yielding the subsample $\{X_{r_j}\}_{j=1}^{m}$ and the random subspace

$$\bar{\mathcal{H}}_m := \left\{ f \in \mathcal{H} \middle| f = \sum_{j=1}^{m} \left(m\alpha_j - \sum_{l=1}^{m} \alpha_l \right) k(\cdot, X_{r_j}) : \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m) \in \mathbb{R} \right\}$$

of $\mathcal{H}$. Nyström KPCA (NY-EKPCA) optimizes the EKPCA objective in (3.21) over $\bar{\mathcal{H}}_m$, that is, NY-EKPCA is the solution to the following problem:

$$\sup \left\{ \left\langle f, \widehat{\Sigma} f \right\rangle_{\mathcal{H}} : f \in \bar{\mathcal{H}}_m,\ \|f\|_{\mathcal{H}} = 1 \right\}. \tag{3.24}$$

The following result, proved in Section 3.3.6, shows that the solution to (3.24) is obtained by solving a finite dimensional linear system, which has better computational complexity than that of EKPCA, provided the subsample size is less than the sample size, $m < n$.

Proposition 3.3. *Define the $m \times m$ matrix $\overline{\mathbf{M}} = \mathbf{K}_{mm}^{-1/2} \mathbf{K}_{mn} \mathbf{H}_n \mathbf{K}_{nm} \mathbf{K}_{mm}^{-1/2}$. The solution to (3.24) is given by*

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

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

The complexity of computing $\overline{\mathbf{M}}$ and its eigendecomposition via the Lanczos method is $O(m^2\ell + nm^2)$; therefore, for $m < n$ the complexity of solving (3.24) scales as $O(nm^2)$, which is a reduction from the $O(n^2\ell)$ complexity of solving EKPCA if $m < \sqrt{n\ell}$. It is worth noting the connection between our Nyström approximation to KPCA and the traditional Nyström approximation to the Gram matrix of Williams and Seeger (2001), given by

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

If we observe

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

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

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

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

Assumption 3.7. *The rank of* $\tilde{\mathbf{K}}\mathbf{H}_n$ *is* m. *The eigenvalues* $(\tilde{\lambda}_i)_{i=1}^{m-1}$ *of* $\frac{1}{n(n-1)}\tilde{\mathbf{K}}\mathbf{H}_n$ *are simple, positive, and w.l.o.g. satisfy a decreasing rearrangement, i.e.,* $\tilde{\lambda}_1 > \tilde{\lambda}_2 > \ldots$.

As shown in the proof of Proposition 3.3, $(\tilde{\phi}_i)_i$ form an orthonormal system. Thus, for some $\ell < m$ the orthogonal projector onto $\operatorname{span}\{(\tilde{\phi}_i)_{i=1}^\ell\}$ is given by

$$P_{nys}^\ell(\widehat{\Sigma}) = \sum_{i=1}^\ell \tilde{\phi}_i \otimes_{\mathcal{H}} \tilde{\phi}_i. \tag{3.25}$$

One may ask if $\tilde{\phi}_1$ is the eigenfunction of some operator. Denoting $\bar{P}_m$ as the orthogonal projector onto $\bar{\mathcal{H}}_m$, we have $\bar{\mathcal{H}}_m = \operatorname{ran}(\tilde{Z}_m^*\mathbf{H}_m)$ and $\mathbf{H}_m\tilde{Z}_m\tilde{Z}_m^*\mathbf{H}_m = \mathbf{H}_m\mathbf{K}_{mm}\mathbf{H}_m$, which

leads to,

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

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

$$\bar{P}_m\widehat{\Sigma}\bar{P}_m\tilde{\phi}_i = \tilde{\lambda}_i\tilde{\phi}_i. \tag{3.26}$$

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

3.2.3 Computational vs. Statistical Trade-Off: $\mathcal{H}-$norm

In a similar fashion to the uncentered variant of Section 3.1.1, we will use reconstruction error to quantify the statistical behavior of NY-KPCA. In contrast to the uncentered case, we now observe

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

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, (3.20), (3.23), and (3.25) as follows:

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

$$R^{\ell}(\widehat{\Sigma}) = \mathbb{E}\left\|\overline{k}(\cdot,x) - P^{\ell}(\widehat{\Sigma})\widetilde{k}(\cdot,X)\right\|_{\mathcal{H}}^2, \tag{3.27}$$

$$R^{\ell}_{nys}(\widehat{\Sigma}) = \mathbb{E}\left\|\overline{k}(\cdot,x) - P^{\ell}_{nys}(\widehat{\Sigma})\widetilde{k}(\cdot,X)\right\|_{\mathcal{H}}^2. \tag{3.28}$$

Here for any $x \in \mathcal{X}$,

$$\overline{k}(\cdot,x) = k(\cdot,x) - \int_{\mathcal{X}} k(\cdot,x)\, d\mathbb{P}(x) \quad \text{and} \quad \widetilde{k}(\cdot,x) = k(\cdot,x) - \frac{1}{n}\sum_{i=1}^{n} k(\cdot,X_i).$$

Since empirical mean is used in empirical linear PCA to find principal components, we used $\widetilde{k}$ in the definition of reconstruction error for EKPCA. The following theorem, proved in Section 3.3.7, 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 3.2. *Suppose Assumptions 3.1, 3.5–3.7 hold. For any $t > 0$ define $\mathcal{N}_{\Sigma}(t) = \mathrm{tr}(\Sigma(\Sigma + tI)^{-1})$ and $\mathcal{N}_{C,\infty}(t) = \sup_{x\in\mathcal{X}} \langle k(\cdot,x), (C+tI)^{-1}k(\cdot,x)\rangle_{\mathcal{H}}$ for the uncentered covariance operator $C = \int_{\mathcal{X}} k(\cdot,x) \otimes_{\mathcal{H}} k(\cdot,x) d\mathbb{P}(x)$. Then the following hold:*

(i)

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

(ii) For any $0 \le \delta \le \frac{1}{2}$ satisfying $n \ge 2\log\frac{2}{\delta}$ 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. $m \geq (67 \vee 5\mathcal{N}_{C,\infty}(t)) \log \frac{4\kappa}{t\delta} \vee \frac{140\kappa}{t} \log \frac{8}{t\delta}$,

3. $n \geq 2 \log \frac{2}{\delta}$.

Remark 3.2. *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 the $(\lambda_i)_i$, which will be presented in the upcoming corollaries. The behavior of the empirical variations 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 the hypothesis space $\mathcal{H}$. 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 3.2, 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 corollary, proved in Section 3.3.8, derives convergence rates from the bounds in Theorem 3.2 under polynomial decay assumptions on the eigenvalues of Σ.*

Corollary 3.3 (Polynomial decay of eigenvalues)**.** *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 the following hold:*

(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)

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

(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 3.3. (i) *Of course, $\alpha > 1$ is required to ensure that Σ is trace class. Observing (*ii*) and (*iii*), 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 using the Lanczos method 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 application.*

3.3 Proofs

3.3.1 Proof of Proposition 3.1

Define the *sampling operator*

$$S : \mathcal{H} \to \mathbb{R}^n,\ f \mapsto \frac{1}{\sqrt{n}}\left(f(X_1), \ldots, f(X_n)\right)^{\top}$$

and *approximate sampling operator*

$$\tilde{Z}_m : \mathcal{H} \to \mathbb{R}^m,\ f \mapsto \Big(f(\tilde{X}_1), \ldots, f(\tilde{X}_n)\Big)^\top.$$

The adjoint of $\tilde{Z}_m$ (Smale and Zhou, 2007) is given by

$$\tilde{Z}_m^* : \mathbb{R}^m \to \mathcal{H},\ \alpha \mapsto \sum_{i=1}^m \alpha_i k(\cdot, \tilde{X}_i).$$

Thus, any $f \in \mathcal{H}_m$ may be written as $\tilde{Z}_m^* \alpha$, for some $\alpha \in \mathbb{R}^m$ and so

$$\Big\langle f, \hat{C} f \Big\rangle_{\mathcal{H}} = \frac{1}{n} \Big\langle \tilde{Z}_m^* \alpha, S^* S \tilde{Z}_m^* \alpha \Big\rangle_{\mathcal{H}} = \frac{1}{n} \alpha^\top \tilde{Z}_m S^* S \tilde{Z}_m^* \alpha,$$

where we used $\frac{1}{n} Z_n^* S = \hat{C}$. It is easy to verify that $S\tilde{Z}_m^* = \mathbf{K}_{nm}$ and $\tilde{Z}_m Z_n^* = \mathbf{K}_{mn}$. Therefore, (3.8) can be written as

$$\arg\sup \left\{ \frac{1}{n} \alpha^\top \mathbf{K}_{mn} \mathbf{K}_{nm} \alpha : \alpha^\top \mathbf{K}_{mm} \alpha = 1 \right\}. \tag{3.29}$$

Letting $\mathbf{u} = \mathbf{K}_{mm}^{1/2} \alpha$ simplifies the constraint in (3.29) to $\mathbf{u}^\top \mathbf{u} = 1$, and we write (3.29) as[1]

$$\arg\sup \left\{ \frac{1}{n} \mathbf{u}^\top \mathbf{K}_{mm}^{-1/2} \mathbf{K}_{mn} \mathbf{K}_{nm} \mathbf{K}_{mm}^{-1/2} \mathbf{u} : \mathbf{u}^\top \mathbf{u} = 1 \right\}.$$

The solution to the above problem is the unit eigenvector of $\frac{1}{n} \mathbf{K}_{mm}^{-1/2} \mathbf{K}_{mn} \mathbf{K}_{nm} \mathbf{K}_{mm}^{-1/2}$ corresponding to its largest eigenvalue. Denoting this eigenvector as $\mathbf{u}_{m,1}$, we obtain a function $\widehat{\phi}_{m,1} \in \mathcal{H}$ solving the NY-KPCA problem in (3.8) via $\widehat{\phi}_{m,1} = \tilde{Z}_m^* \mathbf{K}_{mm}^{-1/2} \mathbf{u}_{m,1}$.

3.3.2 Proof of Theorem 3.1

(i) For $t > 0$, we have

$$R_{\hat{C},\ell}^{nys} = \Big\| (I - P_m^\ell(\hat{C})) C^{1/2} \Big\|_{\mathcal{L}^2(\mathcal{H})}^2 = \Big\| (I - P_m^\ell(\hat{C}))(\hat{C} + tI)^{1/2} (\hat{C} + tI)^{-1/2} C^{1/2} \Big\|_{\mathcal{L}^2(\mathcal{H})}^2$$

[1]The existence of $\mathbf{K}^{-1}$, and by proxy $\mathbf{K}_{mm}^{-1}$, is guaranteed by strict positive definiteness of k, provided all X_i in the training set are unique.

$$\leq \left\|(I - P_m^\ell(\hat{C}))(\hat{C} + tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \left\|(\hat{C} + tI)^{-1/2} C^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})}. \tag{3.30}$$

We now bound the terms in (3.30). First, we have

$$\begin{aligned}
\left\|(\hat{C} + tI)^{-1/2} C^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} &= \left\|(\hat{C} + tI)^{-1/2}(C + tI)^{1/2}(C + tI)^{-1/2} C^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} \\
&\leq \left\|(\hat{C} + tI)^{-1/2}(C + tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \left\|(C + tI)^{-1/2} C^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} \\
&= \underbrace{\left\|(\hat{C} + tI)^{-1/2}(C + tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})}}_{(A)} \mathcal{N}_C(t),
\end{aligned} \tag{3.31}$$

where we used the fact $\left\|(C + tI)^{-1/2} C^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} = \operatorname{tr}(C^{1/2}(C + tI)^{-1} C^{1/2}) = \operatorname{tr}((C + tI)^{-1} C) =: \mathcal{N}_C(t)$. Next, we have

$$\begin{aligned}
\left\|(I - P_m^\ell(\hat{C}))(\hat{C} + tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \leq\ & 2 \underbrace{\left\|(I - P_m)(\hat{C} + tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})}}_{(B)} \\
& + 2 \underbrace{\left\|(P_m - P_m^\ell(\hat{C}))(\hat{C} + tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})}}_{(D)},
\end{aligned} \tag{3.32}$$

where $P_m = Z_m^*(\mathbf{K}_{mm})^{-1} Z_m$ is the orthogonal projector onto $\mathcal{H}_m$. (B) can be bounded as

$$(B) \leq \underbrace{\left\|(I - P_m)(C + tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})}}_{(B_1)} \underbrace{\left\|(C + tI)^{-1/2}(\hat{C} + tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})}}_{(B_2)}, \tag{3.33}$$

and (D) as

$$\begin{aligned}
(D) &\overset{(*)}{=} \left\|(I - P_m^\ell(\hat{C}))P_m(\hat{C} + tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \\
&= \left\|(I - P_m^\ell(\hat{C}))P_m(\hat{C} + tI)P_m(I - P_m^\ell(\hat{C}))\right\|_{\mathcal{L}^\infty(\mathcal{H})} \\
&\leq \left\|(I - P_m^\ell(\hat{C}))P_m\hat{C}P_m(I - P_m^\ell(\hat{C}))\right\|_{\mathcal{L}^\infty(\mathcal{H})} \\
&\quad + t \left\|(I - P_m^\ell(\hat{C}))P_m(I - P_m^\ell(\hat{C}))\right\|_{\mathcal{L}^\infty(\mathcal{H})} \overset{(**)}{\leq} \hat{\lambda}_{m,\ell+1} + t,
\end{aligned} \tag{3.34}$$

where we used the facts that $\mathcal{R}(P_m^\ell(\hat{C})) \subset \mathcal{R}(P_m)$ in $(*)$ and $P_m^\ell(\hat{C})$ projects onto the ℓ-eigenspace of $P_m\hat{C}P_m$ in $(**)$. $\hat{\lambda}_{m,\ell+1}$ can be bounded as

$$\hat{\lambda}_{m,\ell+1} \leq |\hat{\lambda}_{m,\ell+1} - \hat{\lambda}_{\ell+1}| + \hat{\lambda}_{\ell+1} \overset{(\dagger)}{\leq} \frac{1}{n}\left\|\tilde{\mathbf{K}} - \mathbf{K}\right\|_{\mathcal{L}^\infty(\mathbb{R}^n)} + \hat{\lambda}_\ell, \tag{3.35}$$

where (†) follows from the Hoffman-Wiendladt inequality (R. Bhatia, 1994). We may rewrite (3.35) as

$$\begin{aligned}
\frac{1}{n}\left\|\tilde{\mathbf{K}} - \mathbf{K}\right\|_{\mathcal{L}^\infty(\mathbb{R}^n)} &= \|S(I-P_m)S^*\|_{\mathcal{L}^\infty(\mathbb{R}^n)} \\
&= \left\|(I-P_m)\hat{C}(I-P_m)\right\|_{\mathcal{L}^\infty(\mathcal{H})} = \left\|\hat{C}^{1/2}(I-P_m)\hat{C}^{1/2}\right\|_{\mathcal{L}^\infty(\mathcal{H})} \\
&\leq \left\|\hat{C}^{1/2}(C+tI)^{-1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \left\|(C+tI)^{1/2}(I-P_m)\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \\
&\overset{(\ddagger)}{\leq} \left\|(\hat{C}+tI)^{1/2}(C+tI)^{-1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \left\|(C+tI)^{1/2}(I-P_m)\right\|^2_{\mathcal{L}^\infty(\mathcal{H})},
\end{aligned} \tag{3.36}$$

where we used

$$\left\|\hat{C}^{1/2}(C+tI)^{-1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \leq \left\|\hat{C}^{1/2}(\hat{C}+tI)^{-1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \left\|(\hat{C}+tI)^{1/2}(C+tI)^{-1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})}$$

and $\|\hat{C}^{1/2}(\hat{C}+tI)^{-1/2}\|^2_{\mathcal{L}^\infty(\mathcal{H})} \leq 1$ in (‡). The result follows by combining (3.30)–(3.36) and employing Lemmas A.1 and B.1 for (iii).

(ii) The proof follows exactly as in (i); however, we bound $\left\|(I-P_m)(C+tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})}$ with Lemma B.2 with $t_0 = \frac{19\kappa}{n}\log\frac{2n}{\delta}$.

3.3.3 Proof of Corollary 3.1

(i) From Theorem 3.1 (i) we have

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

Similarly, $R_{C,\ell} = \sum_{i>\ell}\lambda_i \gtrsim \sum_{i>\ell} i^{-\alpha} \gtrsim \int_\ell^\infty x^{-\alpha}dx \gtrsim \ell^{1-\alpha} = n^{-\theta(1-\frac{1}{\alpha})}$.

(ii) This is Theorem 3.2 of (Rudi et al., 2015) with $\alpha = \frac{1}{2}$, $r = \alpha$, $p = 2$, and $\ell = n^{\frac{\theta}{\alpha}}$.

(iii) Theorem 3.1 (iii) and Proposition A.1 yield

$$R^{nys}_{\hat{C},\ell} \lesssim_{\mathbb{P}^n} t^{-\frac{1}{\alpha}} n^{-\theta} + t^{1-\frac{1}{\alpha}} \leq \begin{cases} t^{1-\frac{1}{\alpha}}, t \geq n^{-\theta} \\ t^{-\frac{1}{\alpha}} n^{-\theta}, t \leq n^{-\theta} \end{cases},$$

where $\frac{\log n}{n} \lesssim t \leq \lambda_1$ and $m \gtrsim \mathcal{N}_{C,\infty}(t) \log \frac{1}{t}$ with

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

First, consider the case when $t \geq n^{-\theta}$. This means

$$R^{nys}_{\hat{C},\ell} \lesssim \inf \left\{ t^{1-\frac{1}{\alpha}} : t \gtrsim \frac{\log n}{n} \vee n^{-\theta}, m \gtrsim \frac{1}{t} \log \frac{1}{t} \right\}.$$

For $\theta < 1$, we obtain

$$R^{nys}_{\hat{C},\ell} \lesssim \inf \left\{ t^{1-\frac{1}{\alpha}} : t \gtrsim n^{-\theta}, m \gtrsim \frac{1}{t} \log \frac{1}{t} \right\} \leq n^{-\theta\left(1-\frac{1}{\alpha}\right)}$$

if $m \gtrsim n^{\theta} \log n$. For $\theta \geq 1$, we obtain

$$R^{nys}_{\hat{C},\ell} \lesssim \inf \left\{ t^{1-\frac{1}{\alpha}} : t \gtrsim \frac{\log n}{n}, m \gtrsim \frac{1}{t} \log \frac{1}{t} \right\} \leq \left(\frac{\log n}{n} \right)^{\left(1-\frac{1}{\alpha}\right)}$$

if $m \gtrsim \frac{n}{\log n} \log \frac{n}{\log n}$.

Next, consider the case when $t \leq n^{-\theta}$ which means

$$R^{nys}_{\hat{C},\ell} \lesssim \inf \left\{ t^{-\frac{1}{\alpha}} n^{-\theta} : \frac{\log n}{n} \lesssim t \lesssim n^{-\theta}, m \gtrsim \frac{1}{t} \log \frac{1}{t} \right\} \leq n^{-\theta\left(1-\frac{1}{\alpha}\right)}$$

when $\theta < 1$ and $m \gtrsim n^{\theta} \log n$.

(iv) Theorem 3.1(iv) and Proposition A.1 yield

$$R^{nys}_{\hat{C},\ell} \lesssim_{\mathbb{P}^n} t^{-\frac{1}{\alpha}} n^{-\theta} + t^{1-\frac{1}{\alpha}} \leq \begin{cases} t^{1-\frac{1}{\alpha}}, t \geq n^{-\theta} \\ t^{-\frac{1}{\alpha}} n^{-\theta}, t \leq n^{-\theta} \end{cases},$$

where $\frac{\log n}{n} \lesssim t \leq \lambda_1$ and $m \gtrsim \mathcal{N}_C(t) \log n \gtrsim t^{-\frac{1}{\alpha}} \log n$. The result follows by carrying out the analysis as in (iii) for $\theta < 1$ and $\theta \geq 1$.

3.3.4 Proof of Corollary 3.2

(i) From Theorem 3.1 (i) we have

$$R_{C,\ell} = \sum_{i>\ell} \lambda_i \lesssim \sum_{i>\ell} e^{-\tau i} \lesssim \int_{\ell}^{\infty} e^{-\tau x} dx \lesssim e^{-\tau \ell} = n^{-\theta}$$

and

$$R_{C,\ell} = \sum_{i>\ell} \lambda_i \gtrsim \sum_{i>\ell} e^{-\tau i} \gtrsim \int_{\ell+1}^{\infty} e^{-\tau x} dx \gtrsim e^{-\tau(\ell+1)} = e^{-\tau} n^{-\theta}.$$

(ii) Theorem 3.1 (ii) and Proposition A.2 yield

$$R_{\hat{C},\ell} \lesssim_{\mathbb{P}^n} \left(n^{-\theta} + t\right) \log \frac{1}{t} \leq \begin{cases} n^{-\theta} \log \frac{1}{t}, t \leq n^{-\theta} \\ t \log \frac{1}{t}, t \geq n^{-\theta} \end{cases},$$

where $\frac{\log n}{n} \lesssim t \leq \lambda_1$.

For the case of $t \leq n^{-\theta}$, we obtain

$$R_{\hat{C},\ell} \lesssim \inf \left\{ n^{-\theta} \log \frac{1}{t} : \frac{\log n}{n} \lesssim t \leq n^{-\theta} \right\} = n^{-\theta} \log n,$$

where the constraint is only valid for $\theta < 1$.

On the other hand, for $t \geq n^{-\theta}$, we obtain

$$R_{\hat{C},\ell} \lesssim \inf\left\{t \log \frac{1}{t} : t \gtrsim \frac{\log n}{n} \vee n^{-\theta}\right\} = \frac{\log n}{n} \log\left(\frac{n}{\log n}\right) \leq \frac{(\log n)^2}{n},$$

which holds for $\theta \geq 1$.

(iii) Arguing similarly as in (ii), it follows that for $\theta < 1$ and $m \gtrsim n^{\theta} \log n$, we obtain a rate of $n^{-\theta} \log n$ for $R^{nys}_{\hat{C},\ell}$. Similarly for $\theta \geq 1$ and $m \geq \frac{n}{\log n} \log\left(\frac{n}{\log n}\right)$, we obtain a rate of $n^{-1}(\log n)^2$.

(iv) Arguing as in (ii) and enforcing the restriction $m \gtrsim \log n \log \frac{1}{t}$ imposed by Theorem 3.1 (ii) yields the result.

3.3.5 Proof of Proposition 3.2

Define the sampling operator

$$S : \mathcal{H} \to \mathbb{R}^n, \qquad f \mapsto \frac{1}{\sqrt{n}}(f(X_1), \ldots, f(X_n))^\top$$

whose adjoint, called the reconstruction operator can be shown (see Proposition D.1*(i)*) to be

$$S^* : \mathbb{R}^n \to \mathcal{H}, \qquad \boldsymbol{\alpha} \mapsto \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \alpha_i k(\cdot, X_i),$$

where $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_n)^\top$. Define $\tilde{\mathbf{C}}_n = \frac{n}{n-1}\mathbf{C}_n$. It follows from Proposition D.1*(ii)* that $\widehat{\Sigma} = S^* \tilde{\mathbf{C}}_n S$, which implies $(\widehat{\phi}_i)_i$ satisfy

$$S^* \tilde{\mathbf{C}}_n S \widehat{\phi}_i = \widehat{\lambda}_i \widehat{\phi}_i, \tag{3.37}$$

where $\widehat{\lambda}_i \geq 0$. Multiplying both sides of (3.37) on the left by S, we obtain that $(\widehat{\boldsymbol{\alpha}}_i)_i$, $\widehat{\boldsymbol{\alpha}}_i := S\widehat{\phi}_i$, $i \in [n]$ are eigenvectors of $SS^*\tilde{\mathbf{C}}_n = \frac{1}{n}\mathbf{K}\tilde{\mathbf{C}}_n$, i.e., they satisfy the finite dimensional linear system,

$$\mathbf{K}\tilde{\mathbf{C}}_n \widehat{\boldsymbol{\alpha}}_i = n\widehat{\lambda}_i \widehat{\boldsymbol{\alpha}}_i, \tag{3.38}$$

where $\mathbf{K}$ is the Gram matrix, i.e., $(\mathbf{K})_{ij} = k(X_i, X_j)$, $i, j \in [n]$ and the fact that $\mathbf{K} = nSS^*$ follows from Proposition D.1*(iii)*. It is important to note that $(\widehat{\boldsymbol{\alpha}}_i)_i$ do not form an orthogonal system in the usual Euclidean inner product but in the weighted inner product where the weighting matrix is $\tilde{\mathbf{C}}_n$. Indeed, it is easy to verify that

$$\left\langle \widehat{\boldsymbol{\alpha}}_i, \tilde{\mathbf{C}}_n \widehat{\boldsymbol{\alpha}}_j \right\rangle_2 = \left\langle S\widehat{\phi}_i, \tilde{\mathbf{C}}_n S\widehat{\phi}_j \right\rangle_2 = \langle \widehat{\phi}_i, \widehat{\Sigma}\widehat{\phi}_j \rangle_{\mathcal{H}} = \widehat{\lambda}_j \langle \widehat{\phi}_i, \widehat{\phi}_j \rangle_{\mathcal{H}} = \widehat{\lambda}_j \delta_{ij},$$

where δ_{ij} is the Kronecker delta. Having obtained $(\widehat{\alpha}_i)_i$ from (3.38), the eigenfunctions of $\widehat{\Sigma}$ are obtained from (3.37) as

$$\widehat{\phi}_i = \frac{1}{\widehat{\lambda}_i} S^* \tilde{\mathbf{C}}_n \widehat{\boldsymbol{\alpha}}_i,$$

and the result follows.

3.3.6 Proof of Proposition 3.3

Let $\mathbf{C}_m = \mathbf{I}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top$ and $\mathbf{H}_m = m\mathbf{C}_m$. Note that any $f \in \bar{\mathcal{H}}_m$ can be written as $\tilde{Z}_m^* \mathbf{H}_m \boldsymbol{\alpha}$ for some $\boldsymbol{\alpha} \in \mathbb{R}^m$. Thus, we may express (3.24) as

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

where we have used Lemma D.1*(iv)*. Let $\mathbf{u} = \mathbf{K}_{mm}^{1/2}\mathbf{H}_m\boldsymbol{\alpha}$, and the above problem may be written as

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

where $\overline{\mathbf{M}} := \mathbf{K}_{mm}^{-1/2}\mathbf{K}_{mn}\mathbf{H}_n\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1/2}$. The solution to the above problem is the unit eigenvector of $\frac{1}{n(n-1)}\overline{\mathbf{M}}$ 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 $\bar{\mathcal{H}}_m$,

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

solving (3.24). Subsequent eigenfunctions $\tilde{\phi}_i$ may be computed in a similar manner from the eigenvectors of $\frac{1}{n(n-1)}\overline{\mathbf{M}}$, and the orthonormality of the $\{\tilde{\phi}_i\}_i$ follows from the orthonormality of the $\{\mathbf{u}_i\}_i$, i.e.,

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

3.3.7 Proof of Theorem 3.2

(i) From Lemma A.3, 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

$$\begin{aligned} R^\ell(\widehat{\Sigma}) &= \mathbb{E}\left\|(k(\cdot,X) - m_{\mathbb{P}}) - P^\ell(\widehat{\Sigma})(k(\cdot,X) - \widehat{m}_{\mathbb{P}})\right\|_{\mathcal{H}}^2 \\ &= \mathbb{E}\left\|(I - P^\ell(\widehat{\Sigma}))\bar{k}(\cdot,X)\right\|_{\mathcal{H}}^2 + \left\|P^\ell(\widehat{\Sigma})(m_{\mathbb{P}} - \widehat{m}_{\mathbb{P}})\right\|_{\mathcal{H}}^2 \\ &\qquad -2\mathbb{E}\left\langle (I - P^\ell(\widehat{\Sigma}))\bar{k}(\cdot,X), P^\ell(\widehat{\Sigma})(m_{\mathbb{P}} - \widehat{m}_{\mathbb{P}})\right\rangle_{\mathcal{H}} \\ &= \underbrace{\left\|(I - P^\ell(\widehat{\Sigma}))\Sigma^{1/2}\right\|_{\mathcal{L}^2(\mathcal{H})}^2}_{\text{(A)}} + \underbrace{\left\|P^\ell(\widehat{\Sigma})(m_{\mathbb{P}} - \widehat{m}_{\mathbb{P}})\right\|_{\mathcal{H}}^2}_{\text{(B)}}, \end{aligned} \tag{3.39}$$

where the last equality holds because $\mathbb{E}[\bar{k}(\cdot,X)] = 0$ and we have employed Lemma A.3. For any $t > 0$ we have

$$\begin{aligned} \text{(A)} &= \left\|(I - P^\ell(\widehat{\Sigma}))(\widehat{\Sigma} + tI)^{1/2}(\widehat{\Sigma} + tI)^{-1/2}(\Sigma + tI)^{1/2}(\Sigma + tI)^{-1/2}\Sigma^{1/2}\right\|_{\mathcal{L}^2(\mathcal{H})}^2 \\ &\leq \left\|(\Sigma + tI)^{-1/2}\Sigma^{1/2}\right\|_{\mathcal{L}^2(\mathcal{H})}^2 \left\|(\widehat{\Sigma} + tI)^{-1/2}(\Sigma + tI)^{1/2}\right\|_{\mathcal{L}^\infty(\mathcal{H})}^2 \\ &\quad \times \left\|(I - P^\ell(\widehat{\Sigma}))(\widehat{\Sigma} + tI)^{1/2}\right\|_{\mathcal{L}^\infty(\mathcal{H})}^2 \end{aligned}$$

$$\overset{(\dagger)}{\leq} \mathcal{N}_\Sigma(t) \left\| (\widehat{\Sigma} + tI)^{-1/2} (\Sigma + tI)^{1/2} \right\|^2_{\mathcal{L}^\infty(\mathcal{H})} (\widehat{\lambda}_{\ell+1} + t), \tag{3.40}$$

where we have used

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

which holds via invariance of trace under cyclic permutations, in (†). The result follows from applying Lemma A.2 to (3.40) and Lemma C.1 to Ⓑ, noticing that

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

Lower Bound: It is clear from (3.39) that $R^\ell(\widehat{\Sigma}) \geq Ⓐ$. We will show that

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

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 $Ⓐ \geq R^\ell(\Sigma)$. We have

$$\begin{aligned} \left\| (I - P_{\psi,\ell})\, \Sigma^{1/2} \right\|^2_{\mathcal{L}^2(\mathcal{H})} &= \operatorname{Tr}\left[\Sigma^{1/2} (I - P_{\psi,\ell})(I - P_{\psi,\ell}) \Sigma^{1/2} \right] \\ &= \operatorname{Tr}\left[(I - P_{\psi,\ell}) \Sigma \right] = \sum_{i \geq 1} \lambda_i - \langle P_{\psi,\ell}, \Sigma \rangle_{\mathcal{L}^2(\mathcal{H})}. \end{aligned} \tag{3.41}$$

Clearly the l.h.s. of (3.41) 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 decompose the reconstruction error as

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

$$-2\mathbb{E}\left\langle (I - P^{\ell}_{nys}(\widehat{\Sigma}))\bar{k}(\cdot, X), P^{\ell}_{nys}(\widehat{\Sigma})(m_{\mathbb{P}} - \widehat{m}_{\mathbb{P}})\right\rangle_{\mathcal{H}}. \tag{3.42}$$

Now $\mathbb{E}[\bar{k}(\cdot, X)] = 0$ implies the third term in (3.42) 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 C.1, yields

$$\mathbb{P}^n \left\{ \text{(D)} \leq \frac{32\kappa \log \frac{2}{\delta}}{n} \right\} \geq 1 - \delta. \tag{3.43}$$

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

$$\begin{aligned}
\text{(C)} &\overset{(\star)}{=} \left\|(I - P^{\ell}_{nys}(\widehat{\Sigma}))\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} = \left\|(I - P^{\ell}_{nys}(\widehat{\Sigma}))(\widehat{\Sigma} + tI)^{1/2}(\widehat{\Sigma} + tI)^{-1/2}\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} \\
&\leq \left\|(I - P^{\ell}_{nys}(\widehat{\Sigma}))(\widehat{\Sigma} + tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \left\|(\widehat{\Sigma} + tI)^{-1/2}\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} \\
&\leq \left\|(I - P^{\ell}_{nys}(\widehat{\Sigma}))(\widehat{\Sigma} + tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \left\|(\widehat{\Sigma} + tI)^{-1/2}(\Sigma + tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \\
&\times \left\|(\Sigma + tI)^{-1/2}\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} \\
&\overset{(\dagger)}{\leq} 2\mathcal{N}_{\Sigma}(t) \left\|(I - P^{\ell}_{nys}(\widehat{\Sigma}))(\widehat{\Sigma} + tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})},
\end{aligned} \tag{3.44}$$

where we have used Lemma A.3 in $(\star)$ and Lemma A.2 in $(\dagger)$. For convenience, we now let $\widehat{\Sigma}_t = \widehat{\Sigma} + tI$. Observing the last term, we have

$$\begin{aligned}
&\left\|(I - P^{\ell}_{nys}(\widehat{\Sigma}))(\widehat{\Sigma} + tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \\
&\leq 2\left\|\left(I - \bar{P}_m\right)\widehat{\Sigma}_t^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} + 2\left\|\left(\bar{P}_m - P^{\ell}_{nys}(\widehat{\Sigma})\right)\widehat{\Sigma}_t^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \\
&\leq \underbrace{2\left\|(I - \bar{P}_m)\Sigma_t^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \left\|\Sigma_t^{-1/2}\widehat{\Sigma}_t^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})}}_{\text{(C1)}} \\
&\qquad + 2\left\|(I - P^{\ell}_{nys}(\widehat{\Sigma}))\bar{P}_m\widehat{\Sigma}_t^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \\
&= 2\,\text{(C1)} + 2\left\|(I - P^{\ell}_{nys}(\widehat{\Sigma}))\bar{P}_m\widehat{\Sigma}_t\bar{P}_m(I - P^{\ell}_{nys}(\widehat{\Sigma}))\right\|_{\mathcal{L}^\infty(\mathcal{H})} \\
&\leq 2\,\text{(C1)} + 2\left\|(I - P^{\ell}_{nys}(\widehat{\Sigma}))\bar{P}_m\widehat{\Sigma}\bar{P}_m(I - P^{\ell}_{nys}(\widehat{\Sigma}))\right\|_{\mathcal{L}^\infty(\mathcal{H})}
\end{aligned} \tag{3.45}$$

$$
\begin{aligned}
&\quad +2t \left\| (I - P_{nys}^{\ell}(\widehat{\Sigma}))\bar{P}_m(I - P_{nys}^{\ell}(\widehat{\Sigma})) \right\|_{\mathcal{L}^\infty(\mathcal{H})} \\
&\leq 2\textcircled{C1} + 2\left(\tilde{\lambda}_{\ell+1} + t\right), \qquad (3.46)
\end{aligned}
$$

where we have used $\mathrm{ran}(P_{nys}^{\ell}(\widehat{\Sigma})) \subset \mathrm{ran}(\bar{P}_m)$ in (3.45), and (3.46) holds because $P_{nys}^{\ell}(\widehat{\Sigma})$ projects onto the ℓ-eigenspace of $\bar{P}_m\widehat{\Sigma}\bar{P}_m$. Lemmas B.3 and A.2*(iii)* give

$$
\mathbb{P}^n\left\{\textcircled{C1} \leq 3t\right\} \geq 1 - 4\delta. \qquad (3.47)
$$

Continuing, we have

$$
\begin{aligned}
\tilde{\lambda}_{\ell+1} + t &\leq |\tilde{\lambda}_{\ell+1} - \widehat{\lambda}_{\ell+1}| + \widehat{\lambda}_{\ell+1} + t \\
&\overset{(\dagger)}{\leq} \frac{1}{n(n-1)} \left\| \left(\tilde{\mathbf{K}} - \mathbf{K}\right) \mathbf{H}_n \right\|_{\mathcal{L}^\infty(\mathbb{R}^n)} + \widehat{\lambda}_{\ell+1} + t, \qquad (3.48)
\end{aligned}
$$

where (†) uses the Hoffman-Wielandt inequality (R. Bhatia, 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 $\mathrm{span}\{k(\cdot, X_{r_j})|j \in [m]\}$, it is easy to verify that $P_m = \tilde{Z}_m^* \mathbf{K}_{mm}^{-1} \tilde{Z}_m$ (Rudi et al., 2015, Lemma 1). Using $\sqrt{n}S\tilde{Z}_m^* = \mathbf{K}_{nm}$, which follows from Proposition D.1*(iv)*, we have the expression

$$
\tilde{\mathbf{K}} = \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn} = nS\tilde{Z}_m^*\mathbf{K}_{mm}^{-1}\tilde{Z}_m S^* = nSP_mS^*.
$$

Thus, we may write,

$$
\begin{aligned}
\left\| \left(\tilde{\mathbf{K}} - \mathbf{K}\right) \mathbf{H}_n \right\|_{\mathcal{L}^\infty(\mathbb{R}^n)} &\leq \left\| \tilde{\mathbf{K}} - \mathbf{K} \right\|_{\mathcal{L}^\infty(\mathbb{R}^n)} \|\mathbf{H}_n\|_{\mathcal{L}^\infty(\mathbb{R}^n)} \\
&= n^2 \|S(I - P_m)S^*\|_{\mathcal{L}^\infty(\mathbb{R}^n)} \\
&= n^2 \|(I - P_m)S^*S(I - P_m)\|_{\mathcal{L}^\infty(\mathcal{H})} \\
&= n^2 \left\| (I - P_m)\hat{C}(I - P_m) \right\|_{\mathcal{L}^\infty(\mathcal{H})}, \qquad (3.49)
\end{aligned}
$$

where we have used Proposition D.1 *(iv)* in (3.49). Proceeding,

$$
(3.49) = n^2 \left\| \hat{C}^{1/2}(I - P_m)^2 \hat{C}^{1/2} \right\|_{\mathcal{L}^\infty(\mathcal{H})}
$$

$$
\begin{aligned}
&\leq n^2 \left\|\hat{C}^{1/2}(C+tI)^{-1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \left\|(C+tI)^{1/2}(I-P_m)\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \\
&\leq n^2 \left\|\hat{C}^{1/2}(\hat{C}+tI)^{-1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \left\|(\hat{C}+tI)^{1/2}(C+tI)^{-1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \\
&\qquad \times \left\|(C+tI)^{1/2}(I-P_m)\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \\
&\leq n^2 \left\|(\hat{C}+tI)^{1/2}(C+tI)^{-1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \left\|(C+tI)^{1/2}(I-P_m)\right\|^2_{\mathcal{L}^\infty(\mathcal{H})}. \quad (3.50)
\end{aligned}
$$

Applying Lemmas A.1 and B.1 to (3.50) and Lemma A.2*(iv)* in (3.48) gives

$$
\mathbb{P}^n \left\{ \tilde{\lambda}_{\ell+1} + t \leq \frac{9n^2 t}{2n(n-1)} + \frac{3}{2}(\lambda_{\ell+1} + t) \right\} \geq 1 - 4\delta. \quad (3.51)
$$

Combining (3.51) with (3.47) in (3.46) gives

$$
\mathbb{P}^n \left\{ \left\|(I - P^\ell_{nys}(\widehat{\Sigma}))(\widehat{\Sigma} + tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \leq 27t + 3\lambda_{\ell+1} \right\} \geq 1 - 8\delta, \quad (3.52)
$$

where we note that $\frac{1}{n-1} \leq \frac{2}{n}$ for $n \geq 2$. The result follows by combining (3.52), (3.44), and (3.43) in (3.42).

Lower Bound: Using (3.42) we have

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

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.

3.3.8 Proof of Corollary 3.3

(i) From Theorem 3.2*(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,

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

(ii) Theorem 3.2(*ii*) and (Sriperumbudur and Sterge, 2020, Lemma A.8) 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 3.2(*iii*) and (Sriperumbudur and Sterge, 2020, Lemma A.8) 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*).

Chapter 4

Approximate Kernel PCA with Random Features

In this chapter, we present approximate kernel PCA using random features, which we call as RF-KPCA. Throughout this chapter, we assume the following:

Assumption 4.1. *$\mathcal{H}$ is a separable RKHS with reproducing kernel k of the form*

$$k(x,y) = \int_{\Theta} \varphi(x,\theta)\varphi(y,\theta)\, d\Lambda(\theta) = \langle \varphi(x,\cdot), \varphi(y,\cdot) \rangle_{L^2(\Lambda)},$$

where $\varphi : \mathcal{X} \times \Theta \to \mathbb{R}$ is continuous, $\sup_{\theta\in\Theta, x\in\mathcal{X}} |\varphi(x,\theta)| \leq \sqrt{\kappa}$ and Λ is a probability measure on a second countable space $(\Theta, \mathcal{A})$ endowed with Borel σ-algebra $\mathcal{A}$.

The assumption of Λ being a probability measure on Θ is not restrictive as any $\Lambda \in M_+^b(\Theta)$ can be normalized to a probability measure. However, the uniform boundedness of φ over $\mathcal{X} \times \Theta$ is somewhat restrictive as it is sufficient to assume $\varphi(x,\cdot) \in L^2(\mathcal{X}, \Lambda)$, $\forall\, x \in \mathcal{X}$ for k to be well-defined. But the uniform boundedness of φ ensures that k is bounded, as assumed in (A_1). By sampling $(\theta_i)_{i=1}^m \overset{i.i.d.}{\sim} \Lambda$, an approximation to k can be constructed as

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

where $\varphi_i := \frac{1}{\sqrt{m}}\varphi(\cdot,\theta_i)$ and $\Phi_m(x) := (\varphi_1(x), \ldots, \varphi_m(x))^\top \in \mathbb{R}^m$ is the random feature

map. It is easy to verify that k_m is the reproducing kernel of the RKHS

$$\mathcal{H}_m = \left\{ f : f = \sum_{i=1}^{m} \beta_i \varphi_i,\ (\beta_i)_{i=1}^m \subset \mathbb{R} \right\}$$

w.r.t. $\langle \cdot, \cdot \rangle_{\mathcal{H}_m}$ defined as $\langle f, g \rangle_{\mathcal{H}_m} := \sum_{i=1}^m \alpha_i \beta_i$ where $g = \sum_{i=1}^m \alpha_i \varphi_i$. Therefore $\mathcal{H}_m$ is isometrically isomorphic to $\mathbb{R}^m$. We refer the reader to Rudi and Rosasco 2017, Appendix E for examples of φ that yield some widely used reproducing kernels.

Having obtained a random feature map, the idea of RF-KPCA is to perform linear PCA on $\Phi_m(X)$ where $X \sim \mathbb{P}$, i.e., RF-KPCA involves finding a direction $\beta \in \mathbb{R}^m$ such that the variance of $\langle \beta, \Phi_m(X) \rangle_2$ is maximized:

$$\sup_{\|\beta\|_2 = 1} \mathrm{Var}[\langle \beta, \Phi_m(X) \rangle_2] = \sup_{\|\beta\|_2 = 1} \langle \beta, \Omega_m \beta \rangle_2, \tag{4.1}$$

where $\Omega_m := \mathrm{Cov}[\Phi_m(X)] = \mathbb{E}[(\Phi_m(X) - \mathbb{E}[\Phi_m(X)]) \otimes_2 (\Phi_m(X) - \mathbb{E}[\Phi_m(X)])]$ is a self-adjoint positive definite matrix. In fact, it is easy to verify that performing linear PCA on $\Phi_m(X)$ is same as performing KPCA in $\mathcal{H}_m$ since

$$\sup_{\|f\|_{\mathcal{H}_m} = 1} \mathrm{Var}[f(X)] = \sup_{\|\beta\|_2 = 1} \mathrm{Var}[\langle \beta, \Phi_m(X) \rangle_2], \tag{4.2}$$

which follows from $\mathcal{H}_m$ being isometrically isomorphic to $\mathbb{R}^m$ and $f \in \mathcal{H}_m$ has the form $f(x) = \langle \beta, \Phi_m(x) \rangle_2$. Note that

$$\sup_{\|f\|_{\mathcal{H}_m} = 1} \mathrm{Var}[f(X)] = \sup_{\|f\|_{\mathcal{H}_m} = 1} \langle f, \Sigma_m f \rangle_{\mathcal{H}_m},$$

where

$$\begin{aligned} \Sigma_m := \int_{\mathcal{X}} k_m(\cdot, x) \otimes_{\mathcal{H}_m} k_m(\cdot, x)\, d\mathbb{P}(x) \\ - \left(\int k_m(\cdot, x)\, d\mathbb{P}(x) \right) \otimes_{\mathcal{H}_m} \left(\int k_m(\cdot, x)\, d\mathbb{P}(x) \right). \end{aligned} \tag{4.3}$$

It therefore follows from (4.1)–(4.3) that the eigenvalues of Σ_m and Ω_m coincide and the eigenfunctions, $(\phi_{m,i})_{i=1}^m$ of Σ_m and eigenvectors, $(\beta_{m,i})_{i=1}^m$ of Ω_m are related as

$\phi_{m,i}(x) = \langle \beta_{m,i}, \Phi_m(x) \rangle_2$.

The empirical counterpart of RF-KPCA (we call it as RF-EKPCA) is obtained by solving

$$\sup_{\|\beta\|_2=1} \widehat{\mathrm{Var}}[\langle \beta, \Phi_m(X) \rangle_2] = \sup_{\|\beta\|_2=1} \langle \beta, \widehat{\Omega}_m \beta \rangle_2 = \sup_{\|f\|_{\mathcal{H}_m}=1} \langle f, \widehat{\Sigma}_m f \rangle_{\mathcal{H}_m},$$

where

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

is a self-adjoint positive definite operator on $\mathcal{H}_m$ that is equivalent (in the above mentioned sense) to $\widehat{\Omega}_m$, which is a U-statistic estimator of Ω_m. Since Σ_m and $\widehat{\Sigma}_m$ are trace-class (see Proposition D.3*(iii)*) and self-adjoint, spectral theorem (Reed and Simon, 1980, Theorems VI.16, VI.17) yields that

$$\Sigma_m = \sum_{i=1}^{m} \lambda_{m,i} \phi_{m,i} \otimes_{\mathcal{H}_m} \phi_{m,i} \quad \text{and} \quad \widehat{\Sigma}_m = \sum_{i=1}^{m} \widehat{\lambda}_{m,i} \widehat{\phi}_{m,i} \otimes_{\mathcal{H}_m} \widehat{\phi}_{m,i},$$

where $(\lambda_{m,i})_{i=1}^m \subset \mathbb{R}^+$ (*resp.* $(\widehat{\lambda}_{m,i})_{i=1}^m \subset \mathbb{R}^+$) and $(\phi_{m,i})_{i=1}^m$ (*resp.* $(\widehat{\phi}_{m,i})_{i=1}^m$) are the eigenvalues and eigenvectors of Σ_m (*resp.* $\widehat{\Sigma}_m$). We will assume that

Assumption 4.2. *The eigenvalues* $(\lambda_{m,i})_{i=1}^m$ (resp. $(\widehat{\lambda}_{m,i})_{i=1}^m$) *of* Σ_m (resp. $\widehat{\Sigma}_m$) *are simple,* $\mathrm{rank}(\Sigma_m) = m$, $\mathrm{rank}(\widehat{\Sigma}_m) = m$ *and without any loss of generality, they satisfy a decreasing rearrangement, i.e.,* $\lambda_{m,1} > \lambda_{m,2} > \cdots$ (resp. $\widehat{\lambda}_{m,1} > \widehat{\lambda}_{m,2} > \cdots$).

Based on Assumption 4.2, a low-dimensional representation of $X_i \in \mathcal{X}$ can be obtained as

$$\left(\widehat{\phi}_{m,1}(X_i), \ldots, \widehat{\phi}_{m,\ell}(X_i)\right)^\top \in \mathbb{R}^\ell,$$

where $\ell \leq m$ and $i \in [n]$ and the orthogonal projection operators onto the ℓ-eigenspaces of Σ_m and $\widehat{\Sigma}_m$ are given by $P^\ell(\Sigma_m) = \sum_{i=1}^{\ell} \phi_{m,i} \otimes_{\mathcal{H}_m} \phi_{m,i}$ and $P^\ell(\widehat{\Sigma}_m) = \sum_{i=1}^{\ell} \widehat{\phi}_{m,i} \otimes_{\mathcal{H}_m} \widehat{\phi}_{m,i}$, respectively.

Since $(\widehat{\lambda}_{m,i}, \widehat{\phi}_{m,i})_{i=1}^{\ell}$ for $\ell \leq m$ is a subset of the eigensystem of $\widehat{\Sigma}_m$ (which is equivalent

to the $m \times m$ matrix $\widehat{\Omega}_m$), the associated time complexity of finding this set scales as $O(m^2\ell + m^2 n)$, where $O(m^2 n)$ is the complexity of computing $\widehat{\Omega}_m$. This implies that RF-EKPCA is computationally cheaper than EKPCA if $m < \sqrt{n\ell}$ for $\ell \leq n$.

4.1 Computational vs. Statistical Trade-off

In Section 3.1.1 we considered the reconstruction error in KPCA and EKPCA to be

$$\mathbb{E}_{X\sim\mathbb{P}} \left\| \overline{k}(\cdot, X) - \sum_{i=1}^{\ell} \langle \theta(X), \zeta_i \rangle_{\mathcal{H}} \zeta_i \right\|_{\mathcal{H}}^2 \tag{4.4}$$

with $\theta(X) = \overline{k}(\cdot, X)$, $\zeta_i = \phi_i$ for KPCA and $\theta(X) = \widetilde{k}(\cdot, X)$, $\zeta_i = \widehat{\phi}_i$ for EKPCA, where $(\phi_i)_i$ and $(\widehat{\phi}_i)_i$ are the orthonormal eigenfunctions of Σ and $\widehat{\Sigma}$ corresponding to the eigenvalues $(\lambda_i)_i$ and $(\widehat{\lambda}_i)_i$. However, similar definition of reconstruction error as in (4.4) is not possible for RF-KPCA and RF-EKPCA as the orthonormal eigenvectors of Σ_m and $\widehat{\Sigma}_m$ belong to $\mathcal{H}_m$ (isometrically isomorphic to $\mathbb{R}^m$) while $\overline{k}(\cdot, X)$ and $\widetilde{k}(\cdot, X)$ belong to $\mathcal{H}$, which means the notion of respectively projecting $\overline{k}(\cdot, X)$ and $\widetilde{k}(\cdot, X)$ onto $(\phi_{m,i})_i$ and $(\widehat{\phi}_{m,i})_i$ is vacuous. In order to make the comparison between EKPCA and RF-EKPCA possible in terms of reconstruction error, we define certain operators below so that all the objects of interest are embedded into a common space, which we choose to be $L^2(\mathbb{P})$.

Define an inclusion operator (up to a constant)

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

where $f_{\mathbb{P}} := \int_{\mathcal{X}} f(x)\, d\mathbb{P}(x)$. It can be shown (see Proposition D.2) that $\mathfrak{I}^* : L^2(\mathbb{P}) \to \mathcal{H},\ f \mapsto \int_{\mathcal{X}} k(\cdot, x) f(x)\, d\mathbb{P}(x) - m_{\mathbb{P}} f_{\mathbb{P}}$ and $\Sigma = \mathfrak{I}^*\mathfrak{I}$. Similarly, we define an approximation operator

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

where $\varphi_{i,\mathbb{P}} := \int_{\mathcal{X}} \varphi_i(x)\, d\mathbb{P}(x)$. It can be shown (see Proposition D.3) that $\mathfrak{A}^* : L^2(\mathbb{P}) \to \mathcal{H}_m,\ f \mapsto \sum_{i=1}^{m} (\langle f, \varphi_i \rangle_{L^2(\mathbb{P})} - f_{\mathbb{P}} \varphi_{i,\mathbb{P}}) \varphi_i$ and $\Sigma_m = \mathfrak{A}^*\mathfrak{A}$. Based on these operators, we now

redefine the reconstruction error for KPCA, EKPCA, RF-KPCA and RF-EKPCA and present convergence results comparing the statistical behavior of RF-EKPCA with that of EKPCA. This redefinition can be done in two ways: (i) The projections (principal components) are first computed and then embedded in $L^2(\mathbb{P})$ and (ii) the functions are first embedded in $L^2(\mathbb{P})$ and then projected to find the principal components. In the following, we will first elucidate these ideas for KPCA and then define and analyze these different reconstruction errors for EKPCA, RF-KPCA and RF-EKPCA.

Note that $(\phi_i)_i$ and $\left(\frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}}\right)_i$ form orthonormal systems (ONS) in $\mathcal{H}$ and $L^2(\mathbb{P})$ respectively. While the former claim is trivial as $(\phi_i)_i$ form the eigensystem of Σ, the latter is true because $\langle \frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}}, \frac{\mathfrak{I}\phi_j}{\sqrt{\lambda_j}} \rangle_{L^2(\mathbb{P})} = \frac{\langle \mathfrak{I}^*\mathfrak{I}\phi_i, \phi_j \rangle_{\mathcal{H}}}{\sqrt{\lambda_i \lambda_j}} = \frac{\langle \Sigma\phi_i, \phi_j \rangle_{\mathcal{H}}}{\sqrt{\lambda_i \lambda_j}} = \sqrt{\frac{\lambda_i}{\lambda_j}} \langle \phi_i, \phi_j \rangle_{\mathcal{H}} = \delta_{ij}$. Based on this observation, we can define two reconstruction errors for KPCA as follows: *Reconstruct and Embed (R-E)*

$$T^{\ell}(\Sigma) = \mathbb{E}_{X \sim \mathbb{P}} \left\| \mathfrak{I}\overline{k}(\cdot, X) - \mathfrak{I} \left(\sum_{i=1}^{\ell} \left\langle \overline{k}(\cdot, X), \phi_i \right\rangle_{\mathcal{H}} \phi_i \right) \right\|_{L^2(\mathbb{P})}^2,$$

where $\overline{k}(\cdot, X)$ is first projected and reconstructed along $(\phi_i)_{i \in [\ell]}$ and then embedded into $L^2(\mathbb{P})$ through $\mathfrak{I}$; and, *Embed and Reconstruct (E-R)*

$$S^{\ell}(\Sigma) = \mathbb{E}_{X \sim \mathbb{P}} \left\| \mathfrak{I}\overline{k}(\cdot, X) - \sum_{i=1}^{\ell} \left\langle \mathfrak{I}\overline{k}(\cdot, X), \frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}} \right\rangle_{L^2(\mathbb{P})} \frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}} \right\|_{L^2(\mathbb{P})}^2, \tag{4.5}$$

where $\overline{k}(\cdot, X)$ is first embedded into $L^2(\mathbb{P})$ through $\mathfrak{I}$ and then projected and reconstructed along $\left(\frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}}\right)_i$. With this intuition, in the following sections, we define and analyze these reconstruction errors for the above mentioned variants of KPCA.

We would like to highlight the fact while $T^{\ell}(\Sigma) = S^{\ell}(\Sigma)$ (easy to show using $\mathfrak{I}^*\mathfrak{I} = \Sigma$), i.e., the operations of embedding and reconstruction are commutative for KPCA, we show in Theorems 4.1 and 4.2 that it is not the case for EKPCA and RF-EKPCA, i.e., they exhibit different statistical behaviors for R-E and E-R. Also, using $\mathfrak{I}^*\mathfrak{I} = \Sigma$, it can

be shown that

$$T^{\ell}(\Sigma) = \mathbb{E}_{X\sim\mathbb{P}} \left\| \Sigma^{1/2} \left[\overline{k}(\cdot, X) - \sum_{i=1}^{\ell} \left\langle \overline{k}(\cdot, X), \phi_i \right\rangle_{\mathcal{H}} \phi_i \right] \right\|_{\mathcal{H}}^2$$
$$\leq \|\Sigma\|_{\mathcal{L}^{\infty}(\mathcal{H})} \mathbb{E}_{X\sim\mathbb{P}} \left\| \overline{k}(\cdot, X) - \sum_{i=1}^{\ell} \left\langle \overline{k}(\cdot, X), \phi_i \right\rangle_{\mathcal{H}} \phi_i \right\|_{\mathcal{H}}^2,$$

implying that $T^{\ell}(\Sigma)$ (and also $S^{\ell}(\Sigma)$) is a weaker measure of reconstruction error than the one defined in (4.4), with the latter matching with the reconstruction error of linear PCA when $k(x, y) = \langle x, y\rangle_2$. More precisely, in Theorems 4.1 and 4.2, we will show that $T^{\ell}(\Sigma) = S^{\ell}(\Sigma) = \sum_{i>\ell} \lambda_i^2$ while using the same proof technique, the reconstruction error in (4.4) can be shown to be $\sum_{i>\ell} \lambda_i$, which clearly establishes that $T^{\ell}(\Sigma)$ is weaker than the one in (4.4). However, we cannot use the one in (4.4) for the reasons mentioned above, and therefore, in the following, we will investigate EKPCA and RF-EKPCA in R-E and E-R settings. Later, in Section 4.4.2, we will generalize $T^{\ell}(\Sigma)$ such that KPCA's reconstruction error behaves as $\sum_{i>\ell} \lambda_i$.

4.2 Reconstruct and Embed (R-E)

Based on the above intuition, since $(\widehat{\phi}_i)_i$, $(\phi_{m,i})_i$ and $(\widehat{\phi}_{m,i})_i$ form orthonormal systems, we define the reconstruction error for EKPCA, RF-KPCA and RF-EKPCA as

$$T^{\ell}(\widehat{\Sigma}) = \mathbb{E}_{X\sim\mathbb{P}} \left\| \mathfrak{I}\overline{k}(\cdot, X) - \mathfrak{I}P^{\ell}(\widehat{\Sigma})\tilde{k}(\cdot, X) \right\|_{L^2(\mathbb{P})}^2, \tag{4.6}$$

$$T^{\ell}(\Sigma_m) = \mathbb{E}_{X\sim\mathbb{P}} \left\| \mathfrak{I}\overline{k}(\cdot, X) - \mathfrak{A}P^{\ell}(\Sigma_m)\bar{k}_m(\cdot, X) \right\|_{L^2(\mathbb{P})}^2, \tag{4.7}$$

$$T^{\ell}(\widehat{\Sigma}_m) = \mathbb{E}_{X\sim\mathbb{P}} \left\| \mathfrak{I}\overline{k}(\cdot, X) - \mathfrak{A}P^{\ell}(\widehat{\Sigma}_m)\tilde{k}_m(\cdot, X) \right\|_{L^2(\mathbb{P})}^2 \tag{4.8}$$

respectively. Note that the reconstructed functions are computable as they depend only on $(X_i)_i$ while their embedded versions are not because they depend on unknown $\mathbb{P}$. The following result, proved in Section 4.5.1, provides a finite-sample bound on the reconstruction error, using which convergence rates can be obtained.

Theorem 4.1. *Suppose Assumptions 3.1, 3.5, 3.6, 4.1, and 4.2 hold. For any $t > 0$, define $\mathcal{N}_\Sigma(t) = \mathrm{tr}(\Sigma(\Sigma + tI)^{-1})$. 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 $\frac{140\kappa}{n}\log\frac{16\kappa n}{\delta} \leq t \leq \|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})}$,

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

(iii) For any $\delta > 0$ with $m \geq \left(2 \vee \frac{1024\kappa^2}{\sum_{i>\ell}\lambda_i^2}\right)\log\frac{2}{\delta}$,

$$\Lambda^m\left\{(\theta_i)_{i=1}^m : \frac{1}{4}\sum_{i>\ell}\lambda_i^2 \leq T^\ell(\Sigma_m) \leq 4\sum_{i>\ell}\lambda_i^2 + \frac{256\kappa^2\log\frac{2}{\delta}}{m}\right\} \geq 1 - 6\delta.$$

(iv) For any $\delta > 0$ with $n \geq 2\log\frac{2}{\delta}$, $m \geq \left(2 \vee \frac{1024\kappa^2}{\sum_{i>\ell}\lambda_i^2}\right)\log\frac{2}{\delta}$ 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 choice of $\left((X_i)_{i=1}^n, (\theta_j)_{j=1}^m\right)$:

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

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

Since Σ is trace class it is obvious that $\lambda_\ell \to 0$ and $\sum_{i>\ell}\lambda_i^2 \to 0$ as $\ell \to \infty$. It therefore follows that $T^\ell(\Sigma) \to 0$ and $T^\ell(\Sigma_m) \to 0$ as $\ell, m \to \infty$. Further, by assuming a decay rate on $(\lambda_i)_i$, a convergence rate for $T^\ell(\Sigma)$ and $T^\ell(\Sigma_m)$ may be obtained. Note that up to constants, $T^\ell(\Sigma)$ and $T^\ell(\Sigma_m)$ will have the same statistical behavior if m is chosen to be large enough that $\sum_{i>\ell}\lambda_i^2$ dominates $\frac{1}{m}$. As in Theorem 4.1, the behavior of the empirical varieties depend 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), and captures the complexity of $\mathcal{H}$. Since $\mathcal{N}_\Sigma(t) \lesssim \frac{1}{t}$ (better bound can be obtained if a certain decay rate for $(\lambda_i)_i$ is assumed), it is easy to see that $T^\ell(\widehat{\Sigma}) \to 0$ if $\ell, n \to 0$ and $n\lambda_\ell^2 \to \infty$, and $T^\ell(\widehat{\Sigma}_m) \to 0$ if $\ell, m, n \to 0$ and $\lambda_\ell^2(m \wedge n) \to \infty$. However, in order to properly compare the behavior of EKPCA and RF-EKPCA to each other, as well as to their population counterparts, an assumption on the decay rate of $(\lambda_i)_i$ must be made, and the trade-off between t, λ_ℓ and $\mathcal{N}_\Sigma(t)$ must be explored. The following corollaries to Theorem 4.1, proved in Sections 4.5.2 and 4.5.3, investigate the statistical behavior of EKPCA and RF-EKPCA in detail under the polynomial and exponential decay condition on the eigenvalues of Σ.

Corollary 4.1 (Polynomial decay of eigenvalues)**.** *Suppose* $\underline{A} i^{-\alpha} \leq \lambda_i \leq \bar{A} i^{-\alpha}$ *for* $\alpha > 1$ *and* $\underline{A}, \bar{A} \in (0, \infty)$*. Let* $\ell = n^{\frac{\theta}{\alpha}}$*,* $0 < \theta \leq \alpha$*. 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 $\tilde{n} \in \mathbb{N}$ *such that for all* $n > \tilde{n}$*, the following hold:*

(ii)

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

(iii)

$$n^{-2\theta(1-\frac{1}{2\alpha})} \mathbb{1}_{\left\{\gamma \geq \theta\left(2-\frac{1}{\alpha}\right)\right\}} \lesssim_{\Lambda^m} T^\ell(\Sigma_m) \lesssim_{\Lambda^m} \begin{cases} n^{-2\theta(1-\frac{1}{2\alpha})}, & \gamma \geq \theta\left(2-\frac{1}{\alpha}\right),\ \theta \leq \frac{\alpha}{2\alpha-1} \\ n^{-\gamma}, & \gamma \leq 1 \wedge \theta\left(2-\frac{1}{\alpha}\right) \end{cases}$$

with $m = n^\gamma$ *for* $0 < \gamma \leq 1$*;*

(iv)

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

with $m = n^\gamma$ *for* $0 < \gamma \leq 1$.

Remark 4.1. (*i*) *The condition* $\alpha > 1$ *is required to ensure that* Σ *is trace class. Comparing the behavior of* $T^\ell(\widehat{\Sigma})$ *to that of* $T^\ell(\Sigma)$ *it is clear that EKPCA recovers optimal convergence rates (compared to that of KPCA) if* ℓ *grows to infinity not faster than* $n^{1/(2\alpha-1)}$. *Since the reconstruction error is based on* ℓ *eigenfunctions, the computational complexity of EKPCA behaves as* $O(n^2\ell) = O(n^{2+\frac{\theta}{\alpha}})$. *It is important to note that* $0 < \theta \leq \frac{\alpha}{2\alpha-1}$ *is the only useful region both computationally and statistically as* $\theta > \frac{\alpha}{2\alpha-1}$ *does not improve the statistical rates (than that achieved at* $\theta = \frac{\alpha}{2\alpha-1}$*) but increases the computational complexity.*

(*ii*) *Comparing* $T^\ell(\widehat{\Sigma}_m)$ *with* $T^\ell(\widehat{\Sigma})$ *it is clear that if* ℓ *grows to infinity not faster than* $n^{1/(2\alpha-1)}$ *and the number of random features* m *grows sufficiently fast, then RF-EKPCA and EKPCA enjoy the same statistical behavior. The rate at which the number of random features must grow depends on the growth of* ℓ *through* θ *and* α*; the choice of* $1 \geq \gamma \geq \theta\left(2 - \frac{1}{\alpha}\right)$ *yields the same statistical behavior for RF-EKPCA, EKPCA, and KPCA.*

(*iii*) *The computational complexity of RF-EKPCA is given by* $O(m^2\ell + m^2 n) = O(n^{2\gamma+1})$ *which is better than that of EKPCA if* $\gamma < \frac{1}{2} + \frac{\theta}{2\alpha}$. *This means, RF-EKPCA has a lower computational complexity with similar statistical behavior to that of EKPCA if* $2\theta - \frac{\theta}{\alpha} \leq \gamma < \frac{1}{2} + \frac{\theta}{2\alpha}$ *and* $\theta \leq \frac{\alpha}{2\alpha-1}$ *respectively, which implies* $\theta < \frac{\alpha}{4\alpha-3}$. *In other words, if* ℓ *grows at a lower order than* $n^{1/(4\alpha-3)}$ *and the number of random features are larger than* $n^{\theta(2-\frac{1}{\alpha})}$*, then RF-EKPCA enjoys computational superiority with no loss in statistical performance over that of EKPCA.*

On the other hand, if ℓ *grows at an order faster than* $n^{1/(4\alpha-3)}$ *but not faster than* $n^{1/(2\alpha-1)}$*, it results in loss of computational advantage for RF-EKPCA while retaining the same statistical behavior to that of EKPCA—in fact, the rate in this regime is faster than in the previous regime of* $\theta < \frac{\alpha}{4\alpha-3}$.

Corollary 4.2 (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)

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

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)

$$n^{-2\theta} \mathbb{1}_{\{\gamma \geq 2\theta\}} \lesssim_{\Lambda^m} T^{\ell}(\Sigma_m) \lesssim_{\Lambda^m} \begin{cases} n^{-2\theta}, & \gamma \geq 2\theta,\, \theta \leq \frac{1}{2} \\ n^{-\gamma}, & \gamma \leq 1 \wedge 2\theta \end{cases}$$

with $m = n^{\gamma}$ *for* $0 < \gamma \leq 1$*;*

(iv)

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

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

We may draw conclusions similar to Remark 4.1 from Corollary 4.2. The behavior of $T^{\ell}(\Sigma)$ matches that of $T^{\ell}(\widehat{\Sigma})$, up to a $\log n$ factor, if ℓ grows slower than $\log \sqrt{n}$. If $m \geq n^{2\theta}$ with $\theta \leq \frac{1}{2}$, then RF-EKPCA and EKPCA have similar statistical convergence behavior (i.e., no statistical loss) but with RF-EKPCA enjoying a computational edge if $m < \sqrt{n \log n^{\theta}}$, i.e., $\theta \leq \frac{1}{4}$.

4.3 Embed and Reconstruct (E-R)

Based on the idea demonstrated in (4.5) where the objects are first embedded in $L^2(\mathbb{P})$ and then reconstructed in $L^2(\mathbb{P})$, we define the reconstruction error for EKPCA, RF-KPCA

and RF-EKPCA as

$$S^{\ell}(\widehat{\Sigma}) = \mathbb{E}_{X\sim\mathbb{P}}\left\|\mathfrak{I}\overline{k}(\cdot,X) - \sum_{i=1}^{\ell}\left\langle \mathfrak{I}\widetilde{k}(\cdot,X), \frac{\mathfrak{I}\widehat{\phi}_i}{\sqrt{\widehat{\lambda}_i}}\right\rangle_{L^2(\mathbb{P})}\frac{\mathfrak{I}\widehat{\phi}_i}{\sqrt{\widehat{\lambda}_i}}\right\|^2_{L^2(\mathbb{P})}, \tag{4.9}$$

$$S^{\ell}(\Sigma_m) = \mathbb{E}_{X\sim\mathbb{P}}\left\|\mathfrak{I}\overline{k}(\cdot,X) - \sum_{i=1}^{\ell}\left\langle \mathfrak{A}\overline{k}_m(\cdot,X), \frac{\mathfrak{A}\phi_{m,i}}{\sqrt{\lambda_{m,i}}}\right\rangle_{L^2(\mathbb{P})}\frac{\mathfrak{A}\phi_{m,i}}{\sqrt{\lambda_{m,i}}}\right\|^2_{L^2(\mathbb{P})}, \tag{4.10}$$

and

$$S^{\ell}(\widehat{\Sigma}_m) = \mathbb{E}_{X\sim\mathbb{P}}\left\|\mathfrak{I}\overline{k}(\cdot,X) - \sum_{i=1}^{\ell}\left\langle \mathfrak{A}\widetilde{k}_m(\cdot,X), \frac{\mathfrak{A}\widehat{\phi}_{m,i}}{\sqrt{\widehat{\lambda}_{m,i}}}\right\rangle_{L^2(\mathbb{P})}\frac{\mathfrak{A}\widehat{\phi}_{m,i}}{\sqrt{\widehat{\lambda}_{m,i}}}\right\|^2_{L^2(\mathbb{P})} \tag{4.11}$$

respectively. Note that unlike in R-E where the reconstructed functions (before being embedded) are computable (see (4.6)–(4.8)), in the above definitions, the reconstructed functions are not computable because of their dependence on unknown $\mathbb{P}$. Due to this reason, another version of reconstruction error for E-R can be considered where the first argument in $\langle\cdot,\cdot\rangle_{L^2(\mathbb{P})}$ in (4.9)–(4.11) can be replaced by $\mathfrak{I}\overline{k}(\cdot,X)$, which corresponds to reconstructing $\mathfrak{I}\overline{k}(\cdot,X)$ by projecting it along different systems, i.e., $(\mathfrak{I}\phi_i\lambda_i^{-1/2})_i$, $(\mathfrak{A}\phi_{m,i}\lambda_{m,i}^{-1/2})_i$ and $(\mathfrak{A}\widehat{\phi}_{m,i}\widehat{\lambda}_{m,i}^{-1/2})_i$. Clearly, the resulting definitions also yield non-computable reconstructed functions. We discuss this variation in Section 4.4.1. It is important to note that unlike with $(\mathfrak{I}\phi_i\lambda_i^{-1/2})_i$ and $(\mathfrak{A}\phi_{m,i}\lambda_{m,i}^{-1/2})_i$ which form ONS in $L^2(\mathbb{P})$, $(\mathfrak{I}\widehat{\phi}_i\widehat{\lambda}_i^{-1/2})_i$ and $(\mathfrak{A}\widehat{\phi}_{m,i}\widehat{\lambda}_{m,i}^{-1/2})_i$ do not form an ONS in $L^2(\mathbb{P})$. Therefore, one may wonder whether the definitions of $S^{\ell}(\widehat{\Sigma})$ and $S^{\ell}(\widehat{\Sigma}_m)$ make sense. We show below that $S^{\ell}(\widehat{\Sigma})$ and $S^{\ell}(\widehat{\Sigma}_m)$ tend to zero as $n\to\infty$, $\ell\to\infty$ and $m\to\infty$ with appropriate conditions on l, m and n. This means, asymptotically $(\mathfrak{I}\widehat{\phi}_i\widehat{\lambda}_i^{-1/2})_i$ and $(\mathfrak{A}\widehat{\phi}_{m,i}\widehat{\lambda}_{m,i}^{-1/2})_i$ capture all the information about $\mathfrak{I}\overline{k}(\cdot,X)$ by forming an ONS in $L^2(\mathbb{P})$.

The following result provides a finite-sample bound on the reconstruction error, using which convergence rates can be obtained.

Theorem 4.2. *Suppose Assumptions 3.1, 3.5, 3.6, 4.1, and 4.2 hold. For any $t>0$, define $\mathcal{N}_{\Sigma}(t) = \operatorname{tr}(\Sigma(\Sigma+tI)^{-1})$. Then the following hold:*

(i)

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

(ii) For any $\delta > 0$ with $n \geq 2\log\frac{2}{\delta}$ and $\frac{140\kappa}{n}\log\frac{16\kappa n}{\delta} \leq t \leq \frac{\lambda_\ell}{3}$, with probability at least $1 - 11\delta$ over the choice of $(X_i)_{i=1}^n$,

$$\sum_{i>\ell}\lambda_i^2 \leq S^\ell(\widehat{\Sigma}) \lesssim \mathcal{N}_\Sigma(t)(\lambda_{\ell+1}+t)^2 + \kappa^{5/2}\log\frac{2}{\delta}\left[\frac{\mathcal{N}_\Sigma(t)}{n\sqrt{t}} \wedge \frac{\kappa^{3/2}}{nt}\right] + \frac{\kappa^3(\kappa\wedge 1)\log^2\frac{3}{\delta}}{n^2 t} + \frac{\kappa^2\log\frac{2}{\delta}}{n}.$$

(iii) For any $\delta > 0$ with $m \geq \left(2 \vee \frac{1024\kappa^2}{\sum_{i>\ell}\lambda_i^2}\right)\log\frac{2}{\delta}$,

$$\Lambda^m\left\{(\theta_i)_{i=1}^m : \frac{1}{4}\sum_{i>\ell}\lambda_i^2 \leq S^\ell(\Sigma_m) \leq 4\sum_{i>\ell}\lambda_i^2 + \frac{256\kappa^2\log\frac{2}{\delta}}{m}\right\} \geq 1 - 6\delta.$$

(iv) For any $\delta > 0$ with $n \geq 2\log\frac{2}{\delta}$, $m \geq \left(2 \vee \frac{1024\kappa^2}{\sum_{i>\ell}\lambda_i^2}\right)\log\frac{2}{\delta}$ and

$$\frac{140\kappa}{n}\log\frac{16\kappa n}{\delta} \vee \frac{86\kappa}{m}\log\frac{16\kappa m}{\delta} \leq t \leq \frac{\lambda_\ell}{9},$$

with probability at least $1 - 26\delta$ over the choice of $\left((X_i)_{i=1}^n, (\theta_j)_{j=1}^m\right)$:

$$\frac{1}{4}\sum_{i>\ell}\lambda_i^2 \leq S^\ell(\widehat{\Sigma}_m) \lesssim \mathcal{A}_2(t)(\lambda_{\ell+1}+t)^2 + \kappa^{5/2}\log\frac{2}{\delta}\left[\frac{\mathcal{A}_2(t)}{n\sqrt{t}} \wedge \frac{\kappa^{3/2}}{nt}\right] + \frac{\kappa^3(1\wedge\kappa)\log^2\frac{3}{\delta}}{n^2 t} + \frac{\kappa^2\log\frac{2}{\delta}}{n} + \frac{\kappa^2\log\frac{2}{\delta}}{m},$$

where $\mathcal{A}_2(t) := \frac{\kappa\log\frac{2}{\delta}}{tm} + \sqrt{\frac{\kappa\mathcal{N}_\Sigma(t)\log\frac{2}{\delta}}{tm}} + \mathcal{N}_\Sigma(t)$.

Remark 4.2. (i) *By comparing Theorems 4.1 and 4.2 we note that the population reconstruction error and its approximation in R-E (i.e., $T^\ell(\Sigma)$, $T^\ell(\Sigma_m)$) and E-R (i.e., $S^\ell(\Sigma)$ and $S^\ell(\Sigma_m)$) match. However, their sample counterparts in R-E (i.e., $T^\ell(\widehat{\Sigma})$, $T^\ell(\widehat{\Sigma}_m)$) and E-R (i.e., $S^\ell(\widehat{\Sigma})$, $S^\ell(\widehat{\Sigma}_m)$) exhibit different behavior, implying that the*

embedding and projections operations are not commutative on the data.

(ii) *Another key difference of Theorem 4.2 to Theorem 4.1 is the upper bound on t. While t is upper bounded by a constant in Theorem 4.1, t is upper bounded by λ_ℓ (up to constants) in Theorem 4.2, which enforces a lower bound on λ_ℓ. Since $\lambda_\ell \to 0$ as $\ell \to \infty$ and the lower bound on λ_ℓ converges to zero as $n \to \infty$, this enforces a constraint on λ_ℓ to not converge to zero too fast. In other words, it imposes a condition on ℓ to not grow too fast with n.*

(iii) *The explicit universal constants, which are suppressed in* (ii) *and* (iv) *of Theorem 4.2 for brevity, are provided in the proof. It is clear from* (ii) *and* (iv) *of Theorem 4.2 that for m large enough, both EKPCA and RF-EKPCA have similar statistical behavior—a similar observation was made in Theorem 4.1.*

The following corollaries to Theorem 4.2 investigate the statistical behavior of EKPCA and RF-EKPCA in detail under the polynomial and exponential decay condition on the eigenvalues of Σ.

Corollary 4.3 (Polynomial decay of eigenvalues)**.** *Suppose $\underline{A} i^{-\alpha} \leq \lambda_i \leq \bar{A} i^{-\alpha}$ for $\alpha > 1$ and $\underline{A}, \bar{A} \in (0, \infty)$. Let $\ell = n^{\frac{\theta}{\alpha}}$, $0 < \theta \leq \alpha$. Define $\frac{1}{\alpha'} := \left(\frac{1}{\alpha} + \frac{1}{2}\right) \wedge 1$ and $\beta := \frac{1}{2 + \frac{1}{\alpha'} - \frac{1}{\alpha}}$. Then*

(*i*)

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

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

(*ii*)

$$n^{-2\theta(1-\frac{1}{2\alpha})} \lesssim S^\ell(\widehat{\Sigma}) \lesssim_{\mathbb{P}^n} \begin{cases} n^{-2\theta(1-\frac{1}{2\alpha})}, & \theta \leq \beta \\ n^{-\left(1-\frac{\theta}{\alpha'}\right)}, & \beta \leq \theta < 1 \end{cases};$$

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

$$n^{-2\theta(1-\frac{1}{2\alpha})}\mathbb{1}_{\left\{\gamma\geq\theta\left(2-\frac{1}{\alpha}\right)\right\}} \lesssim_{\Lambda^m} S^\ell(\Sigma_m) \lesssim_{\Lambda^m} \begin{cases} n^{-2\theta(1-\frac{1}{2\alpha})}, & \gamma \geq \theta\left(2-\frac{1}{\alpha}\right),\ \theta \leq \frac{\alpha}{2\alpha-1} \\ n^{-\gamma}, & \gamma \leq 1 \wedge \theta\left(2-\frac{1}{\alpha}\right) \end{cases};$$

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

$$n^{-2\theta(1-\frac{1}{2\alpha})}\mathbb{1}_{\left\{\gamma\geq\theta\left(2-\frac{1}{\alpha}\right)\right\}} \lesssim_{\Lambda^m} S^\ell(\widehat{\Sigma}_m) \lesssim_{\mathbb{P}^n\times\Lambda^m} \begin{cases} n^{-2\theta(1-\frac{1}{2\alpha})}, & \gamma \geq \theta\left(2-\frac{1}{\alpha}\right),\ \theta \leq \beta \\ n^{-\left(1-\frac{\theta}{\alpha'}\right)}, & \gamma \geq 1-\frac{\theta}{\alpha'},\ \gamma > \theta,\ \beta \leq \theta < 1 \\ n^{-\gamma}, & \theta < \gamma \leq \left[1 \wedge \theta\left(2-\frac{1}{\alpha}\right) \wedge \left(1-\frac{\theta}{\alpha'}\right)\right] \end{cases}.$$

Corollary 4.4 (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)

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

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

(ii)

$$n^{-2\theta} \lesssim S^\ell(\widehat{\Sigma}) \lesssim_{\mathbb{P}^n} \begin{cases} n^{-2\theta}\log n, & \theta \leq \frac{2}{5} \\ n^{-\left(1-\frac{\theta}{2}\right)}\log n, & \frac{2}{5} \leq \theta < 1 \end{cases};$$

(iii)

$$n^{-2\theta}\mathbb{1}_{\{\gamma\geq 2\theta\}} \lesssim_{\Lambda^m} S^\ell(\Sigma_m) \lesssim_{\Lambda^m} \begin{cases} n^{-2\theta}, & \gamma \geq 2\theta,\ \theta \leq \frac{1}{2} \\ n^{-\gamma}, & \gamma \leq 1 \wedge 2\theta \end{cases}$$

with $m = n^\gamma$ *for* $0 < \gamma \leq 1$;

(*iv*)

$$n^{-2\theta}\mathbb{1}_{\{\gamma\geq 2\theta\}} \lesssim_{\Lambda^m} S^{\ell}(\widehat{\Sigma}_m) \lesssim_{\mathbb{P}^n\times\Lambda^m} \begin{cases} n^{-2\theta}\log n, & \gamma \geq 2\theta,\ \theta \leq \frac{2}{5} \\ n^{-\left(1-\frac{\theta}{2}\right)}\log n, & \gamma \geq 1-\frac{\theta}{2},\ \gamma > \theta,\ \frac{2}{5} \leq \theta < 1 \\ n^{-\gamma}, & \theta < \gamma \leq \left[1 \wedge 2\theta \wedge \left(1-\frac{\theta}{2}\right)\right] \end{cases}$$

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

Remark 4.3. (*i*) *In Corollary 4.3, note that* $\beta = \frac{\alpha}{3\alpha-1}$ *for* $1 < \alpha \leq 2$ *and* $\beta = \frac{2}{5}$ *for* $\alpha \geq 2$ *with the best convergence rate for* $S^{\ell}(\widehat{\Sigma})$ *being attained at* $\theta = \beta$ *as the rate is a convex function of* θ. *Clearly, from both computational and statistical view points, only the range of* $0 < \theta \leq \beta$ *is interesting and useful as* $\theta > \beta$ *yields similar/slower convergence rates with more computational complexity. This means, optimal convergence rates are obtained for EKPCA and RF-EKPCA for* ℓ *not growing faster than* $n^{\theta/\alpha}$, $\theta \leq \beta$ *and* $m \geq n^{\theta(2-\frac{1}{\alpha})}$. *Similar observation holds for Corollary 4.4 as well where* $0 < \theta \leq \frac{2}{5}$ *is the statistically and computationally useful range with EKPCA and RF-EKPCA achieving almost optimal convergence rates (up to a logarithmic term). Arguing as in Remark 4.1*(*iii*), *it can be shown that the computational complexity of RF-EKPCA is better than of EKPCA and with no loss in statistical performance if* $\theta < \frac{\alpha}{4\alpha-3} \wedge \beta$ *and* $\gamma \geq \theta(2-\frac{1}{\alpha})$ *for the case of polynomial decay and if* $\theta < \frac{1}{4}$ *with* $\gamma \geq 2\theta$ *for the exponential decay.*

(*ii*) *More interesting observations can be made by comparing Corollaries 4.1 (*resp. *Corollary 4.2) and 4.3 (*resp. *Corollary 4.4). First, EKPCA and RF-EKPCA have different upper asymptotic behaviors in R-E (Corollaries 4.1, 4.2) and E-R (Corollary 4.3, 4.4). Particularly, while the reconstruction error rate improves with increase in* θ *in both the cases, it saturates beyond a certain* θ *in the case of R-E while it decreases in the case of E-R. This latter behavior is due to the inverse of empirical eigenvalues that appear in* $S^{\ell}(\widehat{\Sigma})$ *and* $S^{\ell}(\widehat{\Sigma}_m)$*—as* ℓ *becomes large, then inverse of the empirical eigenvalues make large contributions to the error, resulting in slower convergence rates. In the regimes of* θ *where R-E and E-R behave similarly (for both EKPCA and RF-EKPCA), we note that* θ *has a larger upper bound (i.e.,* ℓ *can have faster growth) in R-E than in E-R,*

which again relates to the above mentioned issue of the inverse of empirical eigenvalues. Particularly, in the case of E-R, for $\alpha \leq 2$, RF-EKPCA has better computational behavior than EKPCA if $\theta < \beta$ and $\gamma \geq \theta(2 - \frac{1}{\alpha})$ while such a result holds for R-E for a wider range of θ, i.e., $\theta < \frac{\alpha}{4\alpha-3}$, which means faster growth for ℓ is allowable for R-E without losing computational or statistical efficiency. On the other hand, for $\alpha \geq 2$, R-E and E-R behave similarly for $0 < \theta \leq \frac{\alpha}{4\alpha-3}$.

4.4 Extensions

In this section, we discuss other notions of reconstruction error and their computational vs. statistical behavior. First, we discuss a variant of E-R that was introduced in the paragraph following (4.11) and show that it has similar statistical behavior to that of E-R. Next, we consider a variation of R-E (this variation can also be extended to E-R) and discuss its behavior.

4.4.1 A variation of E-R

As mentioned before, a variation of E-R can be obtained by replacing the first argument of $\langle \cdot, \cdot \rangle_{L^2(\mathbb{P})}$ in (4.9)–(4.11) by $\mathfrak{I}\overline{k}(\cdot, X)$, which corresponds to reconstructing $\mathfrak{I}\overline{k}(\cdot, X)$ by projecting it along different systems, i.e., $(\mathfrak{I}\phi_i\lambda_i^{-1/2})_i$, $(\mathfrak{A}\phi_{m,i}\lambda_{m,i}^{-1/2})_i$ and $(\mathfrak{A}\widehat{\phi}_{m,i}\widehat{\lambda}_{m,i}^{-1/2})_i$. By referring to the resulting reconstruction errors as $W^\ell(\widehat{\Sigma})$, $W^\ell(\Sigma_m)$ and $W^\ell(\widehat{\Sigma}_m)$ respectively, we obtain the following result. This result, proved in Section 4.5.7, shows that E-R and its variant are statistically equivalent.

Theorem 4.3. *Under assumptions 3.1, 3.5, 3.6, 4.1, and 4.2, the following hold:*

(i) $\sum_{i>\ell} \lambda_i^2 \leq W^\ell(\widehat{\Sigma}) \lesssim_{\mathbb{P}^n} S^\ell(\widehat{\Sigma}) + \frac{1}{n}$;

(ii) $\sum_{i>\ell} \lambda_i^2 \lesssim_{\Lambda^m} W^\ell(\Sigma_m) \lesssim_{\Lambda^m} S^\ell(\Sigma_m) + \frac{1}{m}$ *for* $m \gtrsim \frac{1}{\sum_{i>\ell}\lambda_i^2}$;

(iii) $\sum_{i>\ell} \lambda_i^2 \lesssim_{\Lambda^m} W^\ell(\widehat{\Sigma}_m) \lesssim_{\mathbb{P}^n \times \Lambda^m} S^\ell(\widehat{\Sigma}_m) + \frac{1}{n} + \frac{1}{m}$ *for* $m \gtrsim \frac{1}{\sum_{i>\ell}\lambda_i^2}$.

4.4.2 Schatten norms

So far, we have seen that the population reconstruction error in E-R and R-E, i.e., $S^\ell(\Sigma)$ and $T^\ell(\Sigma)$ behave as $\sum_{i>\ell}\lambda_i^2$, which is the squared ℓ_2-norm of $\lambda_\ell := (\lambda_{\ell+1}, \lambda_{\ell+2}, \ldots)$. Of course, if we use the population reconstruction error defined in (4.4), it is easy to show that it behaves as $\sum_{i>\ell}\lambda_i$, which is the ℓ_1-norm of λ_ℓ. But the reconstruction error defined in (4.4) is not useful for our purpose because of the aforementioned technical issues and that is why we introduced E-R and R-E in Sections 4.2 and 4.3. In this section, we explore an extension of $T^\ell(\Sigma)$ (similar extension holds for $S^\ell(\Sigma)$ as well) which yields different norms of λ_ℓ. To this end, define

$$T^\ell(\Sigma, s) = \mathbb{E}_{X\sim\mathbb{P}}\left\|(\mathfrak{I}\mathfrak{I}^*)^{-s/2}\left[\mathfrak{I}\overline{k}(\cdot, X) - \mathfrak{I}\left(\sum_{i=1}^{\ell}\left\langle \overline{k}(\cdot, X), \phi_i\right\rangle_{\mathcal{H}}\phi_i\right)\right]\right\|^2_{L^2(\mathbb{P})}, \tag{4.12}$$

where $(\mathfrak{I}\mathfrak{I}^*)^{-1}$ is treated as the inverse of $\mathfrak{I}\mathfrak{I}^*$ restricted to $L^2(\mathbb{P})\backslash\text{Null}(\mathfrak{I}\mathfrak{I}^*)$ where $\text{Null}(\mathfrak{I}\mathfrak{I}^*) = \text{Null}(\mathfrak{I}^*) = \{f \in L^2(\mathbb{P}) : f \text{ is a constant a.s.–}\mathbb{P}\}$. Theorem 4.4 shows $T^\ell(\Sigma, s)$ to behave as a certain p-norm of λ_ℓ, with p being controlled by s.

The empirical counterpart of $T^\ell(\Sigma, s)$, denoted as $T^\ell(\widehat{\Sigma}, s)$ can be defined by replacing $\overline{k}$ and ϕ_i with $\widetilde{k}$ and $\widehat{\phi}_i$ respectively in (4.12). Similarly, the reconstruction error of RF-KPCA can be defined as

$$T^\ell(\Sigma_m, s) = \mathbb{E}_{X\sim\mathbb{P}}\left\|(\mathfrak{I}\mathfrak{I}^*)^{-s/2}\,\mathfrak{I}\overline{k}(\cdot, X) - (\mathfrak{A}\mathfrak{A}^*)^{-s/2}\,\mathfrak{A}\left(\sum_{i=1}^{\ell}\left\langle \overline{k}_m(\cdot, X), \phi_{m,i}\right\rangle_{\mathcal{H}_m}\phi_{m,i}\right)\right\|^2_{L^2(\mathbb{P})}$$

with its empirical counterpart $T^\ell(\widehat{\Sigma}_m, s)$ defined by replacing $\overline{k}_m$ with $\widetilde{k}_m$ and $\phi_{m,i}$ with $\widehat{\phi}_{m,i}$. The following result, proved in Section 4.5.8, provides the probabilistic behavior of these generalized reconstruction errors.

Theorem 4.4. *Suppose Assumptions 3.1, 3.5, 3.6, 4.1, and 4.2 hold. For any $t > 0$, define $\mathcal{N}_\Sigma(t) = \text{tr}(\Sigma(\Sigma + tI)^{-1})$. Then the following hold:*

(i) For any $s \le 1$,

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

(ii) For $\frac{\log n}{n} \lesssim t \lesssim \|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})}$ *and any* $s \leq 1$,

$$\sum_{i>\ell} \lambda_i^{2-s} \leq T^\ell(\widehat{\Sigma}, s) \lesssim_{\mathbb{P}^n} \frac{\mathcal{N}_\Sigma(t)(\lambda_{\ell+1}+t)^2}{t^s} + \frac{1}{n}.$$

(iii) For $m \gtrsim \left[\left(\sum_{i>\ell} \lambda_i^{2-s}\right)^{\frac{2}{s-2}} \vee \left(\sum_{i>\ell} \lambda_i^{2-s}\right)^{\frac{2}{s}}\right] \mathbb{1}_{[-2,0)}(s) + \left(\sum_{i>\ell} \lambda_i^2\right)^{-1} \mathbb{1}_{\{0\}}(s)$,

$$\sum_{i>\ell} \lambda_i^{2-s} \lesssim_{\Lambda^m} T^\ell(\Sigma_m, s) \lesssim_{\Lambda^m} \sum_{i>\ell} \lambda_i^{2-s} + m^{s/2}\mathbb{1}_{[-2,0)}(s) + \frac{1}{m}\mathbb{1}_{\{0\}}(s).$$

(iv) For $m \gtrsim \left[\left(\sum_{i>\ell} \lambda_i^{2-s}\right)^{\frac{2}{s-2}} \vee \left(\sum_{i>\ell} \lambda_i^{2-s}\right)^{\frac{2}{s}}\right] \mathbb{1}_{[-2,0)}(s) + \left(\sum_{i>\ell} \lambda_i^2\right)^{-1} \mathbb{1}_{\{0\}}(s)$ *and* $\frac{\log n}{n} \vee \frac{\log m}{m} \lesssim t \lesssim \|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})}$,

$$\sum_{i>\ell} \lambda_i^{2-s} \lesssim_{\Lambda^m} T^\ell(\widehat{\Sigma}_m, s) \lesssim_{\Lambda^m \times \mathbb{P}^n} \frac{\mathcal{A}_1(t)(\lambda_{\ell+1}+t)^2}{t^s} + \frac{1}{n} + m^{s/2}\mathbb{1}_{[-2,0)}(s) + \frac{1}{m}\mathbb{1}_{\{0\}}(s),$$

where $\mathcal{A}_1(t) := \mathcal{N}_\Sigma(t) + \frac{1}{tm} + \sqrt{\frac{\mathcal{N}_\Sigma(t)}{tm}}$.

Remark 4.4. *(i) The restriction of* $s \leq 1$ *for* $T^\ell(\Sigma, s)$ *appears because* Σ *is a trace class. On the other hand, the bounds for* $T^\ell(\Sigma_m, s)$ *and* $T^\ell(\widehat{\Sigma}_m, s)$ *hold only for* $s \in [-2, 0]$. *This could be an artifact of the analysis as the proof of these bounds involve bounding* $\|(\mathfrak{I}\mathfrak{I}^*)^{-s/2} - (\mathfrak{A}\mathfrak{A}^*)^{-s/2}\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))}$ *by* $\|\mathfrak{I}\mathfrak{I}^* - \mathfrak{A}\mathfrak{A}^*\|^{-s/2}_{\mathcal{L}^\infty(L^2(\mathbb{P}))}$ *by using the operator monotonicity of the map* $x \mapsto x^{-s/2}$ *for* $0 \leq -s/2 \leq 1$, *i.e.,* $s \in [-2, 0]$. *Note that this range of* s *yields weaker notions of reconstruction error as it corresponds to Schatten norms of order greater than 2, with the more interesting range being* $[0, 1]$. *However, the current bounds for* $T^\ell(\Sigma_m, s)$ *and* $T^\ell(\widehat{\Sigma}_m, s)$ *do not hold for* s *taking values in this interesting range. Also note that* $s = 0$ *exactly reduces to Theorem 4.1.*

(ii) As before, assuming $i^{-\alpha} \lesssim \lambda_i \lesssim i^{-\alpha}$ *for* $\alpha > 1$ *and* $\ell = n^{\frac{\theta}{\alpha}}$, $0 < \theta \leq \alpha$, *it follows that* $n^{-2\theta\left(1-\frac{1}{2\alpha}-\frac{s}{2}\right)} \lesssim T^\ell(\Sigma, s) \lesssim n^{-2\theta\left(1-\frac{1}{2\alpha}-\frac{s}{2}\right)}$ *and*

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

which matches with Corollary 4.1(i, ii) *for* $s = 0$. *These show the convergence rate to be slower for* $0 < s \leq 1$ *compared to* $s = 0$, *which is expected as the former corresponds to a stronger notion of reconstruction error. Also* $0 < \theta \leq \frac{\alpha}{2\alpha-1-\alpha s}$ *is the only useful range both computationally and statistically as* $\theta > \frac{\alpha}{2\alpha-1-\alpha s}$ *does not improve the statistical rates but increases the computational complexity. On the other hand for* $-2 \leq s \leq 0$, *it follows that*

$$n^{-2\theta\left(1-\frac{1}{2\alpha}-\frac{s}{2}\right)} \lesssim_{\Lambda^m} T^{\ell}(\widehat{\Sigma}_m, s) \lesssim_{\Lambda^m \times \mathbb{P}^n} n^{-2\theta(1-\frac{1}{2\alpha}-\frac{s}{2})},$$
$$\text{for } m \gtrsim n^{\frac{4\theta}{2-s}(1-\frac{1}{2\alpha}-\frac{s}{2})}, \;\; \theta \leq \frac{\alpha}{2\alpha-1-\alpha s},$$

which implies that for sufficiently large m, *EKPCA and RF-EKPCA have similar statistical behavior as long as* $\theta \leq \frac{\alpha}{2\alpha-1-\alpha s}$. *However, RF-EKPCA is computationally better than EKPCA only when* $0 < \theta < \frac{(2-s)\alpha}{8\alpha-6+s-4s\alpha}$. *Also note that for* $\theta = \frac{\alpha}{2\alpha-1-\alpha s}$, *which is where the best rate of* $\frac{1}{n}$ *is achieved for any* $s \in [-2, 0]$, *we obtain* $m \gtrsim n^{\frac{2}{2-s}}$, *i.e., the requirement on the number of random features is monotonically increasing w.r.t.* $s \in [-2, 0]$. *Since* θ *is an increasing function of* s, *it implies fewer* ℓ *is sufficient for optimal rates for smaller* s. *To elaborate, at the chosen value of* θ, *statistical optimality is conserved for RF-EKPCA at* $s = 0$ *if* $m \gtrsim n$ *while only* $m \gtrsim \sqrt{n}$ *is required at* $s = -2$. *This is understandable as smaller values of* s *result in weaker notions of reconstruction error as explained above.*

4.5 Proofs

4.5.1 Proof of Theorem 4.1

(i) Note that $T^{\ell}(\Sigma)$ can be succinctly written as

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

Since $\Sigma = \mathfrak{I}^*\mathfrak{I}$ and k is continuous and bounded (therefore, Bochner integrable), it follows from Lemma A.4 that

$$\begin{aligned} T^\ell(\Sigma) &= \left\|\Sigma^{1/2}(I - P^\ell(\Sigma))\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} = \|\Sigma - \Sigma_\ell\|^2_{\mathcal{L}^2(\mathcal{H})} \\ &= \operatorname{tr}((\Sigma - \Sigma_\ell)^2) = \operatorname{tr}(\Sigma^2 - \Sigma_\ell^2), \end{aligned}$$

where we used $\Sigma\Sigma_\ell = \Sigma_\ell\Sigma = \Sigma_\ell^2$ with $\Sigma_\ell := \sum_{i=1}^{\ell} \lambda_i \phi_i \otimes_{\mathcal{H}} \phi_i$ and the result follows.

(ii) *Upper bound:* Note that

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

Therefore, adding and subtracting $\mathfrak{I}P^\ell(\widehat{\Sigma})m_{\mathbb{P}}$, we obtain

$$\begin{aligned} T^\ell(\widehat{\Sigma}) = &\underbrace{\mathbb{E}\left\|\mathfrak{I}(I - P^\ell(\widehat{\Sigma}))\overline{k}(\cdot, X)\right\|^2_{L^2(\mathbb{P})}}_{①} + \underbrace{\left\|\mathfrak{I}P^\ell(\widehat{\Sigma})(m_{\mathbb{P}} - \hat{m}_{\mathbb{P}})\right\|^2_{L^2(\mathbb{P})}}_{②} \\ &-2\mathbb{E}\left\langle \mathfrak{I}(I - P^\ell(\widehat{\Sigma}))\overline{k}(\cdot, X), \mathfrak{I}P^\ell(\widehat{\Sigma})(m_{\mathbb{P}} - \hat{m}_{\mathbb{P}})\right\rangle_{L^2(\mathbb{P})}, \end{aligned} \tag{4.13}$$

where the third term is zero since $\mathbb{E}[\overline{k}(\cdot, X)] = 0$. It follows from Lemma A.4 that

$$① = \mathbb{E}\left\|\mathfrak{I}(I - P^\ell(\widehat{\Sigma}))\overline{k}(\cdot, X)\right\|^2_{L^2(\mathbb{P})} = \left\|\Sigma^{1/2}(I - P^\ell(\widehat{\Sigma}))\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})}. \tag{4.14}$$

For any $t > 0$, we have

$$\begin{aligned} &\left\|\Sigma^{1/2}(I - P^\ell(\widehat{\Sigma}))\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} \\ &= \Big\|\Sigma^{1/2}(\Sigma + tI)^{-1/2}(\Sigma + tI)^{1/2}(I - P^\ell(\widehat{\Sigma})) \\ &\qquad\qquad \times (\Sigma + tI)^{1/2}(\Sigma + tI)^{-1/2}\Sigma^{1/2}\Big\|^2_{\mathcal{L}^2(\mathcal{H})} \\ &\leq \left\|\Sigma^{1/2}(\Sigma + tI)^{-1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} \left\|\Sigma^{1/2}(\Sigma + tI)^{-1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \\ &\qquad\qquad \times \left\|(\Sigma + tI)^{1/2}(I - P^\ell(\widehat{\Sigma}))(\Sigma + tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \\ &\overset{(*)}{\leq} \mathcal{N}_\Sigma(t)\left\|(\Sigma + tI)^{1/2}(I - P^\ell(\widehat{\Sigma}))(\Sigma + tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \end{aligned}$$

$$
\begin{aligned}
&\leq \mathcal{N}_\Sigma(t) \left\|(\Sigma+tI)^{1/2}(\widehat{\Sigma}+tI)^{-1/2}\right\|^4_{\mathcal{L}^\infty(\mathcal{H})} \\
&\qquad\qquad \times \left\|(\widehat{\Sigma}+tI)^{1/2}(I-P^\ell(\widehat{\Sigma}))(\widehat{\Sigma}+tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})}, \\
&\leq \mathcal{N}_\Sigma(t)(\widehat{\lambda}_{\ell+1}+t)^2 \left\|(\Sigma+tI)^{1/2}(\widehat{\Sigma}+tI)^{-1/2}\right\|^4_{\mathcal{L}^\infty(\mathcal{H})}, \qquad (4.15)
\end{aligned}
$$

where in $(*)$, we have used $\left\|\Sigma^{1/2}(\Sigma+tI)^{-1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} = \mathrm{tr}((\Sigma+tI)^{-1/2}\Sigma(\Sigma+tI)^{-1/2}) = \mathrm{tr}(\Sigma(\Sigma+tI)^{-1}) = \mathcal{N}_\Sigma(t)$ and $\left\|\Sigma^{1/2}(\Sigma+tI)^{-1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \leq 1$. Applying Lemma A.2*(ii, iv)* to (4.15), we obtain that for any $\delta > 0$ such that $\frac{140\kappa}{n}\log\frac{16\kappa n}{\delta} \leq t \leq \|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})}$, with probability at least $1-2\delta$ over the choice of $(X_i)_{i=1}^n$,

$$
\begin{aligned}
① &= \left\|\Sigma^{1/2}(I-P^\ell(\widehat{\Sigma}))\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} \\
&\leq 4\mathcal{N}_\Sigma(t)(\widehat{\lambda}_{\ell+1}+t)^2 \leq 9\mathcal{N}_\Sigma(t)(\lambda_{\ell+1}+t)^2. \qquad (4.16)
\end{aligned}
$$

We now bound ② as follows.

$$
\begin{aligned}
② &= \left\langle \mathfrak{I}P^\ell(\widehat{\Sigma})(m_{\mathbb{P}}-\hat{m}_{\mathbb{P}}), \mathfrak{I}P^\ell(\widehat{\Sigma})(m_{\mathbb{P}}-\hat{m}_{\mathbb{P}})\right\rangle_{L^2(\mathbb{P})} \\
&= \left\langle \Sigma^{1/2}P^\ell(\widehat{\Sigma})(m_{\mathbb{P}}-\hat{m}_{\mathbb{P}}), \Sigma^{1/2}P^\ell(\widehat{\Sigma})(m_{\mathbb{P}}-\hat{m}_{\mathbb{P}})\right\rangle_{\mathcal{H}} \\
&= \left\|\Sigma^{1/2}P^\ell(\widehat{\Sigma})(m_{\mathbb{P}}-\hat{m}_{\mathbb{P}})\right\|^2_{\mathcal{H}} \\
&\leq \left\|\Sigma^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \left\|P^\ell(\widehat{\Sigma})\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \|m_{\mathbb{P}}-\hat{m}_{\mathbb{P}}\|^2_{\mathcal{H}} \\
&= \|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})} \|m_{\mathbb{P}}-\hat{m}_{\mathbb{P}}\|^2_{\mathcal{H}}, \qquad (4.17)
\end{aligned}
$$

where the last equality uses $\left\|P^\ell(\widehat{\Sigma})\right\|_{\mathcal{L}^\infty(\mathcal{H})} = 1$. The result follows by applying Lemma C.1*(i)* to (4.17), combining it with (4.16) in (4.13), and using $\|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})} \leq 2\kappa$.

<u>*Lower bound:*</u> It follows from (4.13) that $T^\ell(\widehat{\Sigma}) \geq ①$. Also $① \geq T^\ell(\Sigma)$ as we will show below that

$$
T^\ell(\Sigma) = \inf_{(\psi_i)_{i=1}^\ell \in A} \left\|\Sigma^{1/2}\left(I - Q_{\psi,\ell}\right)\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})},
$$

with $A = \left\{(\psi_i)_{i=1}^{\ell} \subset \mathcal{H} : \langle \psi_i, \psi_j \rangle_{\mathcal{H}} = \delta_{ij},\ \forall\, i,j \in [\ell]\right\}$ and $Q_{\psi,\ell} := \sum_{i=1}^{\ell} \psi_i \otimes_{\mathcal{H}} \psi_i$. Note that

$$
\begin{aligned}
&\left\|\Sigma^{1/2}\left(I - Q_{\psi,\ell}\right)\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} \\
&= \left\|\sum_i \lambda_i \phi_i \otimes_{\mathcal{H}} \phi_i - \sum_{i,j} \sqrt{\lambda_i}\sqrt{\lambda_j}\langle \phi_i, Q_{\psi,\ell}\phi_j\rangle_{\mathcal{H}} \phi_i \otimes_{\mathcal{H}} \phi_j \right\|^2_{\mathcal{L}^2(\mathcal{H})} \\
&= \left\|\sum_{i,j} \sqrt{\lambda_i\lambda_j}\delta_{ij}\phi_i \otimes_{\mathcal{H}} \phi_j - \sum_{i,j} \sqrt{\lambda_i\lambda_j}\langle \phi_i, Q_{\psi,\ell}\phi_j\rangle_{\mathcal{H}} \phi_i \otimes_{\mathcal{H}} \phi_j \right\|^2_{\mathcal{L}^2(\mathcal{H})} \\
&= \left\|\sum_{i,j} \sqrt{\lambda_i\lambda_j}\left(\delta_{ij} - \langle \phi_i, Q_{\psi,\ell}\phi_j\rangle_{\mathcal{H}}\right) \phi_i \otimes_{\mathcal{H}} \phi_j \right\|^2_{\mathcal{L}^2(\mathcal{H})} \\
&= \sum_{i,j} \lambda_i\lambda_j \left(\delta_{ij} - \langle \phi_i, Q_{\psi,\ell}\phi_j\rangle_{\mathcal{H}}\right)^2 \\
&= \sum_i \lambda_i^2 \left(1 - \langle \phi_i, Q_{\psi,\ell}\phi_i\rangle_{\mathcal{H}}\right)^2 + \sum_{i\neq j} \lambda_i\lambda_j \langle \phi_i, Q_{\psi,\ell}\phi_j\rangle^2_{\mathcal{H}} \\
&\geq \sum_i \lambda_i^2 \left(1 - \langle \phi_i, Q_{\psi,\ell}\phi_i\rangle_{\mathcal{H}}\right)^2 \qquad (4.18) \\
&= \sum_i \lambda_i^2 \left(1 - \sum_{j=1}^{\ell} \langle \phi_i, \psi_j\rangle^2_{\mathcal{H}}\right)^2. \qquad (4.19)
\end{aligned}
$$

The equality is achieved in (4.18) if $(\psi_i)_{i=1}^{\ell} \in A$ is chosen such that

$$\langle \phi_i, Q_{\psi,\ell}\phi_j\rangle_{\mathcal{H}} = 0,\ \forall\, i \neq j.$$

Note that (4.19) is minimized by making the coefficients of λ_i^2 to be zero. This is achieved by first choosing $\psi_1 = \phi_1$ so that the term involving λ_1^2 is made zero without affecting the other terms. Applying the same argument, we choose $\psi_j = \phi_j$ for $j = 2, \ldots, \ell$ which not only satisfies the constraints but also minimizes (4.19), resulting in the minimum value being $\sum_{i>\ell} \lambda_i^2 = T^{\ell}(\Sigma)$.

(iii) <u>*Upper bound:*</u> Note that $T^{\ell}(\Sigma_m) = \mathbb{E}\left\|\mathfrak{I}\overline{k}(\cdot, X) - \mathfrak{A}P^{\ell}(\Sigma_m)\overline{k}_m(\cdot, X)\right\|^2_{L^2(\mathbb{P})}$ and there-

fore

$$T^{\ell}(\Sigma_m) \leq 2\underbrace{\mathbb{E}\left\|\mathfrak{I}\overline{k}(\cdot,X)-\mathfrak{A}\overline{k}_m(\cdot,X)\right\|^2_{L^2(\mathbb{P})}}_{③} + 2\underbrace{\mathbb{E}\left\|\mathfrak{A}(I-P^{\ell}(\Sigma_m))\overline{k}_m(\cdot,X)\right\|^2_{L^2(\mathbb{P})}}_{④}.$$

It follows from Lemma A.4 that

$$\begin{aligned} ④ &= \left\|\Sigma_m^{1/2}(I-P^{\ell}(\Sigma_m))\Sigma_m^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H}_m)} = \sum_{i=\ell+1}^{m}\lambda_{m,i}^2 \\ &= \sum_{i=\ell+1}^{m}(\lambda_{m,i}-\lambda_i+\lambda_i)^2 \leq 2\sum_{i=\ell+1}^{m}(\lambda_{m,i}-\lambda_i)^2 + 2\sum_{i>\ell}\lambda_i^2 \\ &\overset{(**)}{\leq} 2\left\|\mathfrak{I}\mathfrak{I}^*-\mathfrak{A}\mathfrak{A}^*\right\|^2_{\mathcal{L}^2(L^2(\mathbb{P}))} + 2\sum_{i>\ell}\lambda_i^2, \end{aligned} \tag{4.20}$$

where in $(**)$, we used Hoffman-Wielandt inequality (Bhatia, 1997) along with the fact that $\Sigma = \mathfrak{I}^*\mathfrak{I}$ and $\mathfrak{I}\mathfrak{I}^*$ have same eigenvalues, and similarly $\Sigma_m = \mathfrak{A}^*\mathfrak{A}$ and $\mathfrak{A}\mathfrak{A}^*$. The result follows by applying Lemma C.3 to (4.20) and Lemma C.2 to ③.

<u>*Lower bound:*</u> Note that

$$\begin{aligned} T^{\ell}(\Sigma_m) &= ③ + ④ - 2\mathbb{E}\left\langle \mathfrak{I}\overline{k}(\cdot,X)-\mathfrak{A}\overline{k}_m(\cdot,X), \mathfrak{A}(I-P^{\ell}(\Sigma_m))\overline{k}_m(\cdot,X)\right\rangle_{L^2(\mathbb{P})} \\ &\geq ③ + ④ \\ &\quad -2\mathbb{E}\left[\left\|\mathfrak{I}\overline{k}(\cdot,X)-\mathfrak{A}\overline{k}_m(\cdot,X)\right\|_{L^2(\mathbb{P})}\left\|\mathfrak{A}(I-P^{\ell}(\Sigma_m))\overline{k}_m(\cdot,X)\right\|_{L^2(\mathbb{P})}\right] \\ &\geq ③ + ④ - 2\sqrt{③}\sqrt{④} = \left(\sqrt{④}-\sqrt{③}\right)^2, \end{aligned}$$

where

$$\begin{aligned} \sqrt{④}-\sqrt{③} &= \sqrt{\sum_{i>\ell}\lambda_{m,i}^2} - \sqrt{③} \\ &\geq \left|\sqrt{\sum_{i>\ell}\lambda_i^2} - \sqrt{\sum_{i>\ell}(\lambda_i-\lambda_{m,i})^2}\right| - \sqrt{③} \end{aligned}$$

$$
\begin{aligned}
&\geq \sqrt{\sum_{i>\ell} \lambda_i^2} - \|\mathfrak{I}\mathfrak{I}^* - \mathfrak{A}\mathfrak{A}^*\|_{\mathcal{L}^2(L^2(\mathbb{P}))} - \sqrt{③} \\
&\geq \sqrt{\sum_{i>\ell} \lambda_i^2} - 4\kappa\sqrt{\frac{2\log\frac{2}{\delta}}{m}} - 8\kappa\sqrt{\frac{\log\frac{2}{\delta}}{m}} \\
&\geq \sqrt{\sum_{i>\ell} \lambda_i^2} - 16\kappa\sqrt{\frac{\log\frac{2}{\delta}}{m}} \\
&\geq \frac{1}{2}\sqrt{\sum_{i>\ell} \lambda_i^2} \qquad (4.21)
\end{aligned}
$$

as $\frac{1}{2}\sqrt{\sum_{i>\ell}\lambda_i^2} \geq 16\kappa\sqrt{\frac{\log\frac{2}{\delta}}{m}}$, with (4.21) holding with probability at least $1-3\delta$ over the choice of $(\theta_i)_{i=1}^m$.

(iv) *Upper bound:* $T^\ell(\widehat{\Sigma}_m)$ can be alternately written as

$$
T^\ell(\widehat{\Sigma}_m) = \mathbb{E}\left\|\mathfrak{I}\overline{k}(\cdot, X) - \mathfrak{A}P^\ell(\widehat{\Sigma}_m)(k_m(\cdot, X) - \hat{m}_{\mathbb{P},m})\right\|^2_{L^2(\mathbb{P})},
$$

which can be bounded as

$$
\begin{aligned}
T^\ell(\widehat{\Sigma}_m) &\leq 2\mathbb{E}\left\|\mathfrak{I}\overline{k}(\cdot, X) - \mathfrak{A}P^\ell(\widehat{\Sigma}_m)\overline{k}_m(\cdot, X)\right\|^2_{L^2(\mathbb{P})} \\
&\qquad +2\mathbb{E}\left\|\mathfrak{A}P^\ell(\widehat{\Sigma}_m)(m_{\mathbb{P},m} - \hat{m}_{\mathbb{P},m})\right\|^2_{L^2(\mathbb{P})} \\
&\leq 4\underbrace{\mathbb{E}\left\|\mathfrak{I}\overline{k}(\cdot, X) - \mathfrak{A}\overline{k}_m(\cdot, X)\right\|^2_{L^2(\mathbb{P})}}_{③} + 4\underbrace{\mathbb{E}\left\|\mathfrak{A}(I - P^\ell(\widehat{\Sigma}_m)\overline{k}_m(\cdot, X)\right\|^2_{L^2(\mathbb{P})}}_{⑤} \\
&\qquad +2\underbrace{\left\|\mathfrak{A}P^\ell(\widehat{\Sigma}_m)(m_{\mathbb{P},m} - \hat{m}_{\mathbb{P},m})\right\|^2_{L^2(\mathbb{P})}}_{⑥}. \qquad (4.22)
\end{aligned}
$$

Using Lemma A.4, for $t > 0$, we bound ⑤ as

$$
\begin{aligned}
⑤ &= \left\|\Sigma_m^{1/2}(I - P^\ell(\widehat{\Sigma}_m))\Sigma_m^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H}_m)} \\
&= \left\|\Sigma_m^{1/2}(\Sigma_m + tI)^{-1/2}(\Sigma_m + tI)^{1/2}(I - P^\ell(\widehat{\Sigma}_m))\right. \\
&\qquad \left.\times(\Sigma_m + tI)^{1/2}(\Sigma_m + tI)^{-1/2}\Sigma_m^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H}_m)} \\
&\leq \left\|\Sigma_m^{1/2}(\Sigma_m + tI)^{-1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H}_m)} \left\|(\Sigma_m + tI)^{-1/2}\Sigma_m^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H}_m)}
\end{aligned}
$$

$$
\begin{aligned}
&\times \left\|(\Sigma_m + tI)^{1/2}(I - P^\ell(\widehat{\Sigma}_m))(\Sigma_m + tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H}_m)} \\
&\overset{(*)}{\leq} \mathcal{N}_{\Sigma_m}(t) \left\|(\Sigma_m + tI)^{1/2}(I - P^\ell(\widehat{\Sigma}_m))(\Sigma_m + tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H}_m)} \\
&\leq \mathcal{N}_{\Sigma_m}(t) \left\|(\Sigma_m + tI)^{1/2}(\widehat{\Sigma}_m + tI)^{-1/2}\right\|^4_{\mathcal{L}^\infty(\mathcal{H}_m)} \\
&\qquad \times \left\|(\widehat{\Sigma}_m + tI)^{1/2}(I - P^\ell(\widehat{\Sigma}_m))(\widehat{\Sigma}_m + tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H}_m)} \\
&\leq \mathcal{N}_{\Sigma_m}(t)(\widehat{\lambda}_{m,\ell+1} + t)^2 \left\|(\Sigma_m + tI)^{1/2}(\widehat{\Sigma}_m + tI)^{-1/2}\right\|^4_{\mathcal{L}^\infty(\mathcal{H}_m)}, \qquad (4.23)
\end{aligned}
$$

where we used

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

and $\left\|\Sigma_m^{1/2}(\Sigma_m + tI)^{-1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H}_m)} \leq 1$ in $(*)$. Conditioning on $(\theta_i)_{i=1}^m$ and applying Lemma A.2*(ii, iv)* to (4.23), we obtain that for any $\delta > 0$ and

$$
\frac{140\kappa}{n} \log \frac{16\kappa n}{\delta} \leq t \leq \|\Sigma_m\|_{\mathcal{L}^\infty(\mathcal{H})}, \qquad (4.24)
$$

$$
\mathbb{P}^n_{|(\theta_i)_{i=1}^m} \left\{(X_i)_{i=1}^n : ⑤ \leq 9\mathcal{N}_{\Sigma_m}(t)(\lambda_{m,\ell+1} + t)^2\right\} \geq 1 - 2\delta. \qquad (4.25)
$$

Now, unconditioning w.r.t. $(\theta_i)_{i=1}^m$ and applying Lemma C.4*(ii, iv)* in (4.25), we obtain that for any $\delta > 0$ and $\frac{86\kappa}{m} \log \frac{16\kappa m}{\delta} \leq t \leq \|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})}$,

$$
\begin{aligned}
\Lambda^m \times \mathbb{P}^n \Bigg\{ &((\theta_i)_{i=1}^m, (X_i)_{i=1}^n) : ⑤ \leq \frac{81}{4} \Bigg[\frac{32\kappa \log \frac{2}{\delta}}{tm} \\
&+ \sqrt{\frac{32\kappa \mathcal{N}_\Sigma(t) \log \frac{2}{\delta}}{tm}} + 2\mathcal{N}_\Sigma(t) \Bigg] (\lambda_{\ell+1} + t)^2 \Bigg\} \geq 1 - 5\delta. \qquad (4.26)
\end{aligned}
$$

Note that the upper bound in (4.24) holds because we assumed that $t \leq \frac{1}{3}\|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})}$ which is equivalent to $t \leq \frac{1}{2}(\|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})} - t) \leq \|\Sigma_m\|_{\mathcal{L}^\infty(\mathcal{H})}$ where we used $\frac{1}{2}(\lambda_1 + t) \leq \lambda_{m,1} + t$ from Lemma C.4*(iii)*.

We now bound ⑥ as

$$
⑥ = \left\langle \mathfrak{A}P^\ell(\widehat{\Sigma}_m)(m_{\mathbb{P},m} - \hat{m}_{\mathbb{P},m}), \mathfrak{A}P^\ell(\widehat{\Sigma}_m)(m_{\mathbb{P},m} - \hat{m}_{\mathbb{P},m}) \right\rangle_{L^2(\mathbb{P})}
$$

$$
\begin{aligned}
&= \left\langle \Sigma_m^{1/2} P^\ell(\widehat{\Sigma}_m)(m_{\mathbb{P},m} - \hat{m}_{\mathbb{P},m}), \Sigma_m^{1/2} P^\ell(\widehat{\Sigma}_m)(m_{\mathbb{P},m} - \hat{m}_{\mathbb{P},m}) \right\rangle_{\mathcal{H}_m} \\
&= \left\| \Sigma_m^{1/2} P^\ell(\widehat{\Sigma}_m)(m_{\mathbb{P},m} - \hat{m}_{\mathbb{P},m}) \right\|_{\mathcal{H}_m}^2 \\
&\leq \left\| \Sigma_m^{1/2} \right\|_{\mathcal{L}^\infty(\mathcal{H}_m)}^2 \left\| P^\ell(\widehat{\Sigma}_m) \right\|_{\mathcal{L}^\infty(\mathcal{H}_m)}^2 \left\| m_{\mathbb{P},m} - \hat{m}_{\mathbb{P},m} \right\|_{\mathcal{H}_m}^2 \\
&\leq \lambda_{m,1} \left\| m_{\mathbb{P},m} - \hat{m}_{\mathbb{P},m} \right\|_{\mathcal{H}_m}^2 . \qquad (4.27)
\end{aligned}
$$

Applying Lemma C.1*(ii)* (by conditioning w.r.t. $(\theta_i)_{i=1}^m$) and Lemma C.4*(ii)* (to uncondition $(\theta_i)_{i=1}^m$) to (4.27), for any $\delta > 0$ and $n \geq 2 \log \frac{2}{\delta}$, we obtain

$$
\Lambda^m \times \mathbb{P}^n \left\{ ((\theta_i)_{i=1}^m, (X_i)_{i=1}^n) : ⑥ \leq \frac{320 \kappa^2 \log \frac{2}{\delta}}{3n} \right\} \geq 1 - 2\delta, \qquad (4.28)
$$

where we used $\lambda_{m,1} \leq \frac{3\lambda_1 + t}{2}$ from Lemma C.4*(ii)* and $t \leq \frac{\lambda_1}{3}$ (as per our assumption), resulting in $\lambda_{m,1} \leq \frac{5}{3} \left\| \Sigma \right\|_{\mathcal{L}^\infty(\mathcal{H})} \leq \frac{10\kappa}{3}$. The result therefore follows by applying Lemma C.2 to ③ and combining it with (4.26) and (4.28) in (4.22).

<u>*Lower bound:*</u> As carried out in the proof of the lower bound of *(iii)*, it can be shown that

$$
T^\ell(\widehat{\Sigma}_m) = \mathbb{E} \left\| \mathfrak{I} \overline{k}(\cdot, X) - \mathfrak{A} P^\ell(\widehat{\Sigma}_m) \widetilde{k}_m(\cdot, X) \right\|_{L^2(\mathbb{P})}^2 \geq \left(\sqrt{⑦} - \sqrt{③} \right)^2 ,
$$

where

$$
\begin{aligned}
⑦ &:= \mathbb{E} \left\| \mathfrak{A} \overline{k}_m(\cdot, X) - \mathfrak{A} P^\ell(\widehat{\Sigma}_m) \widetilde{k}_m(\cdot, X) \right\|_{L^2(\mathbb{P})}^2 \\
&= \mathbb{E} \left\| \mathfrak{A}(I - P^\ell(\widehat{\Sigma}_m)) \overline{k}_m(\cdot, X) \right\|_{L^2(\mathbb{P})}^2 + \left\| \mathfrak{A} P^\ell(\widehat{\Sigma}_m)(m_{\mathbb{P},m} - \hat{m}_{\mathbb{P},m}) \right\|_{L^2(\mathbb{P})}^2 \\
&\qquad -2\mathbb{E} \left\langle \mathfrak{A}(I - P^\ell(\widehat{\Sigma}_m)) \overline{k}_m(\cdot, X), \mathfrak{A} P^\ell(\widehat{\Sigma}_m)(m_{\mathbb{P},m} - \hat{m}_{\mathbb{P},m}) \right\rangle_{L^2(\mathbb{P})} \\
&\overset{(\dagger)}{=} ⑤ + ⑥ \overset{(\ddagger)}{\geq} \sum_{i>\ell} \lambda_{m,i}^2 + ⑥ \geq \sum_{i>\ell} \lambda_{m,i}^2,
\end{aligned}
$$

where (†) follows by noting that the term involving the inner product is zero and (‡)

follows by applying the argument used in the lower bound of *(ii)* to show that

$$T^{\ell}(\Sigma_m) = \inf_{(\psi_i)_{i=1}^{\ell} \in A} \left\| \Sigma_m^{1/2} \left(I - \sum_{i=1}^{\ell} \psi_i \otimes_{\mathcal{H}_m} \psi_i \right) \Sigma_m^{1/2} \right\|^2_{\mathcal{L}^2(\mathcal{H}_m)},$$

where $A := \{(\psi_i)_{i=1}^{\ell} \in \mathcal{H}_m^{\ell} : \langle \psi_i, \psi_j \rangle_{\mathcal{H}_m} = \delta_{ij}, \, \forall \, i, j \in [\ell]\}$. The result, therefore, follows from (4.21).

4.5.2 Proof of Corollary 4.1

(i) From Theorem 4.1*(i)* we have

$$T^{\ell}(\Sigma) = \sum_{i>\ell} \lambda_i^2 \lesssim \sum_{i>\ell} i^{-2\alpha} \lesssim \int_{\ell}^{\infty} x^{-2\alpha} dx \lesssim \ell^{1-2\alpha} = n^{-2\theta(1-\frac{1}{2\alpha})}.$$

Similarly,

$$T^{\ell}(\Sigma) = \sum_{i>\ell} \lambda_i^2 \gtrsim \sum_{i>\ell} i^{-2\alpha} \gtrsim \int_{\ell+1}^{\infty} x^{-2\alpha} dx \gtrsim (\ell+1)^{1-2\alpha} \gtrsim n^{-2\theta(1-\frac{1}{2\alpha})}.$$

(ii) From Theorem 4.1*(ii)* we have

$$T^{\ell}(\widehat{\Sigma}) \lesssim_{\mathbb{P}^n} \frac{1}{n} + \mathcal{N}_{\Sigma}(t)(\lambda_{\ell} + t)^2,$$

with $\frac{\log n}{n} \lesssim t \leq \frac{\lambda_1}{3}$. Using $\mathcal{N}_{\Sigma}(t) \lesssim t^{-1/\alpha}$ from Lemma A.1*(i)*, it follows that

$$\begin{aligned} T^{\ell}(\widehat{\Sigma}) &\lesssim_{\mathbb{P}^n} \inf \left\{ t^{-1/\alpha}(n^{-\theta} + t)^2 + n^{-1} : \frac{\log n}{n} \lesssim t \leq \frac{\lambda_1}{3} \right\} \\ &\lesssim \begin{cases} n^{-2\theta(1-\frac{1}{2\alpha})} + \frac{1}{n}, & \theta < 1 \\ \left(\frac{\log n}{n}\right)^{2-\frac{1}{\alpha}} + \frac{1}{n}, & \theta \geq 1 \end{cases}. \end{aligned}$$

Of course, $\left(\frac{\log n}{n}\right)^{2-\frac{1}{\alpha}} \leq \frac{1}{n}$ always holds, and $n^{-2\theta(1-\frac{1}{2\alpha})} \leq \frac{1}{n}$ for $\theta \geq \frac{\alpha}{2\alpha-1}$, yielding the result.

(iii) From Theorem 4.1*(iii)* we have

$$T^{\ell}(\Sigma_m) \lesssim_{\Lambda_m} \frac{1}{m} + \sum_{i>\ell} \lambda_i^2.$$

From *(i)* we have $\sum_{i>\ell} \lambda_i^2 \lesssim n^{-2\theta\left(1-\frac{1}{2\alpha}\right)}$, and the result follows.

(iv) From Theorem 4.1*(iv)* we have

$$T^{\ell}(\widehat{\Sigma}_m) \lesssim_{\mathbb{P}^n \times \Lambda^m} \frac{1}{n} + \frac{1}{m} + \left(\mathcal{N}_{\Sigma}(t) + \sqrt{\frac{\mathcal{N}_{\Sigma}(t)}{tm}} + \frac{1}{tm}\right)(\lambda_{\ell} + t)^2,$$

for $\frac{\log n}{n} \vee \frac{\log m}{m} \lesssim t \lesssim \lambda_1$. Note that $\mathcal{N}_{\Sigma}(t) + \frac{1}{tm} + \sqrt{\frac{\mathcal{N}_{\Sigma}(t)}{tm}} \lesssim t^{-1/\alpha}$, which follows from $\mathcal{N}_{\Sigma}(t) \lesssim t^{-1/\alpha}$ (Lemma A.1*(i)*), $\frac{1}{tm} < t^{-1/\alpha}$ and $\sqrt{\frac{t^{-(1+1/\alpha)}}{m}} \lesssim t^{-1/\alpha}$ since $\frac{1}{m} < \left(\frac{\log n}{n} \vee \frac{\log m}{m}\right)^{1-\frac{1}{\alpha}} \lesssim t^{1-\frac{1}{\alpha}}$. Therefore, we have

$$T^{\ell}(\widehat{\Sigma}_m) \lesssim_{\mathbb{P}^n \times \Lambda^m} n^{-\gamma} + t^{-1/\alpha}(n^{-\theta} + t)^2,$$

for $\frac{\log n}{n^{\gamma}} \lesssim t \lesssim \lambda_1$, where we have used $m = n^{\gamma}$ with $\gamma < 1$ and $\ell = n^{\frac{\theta}{\alpha}}$. This implies

$$\begin{aligned} T^{\ell}(\widehat{\Sigma}_m) &\lesssim_{\mathbb{P}^n \times \Lambda^m} \inf\left\{n^{-\gamma} + t^{-1/\alpha}(n^{-\theta} + t)^2 : \frac{\log n}{n^{\gamma}} \lesssim t \lesssim \lambda_1\right\} \\ &\lesssim \begin{cases} n^{-\gamma} + n^{\frac{\theta}{\alpha} - 2\theta}, & \theta < \gamma \\ n^{-\gamma} + \left(\frac{\log n}{n}\right)^{2-\frac{1}{\alpha}}, & \theta \geq \gamma \end{cases} \end{aligned}$$

and the result follows by considering the cases of $\gamma \geq \theta\left(2 - \frac{1}{\alpha}\right)$ and $\gamma < \theta\left(2 - \frac{1}{\alpha}\right)$.

4.5.3 Proof of Corollary 4.2

(i) From Theorem 4.1*(i)* we have

$$T^{\ell}(\Sigma) = \sum_{i>\ell} \lambda_i^2 \lesssim \sum_{i>\ell} e^{-2\tau i} \lesssim \int_{\ell}^{\infty} e^{-2\tau x} dx \lesssim e^{-2\tau\ell} = n^{-2\theta}$$

and

$$T^{\ell}(\Sigma) = \sum_{i>\ell} \lambda_i^2 \gtrsim \sum_{i>\ell} e^{-2\tau i} \gtrsim \int_{\ell+1}^{\infty} e^{-2\tau x} dx \gtrsim e^{-2\tau(\ell+1)} = e^{-2\tau} n^{-2\theta}.$$

(ii) Using $\mathcal{N}_{\Sigma}(t) \lesssim \log \frac{1}{t}$ from Lemma A.2*(ii)* in Theorem 4.1*(ii)*, we have

$$\begin{aligned} T^{\ell}(\widehat{\Sigma}) &\lesssim_{\mathbb{P}^n} \inf \left\{ (n^{-\theta} + t)^2 \log \frac{1}{t} + n^{-1} : \frac{\log n}{n} \lesssim t \leq \frac{\lambda_1}{3} \right\} \\ &\lesssim \begin{cases} n^{-2\theta} \log n + \frac{1}{n}, & \theta < 1 \\ \frac{\log^3 n}{n^2} + \frac{1}{n}, & \theta \geq 1 \end{cases}, \end{aligned}$$

and the result follows.

(iii) We obtain $T^{\ell}(\Sigma_m) \lesssim_{\Lambda_m} \frac{1}{m} + n^{-2\theta}$ and the result follows.

(iv) Theorem 4.1 *(iv)* yields

$$T^{\ell}(\widehat{\Sigma}_m) \lesssim_{\mathbb{P}^n \times \Lambda^m} n^{-\gamma} + \left(\mathcal{N}_{\Sigma}(t) + \sqrt{\frac{\mathcal{N}_{\Sigma}(t)}{tn^{\gamma}}} + \frac{1}{tn^{\gamma}} \right) (\lambda_{\ell} + t)^2,$$

for $\frac{\log n}{n^{\gamma}} \lesssim t \lesssim \frac{\lambda_1}{3}$. Using $\mathcal{N}_{\Sigma}(t) \lesssim \log \frac{1}{t}$, it is clear that $\frac{1}{tn^{\gamma}} \lesssim \log \frac{1}{t}$ and $\sqrt{\mathcal{N}_{\Sigma}(t) n^{-\gamma}/t} \lesssim \log \frac{1}{t}$ which follows from the constraint on t. Therefore,

$$\begin{aligned} T^{\ell}(\widehat{\Sigma}_m) &\lesssim_{\mathbb{P}^n \times \Lambda^m} \inf \left\{ (n^{-\theta} + t)^2 \log \frac{1}{t} + n^{-\gamma} : \frac{\log n}{n^{\gamma}} \lesssim t \leq \frac{\lambda_1}{3} \right\} \\ &\lesssim \begin{cases} n^{-2\theta} \log n + n^{-\gamma}, & \theta < \gamma \\ \frac{\log^3 n}{n^2} + n^{-\gamma}, & \theta \geq \gamma \end{cases}, \end{aligned}$$

and the result follows.

4.5.4 Proof of Theorem 4.2

(i) By defining $\Sigma_\ell^{-1} := \sum_{i=1}^{\ell} \frac{1}{\lambda_i}\phi_i \otimes_{\mathcal{H}} \phi_i$ and noting that $\Sigma_\ell^{-1}\Sigma = P^\ell(\Sigma)$, we have

$$\begin{aligned} S^\ell(\Sigma) &= \mathbb{E}\left\|(I - \mathfrak{I}\Sigma_\ell^{-1}\mathfrak{I}^*)\mathfrak{I}\overline{k}(\cdot, X)\right\|^2_{L^2(\mathbb{P})} = \mathbb{E}\left\|\mathfrak{I}(I - \Sigma_\ell^{-1}\Sigma)\overline{k}(\cdot, X)\right\|^2_{L^2(\mathbb{P})} \\ &\overset{(\dagger)}{=} \left\|\Sigma^{1/2}(I - \Sigma_\ell^{-1}\Sigma)\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} = \left\|\Sigma^{1/2}(I - P^\ell(\Sigma))\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} \\ &= \|\Sigma - \Sigma_\ell\|^2_{\mathcal{L}^2(\mathcal{H})} = \operatorname{tr}\left((\Sigma - \Sigma_\ell)^2\right) = \sum_{i>\ell} \lambda_i^2, \end{aligned}$$

where (†) follows from Lemma A.4.

(ii) *Upper bound:* By defining $\widehat{\Sigma}_\ell^{-1} := \sum_{i=1}^{\ell} \frac{1}{\widehat{\lambda}_i}\widehat{\phi}_i \otimes_{\mathcal{H}} \widehat{\phi}_i$, we have,

$$\begin{aligned} S^\ell(\widehat{\Sigma}) &= \mathbb{E}\left\|\mathfrak{I}\overline{k}(\cdot, X) - \mathfrak{I}\widehat{\Sigma}_\ell^{-1}\mathfrak{I}^*\mathfrak{I}(k(\cdot, X) - \widehat{m}_{\mathbb{P}})\right\|^2_{L^2(\mathbb{P})} \\ &\overset{(\star)}{=} \mathbb{E}\left\|\mathfrak{I}(I - \widehat{\Sigma}_\ell^{-1}\Sigma)\overline{k}(\cdot, X)\right\|^2_{L^2(\mathbb{P})} + \left\|\mathfrak{I}\widehat{\Sigma}_\ell^{-1}\Sigma(m_{\mathbb{P}} - \widehat{m}_{\mathbb{P}})\right\|^2_{L^2(\mathbb{P})} \\ &\overset{(\dagger)}{=} \left\|\Sigma^{1/2}(I - \widehat{\Sigma}_\ell^{-1}\Sigma)\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} + \left\|\Sigma^{1/2}\widehat{\Sigma}_\ell^{-1}\Sigma(m_{\mathbb{P}} - \widehat{m}_{\mathbb{P}})\right\|^2_{\mathcal{H}}, \end{aligned} \tag{4.29}$$

where we have used $\mathfrak{I}^*\mathfrak{I} = \Sigma$ (see Proposition D.2*(iii)*) in $(\star)$ and Lemma A.4 in (†). We can decompose the first term of (4.29) as

$$\begin{aligned} \left\|\Sigma^{1/2}(I - \widehat{\Sigma}_\ell^{-1}\Sigma)\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} \le 2\underbrace{\left\|\Sigma^{1/2}(I - P^\ell(\widehat{\Sigma}))\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})}}_{\text{ⓐ}} \\ + 2\underbrace{\left\|\Sigma^{1/2}(P^\ell(\widehat{\Sigma}) - \widehat{\Sigma}_\ell^{-1}\Sigma)\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})}}_{\text{ⓑ}}. \end{aligned} \tag{4.30}$$

Note that ⓐ is same as ① in (4.14) and therefore,

$$\mathbb{P}^n\left\{(X_i)_{i=1}^n : \text{ⓐ} \le 9\mathcal{N}_\Sigma(t)(\lambda_{\ell+1} + t)^2\right\} \ge 1 - 2\delta, \tag{4.31}$$

where $\frac{140\kappa}{n}\log\frac{16\kappa n}{\delta} \le t \le \|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})}$. For ⓑ we write,

$$\text{ⓑ} = \left\|\Sigma^{1/2}\widehat{\Sigma}_\ell^{-1}(\widehat{\Sigma} - \Sigma)\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})}$$

$$
\begin{aligned}
&\leq \left\|\Sigma^{1/2}(\Sigma+tI)^{-1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \left\|(\Sigma+tI)^{1/2}(\widehat{\Sigma}+tI)^{-1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \\
&\qquad \times \left\|(\widehat{\Sigma}+tI)^{1/2}\widehat{\Sigma}_\ell^{-1}(\widehat{\Sigma}+tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \\
&\qquad \times \left\|(\widehat{\Sigma}+tI)^{-1/2}(\Sigma+tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \\
&\qquad \times \left\|(\Sigma+tI)^{-1/2}(\widehat{\Sigma}-\Sigma)\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} \\
&\overset{(*)}{\leq} 4\sup_{i\leq\ell}\left(\frac{\widehat{\lambda}_i+t}{\widehat{\lambda}_i}\right)^2 \times \textcircled{c} = 4\left(\frac{\widehat{\lambda}_\ell+t}{\widehat{\lambda}_\ell}\right)^2 \times \textcircled{c} \\
&\overset{(\ddagger)}{\leq} 36\left(\frac{\lambda_\ell+t}{\lambda_\ell-t}\right)^2 \times \textcircled{c} \overset{(\star)}{\leq} 144 \times \textcircled{c},
\end{aligned} \tag{4.32}
$$

which holds with probability at least $1-2\delta$ over the choice of $(X_i)_{i=1}^n$, where we used Lemma A.2*(ii)* in $(*)$, Lemma A.2*(iv, v)* in $(\ddagger)$ and the assumption that $t \leq \frac{1}{3}\lambda_\ell$ in $(\star)$, with

$$
\textcircled{c} := \left\|(\Sigma+tI)^{-1/2}(\widehat{\Sigma}-\Sigma)\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})}.
$$

In the following, we will obtain two different bounds for $\textcircled{c}$ based on different decompositions, which we then combine by choosing the minimum of them. Applying (A.3) to $\textcircled{c}$ yields

$$
\mathbb{P}^n\left\{(X_i)_{i=1}^n : \textcircled{c} \leq \frac{128\kappa^{5/2}\mathcal{N}_\Sigma(t)\log\frac{2}{\delta}}{n\sqrt{t}} + \frac{4096\kappa^3\log^2\frac{3}{\delta}}{n^2 t}\right\} \geq 1-2\delta. \tag{4.33}
$$

$\textcircled{c}$ can be alternately bounded as

$$
\textcircled{c} \leq \left\|(\Sigma+tI)^{-1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \left\|\widehat{\Sigma}-\Sigma\right\|^2_{\mathcal{L}^2(\mathcal{H})} \|\Sigma\|^2_{\mathcal{L}^\infty(\mathcal{H})},
$$

yielding

$$
\mathbb{P}^n\left\{(X_i)_{i=1}^n : \textcircled{c} \leq \frac{256\kappa^4\log\frac{2}{\delta}}{nt} + \frac{8192\kappa^4\log^2\frac{3}{\delta}}{n^2 t}\right\} \geq 1-2\delta \tag{4.34}
$$

through an application of Theorem E.3*(ii)*. Combining (4.33) and (4.34) provides

$$
\mathbb{P}^n\left\{(X_i)_{i=1}^n : \textcircled{c} \leq 256\kappa^{5/2}\log\frac{2}{\delta}\left[\frac{\mathcal{N}_\Sigma(t)}{n\sqrt{t}} \wedge \frac{\kappa^{3/2}}{nt}\right]\right.
$$

$$+\frac{8192\kappa^3(\kappa\wedge 1)\log^2\frac{3}{\delta}}{n^2 t}\Bigg\}\geq 1-4\delta,$$

using which in (4.32) yields

$$\mathbb{P}^n\Bigg\{(X_i)_{i=1}^n : \textcircled{b}\leq 144\Bigg[256\kappa^{5/2}\log\frac{2}{\delta}\left[\frac{\mathcal{N}_\Sigma(t)}{n\sqrt{t}}\wedge\frac{\kappa^{3/2}}{nt}\right] +\frac{8192\kappa^3(\kappa\wedge 1)\log^2\frac{3}{\delta}}{n^2 t}\Bigg]\Bigg\}\geq 1-6\delta. \tag{4.35}$$

To bound the second term of (4.29) we have

$$\begin{aligned}\left\|\Sigma^{1/2}\widehat{\Sigma}_\ell^{-1}\Sigma(m_{\mathbb{P}}-\widehat{m}_{\mathbb{P}})\right\|_{\mathcal{H}}^2 &\leq \left\|\Sigma^{1/2}\widehat{\Sigma}_\ell^{-1}\Sigma^{1/2}\right\|_{\mathcal{L}^\infty(\mathcal{H})}^2 \\ &\quad\times\left\|\Sigma^{1/2}(m_{\mathbb{P}}-\widehat{m}_{\mathbb{P}})\right\|_{\mathcal{H}}^2 \\ &\leq \left\|\Sigma^{1/2}(\Sigma+tI)^{-1/2}\right\|_{\mathcal{L}^\infty(\mathcal{H})}^4\left\|(\Sigma+tI)^{1/2}(\widehat{\Sigma}+tI)^{-1/2}\right\|_{\mathcal{L}^\infty(\mathcal{H})}^4 \\ &\quad\times\left\|(\widehat{\Sigma}+tI)^{1/2}\widehat{\Sigma}_\ell^{-1}(\widehat{\Sigma}+tI)^{1/2}\right\|_{\mathcal{L}^\infty(\mathcal{H})}^2\left\|\Sigma^{1/2}(m_{\mathbb{P}}-\widehat{m}_{\mathbb{P}})\right\|_{\mathcal{H}}^2 \\ &\overset{(\diamond)}{\leq} 144\|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})}\|m_{\mathbb{P}}-\widehat{m}_{\mathbb{P}}\|_{\mathcal{H}}^2\leq 288\kappa\|m_{\mathbb{P}}-\widehat{m}_{\mathbb{P}}\|_{\mathcal{H}}^2,\end{aligned} \tag{4.36}$$

which holds with probability at least $1-2\delta$ over the choice of $(X_i)_{i=1}^n$, wherein we have used Lemma A.2*(ii)* and the bound in (4.32) on

$$\left\|(\widehat{\Sigma}+tI)^{1/2}\widehat{\Sigma}_\ell^{-1}(\widehat{\Sigma}+tI)^{1/2}\right\|_{\mathcal{L}^\infty(\mathcal{H})}^2$$

in $(\diamond)$. Applying Lemma C.1*(i)* to (4.36), we obtain

$$\mathbb{P}^n\left\{(X_i)_{i=1}^n : \left\|\Sigma^{1/2}\widehat{\Sigma}_\ell^{-1}\Sigma(m_{\mathbb{P}}-\widehat{m}_{\mathbb{P}})\right\|_{\mathcal{H}}^2\leq\frac{9216\kappa^2\log\frac{2}{\delta}}{n}\right\}\geq 1-3\delta. \tag{4.37}$$

Combining (4.29)–(4.31), (4.35) and (4.37), yields the result.

<u>*Lower bound:*</u> It follows from (4.29) that

$$S^\ell(\widehat{\Sigma})\geq\left\|\Sigma^{1/2}(I-\widehat{\Sigma}_\ell^{-1}\Sigma)\Sigma^{1/2}\right\|_{\mathcal{L}^2(\mathcal{H})}^2.$$

As in the proof of the lower bound of $R_{\widehat{\Sigma},\ell}$ in Theorem 4.1*(ii)*, we will show that $\|\Sigma^{1/2}(I - \widehat{\Sigma}_\ell^{-1}\Sigma)\Sigma^{1/2}\|^2_{\mathcal{L}^2(\mathcal{H})} \geq S^\ell(\Sigma)$ by establishing that $S^\ell(\Sigma) = \inf\{\Sigma^{1/2}(I - Q_{\psi,\alpha,\ell}\Sigma)\Sigma^{1/2}\|^2_{\mathcal{L}^2(\mathcal{H})} : (\psi_i)_{i=1}^\ell \in B,\, (\alpha_i)_{i=1}^\ell \subset (0,\infty)\}$ where $Q_{\psi,\alpha,\ell} := \sum_{i=1}^\ell \frac{1}{\alpha_i}\psi_i \otimes_{\mathcal{H}} \psi_i$ and $B := \{(\psi_i)_{i=1}^\ell \subset \mathcal{H} : \langle\psi_i, \psi_j\rangle_{\mathcal{H}} = \delta_{ij}\, \forall\, i,j \in \{1,\ldots,\ell\}\}$. Using the same idea as in the proof of the lower bound of $R_{\widehat{\Sigma},\ell}$ in Theorem 4.1*(ii)*, it is easy to show that

$$\begin{aligned}
\left\|\Sigma^{1/2}(I - Q_{\psi,\alpha,\ell}\Sigma)\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} &= \sum_i \lambda_i^2 \left(1 - \lambda_i \langle\phi_i, Q_{\psi,\alpha,\ell}\phi_i\rangle_{\mathcal{H}}\right)^2 \\
&\quad + \sum_{i\neq j} \lambda_i\lambda_j^3 \langle\phi_i, Q_{\psi,\alpha,\ell}\phi_j\rangle^2_{\mathcal{H}} \\
&\geq \sum_i \lambda_i^2 \left(1 - \lambda_i \langle\phi_i, Q_{\psi,\alpha,\ell}\phi_i\rangle_{\mathcal{H}}\right)^2 \\
&= \sum_i \lambda_i^2 \left(1 - \lambda_i \sum_{j=1}^\ell \frac{1}{\alpha_j}\langle\phi_i, \psi_j\rangle^2_{\mathcal{H}}\right)^2, \qquad (4.38)
\end{aligned}$$

where strict equality holds in the penultimate expression if and only if $\langle\phi_i, Q_{\psi,\alpha,\ell}\phi_i\rangle_{\mathcal{H}} = 0$ for all $i \neq j$. Note that the quantity in (4.38) is minimized by making the coefficients that correspond to large λ_i to be zero. This is achieved when $\psi_j = \phi_j$ and $\alpha_j = \lambda_j$ for $j = 1,\ldots,\ell$, which therefore yields the result.

(iii) Define $\Sigma_{m,\ell}^{-1} = \sum_{i=1}^\ell \frac{1}{\lambda_{i,m}}\phi_{i,m} \otimes_{\mathcal{H}_m} \phi_{i,m}$. Then

$$\begin{aligned}
S^\ell(\Sigma_m) &= \mathbb{E}\left\|\mathfrak{I}\overline{k}(\cdot, X) - \mathfrak{A}\Sigma_{m,\ell}^{-1}\mathfrak{A}^*\mathfrak{A}\overline{k}_m(\cdot, X)\right\|^2_{L^2(\mathbb{P})} \\
&= \mathbb{E}\left\|\mathfrak{I}\overline{k}(\cdot, X) - \mathfrak{A}\Sigma_{m,\ell}^{-1}\Sigma_m\overline{k}_m(\cdot, X)\right\|^2_{L^2(\mathbb{P})} \\
&= \mathbb{E}\left\|\mathfrak{I}\overline{k}(\cdot, X) - \mathfrak{A}P^\ell(\Sigma_m)\overline{k}_m(\cdot, X)\right\|^2_{L^2(\mathbb{P})} = R_{\Sigma_m,\ell}
\end{aligned}$$

and the result follows from Theorem 4.1*(iii)*.

(iv) <u>*Upper bound:*</u> Define $\widehat{\Sigma}_{m,\ell}^{-1} = \sum_{i=1}^\ell \frac{1}{\widehat{\lambda}_{i,m}}\widehat{\phi}_{i,m} \otimes_{\mathcal{H}_m} \widehat{\phi}_{i,m}$. Then

$$\begin{aligned}
S^\ell(\widehat{\Sigma}_m) &= \mathbb{E}\left\|\mathfrak{I}\overline{k}(\cdot, X) - \mathfrak{A}\widehat{\Sigma}_{m,\ell}^{-1}\mathfrak{A}^*\mathfrak{A}(k_m(\cdot, X) - \widehat{m}_{\mathbb{P},m})\right\|^2_{L^2(\mathbb{P})} \\
&\leq \underbrace{3\,\mathbb{E}\left\|\mathfrak{I}\overline{k}(\cdot, X) - \mathfrak{A}\overline{k}_m(\cdot, X)\right\|^2_{L^2(\mathbb{P})}}_{\text{(d)}} + \underbrace{3\,\mathbb{E}\left\|(I - \mathfrak{A}\widehat{\Sigma}_{m,\ell}^{-1}\mathfrak{A}^*)\mathfrak{A}\overline{k}_m(\cdot, X)\right\|^2_{L^2(\mathbb{P})}}_{\text{(e)}}
\end{aligned}$$

$$+ 3 \underbrace{\left\| \mathfrak{A} \widehat{\Sigma}_{m,\ell}^{-1} \mathfrak{A}^* \mathfrak{A} (m_{\mathbb{P},m} - \widehat{m}_{\mathbb{P},m}) \right\|_{L^2(\mathbb{P})}^2}_{\text{(f)}}. \tag{4.39}$$

Note that (d) = (3) which can be bounded using Lemma C.2 and

$$\begin{aligned}
\text{(e)} &= \mathbb{E} \left\| \mathfrak{A} (I - \widehat{\Sigma}_{m,\ell}^{-1} \Sigma_m) \overline{k}_m(\cdot, X) \right\|_{L^2(\mathbb{P})}^2 \\
&\overset{(\clubsuit)}{=} \left\| \Sigma_m^{1/2} (I - \widehat{\Sigma}_{m,\ell}^{-1} \Sigma_m) \Sigma_m^{1/2} \right\|_{\mathcal{L}^2(\mathcal{H}_m)}^2 \\
&\leq 2 \underbrace{\left\| \Sigma_m^{1/2} (I - P^\ell(\widehat{\Sigma}_m)) \Sigma_m^{1/2} \right\|_{\mathcal{L}^2(\mathcal{H}_m)}^2}_{\text{(e1)}} \\
&\quad + 2 \underbrace{\left\| \Sigma_m^{1/2} (P^\ell(\widehat{\Sigma}_m) - \widehat{\Sigma}_{m,\ell}^{-1} \Sigma_m) \Sigma_m^{1/2} \right\|_{\mathcal{L}^2(\mathcal{H}_m)}^2}_{\text{(e2)}},
\end{aligned} \tag{4.40}$$

where we have used Lemma A.4 in (♣). Note that (e1) = (5), whose bound follows from (4.26).

By handling (e2) in a similar manner as (b) in *(ii)*, by conditioning on $(\theta_i)_{i=1}^m$, for $\frac{140\kappa}{n} \log \frac{16\kappa n}{\delta} \leq t \leq \|\Sigma_m\|_{\mathcal{L}^\infty(\mathcal{H}_m)}$, with probability at least $1 - 4\delta$ over the choice of $(X_i)_{i=1}^n$, we obtain

$$\begin{aligned}
\text{(e2)} &= \left\| \Sigma_m^{1/2} \widehat{\Sigma}_{m,\ell}^{-1} (\widehat{\Sigma}_m - \Sigma_m) \Sigma_m^{1/2} \right\|_{\mathcal{L}^2(\mathcal{H}_m)}^2 \\
&\leq 36 \left(\frac{\lambda_{m,\ell} + t}{\lambda_{m,\ell} - t} \right)^2 \left[\frac{128 \kappa^{5/2} \mathcal{N}_{\Sigma_m}(t) \log \frac{2}{\delta}}{n\sqrt{t}} + \frac{4096 \kappa^3 \log^2 \frac{3}{\delta}}{n^2 t} \right].
\end{aligned} \tag{4.41}$$

By unconditioning w.r.t. $(\theta_i)_{i=1}^m$ in the above inequality, for $\frac{86\kappa}{m} \log \frac{16\kappa m}{\delta} \leq t \leq \|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})}$, with probability at least $1 - 7\delta$ jointly over the choice of $((\theta_i)_{i=1}^m, (X_i)_{i=1}^n)$, we obtain

$$\begin{aligned}
\text{(e2)} &\leq 36 \left(\frac{3\lambda_\ell + 3t}{\lambda_\ell - 3t} \right)^2 \left[\frac{4096 \kappa^3 \log^2 \frac{3}{\delta}}{n^2 t} + \frac{256 \kappa^{5/2} \mathcal{D}(t) \log \frac{2}{\delta}}{n\sqrt{t}} \right] \\
&\leq 900 \left[\frac{4096 \kappa^3 \log^2 \frac{3}{\delta}}{n^2 t} + \frac{256 \kappa^{5/2} \mathcal{D}(t) \log \frac{2}{\delta}}{n\sqrt{t}} \right],
\end{aligned} \tag{4.42}$$

where the first inequality follows from applying Lemma C.4*(ii)–(iv)* to (4.41), the second

inequality follows by using $t \le \frac{\lambda_\ell}{9}$ and we used $\mathcal{D}(t) := \frac{16\kappa \log \frac{2}{\delta}}{tm} + \sqrt{\frac{8\kappa \mathcal{N}_\Sigma(t) \log \frac{2}{\delta}}{tm}} + \mathcal{N}_\Sigma(t)$. (e2) can be alternately bounded as follows. By conditioning on $(\theta_i)_{i=1}^m$, for $\frac{140\kappa}{n} \log \frac{16\kappa n}{\delta} \le t \le \|\Sigma_m\|_{\mathcal{L}^\infty(\mathcal{H}_m)}$, with probability at least $1 - 4\delta$ over the choice of $(X_i)_{i=1}^n$, we obtain

$$\text{(e2)} \le \frac{36\lambda_{m,1}^2}{t} \left(\frac{\lambda_{m,\ell} + t}{\lambda_{m,\ell} - t} \right)^2 \left[\frac{64\kappa^2 \log \frac{2}{\delta}}{n} + \frac{2048\kappa^2 \log^2 \frac{3}{\delta}}{n^2} \right].$$

By unconditioning w.r.t. $(\theta_i)_{i=1}^m$ in the above inequality, for $\frac{86\kappa}{m} \log \frac{16\kappa m}{\delta} \le t \le \|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})}$, with probability at least $1 - 5\delta$ jointly over the choice of $((\theta_i)_{i=1}^m, (X_i)_{i=1}^n)$, we obtain

$$\begin{aligned} \text{(e2)} &\le \frac{400\kappa^4}{t} \left(\frac{3\lambda_\ell + 3t}{\lambda_\ell - 3t} \right)^2 \left[\frac{64 \log \frac{2}{\delta}}{n} + \frac{2048 \log^2 \frac{3}{\delta}}{n^2} \right] \\ &\le 10^4 \kappa^4 \left[\frac{64 \log \frac{2}{\delta}}{nt} + \frac{2048 \log^2 \frac{3}{\delta}}{n^2 t} \right]. \end{aligned} \tag{4.43}$$

Combining (4.42) and (4.43), with probability at least $1 - 12\delta$ jointly over the choice of $((\theta_i)_{i=1}^m, (X_i)_{i=1}^n)$, we have

$$\begin{aligned} \text{(e2)} \le 24 \times 10^4 &\left[\left(\frac{\kappa^4 \log \frac{2}{\delta}}{nt} \right) \wedge \left(\frac{\kappa^{5/2} \mathcal{D}(t) \log \frac{2}{\delta}}{n\sqrt{t}} \right) \right] \\ &+ \frac{8 \times 10^6 \kappa^3 (1 \wedge k) \log^2 \frac{3}{\delta}}{n^2 t}. \end{aligned} \tag{4.44}$$

(f) can be bounded as

$$\begin{aligned} \text{(f)} &= \left\| \Sigma_m^{1/2} \widehat{\Sigma}_{m,\ell}^{-1} \Sigma_m (m_{\mathbb{P},m} - \widehat{m}_{\mathbb{P},m}) \right\|_{\mathcal{H}_m}^2 \\ &\le \left\| \Sigma_m^{1/2} \widehat{\Sigma}_{m,\ell}^{-1} \Sigma_m^{1/2} \right\|_{\mathcal{L}^\infty(\mathcal{H}_m)}^2 \left\| \Sigma_m^{1/2} (m_{\mathbb{P},m} - \widehat{m}_{\mathbb{P},m}) \right\|_{\mathcal{H}_m}^2 \\ &\le \lambda_{m,1} \left\| \Sigma_m^{1/2} (\Sigma_m + tI)^{-1/2} \right\|_{\mathcal{L}^\infty(\mathcal{H}_m)}^4 \|m_{\mathbb{P},m} - \widehat{m}_{\mathbb{P},m}\|_{\mathcal{H}_m}^2 \\ &\quad \times \left\| (\Sigma_m + tI)^{1/2} (\widehat{\Sigma}_m + tI)^{-1/2} \right\|_{\mathcal{L}^\infty(\mathcal{H}_m)}^4 \\ &\quad\quad \times \left\| (\widehat{\Sigma}_m + tI)^{1/2} \widehat{\Sigma}_{m,\ell}^{-1} (\widehat{\Sigma}_m + tI)^{1/2} \right\|_{\mathcal{L}^\infty(\mathcal{H}_m)}^2. \end{aligned} \tag{4.45}$$

By conditioning on $(\theta_i)_{i=1}^m$, for $\frac{140\kappa}{n} \log \frac{16\kappa n}{\delta} \le t \le \|\Sigma_m\|_{\mathcal{L}^\infty(\mathcal{H}_m)}$, with probability at least

$1-3\delta$ over the choice of $(X_i)_{i=1}^n$, we obtain

$$\textcircled{f} \leq \frac{3840\kappa^2 \log\frac{2}{\delta}}{n}\left(\frac{\lambda_{m,\ell}+t}{\lambda_{m,\ell}-t}\right)^2,$$

where we used the fact that $\lambda_{m,1} \leq \frac{10\kappa}{3}$ (see the proof of Theorem 4.1*(iv)*) and employed Lemmas A.2*(ii, iv, v)* and C.1*(ii)*. By unconditioning w.r.t. $(\theta_i)_{i=1}^m$ in the above inequality, with probability at least $1-4\delta$ jointly over the choice of $((\theta_i)_{i=1}^m, (X_i)_{i=1}^n)$, we obtain

$$\textcircled{f} \leq \frac{3840\kappa^2 \log\frac{2}{\delta}}{n}\left(\frac{3\lambda_\ell+3t}{\lambda_\ell-3t}\right)^2 \leq \frac{96000\kappa^2 \log\frac{2}{\delta}}{n}, \tag{4.46}$$

where we applied Lemma C.4*(ii, iii)* and $\frac{86\kappa}{m}\log\frac{16\kappa m}{\delta} \leq t \leq \|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})}$. The result therefore follows by combining (4.39), (4.40), (4.44) and (4.46) under the condition that $\frac{140\kappa}{n}\log\frac{8n}{\delta} \vee \frac{86\kappa}{m}\log\frac{8m}{\delta} \leq t \leq \frac{\lambda_\ell}{9}$ and $n \wedge m \geq 8\log\frac{1}{\delta}$.

<u>*Lower bound:*</u> As carried out in the proof of the lower bound of *(iii)*, it can be shown that

$$\begin{aligned}
S^\ell(\widehat{\Sigma}_m) &= \mathbb{E}\left\|\mathfrak{I}\overline{k}(\cdot,X) - \mathfrak{A}\widehat{\Sigma}_{m,\ell}^{-1}\Sigma_m\widetilde{k}_m(\cdot,X)\right\|_{L^2(\mathbb{P})}^2 \\
&\geq \left(\sqrt{\textcircled{e}+\textcircled{f}} - \sqrt{\textcircled{3}}\right)^2 \geq \left(\sqrt{\textcircled{e}} - \sqrt{\textcircled{3}}\right)^2 \\
&\overset{(*)}{\geq} \left(\sqrt{\sum_{i>\ell}\lambda_{m,i}^2} - \sqrt{\textcircled{3}}\right)^2,
\end{aligned}$$

where $(*)$ follows from applying the argument used in the proof of the lower bound of $S^\ell(\widehat{\Sigma})$ to $\textcircled{e}$. The result, therefore, follows from (4.21).

4.5.5 Proof of Corollary 4.3

(i) and *(iii)* are exactly same as that of the proof of Corollary 4.1.

(ii) From Theorem 4.2*(ii)* we have

$$S^\ell(\widehat{\Sigma}) \lesssim_{\mathbb{P}^n} \frac{1}{n} + \frac{1}{n^2 t} + \left(\frac{\mathcal{N}_\Sigma(t)}{n\sqrt{t}} \wedge \frac{1}{nt}\right) + \mathcal{N}_\Sigma(t)(\lambda_\ell + t)^2,$$

with $\frac{\log n}{n} \lesssim t \leq \frac{\lambda_\ell}{3}$. Clearly $\frac{1}{n^2 t} \lesssim \frac{1}{n}$. Using $\mathcal{N}_\Sigma(t) \lesssim t^{-1/\alpha}$ from Lemma A.1*(i)*, it follows that

$$S^\ell(\widehat{\Sigma}) \lesssim_{\mathbb{P}^n} \inf\left\{ t^{-1/\alpha}(n^{-\theta}+t)^2 + \left(\frac{t^{-(\frac{1}{\alpha}+\frac{1}{2})}}{n} \wedge \frac{1}{nt}\right) + \frac{1}{n} : \frac{\log n}{n} \lesssim t \lesssim n^{-\theta} \right\}.$$

It is clear that both $\frac{t^{-(\frac{1}{\alpha}+\frac{1}{2})}}{n}$ and $\frac{1}{nt}$ dominate n^{-1} and using $t \lesssim n^{-\theta}$ in the first term, we obtain

$$\begin{aligned} S^\ell(\widehat{\Sigma}) &\lesssim_{\mathbb{P}^n} \inf\left\{ t^{-1/\alpha} n^{-2\theta} + \frac{t^{-\frac{1}{\alpha'}}}{n} : \frac{\log n}{n} \lesssim t \lesssim n^{-\theta} \right\} \\ &= n^{-2\theta\left(1-\frac{1}{2\alpha}\right)} + n^{-\left(1-\frac{\theta}{\alpha'}\right)}, \end{aligned}$$

where $\frac{1}{\alpha'} := \left(\frac{1}{\alpha}+\frac{1}{2}\right) \wedge 1$ and the result follows.

(iv) From Theorem 4.2*(iv)* we have

$$S^\ell(\widehat{\Sigma}_m) \lesssim_{\mathbb{P}^n \times \Lambda^m} \frac{1}{n} + \frac{1}{m} + \frac{1}{n^2 t} + \left(\frac{\mathcal{A}(t)}{n\sqrt{t}} \wedge \frac{1}{nt}\right) + \mathcal{A}(t)(\lambda_\ell + t)^2$$

for $\frac{\log n}{n} \vee \frac{\log m}{m} \lesssim t \lesssim \lambda_\ell$, where $\mathcal{A}(t) = \mathcal{N}_\Sigma(t) + \sqrt{\frac{\mathcal{N}_\Sigma(t)}{tm}} + \frac{1}{tm}$. From the proof of Corollary 4.1*(iv)*, we have $\mathcal{A}(t) \lesssim t^{-1/\alpha}$. Also it is obvious that $\frac{1}{n^2 t} \lesssim \frac{1}{n}$. Therefore,

$$\begin{aligned} S^\ell(\widehat{\Sigma}_m) &\lesssim_{\mathbb{P}^n \times \Lambda^m} \inf\left\{ \frac{1}{n^\gamma} + t^{-1/\alpha} n^{-2\theta} + \frac{t^{-\frac{1}{\alpha'}}}{n} : \frac{\log n}{n^\gamma} \lesssim t \lesssim n^{-\theta} \right\} \\ &= n^{-\gamma} + n^{-2\theta\left(1-\frac{1}{2\alpha}\right)} + n^{-\left(1-\frac{\theta}{\alpha'}\right)} \end{aligned}$$

and the result follows by imposing $\theta < \gamma$ to ensure the constraint $\frac{\log n}{n^\gamma} \lesssim t \lesssim n^{-\theta}$ is satisfied.

4.5.6 Proof of Corollary 4.4

(i) and *(iii)* are exactly same as that of the proof of Corollary 4.2.

(ii) Using $\mathcal{N}_\Sigma(t) \lesssim \log \frac{1}{t}$ from Lemma A.2*(ii)* in Theorem 4.2*(ii)*, we have $\frac{\mathcal{N}_\Sigma(t)}{n\sqrt{t}} \wedge \frac{1}{nt} \lesssim$

$\frac{\mathcal{N}_\Sigma(t)}{n\sqrt{t}} \lesssim \frac{\log \frac{1}{t}}{n\sqrt{t}}$, which implies

$$\begin{aligned} S^\ell(\widehat{\Sigma}) &\lesssim_{\mathbb{P}^n} \inf\left\{ n^{-2\theta} \log \frac{1}{t} + \frac{\log \frac{1}{t}}{n\sqrt{t}} : \frac{\log n}{n} \lesssim t \lesssim n^{-\theta} \right\} \\ &\lesssim \left(n^{-2\theta} + n^{-\left(1-\frac{\theta}{2}\right)} \right) \log n \end{aligned}$$

and the result follows.

(iv) By noting that $\mathcal{A}(t) \lesssim \mathcal{N}_\Sigma(t) \lesssim \log \frac{1}{t}$, we obtain

$$\begin{aligned} S^\ell(\widehat{\Sigma}_m) &\lesssim_{\mathbb{P}^n} \inf\left\{ n^{-\gamma} + n^{-2\theta} \log \frac{1}{t} + \frac{\log \frac{1}{t}}{n\sqrt{t}} : \frac{\log n}{n} \lesssim t \lesssim n^{-\theta} \right\} \\ &\lesssim n^{-\gamma} + \left(n^{-2\theta} + n^{-\left(1-\frac{\theta}{2}\right)} \right) \log n, \end{aligned}$$

which yields the result.

4.5.7 Proof of Theorem 4.3

As mentioned in Section 4.4.1, $W^\ell(\widehat{\Sigma})$, $W^\ell(\Sigma_m)$ and $W^\ell(\widehat{\Sigma}_m)$ are defined as

$$W^\ell(\widehat{\Sigma}) = \mathbb{E}_{X\sim\mathbb{P}} \left\| \mathfrak{I}\overline{k}(\cdot, X) - \sum_{i=1}^{\ell} \left\langle \mathfrak{I}\overline{k}(\cdot, X), \frac{\mathfrak{I}\widehat{\phi}_i}{\sqrt{\widehat{\lambda}_i}} \right\rangle_{L^2(\mathbb{P})} \frac{\mathfrak{I}\widehat{\phi}_i}{\sqrt{\widehat{\lambda}_i}} \right\|^2_{L^2(\mathbb{P})},$$

$$W^\ell(\Sigma_m) = \mathbb{E}_{X\sim\mathbb{P}} \left\| \mathfrak{I}\overline{k}(\cdot, X) - \sum_{i=1}^{\ell} \left\langle \mathfrak{I}\overline{k}(\cdot, X), \frac{\mathfrak{A}\phi_{m,i}}{\sqrt{\lambda_{m,i}}} \right\rangle_{L^2(\mathbb{P})} \frac{\mathfrak{A}\phi_{m,i}}{\sqrt{\lambda_{m,i}}} \right\|^2_{L^2(\mathbb{P})},$$

and

$$W^\ell(\widehat{\Sigma}_m) = \mathbb{E}_{X\sim\mathbb{P}} \left\| \mathfrak{I}\overline{k}(\cdot, X) - \sum_{i=1}^{\ell} \left\langle \mathfrak{I}\overline{k}(\cdot, X), \frac{\mathfrak{A}\widehat{\phi}_{m,i}}{\sqrt{\widehat{\lambda}_{m,i}}} \right\rangle_{L^2(\mathbb{P})} \frac{\mathfrak{A}\widehat{\phi}_{m,i}}{\sqrt{\widehat{\lambda}_{m,i}}} \right\|^2_{L^2(\mathbb{P})}.$$

respectively.

(i) By adding and subtracting $\mathfrak{I}\widetilde{k}(\cdot, X)$ to the first argument of the inner product in

$W^\ell(\widehat{\Sigma})$, we obtain

$$\begin{aligned} W^\ell(\widehat{\Sigma}) &\leq 2S^\ell(\widehat{\Sigma}) + 2\mathbb{E}\left\|\sum_{i=1}^{\ell}\left\langle \mathfrak{I}\overline{k}(\cdot,X) - \mathfrak{I}\widetilde{k}(\cdot,X), \frac{\mathfrak{I}\widehat{\phi}_i}{\sqrt{\widehat{\lambda}_i}}\right\rangle_{L^2(\mathbb{P})} \frac{\mathfrak{I}\widehat{\phi}_i}{\sqrt{\widehat{\lambda}_i}}\right\|^2_{L^2(\mathbb{P})} \\ &= 2S^\ell(\widehat{\Sigma}) + 2\left\|\mathfrak{I}\widehat{\Sigma}_\ell^{-1}\Sigma(m_{\mathbb{P}} - \widehat{m}_{\mathbb{P}})\right\|^2_{L^2(\mathbb{P})} \\ &= 2S^\ell(\widehat{\Sigma}) + 2\left\|\Sigma^{1/2}\widehat{\Sigma}_\ell^{-1}\Sigma(m_{\mathbb{P}} - \widehat{m}_{\mathbb{P}})\right\|^2_{\mathcal{H}} \end{aligned}$$

and the result follows from (4.37). Also note that $W^\ell(\widehat{\Sigma}) = \|\Sigma^{1/2}(I - \widehat{\Sigma}_\ell^{-1}\Sigma)\Sigma^{1/2}\|^2_{\mathcal{L}^2(\mathcal{H})}$ and therefore the lower bound follows from the proof of the lower bound of $S^\ell(\widehat{\Sigma})$.

(ii) As above, adding and subtracting $\mathfrak{A}\overline{k}_m(\cdot,X)$ to the first argument of the inner product in $W^\ell(\Sigma_m)$, we obtain

$$\begin{aligned} W^\ell(\Sigma_m) &\leq 2S^\ell(\Sigma_m) + 2\mathbb{E}\left\|\sum_{i=1}^{\ell}\left\langle \mathfrak{I}\overline{k}(\cdot,X) - \mathfrak{A}\overline{k}_m(\cdot,X), \frac{\mathfrak{A}\phi_{m,i}}{\sqrt{\lambda_{m,i}}}\right\rangle_{L^2(\mathbb{P})} \frac{\mathfrak{A}\phi_{m,i}}{\sqrt{\lambda_{m,i}}}\right\|^2_{L^2(\mathbb{P})} \\ &= 2S^\ell(\Sigma_m) + 2\mathbb{E}\left\|\mathfrak{A}\Sigma_{m,\ell}^{-1}\mathfrak{A}^*(\mathfrak{I}\overline{k}(\cdot,X) - \mathfrak{A}\overline{k}_m(\cdot,X))\right\|^2_{L^2(\mathbb{P})} \\ &\leq 2S^\ell(\Sigma_m) + 2\left\|\Sigma_m\Sigma_{m,\ell}^{-1}\right\|^2_{\mathcal{L}^\infty(\mathcal{H}_m)}\mathbb{E}\left\|\mathfrak{I}\overline{k}(\cdot,X) - \mathfrak{A}\overline{k}_m(\cdot,X)\right\|^2_{L^2(\mathbb{P})} \end{aligned}$$

and the result follows by noting that $\Sigma_m\Sigma_{m,\ell}^{-1} = P^\ell(\Sigma_m)$, $\|P^\ell(\Sigma_m)\|_{\mathcal{L}^\infty(\mathcal{H}_m)} = 1$ and applying Lemma C.2. For the lower bound, note that

$$\begin{aligned} W^\ell(\Sigma_m) &= \mathbb{E}\|\mathfrak{I}\overline{k}(\cdot,X) - \mathfrak{A}\Sigma_{m,\ell}^{-1}\mathfrak{A}^*\mathfrak{I}\overline{k}(\cdot,X)\|^2_{L^2(\mathbb{P})} \\ &= S^\ell(\Sigma_m) + \mathbb{E}\left\|\mathfrak{A}\Sigma_{m,\ell}^{-1}\mathfrak{A}^*\left[\mathfrak{I}\overline{k}(\cdot,X) - \mathfrak{A}\overline{k}_m(\cdot,X)\right]\right\|^2_{L^2(\mathbb{P})} \\ &\quad - 2\mathbb{E}\left\langle \mathfrak{I}\overline{k}(\cdot,X) - \mathfrak{A}\Sigma_{m,\ell}^{-1}\mathfrak{A}^*\mathfrak{A}\overline{k}_m(\cdot,X), \mathfrak{A}\Sigma_{m,\ell}^{-1}\mathfrak{A}^*\left[\mathfrak{I}\overline{k}(\cdot,X) - \mathfrak{A}\overline{k}_m(\cdot,X)\right]\right\rangle_{L^2(\mathbb{P})} \\ &\geq S^\ell(\Sigma_m) + \textcircled{g} - 2\sqrt{S^\ell(\Sigma_m)}\sqrt{\textcircled{g}} = \left(\sqrt{S^\ell(\Sigma_m)} - \sqrt{\textcircled{g}}\right)^2 \\ &\geq \left(\sqrt{S^\ell(\Sigma_m)} - \sqrt{\textcircled{3}}\right)^2, \end{aligned}$$

where we used

$$\begin{aligned}\textcircled{g} :&= \mathbb{E}\left\|\mathfrak{A}\Sigma_{m,\ell}^{-1}\mathfrak{A}^*\left[\mathfrak{I}\overline{k}(\cdot,X)-\mathfrak{A}\overline{k}_m(\cdot,X)\right]\right\|_{L^2(\mathbb{P})}^2 \leq \textcircled{3}\times\|\mathfrak{A}\Sigma_{m,\ell}^{-1}\mathfrak{A}^*\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))}^2\\ &= \textcircled{3}\times\|\Sigma_m\Sigma_{m,\ell}^{-1}\|_{\mathcal{L}^\infty(\mathcal{H}_m)}^2 = \textcircled{3}\end{aligned}$$

Since $\sqrt{S^\ell(\Sigma_m)} \geq \sqrt{\textcircled{4}} - \sqrt{\textcircled{3}}$ we have $W^\ell(\Sigma_m) \geq \left(\sqrt{\textcircled{4}} - 2\times\sqrt{\textcircled{3}}\right)^2$. Based on the calculation we made for the lower bound of $R_{\Sigma_m,\ell}$, for $\frac{1}{2}\sqrt{\sum_{i>\ell}\lambda_i^2} \geq 20\kappa\sqrt{\frac{\log\frac{2}{\delta}}{m}}$, we obtain $W^\ell(\Sigma_m) \geq \frac{1}{4}\sum_{i>\ell}\lambda_i^2$, which holds with probability at least $1-3\delta$ over the choice of $(\theta_i)_{i=1}^m$.

(iii) Doing as above, we obtain

$$\begin{aligned}W^\ell(\widehat{\Sigma}_m) &\leq 2S^\ell(\widehat{\Sigma}_m) + 2\mathbb{E}\left\|\sum_{i=1}^{\ell}\left\langle\mathfrak{I}\overline{k}(\cdot,X)-\mathfrak{A}\widetilde{k}_m(\cdot,X),\frac{\mathfrak{A}\phi_{m,i}}{\sqrt{\lambda_{m,i}}}\right\rangle_{L^2(\mathbb{P})}\frac{\mathfrak{A}\phi_{m,i}}{\sqrt{\lambda_{m,i}}}\right\|_{L^2(\mathbb{P})}^2\\ &= 2S^\ell(\widehat{\Sigma}_m) + 2\mathbb{E}\left\|\mathfrak{A}\widehat{\Sigma}_{m,\ell}^{-1}\mathfrak{A}^*(\mathfrak{I}\overline{k}(\cdot,X)-\mathfrak{A}\widetilde{k}_m(\cdot,X))\right\|_{L^2(\mathbb{P})}^2\\ &\leq 2S^\ell(\widehat{\Sigma}_m) + 4\left\|\mathfrak{A}\widehat{\Sigma}_{m,\ell}^{-1}\mathfrak{A}^*\right\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))}^2\mathbb{E}\left\|\mathfrak{I}\overline{k}(\cdot,X)-\mathfrak{A}\overline{k}_m(\cdot,X)\right\|_{L^2(\mathbb{P})}^2\\ &\qquad +4\mathbb{E}\left\|\mathfrak{A}\widehat{\Sigma}_{m,\ell}^{-1}\mathfrak{A}^*(\mathfrak{A}\overline{k}(\cdot,X)-\mathfrak{A}\widetilde{k}(\cdot,X))\right\|_{L^2(\mathbb{P})}^2\\ &\leq 2S^\ell(\widehat{\Sigma}_m) + 4\left\|\Sigma_m^{1/2}\widehat{\Sigma}_{m,\ell}^{-1}\Sigma_m^{1/2}\right\|_{\mathcal{L}^\infty(\mathcal{H}_m)}^2\mathbb{E}\left\|\mathfrak{I}\overline{k}(\cdot,X)-\mathfrak{A}\overline{k}_m(\cdot,X)\right\|_{L^2(\mathbb{P})}^2\\ &\qquad +4\left\|\Sigma_m^{1/2}\widehat{\Sigma}_{m,\ell}^{-1}\Sigma_m(m_{\mathbb{P},m}-\widehat{m}_{\mathbb{P},m})\right\|_{\mathcal{H}_m}^2.\end{aligned}$$

By applying Lemma C.2, the result follows from (4.45)—see the second line in the chain of equations leading to (4.45)—and (4.46). For the lower bound, note that $W^\ell(\widehat{\Sigma}_m) \geq \left(\sqrt{S^\ell(\widehat{\Sigma}_m)} - \sqrt{\textcircled{h}}\right)^2 \geq \left(\sqrt{\sum_{i>\ell}\lambda_{m,i}^2} - \sqrt{\textcircled{3}} - \sqrt{\textcircled{h}}\right)^2$, where

$$\begin{aligned}\textcircled{h} &:= \mathbb{E}\left\|\mathfrak{A}\widehat{\Sigma}_{m,\ell}^{-1}\mathfrak{A}^*\left[\mathfrak{I}\overline{k}(\cdot,X)-\mathfrak{A}\overline{k}_m(\cdot,X)\right]\right\|_{L^2(\mathbb{P})}^2 \leq \left\|\mathfrak{A}\widehat{\Sigma}_{m,\ell}^{-1}\mathfrak{A}^*\right\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))}^2\times\textcircled{3}\\ &= \left\|\Sigma_m^{1/2}\widehat{\Sigma}_{m,\ell}^{-1}\Sigma_m^{1/2}\right\|_{\mathcal{L}^\infty(\mathcal{H}_m)}^2\times\textcircled{3} \lesssim_{\Lambda^m}\textcircled{3}.\end{aligned}$$

The result therefore follows by choosing m sufficiently larger than $\frac{1}{\sum_{i>\ell}\lambda_i^2}$.

4.5.8 Proof of Theorem 4.4

(i) Note that $(\mathfrak{I}\mathfrak{I}^*)^a\mathfrak{I} = \mathfrak{I}(\mathfrak{I}^*\mathfrak{I})^a$ for any $a \in \mathbb{R}$. This follows by observing that for any $f \in \mathcal{H}$,

$$(\mathfrak{I}\mathfrak{I}^*)^a\mathfrak{I}f = \sum_i \lambda_i^a \left(\frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}} \otimes_{L^2(\mathbb{P})} \frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}}\right)\mathfrak{I}f = \mathfrak{I}\sum_i \lambda_i^{a-1}(\phi_i \otimes_{\mathcal{H}} \phi_i)\Sigma f$$
$$= \mathfrak{I}\sum_i \lambda_i^{a-1}\phi_i\langle\phi_i, \Sigma f\rangle_{\mathcal{H}} = \mathfrak{I}\sum_i \lambda_i^a\phi_i\langle\phi_i, f\rangle_{\mathcal{H}} = \mathfrak{I}\Sigma^a f = \mathfrak{I}(\mathfrak{I}^*\mathfrak{I})^a f. \quad (4.47)$$

Therefore,

$$\begin{aligned} T^\ell(\Sigma, s) &= \mathbb{E}\left\|(\mathfrak{I}\mathfrak{I}^*)^{-s/2}\,\mathfrak{I}\left(I - P^\ell(\Sigma)\right)\overline{k}(\cdot, X)\right\|^2_{L^2(\mathbb{P})} \\ &\overset{(4.47)}{=} \mathbb{E}\left\|\mathfrak{I}\Sigma^{-s/2}\left(I - P^\ell(\Sigma)\right)\overline{k}(\cdot, X)\right\|^2_{L^2(\mathbb{P})} \\ &\overset{(\dagger)}{=} \left\|\Sigma^{(1-s)/2}\left(I - P^\ell(\Sigma)\right)\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} = \left\|\Sigma^{(2-s)/2} - \Sigma_\ell^{(2-s)/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} \\ &= \mathrm{tr}\left[\Sigma^{2-s} - \Sigma_\ell^{2-s}\right], \end{aligned}$$

where $(\dagger)$ follows from Lemma A.4.

(ii) Along the lines of the proof of Theorem 4.1*(ii)* and using (4.47), it is easy to show that

$$T^\ell(\widehat{\Sigma}, s) = \left\|\Sigma^{(1-s)/2}(I - P^\ell(\widehat{\Sigma}))\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} + \left\|\Sigma^{(1-s)/2}P^\ell(\widehat{\Sigma})(m_{\mathbb{P}} - \hat{m}_{\mathbb{P}})\right\|^2_{\mathcal{H}},$$

where the second term can be bounded as

$$\|\Sigma\|^{1-s}_{\mathcal{L}^\infty(\mathcal{H})}\|P^\ell(\widehat{\Sigma})\|^2_{\mathcal{L}^\infty(\mathcal{H})}\|m_{\mathbb{P}} - \hat{m}_{\mathbb{P}}\|^2_{\mathcal{H}} \lesssim_{\mathbb{P}^n} \frac{1}{n},$$

through an application of Lemma C.1*(i)*. For the first term, employing the strategy used for bounding ①, for any $t > 0$, we obtain

$$\left\|\Sigma^{(1-s)/2}(I - P^\ell(\widehat{\Sigma}))\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})}$$
$$= \left\|\Sigma^{(1-s)/2}(\Sigma + tI)^{-1/2}(\Sigma + tI)^{1/2}(I - P^\ell(\widehat{\Sigma}))(\Sigma + tI)^{1/2}(\Sigma + tI)^{-1/2}\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})}$$

$$
\begin{aligned}
&\leq \left\|(\Sigma+tI)^{-1/2}\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} \left\|\Sigma^{(1-s)/2}(\Sigma+tI)^{-1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \\
&\qquad\qquad \times \left\|(\Sigma+tI)^{1/2}(I-P^\ell(\widehat{\Sigma}))(\Sigma+tI)^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} \\
&\leq \mathcal{N}_\Sigma(t)(\widehat{\lambda}_{\ell+1}+t)^2 \left\|\Sigma^{(1-s)/2}(\Sigma+tI)^{-1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})},
\end{aligned}
$$

where

$$
\left\|\Sigma^{(1-s)/2}(\Sigma+tI)^{-1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})} = \sup_i \frac{\lambda_i^{1-s}}{\lambda_i+t} = \sup_i \left(\frac{\lambda_i}{\lambda_i+t}\right)^{1-s} \frac{1}{(\lambda_i+t)^s} \leq \frac{1}{t^s}
$$

for $s \leq 1$. The result is completed by bounding $(\widehat{\lambda}_{\ell+1}+t)^2$ as in the proof of the upper bound in Theorem 4.1*(ii)*. The lower bound is obtained by following exactly the same idea as in the proof of the lower bound of Theorem 4.1*(ii)* by noting that

$$
\left\|\Sigma^{(1-s)/2}(I-Q_{\psi,\ell})\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} = \sum_i \lambda_i^{2-s}\left(1-\sum_{j=1}^{\ell}\langle\phi_i,\psi_j\rangle^2_{\mathcal{H}}\right)^2.
$$

(*iii*) Note that

$$
\begin{aligned}
T^\ell(\Sigma_m,s) &= \mathbb{E}_{X\sim\mathbb{P}}\left\|(\mathfrak{I}\mathfrak{I}^*)^{-s/2}\,\mathfrak{I}\overline{k}(\cdot,X) - (\mathfrak{A}\mathfrak{A}^*)^{-s/2}\,\mathfrak{A}P^\ell(\Sigma_m)\overline{k}_m(\cdot,X)\right\|^2_{L^2(\mathbb{P})} \\
&\leq 2\left(⑧+⑨\right), \qquad (4.48)
\end{aligned}
$$

where $⑧ := \mathbb{E}_{X\sim\mathbb{P}}\left\|(\mathfrak{I}\mathfrak{I}^*)^{-s/2}\,\mathfrak{I}\overline{k}(\cdot,X) - (\mathfrak{A}\mathfrak{A}^*)^{-s/2}\,\mathfrak{A}\overline{k}_m(\cdot,X)\right\|^2_{L^2(\mathbb{P})}$ and

$$
⑨ := \mathbb{E}_{X\sim\mathbb{P}}\left\|(\mathfrak{A}\mathfrak{A}^*)^{-s/2}\,\mathfrak{A}(I-P^\ell(\Sigma_m))\overline{k}_m(\cdot,X)\right\|^2_{L^2(\mathbb{P})} = \sum_{i>\ell}\lambda_{m,i}^{2-s}, \qquad (4.49)
$$

which follows by replicating the analysis in (i) for $\mathfrak{A}\mathfrak{A}^*$. To bound ⑧, we adapt the proof idea of Lemma C.2. Similar to (C.1), it can be shown that

$$
\begin{aligned}
⑧ = \|\,(\mathfrak{I}\mathfrak{I}^*)^{(2-s)/2} - (\mathfrak{A}\mathfrak{A}^*)^{(2-s)/2}\,\|^2_{\mathcal{L}^2(L^2(\mathbb{P}))} + 2\Big[&\mathrm{tr}\left((\mathfrak{I}\mathfrak{I}^*)^{-s/2}\,\mathfrak{I}\mathfrak{I}^*\mathfrak{A}\mathfrak{A}^*\,(\mathfrak{A}\mathfrak{A}^*)^{-s/2}\right) \\
&-\mathbb{E}\langle(\mathfrak{I}\mathfrak{I}^*)^{-s/2}\,\mathfrak{I}\overline{k}(\cdot,X),(\mathfrak{A}\mathfrak{A}^*)^{-s/2}\,\mathfrak{A}\overline{k}_m(\cdot,X)\rangle_{L^2(\mathbb{P})}\Big]. \qquad (4.50)
\end{aligned}
$$

Similar to (C.2) and (C.3), we obtain

$$\mathbb{E}\langle(\mathfrak{I}\mathfrak{I}^*)^{-s/2}\,\mathfrak{I}\bar{k}(\cdot,X),(\mathfrak{A}\mathfrak{A}^*)^{-s/2}\,\mathfrak{A}\bar{k}_m(\cdot,X)\rangle_{L^2(\mathbb{P})}=\int_\Theta\sum_{i=1}^m B_i(\theta)\,\langle\varphi(\cdot,\theta),\varphi_i\rangle_{L^2(\mathbb{P})}\,d\Lambda(\theta)$$

and

$$\operatorname{tr}\left((\mathfrak{I}\mathfrak{I}^*)^{-s/2}\,\mathfrak{I}\mathfrak{I}^*\mathfrak{A}\mathfrak{A}^*\,(\mathfrak{A}\mathfrak{A}^*)^{-s/2}\right)=\int_\Theta\sum_{i=1}^m B_i(\theta)\left[\langle\varphi(\cdot,\theta),\varphi_i\rangle_{L^2(\mathbb{P})}-\varphi_{\mathbb{P}}(\theta)\varphi_{i,\mathbb{P}}\right]d\Lambda(\theta),$$

where $B_i(\theta):=\left\langle(\mathfrak{I}\mathfrak{I}^*)^{-s/2}\left(\varphi(\cdot,\theta)-\varphi_{\mathbb{P}}(\theta)\right),(\mathfrak{A}\mathfrak{A}^*)^{-s/2}\left(\varphi_i-\varphi_{i,\mathbb{P}}\right)\right\rangle_{L^2(\mathbb{P})}$. This implies

$$\begin{aligned}
&\operatorname{tr}\left((\mathfrak{I}\mathfrak{I}^*)^{-s/2}\,\mathfrak{I}\mathfrak{I}^*\mathfrak{A}\mathfrak{A}^*\,(\mathfrak{A}\mathfrak{A}^*)^{-s/2}\right)\\
&\qquad-\mathbb{E}\langle(\mathfrak{I}\mathfrak{I}^*)^{-s/2}\,\mathfrak{I}\bar{k}(\cdot,X),(\mathfrak{A}\mathfrak{A}^*)^{-s/2}\,\mathfrak{A}\bar{k}_m(\cdot,X)\rangle_{L^2(\mathbb{P})}\\
&=-\left\langle(\mathfrak{I}\mathfrak{I}^*)^{-s/2}\int_\Theta A(\theta)\,d\Lambda(\theta),(\mathfrak{A}\mathfrak{A}^*)^{-s/2}\,\frac{1}{m}\sum_{i=1}^m A(\theta_i)\right\rangle_{L^2(\mathbb{P})}\\
&\overset{(*)}{\leq}\frac{1}{4}\left\|(\mathfrak{I}\mathfrak{I}^*)^{-s/2}\int_\Theta A(\theta)\,d\Lambda(\theta)-(\mathfrak{A}\mathfrak{A}^*)^{-s/2}\,\frac{1}{m}\sum_{i=1}^m A(\theta_i)\right\|^2_{L^2(\mathbb{P})}\\
&\leq\frac{1}{2}\left\|(\mathfrak{I}\mathfrak{I}^*)^{-s/2}-(\mathfrak{A}\mathfrak{A}^*)^{-s/2}\right\|^2_{\mathcal{L}^\infty(L^2(\mathbb{P}))}\left\|\int_\Theta A(\theta)\,d\Lambda(\theta)\right\|^2_{L^2(\mathbb{P})}\\
&\qquad+\left\|(\mathfrak{I}\mathfrak{I}^*)^{-s/2}-(\mathfrak{A}\mathfrak{A}^*)^{-s/2}\right\|^2_{\mathcal{L}^\infty(L^2(\mathbb{P}))}\left\|\int_\Theta A(\theta)\,d\Lambda(\theta)-\frac{1}{m}\sum_{i=1}^m A(\theta_i)\right\|^2_{L^2(\mathbb{P})}\\
&\qquad+\left\|(\mathfrak{I}\mathfrak{I}^*)^{-s/2}\right\|^2_{\mathcal{L}^\infty(L^2(\mathbb{P}))}\left\|\int_\Theta A(\theta)\,d\Lambda(\theta)-\frac{1}{m}\sum_{i=1}^m A(\theta_i)\right\|^2_{L^2(\mathbb{P})}
\end{aligned}$$

where $(*)$ follows from the parallelogram identity with $A(\theta):=\varphi(\cdot,\theta)\varphi_{\mathbb{P}}(\theta)-\varphi^2_{\mathbb{P}}(\theta)$. It therefore follows from Theorem E.1 that

$$\begin{aligned}
&\operatorname{tr}\left((\mathfrak{I}\mathfrak{I}^*)^{-s/2}\,\mathfrak{I}\mathfrak{I}^*\mathfrak{A}\mathfrak{A}^*\,(\mathfrak{A}\mathfrak{A}^*)^{-s/2}\right)-\mathbb{E}\langle(\mathfrak{I}\mathfrak{I}^*)^{-s/2}\,\mathfrak{I}\bar{k}(\cdot,X),(\mathfrak{A}\mathfrak{A}^*)^{-s/2}\,\mathfrak{A}\bar{k}_m(\cdot,X)\rangle_{L^2(\mathbb{P})}\\
\lesssim_{\Lambda^m}&\left\|(\mathfrak{I}\mathfrak{I}^*)^{-s/2}-(\mathfrak{A}\mathfrak{A}^*)^{-s/2}\right\|^2_{\mathcal{L}^\infty(L^2(\mathbb{P}))}+\frac{1}{m}\left\|(\mathfrak{I}\mathfrak{I}^*)^{-s/2}\right\|^2_{\mathcal{L}^\infty(L^2(\mathbb{P}))}.
\end{aligned}\tag{4.51}$$

Combining (4.48)–(4.51), we obtain

$$T^\ell(\Sigma_m,s)\lesssim_{\Lambda^m}\sum_{i>\ell}\lambda^{2-s}_{m,i}+\left\|(\mathfrak{I}\mathfrak{I}^*)^{(2-s)/2}-(\mathfrak{A}\mathfrak{A}^*)^{(2-s)/2}\right\|^2_{\mathcal{L}^2(L^2(\mathbb{P}))}$$

$$+ \quad \left\|(\mathfrak{I}\mathfrak{I}^*)^{-s/2} - (\mathfrak{A}\mathfrak{A}^*)^{-s/2}\right\|^2_{\mathcal{L}^\infty(L^2(\mathbb{P}))} + \frac{1}{m}\left\|(\mathfrak{I}\mathfrak{I}^*)^{-s/2}\right\|^2_{\mathcal{L}^\infty(L^2(\mathbb{P}))}, \quad (4.52)$$

where for $s \leq 1$, $\sum_{i>\ell}\lambda_{m,i}^{2-s} \lesssim \sum_{i>\ell}|\lambda_{m,i}-\lambda_i|^{2-s} + \sum_{i>\ell}\lambda_i^{2-s} \overset{(*)}{\lesssim} \|\mathfrak{I}\mathfrak{I}^* - \mathfrak{A}\mathfrak{A}^*\|^{2-s}_{2-s} + \sum_{i>\ell}\lambda_i^{2-s}$, with $(*)$ following from (Kato, 1987, Theorem II). Here $\|\cdot\|_{2-s}$ denotes the $(2-s)$-Schatten norm. Since $t \mapsto t^\alpha$ is Lipschitz on a bounded subset of $(0,\infty)$ for $\alpha \geq 1$, it follows from (De Vito et al., 2014, Lemma 7) that the second term of (4.52) is bounded (up to constants) by $\|\mathfrak{I}\mathfrak{I}^* - \mathfrak{A}\mathfrak{A}^*\|^2_{\mathcal{L}^2(L^2(\mathbb{P}))}$ for $s \leq 0$. Using the fact that $t \mapsto t^\alpha$ is operator monotone on $(0,\infty)$ for $0 \leq \alpha \leq 1$, the third term in (4.52) is bounded by $\|\mathfrak{I}\mathfrak{I}^* - \mathfrak{A}\mathfrak{A}^*\|^{-s}_{\mathcal{L}^\infty(L^2(\mathbb{P}))}$ for $-2 \leq s \leq 0$ (follows from Bhatia, 1997, Theorem X.1.1). Therefore for $-2 \leq s \leq 0$, (4.52) reduces to

$$\begin{aligned} T^\ell(\Sigma_m, s) \lesssim_{\Lambda^m} & \sum_{i>\ell}\lambda_i^{2-s} + \|\mathfrak{I}\mathfrak{I}^* - \mathfrak{A}\mathfrak{A}^*\|^{2-s}_{2-s} + \|\mathfrak{I}\mathfrak{I}^* - \mathfrak{A}\mathfrak{A}^*\|^2_{\mathcal{L}^2(L^2(\mathbb{P}))} \\ & + \|\mathfrak{I}\mathfrak{I}^* - \mathfrak{A}\mathfrak{A}^*\|^{-s}_{\mathcal{L}^\infty(L^2(\mathbb{P}))} + \frac{1}{m}. \end{aligned}$$

The result follows by applying Lemma C.3 to $\|\mathfrak{I}\mathfrak{I}^* - \mathfrak{A}\mathfrak{A}^*\|_{\mathcal{L}^2(L^2(\mathbb{P}))}$ and noting that $\|\mathfrak{I}\mathfrak{I}^* - \mathfrak{A}\mathfrak{A}^*\|_{2-s} \leq \|\mathfrak{I}\mathfrak{I}^* - \mathfrak{A}\mathfrak{A}^*\|_{\mathcal{L}^2(L^2(\mathbb{P}))}$ since $-2 \leq s \leq 0$.

The lower bound follows the idea in the proof of lower bound on $T^\ell(\Sigma_m)$ by noticing that $T^\ell(\Sigma_m, s) \geq \left(\sqrt{⑨} - \sqrt{⑧}\right)^2$ where $⑧ \lesssim m^{s/2}$ for $s \in [-2, 0)$ and $⑧ \lesssim \frac{1}{m}$ for $s = 0$. Considering $⑨ = \sum_{i>\ell}\lambda_{m,i}^{2-s}$, we have

$$\begin{aligned} \left(\sum_{i>\ell}\lambda_{m,i}^{2-s}\right)^{\frac{1}{2-s}} &\geq \left|\left(\sum_{i>\ell}\lambda_i^{2-s}\right)^{\frac{1}{2-s}} - \left(\sum_{i>\ell}|\lambda_{m,i}-\lambda_i|^{2-s}\right)^{\frac{1}{2-s}}\right| \\ &\geq \left(\sum_{i>\ell}\lambda_i^{2-s}\right)^{\frac{1}{2-s}} - \|\mathfrak{I}\mathfrak{I}^* - \mathfrak{A}\mathfrak{A}^*\|_{2-s} \gtrsim_{\Lambda^m} \left(\sum_{i>\ell}\lambda_i^{2-s}\right)^{\frac{1}{2-s}} - \frac{1}{\sqrt{m}} \\ &\gtrsim \left(\sum_{i>\ell}\lambda_i^{2-s}\right)^{\frac{1}{2-s}}, \end{aligned}$$

since $m \gtrsim \left(\sum_{i>\ell}\lambda_i^{2-s}\right)^{\frac{2}{s-2}}$. Since $m \gtrsim \left(\sum_{i>\ell}\lambda_i^{2-s}\right)^{\frac{2}{s}}$ for $s \in [-2, 0)$, it also follows that $\sqrt{⑨} - \sqrt{⑧} \gtrsim_{\Lambda^m} \left(\sum_{i>\ell}\lambda_i^{2-s}\right)^{\frac{1}{2}}$ and the result follows.

(*iv*) We skip the proof as it follows the ideas in the proof of $T^{\ell}(\widehat{\Sigma})$ combined with the bounds on ⑧ and ⑨.

Chapter 5

Nyström vs. Random Features: Which is the Best?

Direct comparison of NY-KPCA (Theorem 3.2) and RF-KCPA (Theorems 4.1 and 4.2) is not possible, as the reconstruction errors of NY-KPCA and RF-KPCA are analyzed in $\mathcal{H}$ and $L^2(\mathbb{P})$ respectively. As mentioned in Section 4.1, it is not possible to compute the reconstruction error of RF-KPCA in $\mathcal{H}$. Thus, to determine whether Nyström or random features is a better approximation method for kernel PCA, we must first establish convergence results for NY-KPCA in $L^2(\mathbb{P})$. In this Chapter, we analyze NY-EKPCA in $L^2(\mathbb{P})$ norm with respect to the R-E framework as the E-R reconstruction error is less user-friendly with the principal components being non-computable due of their dependence on $\mathbb{P}$ through $\mathfrak{I}$. The reconstruction error for NY-KPCA in the R-E framework is

$$T_{nys}^{\ell}(\widehat{\Sigma}) = \mathbb{E}\left\|\mathfrak{I}\bar{k}(\cdot, X) - \mathfrak{I}P_{nys}^{\ell}(\widehat{\Sigma})\tilde{k}(\cdot, X)\right\|_{L^2(\mathbb{P})}^2.$$

Of course, $\|\cdot\|_{L^2(\mathbb{P})}$ is weaker than $\|\cdot\|_{\mathcal{H}}$, so naturally we can expect better error behavior of $T_{nys}^{\ell}(\widehat{\Sigma})$ when compared with $R_{nys}^{\ell}(\widehat{\Sigma})$. In light of this, we now present an analogous result, proved in Section 5.1.1, for NY-EKPCA in $L^2(\mathbb{P})$-norm.

Theorem 5.1. *Under the same assumptions as in Theorem 3.2,*

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

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 5.1. (i) *Comparing Theorem 5.1 with Theorem 4.1*(ii)*, we note that the bounds for EKPCA and NY-EKPCA are identical up to constants, similar to the case in Theorem 3.2. Compared to their counterparts in Theorem 3.2, $T^{\ell}(\widehat{\Sigma})$ 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 3.2. 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 5.1. 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 $T^{\ell}(\widehat{\Sigma}_m)$ 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 corollary, proved in Section 5.1.2, examines the statisitical convergence rate of $T_{nys}^{\ell}(\widehat{\Sigma})$ under a polynomial eigenvalue decay assumption. For the reader's convenience, we present analagous results for KPCA ($T^{\ell}(\Sigma)$), EKPCA ($T^{\ell}(\widehat{\Sigma})$), and RF-KPCA ($T^{\ell}(\widehat{\Sigma}_m)$), originally given in Corollary 4.1.

Corollary 5.1. *Under the same assumptions as in Corollary 3.3, the following hold:*

(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)

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

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

Remark 5.2. *Results* (i), (ii), *and* (iv) *are quoted from Sriperumbudur and Sterge (2020, Corollary 3).*

(i) *We first highlight the difference between the $L^2(\mathbb{P})$ and $\mathcal{H}$ norms through the comparison of Corollaries 3.3 and 5.1. $T^{\ell}(\Sigma)$ decays at a rate of $n^{-2\theta(1-\frac{1}{2\alpha})}$, which is faster than that of its analog 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 3.3, NY-EKPCA and EKPCA have similar convergence behavior, provided m is large enough. Therefore, it follows from Remark 3.3, 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 the optimal 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 $O(n^{1+\frac{2\alpha}{2\alpha-1}})$.*

5.1 Proofs

5.1.1 Proof of Theorem 5.1

Upper Bound: We have

$$\begin{aligned} T^{\ell}_{nys}(\widehat{\Sigma}) &= \mathbb{E}\left\|\mathfrak{I}\bar{k}(\cdot,X) - \mathfrak{I}P^{\ell}_{nys}(\widehat{\Sigma})\tilde{k}(\cdot,X)\right\|^2_{L^2(\mathbb{P})} \\ &= \underbrace{\mathbb{E}\left\|\mathfrak{I}(I - P^{\ell}_{nys}(\widehat{\Sigma}))\bar{k}(\cdot,X)\right\|^2_{L^2(\mathbb{P})}}_{\text{(E)}} + \underbrace{\left\|\mathfrak{I}P^{\ell}_{nys}(\widehat{\Sigma})(m_{\mathbb{P}} - \widehat{m}_{\mathbb{P}})\right\|^2_{L^2(\mathbb{P})}}_{\text{(F)}} \\ &\qquad -2\mathbb{E}\left\langle \mathfrak{I}(I - P^{\ell}_{nys}(\widehat{\Sigma}))\bar{k}(\cdot,X), \mathfrak{I}P^{\ell}_{nys}(\widehat{\Sigma})(m_{\mathbb{P}} - \widehat{m}_{\mathbb{P}})\right\rangle_{L^2(\mathbb{P})}. \end{aligned} \tag{5.1}$$

The third term of (5.1) is 0, because $\mathbb{E}[\bar{k}(\cdot, X)] = 0$. Using $\Sigma = \mathfrak{I}^*\mathfrak{I}$ (Sriperumbudur and Sterge, 2020, Proposition B.2 (*iii*)) we write

$$\begin{aligned}
\text{(E)} &= \mathbb{E}\left\|\mathfrak{I}(I - P^{\ell}_{nys}(\widehat{\Sigma}))\bar{k}(\cdot, X)\right\|^2_{L^2(\mathbb{P})} = \left\langle \Sigma(I - P^{\ell}_{nys}(\widehat{\Sigma}))\bar{k}(\cdot, X), (I - P^{\ell}_{nys}(\widehat{\Sigma}))\bar{k}(\cdot, X)\right\rangle_{\mathcal{H}} \\
&\overset{(\dagger)}{=} \left\|\Sigma^{1/2}(I - P^{\ell}_{nys}(\widehat{\Sigma}))\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})}, \qquad (5.2)
\end{aligned}$$

where (†) follows from Lemma A.3. Now (5.2) is similar to (C) in the proof of Theorem 3.2(*iii*), and the proof will proceed similarly. Using a similar argument to that in (3.44), and the idempotency of $I - P^{\ell}_{nys}(\widehat{\Sigma})$, we have

$$\begin{aligned}
\left\|\Sigma^{1/2}(I - P^{\ell}_{nys}(\widehat{\Sigma}))\Sigma^{1/2}\right\|^2_{\mathcal{L}^2(\mathcal{H})} &\leq \mathcal{N}_{\Sigma}(t)\left\|\Sigma_t^{-1/2}\Sigma^{1/2}\right\|^2_{\mathcal{L}^\infty(\mathcal{H})}\left\|\Sigma_t^{1/2}(I - P^{\ell}_{nys}(\widehat{\Sigma}))\right\|^4_{\mathcal{L}^\infty(\mathcal{H})} \\
&\leq \mathcal{N}_{\Sigma}(t)\left\|\widehat{\Sigma}_t^{-1/2}\Sigma_t^{1/2}\right\|^4_{\mathcal{L}^\infty(\mathcal{H})}\left\|(I - P^{\ell}_{nys}(\widehat{\Sigma}))\widehat{\Sigma}_t^{1/2}\right\|^4_{\mathcal{L}^\infty(\mathcal{H})} \qquad (5.3)
\end{aligned}$$

The last term in (5.3) is simply the square of the last term of (3.44); therefore, we simply apply the result from (3.52), yielding

$$\mathbb{P}^n\left\{\left\|\widehat{\Sigma}_t^{1/2}(I - P^{\ell}_{nys}(\widehat{\Sigma}))\right\|^4_{\mathcal{L}^\infty(\mathcal{H})} \leq 9(9t + \lambda_{\ell+1})^2\right\} \geq 1 - 8\delta. \qquad (5.4)$$

Continuing,

$$\begin{aligned}
\text{(F)} &= \left\|\Sigma^{1/2}P^{\ell}_{nys}(\widehat{\Sigma})(m_{\mathbb{P}} - \widehat{m}_{\mathbb{P}})\right\|^2_{\mathcal{H}} \leq \|\Sigma\|^2_{\mathcal{L}^\infty(\mathcal{H})}\left\|P^{\ell}_{nys}(\widehat{\Sigma})\right\|^2_{\mathcal{L}^\infty(\mathcal{H})}\|m_{\mathbb{P}} - \widehat{m}_{\mathbb{P}}\|^2_{\mathcal{H}} \\
&\leq \kappa\,\|m_{\mathbb{P}} - \widehat{m}_{\mathbb{P}}\|^2_{\mathcal{H}} \leq \frac{32\kappa^2\log\frac{2}{\delta}}{n}, \qquad (5.5)
\end{aligned}$$

where last inequality holds with probability at least $1 - \delta$ from Lemma C.1. The result follows by applying Lemma A.2(*ii*) to the middle term in (5.3) and combining with (5.4) and (5.5).

Lower Bound: The proof of Sriperumbudur and Sterge (2020, Theorem 2(*ii*)) gives

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

The result therefore follows by noticing that

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

5.1.2 Proof of Corollary 5.1

(i),(ii) are omitted due to their similarity with analogous results in Corollary 3.1.

(iii) The lower bound follows immediately from previous results. For the upper bound, Theorem 5.1 and the proof of Corollary 3.3 *(ii)* yield

$$T^{\ell}_{nys}(\widehat{\Sigma}) \lesssim_{\mathbb{P}^n} t^{-1/\alpha}(t + n^{-\theta})^2 + \frac{1}{n}, \tag{5.6}$$

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 (5.6) 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.

Chapter 6

Discussion and Future Work

To summarize, we investigated the computational vs. statistical trade-off in the problem of approximating kernel PCA using Nyström subsampling and random features. While it is clear that Nyström kernel PCA will offer a computational advantage for Nyström subsamples $m < n$, we showed the error in reconstructing $k(\cdot, X)$ in $\mathcal{H}$ using ℓ-eigenfunctions in Nyström kernel PCA to be statistically optimal when compared to standard kernel PCA, provided m is large enough, but still $m < n$, 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. Additionally, unlike several existing theoretical works on kernel PCA, we derived these results by not assuming the mean element of k to be zero. Similarly, in the case of random features, while it is obvious that approximate kernel PCA using m random features has lower computational complexity than kernel PCA when $m < n$ with n being the number of samples, it is not obvious that this computational gain is not achieved at the cost of statistical efficiency. Through inclusion and approximation operators, we explored various notions of reconstructing a kernel function using ℓ eigenfunctions, wherein we showed that approximate kernel PCA has computational advantage with no loss in statistical optimality as long as m is large enough (but still $m < n$) and ℓ is small enough with m depending on the number of eigenfunctions ℓ being considered. If ℓ is large, then more features are needed to maintain the statistical behavior, thereby resulting in the loss of computational advantage.

Further, we derived results for Nyström kernel PCA in $L^2(\mathbb{P})$ to allow for comparison between Nyström kernel PCA and 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 that the reconstruction error of Nyström KPCA converges to zero faster than random features with less computational complexity.

There are few open questions in this topic which may be of interest to address.

(i) In the case of random-features, as mentioned in Section 4.1, it is not possible to compute the reconstruction error directly in $\mathcal{H}$. Thus, we lift the reconstructions into $L^2(\mathbb{P})$ to compute the error. Though the reconstruction error considered is analytically convenient, and makes sense theoretically, there is no obvious, natural choice. A detailed empirical study of the errors considered could help in justifying the choice of error criterion.

(ii) In contrast to the setting of this paper where ℓ grows with n, it may be of interest to consider asymptotics when ℓ is fixed but $n \to \infty$. In such a setting, one may investigate E-R, R-E and their variations/generalizations. For example, in R-E, we can compare EKPCA and RF-EKPCA by comparing $T^{\ell}(\widehat{\Sigma}) - T^{\ell}(\Sigma)$ and $T^{\ell}(\widehat{\Sigma}_m) - T^{\ell}(\Sigma)$. While Theorems 4.1, 4.2, and 4.4 do not directly specialize to the setting of fixed ℓ, using ideas employed in their proofs, upper bounds can be derived on $T^{\ell}(\widehat{\Sigma}) - T^{\ell}(\Sigma)$ and $T^{\ell}(\widehat{\Sigma}_m) - T^{\ell}(\Sigma)$. However, lower bounds are needed to establish the sharpness of these upper bounds so that these excess errors can be matched for a certain choice of m.

(iii) Apart from reconstruction error, one may compare ℓ-eigenspaces (for fixed ℓ) associated with EKPCA and RF-EKPCA by comparing the corresponding projection operators through their embeddings as bounded operators on $L^2(\mathbb{P})$. Ullah et al. (2018) investigated this direction 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). Different

meaningful notions of comparing the projection operators can be explored and upper convergence rates can be derived using the perturbation theory for self-adjoint operators. However, as above, developing lower bounds will be critical to establish the sharpness of the upper bounds, thereby facilitating a meaningful comparison of the statistical performances of EKPCA and RF-EKPCA.

Though supervised kernel learning problems have received a theoretical treatment in a large number of works as mentioned in Section 2, this work establishes some of the first rigorous theoretical studies in the unsupervised case. Among further interest in the unsupervised setting is the application of approximation methods to kernel canonical correlation analysis (KCCA), which is a close relative of KPCA. For random variables X and Y, and reproducing kernel Hilbert spaces $\mathcal{H}$ and $\mathcal{G}$, KCCA finds functions $f \in \mathcal{H}$ and $g \in \mathcal{G}$ such that the correlation between $f(X)$ and $g(Y)$ is maximized. More formally, in the population KCCA solves

$$\arg \sup_{\substack{\langle f, \Sigma_X f \rangle_{\mathcal{H}} = 1 \\ \langle g, \Sigma_Y g \rangle_{\mathcal{G}} = 1}} \langle f, \Sigma_{XY} g \rangle_{\mathcal{H}} = \mathrm{Cov}(f(X), g(Y)), \tag{6.1}$$

where Σ_X and Σ_Y are the covariance operators of X and Y respectively, and Σ_{XY} is the *cross-covariance operator*. Of course, when $Y = X$ and $\mathcal{G} = \mathcal{H}$ the problem exactly reduces to KPCA. The theoretical properties of kernel CCA have been well established (Gao et al., 2015); however, like KPCA they suffer from similar computational issues. It is likely that random features and the Nyström approximation can provide a similar benefit in KCCA as in KPCA. The interplay between Σ_X and Σ_Y in terms of their effect on the computational and statistical trade-off is of specific interest.

Extending this work, it becomes interesting to consider the effectiveness of these approximation methods in nonparametric testing through kernel embeddings of distributions. The mean element $m_{\mathbb{P}} = \int k(\cdot, x) d\mathbb{P}(x)$ maps the probability distribution $\mathbb{P}$ into the RKHS associated with k. The distance between probability distributions $\mathbb{P}, \mathbb{Q}$ can thus be measured by $\|m_{\mathbb{P}} - \hat{m}_{\mathbb{Q}}\|_{\mathcal{H}}$. Given samples from $\mathbb{P}, \mathbb{Q}$, the difference in their mean-embeddings can be estimated through $\|\hat{m}_{\mathbb{P}} - \hat{m}_{\mathbb{Q}}\|_{\mathcal{H}}$, where $\hat{m}_{\mathbb{P}} = \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i)$.

This idea of kernel mean embedding can be used to develop statistical procedures such as testing. For example, given a random sample $(X_i)_{i=1}^n$ from a distribution $\mathbb{P}$, a goodness-of-fit test would check if the sample came from a certain distribution P_0. A kernel goodness-of-fit considers the null hypothesis $H_0 : \mathbb{P} = \mathbb{P}_0$, rejecting H_0 if $\|\hat{m}_{\mathbb{P}} - m_{\mathbb{Q}}\|_{\mathcal{H}}$ is sufficiently large. Similar tests for independence and homogeneity, often called two-sample testing, may be constructed using kernel mean embeddings. In a similar vein to the contributions of this dissertation, the question using kernel approximation methods to lessen the computational burden in tests of these types is of interest. However, there is an issue from the onset. Little is known about the statistical performance of these types of tests. Balasubramanian et al. (2021) shows that goodness-of-fit tests using kernel mean embeddings for a fixed kernel are suboptimal and proposes an alternative embedding method to lead to more powerful tests. The statistical properties of tests for goodness-of-fit, homogeneity, and independence using Gaussian kernel mean embeddings have been studied by Li and Yuan (2019), where it is shown that such tests are minimax optimal against smooth alternatives. This existing work motivates two future directions for research: Generalizing the existing statistical analysis from the Gaussian case to a more broad class of kernels. Investigating whether approximation methods such as Nyström and random features can be used to create approximate kernel embedding based tests which have lesser computational complexity than traditional tests, but with no statistical loss.

Appendix A

Technical Results

Lemma A.1. *Define* $C := \int_{\mathcal{X}} k(\cdot, x) \otimes_{\mathcal{H}} k(\cdot, x) d\mathbb{P}(x)$, $\hat{C} := \frac{1}{n}\sum_{i=1}^{n} k(\cdot, X_i) \otimes_{\mathcal{H}} k(\cdot, X_i)$, $\lambda_1 = \|C\|_{\mathcal{L}^\infty(\mathcal{H})}$, $\lambda_\ell := \lambda_\ell(C)$, *and* $\widehat{\lambda}_\ell := \lambda_\ell(\hat{C})$. *For* $\delta > 0$, *suppose* $\frac{9\kappa}{n}\log\frac{n}{\delta} \le t \le \lambda_1$. *Then the following hold:*

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

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

(iii) $\mathbb{P}^n\left\{\widehat{\lambda}_\ell + t \le \frac{3}{2}(\lambda_\ell + t)\right\} \ge 1-\delta.$

(iv) $\mathbb{P}^n\left\{\lambda_\ell + t \le 2(\widehat{\lambda}_\ell + t)\right\} \ge 1-\delta.$

Proof. (i) The result is quoted from Lemma 3.6 of (Rudi et al., 2013) with $\alpha = \frac{1}{2}$.
(ii) This is a slight variation of (i) and the proof idea follows that of Lemma 3.6 of (Rudi et al., 2013) with $\alpha = \frac{1}{2}$. Note that

$$\left\|(C+tI)^{-1/2}(\hat{C}+tI)^{1/2}\right\|_{\mathcal{L}^\infty(\mathcal{H})} = \left\|(C+tI)^{-1/2}(\hat{C}+tI)(C+tI)^{-1/2}\right\|^{1/2}_{\mathcal{L}^\infty(\mathcal{H})}.$$

By defining $B_n = (C+tI)^{-1/2}(C-\hat{C})(C+tI)^{-1/2}$, we have

$$I - B_n = (C+tI)^{-1/2}\left((C+tI) - C + \hat{C}\right)(C+tI)^{-1/2} = (C+tI)^{-1/2}(\hat{C}+tI)(C+tI)^{-1/2}$$

and therefore

$$\left\|(C+tI)^{-1/2}(\hat{C}+tI)^{1/2}\right\|_{\mathcal{L}^\infty(\mathcal{H})} = \|I-B_n\|^{1/2}_{\mathcal{L}^\infty(\mathcal{H})} \leq \left(1+\|B_n\|_{\mathcal{L}^\infty(\mathcal{H})}\right)^{1/2}. \tag{A.1}$$

It follow from the proof of Lemma 3.6 of (Rudi et al., 2013) that for $\frac{9\kappa}{n}\log\frac{n}{\delta} \leq t$,

$$\mathbb{P}^n\left\{\|B_n\|_{\mathcal{L}^\infty(\mathcal{H})} \leq \frac{1}{2}\right\} \geq 1-\delta. \tag{A.2}$$

Combining (A.1) and (A.2) completes the proof.
(iii) Since $\sqrt{\frac{2}{3}} \leq \left\|(C+tI)^{1/2}(\hat{C}+tI)^{-1/2}\right\|_{\mathcal{L}^\infty(\mathcal{H})}$ as obtained in (i), it is equivalent (see Rudi et al. 2013, Lemmas B.2 and 3.5) to $\hat{C}+tI \preceq \frac{3}{2}(C+tI)$. This implies (see Gohberg and Goldberg, 2003) that $\widehat{\lambda}_k + t \leq \frac{3}{2}(\lambda_k + t)$ for all $k \geq 1$. *(iv)* follows similarly. □

Lemma A.2 (Sriperumbudur and Sterge, 2020, Lemma A.1)**.** *Let H be a separable Hilbert space and $\mathcal{Y}$ be a separable topological space. Define*

$$\mathfrak{C} = \frac{1}{2}\int_{\mathcal{Y}}\int_{\mathcal{Y}} (s(x)-s(y)) \otimes_H (s(x)-s(y))\, dP(x)\, dP(y)$$

where $s : \mathcal{Y} \to H$ is a Bochner-measurable function with $\sup_{x\in\mathcal{Y}} \|s(x)\|_H^2 = \kappa$. Given $(Y_i)_{i=1}^r \overset{i.i.d.}{\sim} P$ with $r \geq 2$, define

$$\widehat{\mathfrak{C}} = \frac{1}{2r(r-1)}\sum_{i\neq j}^{r} (s(Y_i)-s(Y_j)) \otimes_H (s(Y_i)-s(Y_j)).$$

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

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

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

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

(iv) $P^r\left\{(Y_i)_{i=1}^r : \lambda_\ell(\widehat{\mathfrak{C}}) + t \leq \frac{3}{2}(\lambda_\ell(\mathfrak{C})+t)\right\} \geq 1-2\delta$ *for all* $\ell \geq 1;$

(v) $P^r\left\{(Y_i)_{i=1}^r : \lambda_\ell(\mathfrak{C}) + t \leq 2(\lambda_\ell(\widehat{\mathfrak{C}})+t)\right\} \geq 1-2\delta$ *for all* $\ell \geq 1.$

In addition, for any $0 < t \leq \|\mathfrak{C}\|_{\mathcal{L}^\infty(H)}$,

$$P^r \left\{ (Y_i)_{i=1}^r : \left\| \mathfrak{C}_t^{-1/2} (\widehat{\mathfrak{C}} - \mathfrak{C}) \mathfrak{C}^{1/2} \right\|_{\mathcal{L}^2(H)} \leq \sqrt{\frac{64\kappa^{5/2} \mathcal{N}_{\mathfrak{C}}(t) \log \frac{2}{\delta}}{r\sqrt{t}}} + \frac{32\sqrt{2}\kappa^{3/2} \log \frac{3}{\delta}}{r\sqrt{t}} \right\} \geq 1 - 2\delta, \tag{A.3}$$

where $\mathfrak{C}_t := (\mathfrak{C} + tI)$ *and* $\mathcal{N}_{\mathfrak{C}}(t) = \mathrm{tr}(\mathfrak{C}_t^{-1}\mathfrak{C})$.

Proof. (i) Define $A(x, y) := \frac{1}{\sqrt{2}}(s(x) - s(y))$,

$$U(x, y) := (\mathfrak{C} + tI)^{-1/2} A(x, y) \in H$$

and $Z(x, y) := U(x, y) \otimes_H U(x, y)$. Clearly $Z(x, y) = Z(y, x)$ and

$$(\mathfrak{C} + tI)^{-1/2} (\widehat{\mathfrak{C}} - \mathfrak{C})(\mathfrak{C} + tI)^{-1/2} = \frac{1}{r(r-1)} \sum_{i \neq j}^{r} Z(Y_i, Y_j) - \mathbb{E}[Z(X, Y)].$$

Also

$$\begin{aligned} \sup_{x,y \in \mathcal{Y}} \|Z(x, y)\|_{\mathcal{L}^2(H)} &\overset{(*)}{=} \sup_{x,y \in \mathcal{Y}} \|U(x, y)\|_H^2 \\ &\leq \frac{1}{2} \sup_{x,y \in \mathcal{Y}} \|(\mathfrak{C} + tI)^{-1/2} (s(x) - s(y))\|_H^2 = \frac{2\kappa}{t}, \end{aligned}$$

where $(*)$ follows from Lemma A.5. Define $\psi(x) := \mathbb{E}_Y[Z(x, Y)] = \mathbb{E}_Y[U(x, Y) \otimes_H U(x, Y)]$. Clearly,

$$\sup_{x \in \mathcal{Y}} \|\psi(x)\|_{\mathcal{L}^\infty(H)} \leq \sup_{x,y \in \mathcal{Y}} \|U(x, y) \otimes_H U(x, y)\|_{\mathcal{L}^\infty(H)} \overset{(\dagger)}{=} \sup_{x,y \in \mathcal{Y}} \|U(x, y)\|_H^2 \leq \frac{2\kappa}{t},$$

where $(\dagger)$ follows from Lemma A.5. Since $\mathbb{E}[\psi(X)] = \mathbb{E}[Z(X, Y)]$, and

$$\mathbb{E}[(\psi(X) - \mathbb{E}[Z(X, Y)])^2] = \mathbb{E}[\psi^2(X)] - \mathbb{E}^2[Z(X, Y)] \preceq \mathbb{E}[\psi^2(X)].$$

By defining $\mathfrak{C}_t = \mathfrak{C} + tI$, we have

$$\begin{aligned}
&\mathbb{E}[\psi^2(X)] = \mathbb{E}[\mathbb{E}_Y^2[U(X,Y) \otimes_H U(X,Y)]] \\
&= \mathbb{E}\left[\mathfrak{C}_t^{-1/2}\mathbb{E}_Y[A(X,Y) \otimes_H A(X,Y)]\mathfrak{C}_t^{-1}\mathbb{E}_Y[A(X,Y) \otimes_H A(X,Y)]\mathfrak{C}_t^{-1/2}\right] \\
&\preceq \sup_{x \in \mathcal{Y}} \|\mathfrak{C}_t^{-1/2}\mathbb{E}_Y[A(x,Y) \otimes_H A(x,Y)]\mathfrak{C}_t^{-1/2}\|_{\mathcal{L}^\infty(H)} \\
&\qquad\qquad \times \mathbb{E}\left[\mathfrak{C}_t^{-1/2}\mathbb{E}_Y[A(X,Y) \otimes_H A(X,Y)]\mathfrak{C}_t^{-1/2}\right] \\
&\preceq \frac{2\kappa}{t}(\mathfrak{C} + tI)^{-1/2}\mathfrak{C}(\mathfrak{C} + tI)^{-1/2} =: S.
\end{aligned}$$

Note that $\|S\|_{\mathcal{L}^\infty(H)} \leq \frac{2\kappa}{t}$ and

$$d := \frac{\|S\|_{\mathcal{L}^1(H)}}{\|S\|_{\mathcal{L}^\infty(H)}} = \frac{\mathrm{tr}(\mathfrak{C}_t^{-1}\mathfrak{C})}{\|\mathfrak{C}_t^{-1}\mathfrak{C}\|_{\mathcal{L}^\infty(H)}} \leq \frac{(\|\mathfrak{C}\|_{\mathcal{L}^\infty(H)} + t)\mathrm{tr}(\mathfrak{C}_t^{-1}\mathfrak{C})}{\|\mathfrak{C}\|_{\mathcal{L}^\infty(H)}}.$$

Therefore, applying Theorem E.3 yields that for $0 < \delta \leq d$ with probability at least $1 - 2\delta$,

$$\begin{aligned}
\left\|(\mathfrak{C} + tI)^{-1/2}(\widehat{\mathfrak{C}} - \mathfrak{C})(\mathfrak{C} + tI)^{-1/2}\right\|_{\mathcal{L}^\infty(H)} &\leq \frac{4\kappa\beta}{rt} + \sqrt{\frac{24\kappa\beta}{rt}} + \frac{16\kappa\log\frac{3}{\delta}}{rt} \\
&\leq \frac{4\kappa\beta}{rt} + \sqrt{\frac{24\kappa\beta}{rt}} + \frac{24\kappa\beta}{rt} \\
&= \frac{28\kappa\beta}{rt} + \sqrt{\frac{24\kappa\beta}{rt}}, \qquad \text{(A.4)}
\end{aligned}$$

where $\beta = \frac{2}{3}\log\frac{4d}{\delta}$ and we used that fact that $d > 1$ in the second line. Since $t \geq \frac{140\kappa}{r}\log\frac{16\kappa r}{\delta}$, it follows that $t \geq \frac{140\kappa}{r}\log\frac{4d}{\delta}$ as $\frac{140\kappa}{r}\log\frac{16\kappa r}{\delta} \geq \frac{140\kappa}{r}\log\frac{16\kappa}{t\delta} \geq \frac{140\kappa}{r}\log\frac{4d}{\delta}$ where we use the fact that $d \leq \frac{4\kappa}{t}$ which follows from $t \leq \|\mathfrak{C}\|_{\mathcal{L}^\infty(H)}$ and $\mathrm{tr}(\mathfrak{C}_t^{-1}\mathfrak{C}) \leq \frac{\mathrm{tr}(\mathfrak{C})}{t} \leq \frac{2\kappa}{t}$. This implies $t \geq \frac{210\kappa\beta}{r}$ or $\frac{\kappa\beta}{rt} \leq \frac{1}{210}$. Using this in (A.4) yields the result.

(ii) By defining $B_n = (\mathfrak{C} + tI)^{-1/2}(\mathfrak{C} - \widehat{\mathfrak{C}})(\mathfrak{C} + tI)^{-1/2}$, we have

$$\begin{aligned}
&\left\|(\mathfrak{C} + tI)^{1/2}(\widehat{\mathfrak{C}} + tI)^{-1/2}\right\|_{\mathcal{L}^\infty(H)} \\
&= \left\|(\widehat{\mathfrak{C}} + tI)^{-1/2}(\mathfrak{C} + tI)(\widehat{\mathfrak{C}} + tI)^{-1/2}\right\|_{\mathcal{L}^\infty(H)}^{1/2} \\
&= \left\|(\mathfrak{C} + tI)^{1/2}(\widehat{\mathfrak{C}} + tI)^{-1}(\mathfrak{C} + tI)^{1/2}\right\|_{\mathcal{L}^\infty(H)}^{1/2}
\end{aligned}$$

$$= \left\|(I - B_n)^{-1}\right\|^{1/2}_{\mathcal{L}^\infty(H)} \leq (1 - \|B_n\|_{\mathcal{L}^\infty(H)})^{-1/2},$$

where the last inequality holds whenever $\|B_n\|_{\mathcal{L}^\infty(H)} < 1$. Similarly,

$$\begin{aligned}\left\|(\mathfrak{C} + tI)^{1/2}(\widehat{\mathfrak{C}} + tI)^{-1/2}\right\|_{\mathcal{L}^\infty(H)} &= \left\|(I + (-B_n))^{-1}\right\|^{1/2}_{\mathcal{L}^\infty(H)} \\ &\geq (1 + \|B_n\|_{\mathcal{L}^\infty(H)})^{-1/2}.\end{aligned}$$

The result therefore follows from (i).

(iii) Since

$$\left\|(\mathfrak{C} + tI)^{-1/2}(\widehat{\mathfrak{C}} + tI)^{1/2}\right\|_{\mathcal{L}^\infty(H)} = \|I - B_n\|^{1/2}_{\mathcal{L}^\infty(H)} \leq (1 + \|B_n\|_{\mathcal{L}^\infty(H)})^{1/2},$$

the result follows from (i).

(iv) Since $\sqrt{\frac{2}{3}} \leq \left\|(\mathfrak{C} + tI)^{1/2}(\widehat{\mathfrak{C}} + tI)^{-1/2}\right\|_{\mathcal{L}^\infty(H)} \leq \sqrt{2}$ as obtained in (i), it follows that $\widehat{\mathfrak{C}} + tI \preceq \frac{3}{2}(\mathfrak{C} + tI)$ (see Rudi et al., 2013, Lemmas B.2 and 3.5). This implies (see Gohberg and Goldberg, 2003) that $\lambda_\ell(\widehat{\mathfrak{C}}) + t \leq \frac{3}{2}(\lambda_\ell(\mathfrak{C}) + t)$ for all $\ell \geq 1$. (v) follows similarly.

Proof of (A.3)*:* Define $Z(x, y) := \mathfrak{C}_t^{-1/2}(A(x, y) \otimes_H A(x, y))\mathfrak{C}^{1/2}$ so that $Z(x, y) = Z(y, x)$ and

$$\mathfrak{C}_t^{-1/2}(\hat{\mathfrak{C}} - \mathfrak{C})\mathfrak{C}^{1/2} = \frac{1}{r(r-1)} \sum_{i \neq j}^{r} Z(X_i, X_j) - \mathbb{E}[Z(X, Y)].$$

We have

$$\sup_{x,y \in \mathcal{X}} \|Z(x, y)\|_{\mathcal{L}^2(H)} \leq \|\mathfrak{C}_t^{-1/2}\|_{\mathcal{L}^\infty(H)} \left\|\mathfrak{C}^{1/2}\right\|_{\mathcal{L}^\infty(H)} \|A(x, y)\|^2_H \leq \frac{(2\kappa)^{3/2}}{\sqrt{t}} := M.$$

By defining $\psi(x) := \mathbb{E}_Y[Z(x, Y)]$, we have

$$\begin{aligned}&\mathbb{E}\|\psi(X) - \mathfrak{C}\|^2_{\mathcal{L}^2(H)} = \mathbb{E}\|\psi(X)\|^2_{\mathcal{L}^2(H)} - \|\mathfrak{C}\|^2_{\mathcal{L}^2(H)} \leq \mathbb{E}\|\psi(X)\|^2_{\mathcal{L}^2(H)} \\ &= \mathbb{E}\left\|\mathfrak{C}_t^{-1/2}\mathbb{E}_Y[A(X, Y) \otimes_H A(X, Y)]\mathfrak{C}^{1/2}\right\|^2_{\mathcal{L}^2(H)} \\ &= \mathbb{E}\,\mathrm{tr}\left[\mathfrak{C}^{1/2}\mathbb{E}_Y[A(X, Y) \otimes_H A(X, Y)]\mathfrak{C}_t^{-1}\mathbb{E}_Y[A(X, Y) \otimes_H A(X, Y)]\mathfrak{C}^{1/2}\right] \\ &= \mathbb{E}\,\mathrm{tr}\left[\mathfrak{C}_t^{-1/2}\mathbb{E}_Y[A(X, Y) \otimes_H A(X, Y)]\mathfrak{C}_t^{-1}\mathbb{E}_Y[A(X, Y) \otimes_H A(X, Y)]\mathfrak{C}\mathfrak{C}_t^{1/2}\right]\end{aligned}$$

$$
\begin{aligned}
&\leq \sup_{x\in\mathcal{X}} \left\| \mathfrak{C}_t^{-1/2} \mathbb{E}_Y[A(X,Y)\otimes_H A(X,Y)]\mathfrak{C}\mathfrak{C}_t^{1/2} \right\|_{\mathcal{L}^\infty(H)} \\
&\qquad\qquad \times \mathbb{E}\,\mathrm{tr}\left[\mathfrak{C}_t^{-1/2}\mathbb{E}_Y[A(X,Y)\otimes_H A(X,Y)]\mathfrak{C}_t^{-1/2}\right] \\
&\leq \left\|\mathfrak{C}_t^{-1/2}\right\|_{\mathcal{L}^\infty(H)} \|\mathfrak{C}\|_{\mathcal{L}^\infty(H)} \left\|\mathfrak{C}_t^{1/2}\right\|_{\mathcal{L}^\infty(H)} \mathrm{tr}\left[\mathfrak{C}_t^{-1}\mathfrak{C}\right] \\
&\qquad\qquad \times \sup_{x,y\in\mathcal{X}} \|A(x,y)\otimes_H A(x,y)\|_{\mathcal{L}^\infty(H)} \\
&\leq \sqrt{2\kappa+t}(2\kappa)^2 \frac{\mathcal{N}_{\mathfrak{C}}(t)}{\sqrt{t}} \leq \sqrt{2}(2\kappa)^{5/2}\frac{\mathcal{N}_{\mathfrak{C}}(t)}{\sqrt{t}},
\end{aligned}
$$

where we used $t \leq \|\mathfrak{C}\|_{\mathcal{L}^\infty(H)} \leq 2\kappa$ in the last inequality. The result follows by applying Theorem E.3 *(ii)*. □

Lemma A.3. *For any orthogonal projector $P : \mathcal{H} \to \mathcal{H}$, the following holds:*

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

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

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

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

$$
\begin{aligned}
\mathbb{E}\left\langle I-P, \bar{k}(\cdot,X)\otimes_{\mathcal{H}}\bar{k}(\cdot,X)\right\rangle_{\mathcal{L}^2(\mathcal{H})} &= \left\langle I-P, \mathbb{E}[\bar{k}(\cdot,X)\otimes_{\mathcal{H}}\bar{k}(\cdot,X)]\right\rangle_{\mathcal{L}^2(\mathcal{H})} \\
&= \langle I-P,\Sigma\rangle_{\mathcal{L}^2(\mathcal{H})} = \mathrm{tr}\left((I-P)\Sigma\right) \\
&= \mathrm{tr}\left(\Sigma^{1/2}(I-P)^2\Sigma^{1/2}\right). \qquad \text{(A.5)}
\end{aligned}
$$

The proof is completed by using $\left\|(I-P)\Sigma^{1/2}\right\|_{\mathcal{L}^2(\mathcal{H})}^2 = \mathrm{tr}\left(\Sigma^{1/2}(I-P)^2\Sigma^{1/2}\right)$ in (A.5). □

Lemma A.4. *Let X be a separable topological space, H be a separable Hilbert space and ρ be a probability measure on X. Suppose $v : X \to H$ is Bochner-measurable and $\mathbb{E}_\rho \|v\|_H^2 :=$*

$\int_X \|v(x)\|_H^2 \, d\rho(x) < \infty$. *Define* $A = B^*B = \int_X v(x) \otimes_H v(x) \, d\rho(x) =: \mathbb{E}_\rho[v \otimes_H v]$ *where* $B : H \to G$ *and* G *is a separable Hilbert space. Then for any* $Q : H \to H$,

$$\mathbb{E}_\rho \|BQv\|_G^2 = \left\| A^{1/2} Q A^{1/2} \right\|_{\mathcal{L}^2(H)}^2 .$$

Proof. Note that

$$\begin{aligned} \mathbb{E}_\rho \|BQv\|_G^2 = \mathbb{E}_\rho \langle BQv, BQv \rangle_G &= \mathbb{E}_\rho \langle Q^*AQv, v \rangle_H \\ &= \mathbb{E}_\rho \langle Q^*AQ, v \otimes_H v \rangle_{\mathcal{L}^2(H)} . \end{aligned}$$

Since v is Bochner-measurable and $\mathbb{E}_\rho \|v\|_H^2 < \infty$, it is Bochner integrable, which yields

$$\mathbb{E}_\rho \langle Q^*AQ, v \otimes_H v \rangle_{\mathcal{L}^2(H)} = \langle Q^*AQ, \mathbb{E}_\rho[v \otimes_H v] \rangle_{\mathcal{L}^2(H)} = \langle Q^*AQ, A \rangle_{\mathcal{L}^2(H)} .$$

The result follows by noting that

$$\begin{aligned} \langle Q^*AQ, A \rangle_{\mathcal{L}^2(H)} = \operatorname{tr}(Q^*AQA) &= \operatorname{tr}\left(A^{1/2} Q^* A^{1/2} A^{1/2} Q A^{1/2} \right) \\ &= \left\| A^{1/2} Q A^{1/2} \right\|_{\mathcal{L}^2(H)}^2 , \end{aligned}$$

where we have used invariance of the trace under cyclic permutations. □

Lemma A.5. *Define* $B = f \otimes_H f$ *where* H *is a separable Hilbert space and* $f \in H$. *Then* $\|B\|_{\mathcal{L}^\infty(H)} = \|B\|_{\mathcal{L}^2(H)} = \|B\|_{\mathcal{L}^1(H)} = \|f\|_H^2$.

Proof. Since B is self-adjoint,

$$\|B\|_{\mathcal{L}^\infty(H)} = \lambda_1(B) = \sup_{\|g\|_H = 1} \langle g, Bg \rangle_H = \sup_{\|g\|_H = 1} \langle f, g \rangle_H^2 = \|f\|_H^2 .$$

Note that $\|B\|_{\mathcal{L}^1(H)} = \sum_j \langle e_j, (f \otimes_H f) e_j \rangle_H = \sum_j \langle f, e_j \rangle_H^2 = \|f\|_H^2$ for any orthonormal basis $(e_j)_j$ in H. □

Proposition A.1. *Suppose* $\underline{A} i^{-\alpha} \le \lambda_i(\mathfrak{C}) \le \bar{A} i^{-\alpha}$ *for* $\alpha > 1$ *and* $\underline{A}, \bar{A} \in (0, \infty)$. *Define*

$\mathcal{N}_{\mathfrak{C}}(t) := tr((\mathfrak{C}+tI)^{-1}\mathfrak{C})$ *The following holds:*

$$\mathcal{N}_{\mathfrak{C}}(t) \lesssim t^{-1/\alpha}.$$

Proof. We have

$$\mathrm{tr}\left((\mathfrak{C}+tI)^{-1}\mathfrak{C}\right) = \sum_{i\geq 1}\frac{\lambda_i(\mathfrak{C})}{\lambda_i(\mathfrak{C})+t} \leq \sum_{i\geq 1}\frac{\bar{A}i^{-\alpha}}{\underline{A}i^{-\alpha}+t} = \frac{\bar{A}}{\underline{A}}\sum_{i\geq 1}\frac{i^{-\alpha}}{i^{-\alpha}+t\underline{A}^{-1}}.$$

Let $u = t^{1/\alpha}\underline{A}^{-1/\alpha}x \implies u^\alpha = t\underline{A}^{-1}x^\alpha$ and $dx = t^{-1/\alpha}\underline{A}^{1/\alpha}du$. Therefore,

$$\sum_{i\geq 1}\frac{i^{-\alpha}}{i^{-\alpha}+t\underline{A}^{-1}} \leq \int_0^\infty \frac{x^{-\alpha}}{x^{-\alpha}+t\underline{A}^{-1}}\,dx = \int_0^\infty \frac{1}{1+t\underline{A}^{-1}x^\alpha}dx = \left(\frac{\underline{A}}{t}\right)^{1/\alpha}\int_0^\infty \frac{1}{1+u^\alpha}du.$$

Since $\frac{1}{1+u^\alpha}$ is decreasing in α on $u\in(0,\infty)$, we have

$$\frac{1}{1+u^\alpha} \leq \frac{1}{1+u^2}, \quad \text{if} \quad \alpha \geq 2.$$

So for $\alpha \geq 2$,

$$\left(\frac{\underline{A}}{t}\right)^{1/\alpha}\int_0^\infty \frac{1}{1+u^\alpha}du \lesssim t^{-1/\alpha}\int_0^\infty \frac{1}{1+u^2}du = t^{-1/\alpha}\left[\tan^{-1}(u)|_0^\infty\right] = \frac{\pi}{2}t^{-1/\alpha},$$

implying $\mathcal{N}_{\mathfrak{C}}(t) \lesssim t^{-1/\alpha}$. For $1<\alpha<2$, we obtain

$$t^{-1/\alpha}\int_0^\infty \frac{1}{1+u^\alpha}du \leq t^{-1/\alpha}\sum_{k=0}^\infty \frac{1}{1+k^\alpha} \leq t^{-1/\alpha}\left(1+\sum_{k=1}^\infty \frac{1}{k^\alpha}\right).$$

Since $1+\sum_{k=1}^\infty \frac{1}{k^\alpha}$ converges for $\alpha>1$, we obtain $\mathcal{N}_{\mathfrak{C}}(t) \lesssim t^{-1/\alpha}$. □

Proposition A.2. *Suppose* $\underline{B}e^{-\tau i} \leq \lambda_i(\mathfrak{C}) \leq \bar{B}e^{-\tau i}$ *for* $\tau>0$ *and* $\underline{B},\bar{B}\in(0,\infty)$. *Let* $\ell = \frac{1}{\tau}\log n^\theta$, $\theta>0$. *Then*

$$\mathcal{N}_{\mathfrak{C}}(t) \lesssim \log\left(\frac{1}{t}\right).$$

Proof. We have

$$\mathcal{N}_{\mathfrak{C}}(t) = \operatorname{tr}\left((\mathfrak{C}+tI)^{-1}\mathfrak{C}\right) = \sum_{i\geq 1}\frac{\lambda_i(\mathfrak{C})}{\lambda_i(\mathfrak{C})+t} \leq \frac{\bar{B}e^{-\tau i}}{\underline{\mathrm{B}}e^{-\tau i}+t} = \frac{\bar{B}}{\underline{\mathrm{B}}}\sum_{i\geq 1}\frac{1}{1+t\underline{\mathrm{B}}^{-1}e^{\tau i}}$$

$$\lesssim \int_0^\infty \frac{1}{1+t\underline{\mathrm{B}}^{-1}e^{\tau x}}dx = \left[x - \frac{1}{\tau}\log\left(t\underline{\mathrm{B}}^{-1}\,e^{\tau x}+1\right)\right]\Bigg|_0^\infty.$$

Now

$$x - \frac{1}{\tau}\log\left(t\underline{\mathrm{B}}^{-1}\,e^{\tau x}+1\right) = \frac{1}{\tau}\left(\log(e^{\tau x}) - \log\left(t\underline{\mathrm{B}}^{-1}\,e^{\tau x}+1\right)\right) = \frac{1}{\tau}\log\left(t^{-1}\underline{\mathrm{B}}\frac{e^{\tau x}}{e^{\tau x}+t^{-1}\underline{\mathrm{B}}}\right).$$

Evaluating

$$\frac{1}{\tau}\log\left(t^{-1}\underline{\mathrm{B}}\frac{e^{\tau x}}{e^{\tau x}+t^{-1}\underline{\mathrm{B}}}\right)\Bigg|_0^\infty$$

yields the result. □

Appendix B

Supplementary Results: Nyström

Lemma B.1 (Rudi et al. (2015), Lemma 6)**.** *Suppose Assumption 3.1 holds, and suppose for some $m < n$, the set $\{\tilde{X}_j\}_{j=1}^m$ is drawn uniformly from the set of all partitions of size m of the training data, $\{X_i\}_{i=1}^n$. For $t > 0$ and any $\delta > 0$ such that $m \geq (67 \vee 5\mathcal{N}_{C,\infty}(t)) \log \frac{4\kappa}{t\delta}$, we have*

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

where P_m is the orthogonal projector onto $\mathcal{H}_m = \text{span}\{k(\cdot, \tilde{X}_j) | j \in [m]\}$.

Lemma B.2 (Rudi et al. (2015), Lemma 7)**.** *Suppose Assumption 3.1 holds. Let $(\hat{l}_i(s))_{i=1}^n$ be the collection of approximate leverage scores. Letting $N := \{1, ..., n\}$, for $t > 0$ define p_t as the distribution over N with probabilities $p_t(i) = \hat{l}_i(t) / \sum_{j=1}^n \hat{l}_j(t)$. Let $\mathcal{I}_m = \{i_1, ..., i_m\} \subset N$ be a collection of indices independently sampled from p_t with replacement. Let P_m be the orthogonal projector onto $\mathcal{H}_m = \text{span}\{k(\cdot, \tilde{X}_j) | j \in \mathcal{I}_m\}$. Additionally, for any $\delta > 0$, suppose the following hold:*

1. *There exists $T \geq 1$ and $t_0 > 0$ such that for any $s \geq t_0$, $(\hat{l}_i(s))_{i=1}^n$ are $T-$approximate leverage scores with confidence δ,*
2. *$n \geq 1655\kappa + 223\kappa \log \frac{2\kappa}{\delta}$,*
3. *$t_0 \vee \frac{19\kappa}{n} \log \frac{2n}{\delta} \leq t \leq \lambda_1$,*

4. $m \geq 334 \log \frac{8n}{\delta} \vee 78T^2 \mathcal{N}_C(t) \log \frac{8n}{\delta}$.

Then

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

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

Lemma B.3. *Suppose Assumption 3.1 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* $\{1, 2, ..., n\}$*, yielding the subsample* $(X_{i_j})_{j=1}^m$*. Define* $\Sigma := \int_{\mathcal{X}} k(\cdot, x) \otimes_{\mathcal{H}} k(\cdot, x) \mathbb{P}(x) - m_{\mathbb{P}} \otimes_{\mathcal{H}} m_{\mathbb{P}}$*. Then, for* $0 \leq \delta \leq \frac{1}{2}$*,* $0 < t \leq \|\Sigma\|_{\mathcal{L}^\infty(H)}$ *and* $m \geq \frac{140\kappa}{t} \log \frac{8}{t\delta}$*, we have*

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

where $\bar{P}_m$ *is the orthogonal projector onto*

$$\bar{\mathcal{H}}_m := \left\{ f \in \mathcal{H} \middle| f = \sum_{j=1}^m \left(m\boldsymbol{\alpha}_j - \sum_{l=1}^m \boldsymbol{\alpha}_l \right) k(\cdot, X_{i_j}) : \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m) \in \mathbb{R} \right\}.$$

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 $\widehat{\Sigma}_m = \frac{1}{2m(m-1)} \tilde{Z}_m^* \mathbf{H}_m \tilde{Z}_m = \frac{1}{2(m-1)} \tilde{Z}_m^* \mathbf{C}_m^2 \tilde{Z}_m = Z^* Z$, where $Z^* = \frac{1}{\sqrt{2(m-1)}} \tilde{Z}_m^* \mathbf{C}_m$. Note that Z^* has range $\bar{\mathcal{H}}_m$, and so $\operatorname{ran}(\bar{P}_m) = \operatorname{ran}(Z^*)$. 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.2. □

Appendix C

Supplementary Results: Random Features

Lemma C.1. *Suppose Assumptions 3.1 and 4.1 hold. For any* $0 < \delta < 1$ *with* $n \geq 2\log\frac{2}{\delta}$, *then the following hold:*

(i) $\mathbb{P}^n\left\{(X_i)_{i=1}^n : \|m_{\mathbb{P}} - \widehat{m}_{\mathbb{P}}\|_{\mathcal{H}}^2 \leq \frac{32\kappa\log\frac{2}{\delta}}{n}\right\} \geq 1-\delta;$

(ii) $\mathbb{P}^n\left\{(X_i)_{i=1}^n : \|m_{\mathbb{P},m} - \widehat{m}_{\mathbb{P},m}\|_{\mathcal{H}_m}^2 \leq \frac{32\kappa\log\frac{2}{\delta}}{n}\middle|(\theta_i)_{i=1}^m\right\} \geq 1-\delta.$

Proof. Define $\xi_i = k(\cdot, X_i) - \int_{\mathcal{X}} k(\cdot, x)d\mathbb{P}(x)$. Clearly $\frac{1}{n}\sum_{i=1}^n \xi_i = \widehat{m}_{\mathbb{P}} - m_{\mathbb{P}}$. Note that $\|\xi_i\|_{\mathcal{H}} \leq 2\sqrt{\kappa}$ for all i. The result therefore follows by applying Theorem E.1 with $B = \theta = 2\sqrt{\kappa}$. Conditioned on $(\theta_i)_{i=1}^m$, the second result follows exactly the first one with k replaced by k_m. □

Lemma C.2. *Suppose Assumptions 3.1 and 4.1 hold. For any* $\delta > 0$ *with* $m \geq 2\log\frac{2}{\delta}$,

$$\Lambda^m\left\{\mathbb{E}\left\|\mathfrak{I}\bar{k}(\cdot, X) - \mathfrak{A}\bar{k}_m(\cdot, X)\right\|_{L^2(\mathbb{P})}^2 \leq \frac{64\kappa^2\log\frac{2}{\delta}}{m}\right\} \geq 1-2\delta.$$

Proof. Note that

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

$$
\begin{aligned}
&\overset{(\dagger)}{=} \|\Sigma\|^2_{\mathcal{L}^2(\mathcal{H})} + \|\Sigma_m\|^2_{\mathcal{L}^2(\mathcal{H}_m)} - 2\mathbb{E}\langle \mathfrak{I}\bar{k}(\cdot, X), \mathfrak{A}\bar{k}_m(\cdot, X)\rangle_{L^2(\mathbb{P})} \\
&= \|\mathfrak{I}\mathfrak{I}^*\|^2_{\mathcal{L}^2(L^2(\mathbb{P}))} + \|\mathfrak{A}\mathfrak{A}^*\|^2_{\mathcal{L}^2(L^2(\mathbb{P}))} - 2\mathbb{E}\langle \mathfrak{I}\bar{k}(\cdot, X), \mathfrak{A}\bar{k}_m(\cdot, X)\rangle_{L^2(\mathbb{P})} \\
&= \|\mathfrak{I}\mathfrak{I}^* - \mathfrak{A}\mathfrak{A}^*\|^2_{\mathcal{L}^2(L^2(\mathbb{P}))} + 2\left[\mathrm{tr}(\mathfrak{I}\mathfrak{I}^*\mathfrak{A}\mathfrak{A}^*) - \mathbb{E}\langle \mathfrak{I}\bar{k}(\cdot, X), \mathfrak{A}\bar{k}_m(\cdot, X)\rangle_{L^2(\mathbb{P})}\right], \quad \text{(C.1)}
\end{aligned}
$$

where we used Lemma A.4 in (†). We will now focus on computing

$$
\mathbb{E}\langle \mathfrak{I}\bar{k}(\cdot, X), \mathfrak{A}\bar{k}_m(\cdot, X)\rangle_{L^2(\mathbb{P})} \quad \text{and} \quad \mathrm{tr}(\mathfrak{I}\mathfrak{I}^*\mathfrak{A}\mathfrak{A}^*).
$$

Note that $k(\cdot, x) = \int_\Theta \varphi(\cdot, \theta)\varphi(x, \theta)\, d\Lambda(\theta)$ and $k_m(\cdot, x) = \sum_{i=1}^m \varphi_i(x)\varphi_i$. Define $\varphi_{\mathbb{P}}(\theta) := \int_{\mathcal{X}} \varphi(x, \theta)\, d\mathbb{P}(x)$ and $\varphi_{i,\mathbb{P}} := \int_{\mathcal{X}} \varphi_i(x)\, d\mathbb{P}(x)$. Therefore,

$$
\begin{aligned}
&\mathbb{E}\langle \mathfrak{I}\bar{k}(\cdot, X), \mathfrak{A}\bar{k}_m(\cdot, X)\rangle_{L^2(\mathbb{P})} \\
&= \int_{\mathcal{X}} \langle \mathfrak{I}\bar{k}(\cdot, x), \mathfrak{A}\bar{k}_m(\cdot, x)\rangle_{L^2(\mathbb{P})}\, d\mathbb{P}(x) \\
&\overset{(*)}{=} \int_{\mathcal{X}} \left\langle \int_\Theta \left(\varphi(\cdot, \theta) - \varphi_{\mathbb{P}}(\theta)\right) \varphi(x, \theta)\, d\Lambda(\theta), \sum_{i=1}^m \varphi_i(x)\left(\varphi_i - \varphi_{i,\mathbb{P}}\right) \right\rangle_{L^2(\mathbb{P})} d\mathbb{P}(x) \\
&= \int_{\mathcal{X}} \int_\Theta \sum_{i=1}^m \langle \varphi(\cdot, \theta) - \varphi_{\mathbb{P}}(\theta), \varphi_i - \varphi_{i,\mathbb{P}}\rangle_{L^2(\mathbb{P})}\, \varphi_i(x)\varphi(x, \theta)\, d\Lambda(\theta)\, d\mathbb{P}(x) \\
&= \int_\Theta \sum_{i=1}^m \langle \varphi(\cdot, \theta) - \varphi_{\mathbb{P}}(\theta), \varphi_i - \varphi_{i,\mathbb{P}}\rangle_{L^2(\mathbb{P})} \langle \varphi_i, \varphi(\cdot, \theta)\rangle_{L^2(\mathbb{P})}\, d\Lambda(\theta), \quad \text{(C.2)}
\end{aligned}
$$

where the penultimate and last equalities follow by employing Fubini's theorem and $(*)$ follows from Propositions D.2 and D.3. On the other hand,

$$
\begin{aligned}
&\mathrm{tr}(\mathfrak{I}\mathfrak{I}^*\mathfrak{A}\mathfrak{A}^*) \\
&\overset{(\ddagger)}{=} \mathrm{tr}\left[\int_\Theta \left[\varphi(\cdot, \theta) - \left(1 \otimes_{L^2(\mathbb{P})} 1\right)\varphi(\cdot, \theta)\right] \otimes_{L^2(\mathbb{P})} \left[\varphi(\cdot, \theta) - \left(1 \otimes_{L^2(\mathbb{P})} 1\right)\varphi(\cdot, \theta)\, d\Lambda(\theta)\right]\right. \\
&\qquad \left. \sum_{i=1}^m \left[\varphi_i - \left(1 \otimes_{L^2(\mathbb{P})} 1\right)\varphi_i\right] \otimes_{L^2(\mathbb{P})} \left[\varphi_i - \left(1 \otimes_{L^2(\mathbb{P})} 1\right)\varphi_i\right]\right] \\
&= \mathrm{tr}\left[\int_\Theta \sum_{i=1}^m \langle \varphi(\cdot, \theta) - \varphi_{\mathbb{P}}(\theta), \varphi_i - \varphi_{i,\mathbb{P}}\rangle_{L^2(\mathbb{P})} \left[\varphi(\cdot, \theta) - \varphi_{\mathbb{P}}(\theta)\right] \otimes_{L^2(\mathbb{P})} \left[\varphi_i - \varphi_{i,\mathbb{P}}\right]\, d\Lambda(\theta)\right] \\
&= \int_\Theta \sum_{i=1}^m \langle \varphi(\cdot, \theta) - \varphi_{\mathbb{P}}(\theta), \varphi_i - \varphi_{i,\mathbb{P}}\rangle_{L^2(\mathbb{P})} \langle \varphi(\cdot, \theta) - \varphi_{\mathbb{P}}(\theta), \varphi_i - \varphi_{i,\mathbb{P}}\rangle_{L^2(\mathbb{P})}\, d\Lambda(\theta)
\end{aligned}
$$

$$= \int_\Theta \sum_{i=1}^m \langle \varphi(\cdot,\theta) - \varphi_{\mathbb{P}}(\theta), \varphi_i - \varphi_{i,\mathbb{P}} \rangle_{L^2(\mathbb{P})} \left[\langle \varphi(\cdot,\theta), \varphi_i \rangle_{L^2(\mathbb{P})} - \varphi_{\mathbb{P}}(\theta)\varphi_{i,\mathbb{P}} \right] d\Lambda(\theta), \qquad \text{(C.3)}$$

where we used Propositions D.2*(iv)* and D.3*(iv)* in (‡). It follows from (C.2) and (C.3) that

$$\operatorname{tr}(\mathfrak{I}\mathfrak{I}^*\mathfrak{A}\mathfrak{A}^*) = \mathbb{E}\langle \mathfrak{I}\bar{k}(\cdot,X), \mathfrak{A}\bar{k}_m(\cdot,X)\rangle_{L^2(\mathbb{P})} - \left\langle \int_\Theta A(\theta)\, d\Lambda(\theta), \frac{1}{m}\sum_{i=1}^m A(\theta_i) \right\rangle_{L^2(\mathbb{P})}, \quad \text{(C.4)}$$

where $A(\theta) = \varphi(\cdot,\theta)\varphi_{\mathbb{P}}(\theta) - \varphi^2_{\mathbb{P}}(\theta)$. We remind the reader that $\varphi_i = \frac{1}{\sqrt{m}}\varphi(\cdot,\theta_i)$ with $(\theta_i)_{i=1}^m \overset{i.i.d.}{\sim} \Lambda$. Define Λ_m to be the empirical measure based on $(\theta_i)_{i=1}^m$. Then, (C.4) can be written as

$$\begin{aligned}
\operatorname{tr}(\mathfrak{I}\mathfrak{I}^*\mathfrak{A}\mathfrak{A}^*) &= \mathbb{E}\langle \mathfrak{I}\bar{k}(\cdot,X), \mathfrak{A}\bar{k}_m(\cdot,X)\rangle_{L^2(\mathbb{P})} + \frac{1}{2}\left\| \int_\Theta A(\theta)\, d(\Lambda_m - \Lambda)(\theta) \right\|^2_{L^2(\mathbb{P})} \\
&\quad - \frac{1}{2}\left\| \int_\Theta A(\theta)\, d\Lambda(\theta) \right\|^2_{L^2(\mathbb{P})} - \frac{1}{2}\left\| \int_\Theta A(\theta)\, d\Lambda_m(\theta) \right\|^2_{L^2(\mathbb{P})} \\
&\leq \mathbb{E}\langle \mathfrak{I}\bar{k}(\cdot,X), \mathfrak{A}\bar{k}_m(\cdot,X)\rangle_{L^2(\mathbb{P})} + \frac{1}{2}\left\| \int_\Theta A(\theta)\, d(\Lambda_m - \Lambda)(\theta) \right\|^2_{L^2(\mathbb{P})}, \quad \text{(C.5)}
\end{aligned}$$

which holds Λ-a.s. Using (C.5) in (C.1), we obtain

$$\mathbb{E}\left\| \mathfrak{I}\bar{k}(\cdot,X) - \mathfrak{A}\bar{k}_m(\cdot,X) \right\|^2_{L^2(\mathbb{P})} \leq \|\mathfrak{I}\mathfrak{I}^* - \mathfrak{A}\mathfrak{A}^*\|^2_{\mathcal{L}^2(L^2(\mathbb{P}))} + \left\| \int_\Theta A(\theta)\, d(\Lambda_m - \Lambda)(\theta) \right\|^2_{L^2(\mathbb{P})}, \quad \text{(C.6)}$$

which holds Λ-a.s. The result follows by applying Lemma C.3 and Theorem E.1 to (C.6) by noting that $\sup_{\theta\in\Theta} \|A(\theta)\|_{L^2(\mathbb{P})} \leq 2\kappa$ and $\mathbb{E}_{\theta\sim\Lambda} \|A(\theta)\|^2_{L^2(\mathbb{P})} \leq 4\kappa^2$. □

Lemma C.3. *Suppose Assumptions 3.1 and 4.1 hold. Then for any* $0 < \delta < 1$ *and* $m \geq 2\log\frac{2}{\delta}$,

$$\Lambda^m \left\{ (\theta_i)_{i=1}^m : \|\mathfrak{A}\mathfrak{A}^* - \mathfrak{I}\mathfrak{I}^*\|_{\mathcal{L}^2(L^2(\mathbb{P}))} \leq 4\kappa\sqrt{\frac{2\log\frac{2}{\delta}}{m}} \right\} \geq 1 - \delta.$$

Proof. From Proposition D.2*(iv)*, Lemma D.1 and Proposition D.3*(iv)*, we have

$$\mathfrak{I}\mathfrak{I}^* = \Upsilon - (1 \otimes_{L^2(\mathbb{P})} 1)\Upsilon - \Upsilon(1 \otimes_{L^2(\mathbb{P})} 1) + (1 \otimes_{L^2(\mathbb{P})} 1)\Upsilon(1 \otimes_{L^2(\mathbb{P})} 1)$$

and

$$\mathfrak{A}\mathfrak{A}^* = \Pi - (1 \otimes_{L^2(\mathbb{P})} 1)\Pi - \Pi(1 \otimes_{L^2(\mathbb{P})} 1) + (1 \otimes_{L^2(\mathbb{P})} 1)\Pi(1 \otimes_{L^2(\mathbb{P})} 1)$$

where

$$\Upsilon := \int_\Theta \varphi(\cdot,\theta) \otimes_{L^2(\mathbb{P})} \varphi(\cdot,\theta)\, d\Lambda(\theta) \text{ and } \Pi := \sum_{i=1}^m \varphi_i \otimes_{L^2(\mathbb{P})} \varphi_i = \frac{1}{m}\sum_{i=1}^m \varphi(\cdot,\theta_i) \otimes_{L^2(\mathbb{P})} \varphi(\cdot,\theta_i).$$

Define $A_i := \varphi(\cdot,\theta_i) - (1 \otimes_{L^2(\mathbb{P})} 1)\varphi(\cdot,\theta_i)$ and $D_i := A_i \otimes_{L^2(\mathbb{P})} A_i$. Then it follows that $\mathfrak{A}\mathfrak{A}^* = \frac{1}{m}\sum_{i=1}^m D_i$ and $\mathfrak{I}\mathfrak{I}^* = \mathbb{E}[\mathfrak{A}\mathfrak{A}^*]$. The result follows by applying Theorem E.1 with $B = \theta = \sup_{\theta_1} \left\| A_1 \otimes_{L^2(\mathbb{P})} A_1 \right\|_{\mathcal{L}^2(L^2(\mathbb{P}))} = \sup_{\theta_1} \|A_1\|^2_{L^2(\mathbb{P})} \leq 2\sup_{\theta_1} \|\varphi(\cdot,\theta_1)\|^2_{L^2(\mathbb{P})} \leq 2\kappa$ and noting that $\mathcal{L}^2(L^2(\mathbb{P}))$ is a separable Hilbert space since $L^2(\mathbb{P})$ is separable. □

Lemma C.4. *Suppose Assumptions 3.1, 3.5, 4.1 and 4.2 hold. For $t > 0$, define $\mathcal{N}_\Sigma(t) = \operatorname{tr}(\Sigma(\Sigma + tI)^{-1})$ and $\mathcal{N}_{\Sigma_m}(t) = \operatorname{tr}(\Sigma_m(\Sigma_m + tI)^{-1})$. For $\delta > 0$ and $\frac{86\kappa}{m}\log\frac{16\kappa m}{\delta} \leq t \leq \|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})}$, the following hold:*

(i) $\Lambda^m\left\{(\theta_i)_{i=1}^m : \sqrt{\frac{2}{3}} \leq \left\|(\mathfrak{I}\mathfrak{I}^* + tI)^{1/2}(\mathfrak{A}\mathfrak{A}^* + tI)^{-1/2}\right\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))} \leq \sqrt{2}\right\} \geq 1 - \delta;$

(ii) $\Lambda^m\left\{(\theta_i)_{i=1}^m : \lambda_{m,j} + t \leq \frac{3}{2}(\lambda_j + t)\right\} \geq 1 - \delta$ *for all* $j \geq 1$;

(iii) $\Lambda^m\left\{(\theta_i)_{i=1}^m : \frac{1}{2}(\lambda_j + t) \leq \lambda_{m,j} + t\right\} \geq 1 - \delta$ *for all* $j \geq 1$;

(iv) $\Lambda^m\left\{(\theta_i)_{i=1}^m : \mathcal{N}_{\Sigma_m}(t) \leq \frac{32\kappa\log\frac{2}{\delta}}{tm} + \sqrt{\frac{32\kappa\mathcal{N}_\Sigma(t)\log\frac{2}{\delta}}{tm}} + 2\mathcal{N}_\Sigma(t)\right\} \geq 1 - 2\delta.$

Proof. (i, ii, iii) Define $A_i := \varphi(\cdot,\theta_i) - (1 \otimes_{L^2(\mathbb{P})} 1)\varphi(\cdot,\theta_i)$ and $D_i := A_i \otimes_{L^2(\mathbb{P})} A_i$. Then it follows from Propositions D.2 and D.3 that $\mathfrak{A}\mathfrak{A}^* = \frac{1}{m}\sum_{i=1}^m D_i$ and $\mathfrak{I}\mathfrak{I}^* = \mathbb{E}[\mathfrak{A}\mathfrak{A}^*]$. Define $E_m := (\mathfrak{I}\mathfrak{I}^* + tI)^{-1/2}(\mathfrak{I}\mathfrak{I}^* - \mathfrak{A}\mathfrak{A}^*)(\mathfrak{I}\mathfrak{I}^* + tI)^{-1/2}$. By mimicking the strategy of Lemma A.2*(ii, iii)*, we obtain

$$\begin{aligned}(1 + \|E_m\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))})^{-1/2} &\leq \left\|(\mathfrak{I}\mathfrak{I}^* + tI)^{1/2}(\mathfrak{A}\mathfrak{A}^* + tI)^{-1/2}\right\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))} \\ &\leq (1 - \|E_m\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))})^{-1/2} \end{aligned} \tag{C.7}$$

provided $\|E_m\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))} < 1$. We will now apply Theorem E.2 to bound $\|E_m\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))}$. By defining $Z_i := (\mathfrak{I}\mathfrak{I}^* + tI)^{-1/2}A_i$ and $U_i := Z_i \otimes_{L^2(\mathbb{P})} Z_i$, we obtain $E_m = \frac{1}{m}\sum_{i=1}^m U_i - \mathbb{E}_\Lambda[U_i]$.

Note that

$$\begin{aligned}\|U_i\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))} = \|Z_i\|^2_{L^2(\mathbb{P})} &\leq \frac{1}{t}\|A_i\|^2_{L^2(\mathbb{P})} \\ &\leq \frac{2\|\varphi(\cdot,\theta_i)\|^2_{L^2(\mathbb{P})}}{t}\left(1+\left\|1\otimes_{L^2(\mathbb{P})}1\right\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))}\right) \leq \frac{4\kappa}{t}.\end{aligned}$$

Define $T := E_\Lambda[U_i]$. Then $\mathbb{E}_\Lambda[(U_i - T)^2] = \mathbb{E}_\Lambda[\|Z_i\|^2_{L^2(\mathbb{P})} U_i - T^2] \preceq \mathbb{E}_\Lambda[\|Z_i\|^2_{L^2(\mathbb{P})} U_i] \preceq \frac{4\kappa}{t}T$. Now we set

$$\sigma^2 = \left\|\frac{4\kappa}{t}T\right\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))} \leq \frac{4\kappa}{t} \quad \text{and} \quad d = \frac{\|T\|_{\mathcal{L}^1(L^2(\mathbb{P}))}}{\|T\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))}} \leq \frac{(\lambda_1 + t)\|T\|_{\mathcal{L}^1(L^2(\mathbb{P}))}}{\lambda_1},$$

where $\lambda_1 = \|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})} = \|\mathfrak{I}\mathfrak{I}^*\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))}$. Then Theorem E.2 yields

$$\Lambda^m\left\{\|B_m\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))} \leq \frac{8\beta\kappa}{3tm} + \sqrt{\frac{8\kappa\beta}{tm}}\right\} \leq 1-\delta, \tag{C.8}$$

where $\beta = \log\frac{4d}{\delta}$. Since $t \geq \frac{86\kappa}{m}\log\frac{16\kappa m}{\delta}$, it follows that $t \geq \frac{86\kappa}{m}\log\frac{4d}{\delta}$ as $\frac{86\kappa}{m}\log\frac{16\kappa m}{\delta} \geq \frac{86\kappa}{m}\log\frac{16\kappa}{t\delta} \geq \frac{86\kappa}{m}\log\frac{4d}{\delta}$, where we have used $d \leq \frac{4\kappa}{t}$ which follows from $t \leq \|\Sigma\|_{\mathcal{L}^\infty(\mathcal{H})}$ and $\mathrm{tr}(T) \leq \frac{\mathrm{tr}(\mathfrak{I}\mathfrak{I}^*)}{t} = \frac{\mathrm{tr}(\Sigma)}{t} \leq \frac{2\kappa}{t}$. This implies $t \geq \frac{86\beta\kappa}{m}$. Combining this with (C.8) yields that with probability at least $1-\delta$, $\|B_m\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))} \leq \frac{1}{2}$. (i) follows by using this in (C.7). (ii), (iii) are implied as in (iv), (v) of Lemma A.2.

(iv) Observe that $\mathcal{N}_{\Sigma_m}(t) = \mathrm{tr}(\Sigma_m(\Sigma_m + tI)^{-1}) = \mathrm{tr}(\mathfrak{A}^*\mathfrak{A}(\mathfrak{A}^*\mathfrak{A} + tI)^{-1}) = \mathrm{tr}(\mathfrak{A}(\mathfrak{A}^*\mathfrak{A} + tI)^{-1}\mathfrak{A}^*) = \mathrm{tr}((\mathfrak{A}\mathfrak{A}^* + tI)^{-1}\mathfrak{A}\mathfrak{A}^*)$, where we have used the fact that $\mathfrak{A}(\mathfrak{A}^*\mathfrak{A} + tI)^{-1} = (\mathfrak{A}\mathfrak{A}^* + tI)^{-1}\mathfrak{A}$ and the invariance of trace under cyclic permutations. Similarly, it can be shown that $\mathcal{N}_\Sigma(t) = \mathrm{tr}((\mathfrak{I}\mathfrak{I}^* + tI)^{-1}\mathfrak{I}\mathfrak{I}^*)$. For the ease of notation, define $A := \mathfrak{A}\mathfrak{A}^*$, $B := \mathfrak{I}\mathfrak{I}^*$, $A_t := A + tI$ and $B_t := B + tI$. Then

$$\begin{aligned}A + tI &= (A - B) + (B + tI) \\ &= (B+tI)^{1/2}\left(I + (B+tI)^{-1/2}(A-B)(B+tI)^{-1/2}\right)(B+tI)^{1/2},\end{aligned}$$

implying,

$$A_t^{-1} = B_t^{-1/2}\left(I + B_t^{-1/2}(A-B)B_t^{-1/2}\right)^{-1} B_t^{-1/2}.$$

Therefore,

$$\begin{aligned}
\mathcal{N}_m(t) = \operatorname{tr}(AA_t^{-1}) &= \operatorname{tr}\left[AB_t^{-1/2}\left(I + B_t^{-1/2}(A-B)B_t^{-1/2}\right)^{-1} B_t^{-1/2}\right] \\
&= \operatorname{tr}\left[B_t^{-1/2}AB_t^{-1/2}\left(I + B_t^{-1/2}(A-B)B_t^{-1/2}\right)^{-1}\right] \\
&\leq \left\|(I + B_t^{-1/2}(A-B)B_t^{-1/2})^{-1}\right\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))} \operatorname{tr}(B_t^{-1/2}AB_t^{-1/2}) \\
&= \left\|(I - E_m)^{-1}\right\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))} \operatorname{tr}(B_t^{-1/2}AB_t^{-1/2}),
\end{aligned}$$

where

$E_m := B_t^{-1/2}(B-A)B_t^{-1/2} = (\mathfrak{I}\mathfrak{I}^* + tI)^{-1/2}(\mathfrak{I}\mathfrak{I}^* - \mathfrak{A}\mathfrak{A}^*)(\mathfrak{I}\mathfrak{I}^* + tI)^{-1/2}$. Since we showed in the proof of (i) that with probability at least $1-\delta$, $\|E_m\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))} \leq \frac{1}{2}$, we obtain

$$\mathcal{N}_m(t) \leq 2\operatorname{tr}(B_t^{-1/2}AB_t^{-1/2}), \tag{C.9}$$

where we use $\|(I-E_m)^{-1}\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))} \leq \frac{1}{1-\|E_m\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))}}$.

Next, consider

$$\operatorname{tr}(B_t^{-1/2}AB_t^{-1/2}) = \operatorname{tr}(B_t^{-1}(A-B+B)) = \left\langle B_t^{-1}, A-B\right\rangle_{\mathcal{L}^2(L^2(\mathbb{P}))} + \mathcal{N}_\Sigma(t), \tag{C.10}$$

where $\left\langle B_t^{-1}, A-B\right\rangle_{\mathcal{L}^2(L^2(\mathbb{P}))} = \left\langle (\mathfrak{I}\mathfrak{I}^* + tI)^{-1}, \mathfrak{A}\mathfrak{A}^* - \mathfrak{I}\mathfrak{I}^*\right\rangle_{\mathcal{L}^2(L^2(\mathbb{P}))}$. We now bound this term as follows. Let

$$\zeta_i = \left(\varphi(\cdot,\theta_i) - (1 \otimes_{L^2(\mathbb{P})} 1)\varphi(\cdot,\theta_i)\right) \otimes_{L^2(\mathbb{P})} \left(\varphi(\cdot,\theta_i) - (1 \otimes_{L^2(\mathbb{P})} 1)\varphi(\cdot,\theta_i)\right)$$

so that $\mathbb{E}_\Lambda[\zeta_1] = \mathfrak{I}\mathfrak{I}^*$, $\frac{1}{m}\sum_{i=1}^m \zeta_i = \mathfrak{A}\mathfrak{A}^*$ and

$$\left\langle (\mathfrak{I}\mathfrak{I}^* + tI)^{-1}, \mathfrak{A}\mathfrak{A}^* - \mathfrak{I}\mathfrak{I}^*\right\rangle_{\mathcal{L}^2(L^2(\mathbb{P}))} = \frac{1}{m}\sum_{i=1}^m \left\langle (\mathfrak{I}\mathfrak{I}^* + tI)^{-1}, (\zeta_i - \mathfrak{I}\mathfrak{I}^*)\right\rangle_{\mathcal{L}^2(L^2(\mathbb{P}))}.$$

We will now apply Bernstein's inequality (Theorem E.1). To this end, note that

$$\begin{aligned}
&\left|\left\langle (\mathfrak{I}\mathfrak{I}^* + tI)^{-1}, \zeta_1 - \mathfrak{I}\mathfrak{I}^* \right\rangle_{\mathcal{L}^2(L^2(\mathbb{P}))}\right| \\
&\leq \left\langle (\mathfrak{I}\mathfrak{I}^* + tI)^{-1}, \mathfrak{I}\mathfrak{I}^* \right\rangle_{\mathcal{L}^2(L^2(\mathbb{P}))} + \left\langle (\mathfrak{I}\mathfrak{I}^* + tI)^{-1}, \zeta_i \right\rangle_{\mathcal{L}^2(L^2(\mathbb{P}))} \\
&= \mathcal{N}_\Sigma(t) + \frac{1}{t}\mathrm{tr}\left(\left(\varphi(\cdot,\theta_i) - (1 \otimes_{L^2(\mathbb{P})} 1)\varphi(\cdot,\theta_i)\right) \otimes_{L^2(\mathbb{P})} \left(\varphi(\cdot,\theta_i) - (1 \otimes_{L^2(\mathbb{P})} 1)\varphi(\cdot,\theta_i)\right)\right) \\
&= \mathcal{N}_\Sigma(t) + \frac{1}{t}\left\|\varphi(\cdot,\theta_i) - (1 \otimes_{L^2(\mathbb{P})} 1)\varphi(\cdot,\theta_i)\right\|^2_{L^2(\mathbb{P})} \leq \frac{\|\Sigma\|_{\mathcal{L}^1(\mathcal{H})} + 4\kappa}{t} \leq \frac{8\kappa}{t},
\end{aligned}$$

where we use $\|\Sigma\|_{\mathcal{L}^1(\mathcal{H})} \leq \mathbb{E}\left\|\bar{k}(\cdot,X) \otimes_{\mathcal{H}} \bar{k}(\cdot,X)\right\|_{\mathcal{L}^1(\mathcal{H})} = \mathbb{E}\left\|\bar{k}(\cdot,X)\right\|^2_{\mathcal{H}} \leq 4\kappa$. Also

$$\begin{aligned}
\mathbb{E}_\Lambda \left\langle (\mathfrak{I}\mathfrak{I}^* + tI)^{-1}, \zeta_1 - \mathfrak{I}\mathfrak{I}^* \right\rangle^2_{\mathcal{L}^2(L^2(\mathbb{P}))} &= \mathbb{E}_\Lambda \left[\left(\left\langle (\mathfrak{I}\mathfrak{I}^* + tI)^{-1}, \zeta_1 \right\rangle_{\mathcal{L}^2(L^2(\mathbb{P}))} - \mathcal{N}_\Sigma(t)\right)^2\right] \\
&= \mathbb{E}_\Lambda \left[\left\langle (\mathfrak{I}\mathfrak{I}^* + tI)^{-1}, \zeta_1 \right\rangle^2_{\mathcal{L}^2(L^2(\mathbb{P}))} - \mathcal{N}^2_\Sigma(t)\right] \leq \mathbb{E}_\Lambda \left\langle (\mathfrak{I}\mathfrak{I}^* + tI)^{-1}, \zeta_1 \right\rangle^2_{\mathcal{L}^2(L^2(\mathbb{P}))} \\
&= \mathbb{E}_\Lambda \mathrm{tr}\left[(\mathfrak{I}\mathfrak{I}^* + tI)^{-1}\zeta_1(\mathfrak{I}\mathfrak{I}^* + tI)^{-1}\zeta_1\right] \\
&\leq \sup_{\theta_1} \left\|(\mathfrak{I}\mathfrak{I}^* + tI)^{-1/2}\zeta_1(\mathfrak{I}\mathfrak{I}^* + tI)^{-1/2}\right\|_{\mathcal{L}^\infty(L^2(\mathbb{P}))} \\
&\quad \times \mathbb{E}_\Lambda \left[\mathrm{tr}\left((\mathfrak{I}\mathfrak{I}^* + tI)^{-1/2}\zeta_1(\mathfrak{I}\mathfrak{I}^* + tI)^{-1/2}\right)\right] \\
&\leq \frac{\mathcal{N}_\Sigma(t)}{t} \sup_{\theta_1} \left\|\varphi(\cdot,\theta_1) - (1 \otimes_{L^2(\mathbb{P})} 1)\varphi(\cdot,\theta_1)\right\|^2_{L^2(\mathbb{P})} \leq \frac{4\kappa\mathcal{N}_\Sigma(t)}{t}.
\end{aligned}$$

The result follows by applying Theorem E.1 to $\left\langle B_t^{-1}, (A-B)\right\rangle_{\mathcal{L}^2(L^2(\mathbb{P}))}$ and combining (C.9) and (C.10). □

Appendix D

Sampling, Inclusion and Approximation Operators

In this appendix, we present some technical results related to the properties of sampling, inclusion and approximation operators.

D.1 Properties of the Sampling Operator

The following result presents the properties of the sampling operator, S and its adjoint. While these results are known in the literature (e.g., see Smale and Zhou, 2007), we present it here for completeness.

Proposition D.1. *Let $\mathcal{H}$ be an RKHS of real-valued functions on a non-empty set $\mathcal{X}$ with k as the reproducing kernel. Define $S : \mathcal{H} \to \mathbb{R}^n$, $f \mapsto \frac{1}{\sqrt{n}}(f(X_1), \ldots, f(X_n))^\top$ where $(X_i)_i \subset \mathcal{X}$. Then the following hold:*

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

(ii) $\widehat{\Sigma} = \frac{n}{n-1} S^* \mathbf{C}_n S$ *where* $\widehat{\Sigma}$ *is defined in* (3.21);

(iii) $\mathbf{K} = nSS^*$, *where* $[\mathbf{K}]_{ij} = k(X_i, X_j)$.

(iv) Define the approximate sampling operator $\tilde{Z}_m : \mathcal{H} \to \mathbb{R}^m$, $f \mapsto (f(X_{i_1}), \ldots, f(X_{i_m}))$ *where indices* $(i_j)_{j=1}^m$ *are a subsample of* $\{1, \ldots, n\}$. $\mathbf{K}_{nm} = \sqrt{n} S \tilde{Z}_m^*$ *and* $\mathbf{K}_{mm} = \tilde{Z}_m \tilde{Z}_m^*$.

Proof. (i) For any $g \in \mathcal{H}$ and $\boldsymbol{\alpha} \in \mathbb{R}^n$, we have

$$\langle S^* \alpha, g \rangle_{\mathcal{H}} = \langle \alpha, Sg \rangle_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha_i g(X_i) = \left\langle \frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha_i k(\cdot, X_i), g \right\rangle_{\mathcal{H}},$$

where the last equality follows from the reproducing property.

(ii) For any $f \in \mathcal{H}$,

$$\begin{aligned}
\langle f, \widehat{\Sigma} f \rangle_{\mathcal{H}} &= \frac{1}{2n(n-1)} \sum_{i \neq j}^n (f(X_i) - f(X_j))^2 \\
&= \frac{1}{n} \sum_{i=1}^n f^2(X_i) - \frac{1}{n(n-1)} \sum_{i \neq j} f(X_i) f(X_j) \\
&= \frac{1}{n-1} \sum_{i=1}^n f^2(X_i) - \frac{1}{n-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n f(X_i) \right)^2 \\
&= \frac{n}{n-1} \langle Sf, Sf \rangle_2 - \frac{1}{n-1} \langle \mathbf{1}_n, Sf \rangle_2^2 \\
&= \frac{n}{n-1} \langle f, S^* S f \rangle_{\mathcal{H}} - \frac{1}{n-1} \langle S^* \mathbf{1}_n, f \rangle_{\mathcal{H}}^2 \\
&= \frac{n}{n-1} \langle f, S^* S f \rangle_{\mathcal{H}} - \frac{1}{n-1} \langle f, S^* (\mathbf{1}_n \otimes_2 \mathbf{1}_n) S f \rangle_{\mathcal{H}} \\
&= \frac{n}{n-1} \langle f, S^* \mathbf{C}_n S f \rangle_{\mathcal{H}}.
\end{aligned}$$

(iii) For any $\alpha \in \mathbb{R}^n$,

$$S S^* \alpha = S \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha_i k(\cdot, X_i) \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha_i S k(\cdot, X_i) = \frac{1}{n} \mathbf{K} \alpha,$$

where in the second equality, we used the fact S is a linear operator. *(iv)* This is contained in Rudi et al. (2015, Section B). □

D.2 Properties of the Inclusion Operator

The following result captures the properties of the inclusion operator $\mathfrak{I}$. A variation of the result is known in the literature (e.g., see Steinwart and Christmann, 2008, Theorem 4.26).

Proposition D.2. *Suppose* (A_1) *holds. Define* $\mathfrak{I} : \mathcal{H} \to L^2(\mathbb{P})$, $f \mapsto f - f_{\mathbb{P}}$, *where* $f_{\mathbb{P}} := \int f(x)\, d\mathbb{P}(x)$. *Then the following hold:*

(i) $\mathfrak{I}^* : L^2(\mathbb{P}) \to \mathcal{H}$, $f \mapsto \int_{\mathcal{X}} k(\cdot, x) f(x)\, d\mathbb{P}(x) - m_{\mathbb{P}} f_{\mathbb{P}}$ *where*

$$m_{\mathbb{P}} := \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x).$$

(ii) $\mathfrak{I}$ *and* $\mathfrak{I}^*$ *are Hilbert-Schmidt.*

(iii) $\Sigma = \mathfrak{I}^*\mathfrak{I}$ *is trace-class, where* Σ *is defined in* (3.18).

(iv) $\mathfrak{I}\mathfrak{I}^* = \Upsilon - (1 \otimes_{L^2(\mathbb{P})} 1)\Upsilon - \Upsilon(1 \otimes_{L^2(\mathbb{P})} 1) + (1 \otimes_{L^2(\mathbb{P})} 1)\Upsilon(1 \otimes_{L^2(\mathbb{P})} 1)$ *is trace-class where* $\Upsilon : L^2(\mathbb{P}) \to L^2(\mathbb{P})$, $f \mapsto \int_{\mathcal{X}} k(\cdot, x) f(x)\, d\mathbb{P}(x)$.

Proof. (i) For any $f \in L^2(\mathbb{P})$ and $g \in \mathcal{H}$,

$$\begin{aligned}
\langle \mathfrak{I}^* f, g \rangle_{\mathcal{H}} = \langle f, \mathfrak{I} g \rangle_{L^2(\mathbb{P})} &= \int_{\mathcal{X}} f(x)(\mathfrak{I} g)(x)\, d\mathbb{P}(x) = \int_{\mathcal{X}} f(x)[g(x) - g_{\mathbb{P}}]\, d\mathbb{P}(x) \\
&= \int_{\mathcal{X}} f(x) \langle k(\cdot, x), g \rangle_{\mathcal{H}}\, d\mathbb{P}(x) - \langle m_{\mathbb{P}}, g \rangle_{\mathcal{H}} f_{\mathbb{P}} \\
&= \left\langle \int_{\mathcal{X}} k(\cdot, x) f(x)\, d\mathbb{P}(x), g \right\rangle_{\mathcal{H}} - \langle m_{\mathbb{P}} f_{\mathbb{P}}, g \rangle_{\mathcal{H}}.
\end{aligned}$$

Clearly $f_{\mathbb{P}}$ is well defined as for any $f \in L^2(\mathbb{P})$, $f_{\mathbb{P}} \leq \int |f(x)|\, d\mathbb{P}(x) \leq \|f\|_{L^2(\mathbb{P})} < \infty$ and for $f \in \mathcal{H}$, $f_{\mathbb{P}} = \langle f, m_{\mathbb{P}} \rangle_{\mathcal{H}} \leq \|f\|_{\mathcal{H}} \int \sqrt{k(x, x)}\, d\mathbb{P}(x) < \infty$ and the result therefore follows.
(ii) For any orthonormal basis $(e_j)_j$ in $\mathcal{H}$,

$$\begin{aligned}
\|\mathfrak{I}\|^2_{\mathcal{L}^2(\mathcal{H}, L^2(\mathbb{P}))} = \sum_j \|\mathfrak{I} e_j\|^2_{L^2(\mathbb{P})} &= \sum_j \|e_j - e_{j,\mathbb{P}}\|^2_{L^2(\mathbb{P})} = \sum_j \|e_j\|^2_{L^2(\mathbb{P})} - e^2_{j,\mathbb{P}} \\
&\leq \sum_j \|e_j\|^2_{L^2(\mathbb{P})} = \sum_j \int_{\mathcal{X}} \langle e_j, k(\cdot, x) \rangle^2_{\mathcal{H}}\, d\mathbb{P}(x)
\end{aligned}$$

$$\overset{(\star)}{=} \int_{\mathcal{X}} \sum_j \langle e_j, k(\cdot, x)\rangle_{\mathcal{H}}^2 \, d\mathbb{P}(x) = \int_{\mathcal{X}} k(x,x) \, d\mathbb{P}(x) < \infty,$$

where $(\star)$ follows from monotone convergence theorem. Since

$$\|\mathfrak{I}\|_{\mathcal{L}^2(\mathcal{H}, L^2(\mathbb{P}))} = \|\mathfrak{I}^*\|_{\mathcal{L}^2(L^2(\mathbb{P}), \mathcal{H})},$$

the result follows.

(iii) For any $f \in \mathcal{H}$, $(\mathfrak{I}^*\mathfrak{I})f = \mathfrak{I}^*(f - f_{\mathbb{P}}) = \mathfrak{I}^* f - \mathfrak{I}^* f_{\mathbb{P}} = \mathfrak{I}^* f$, where we use the fact that $\mathfrak{I}^* f_{\mathbb{P}} = 0$ since $f_{\mathbb{P}}$ is a constant function. By using the reproducing property,

$$\begin{aligned}
\mathfrak{I}^*\mathfrak{I} f = \mathfrak{I}^* f &= \int_{\mathcal{X}} f(x) k(\cdot, x) \, d\mathbb{P}(x) - m_{\mathbb{P}} f_{\mathbb{P}} \\
&= \int_{\mathcal{X}} k(\cdot, x) \langle k(\cdot, x), f\rangle_{\mathcal{H}} \, d\mathbb{P} - m_{\mathbb{P}} \langle m_{\mathbb{P}}, f\rangle_{\mathcal{H}} \\
&= \int_{\mathcal{X}} (k(\cdot, x) \otimes_{\mathcal{H}} k(\cdot, x)) f \, d\mathbb{P}(x) - (m_{\mathbb{P}} \otimes_{\mathcal{H}} m_{\mathbb{P}}) f = \Sigma f
\end{aligned}$$

and the result follows. Since $\|\mathfrak{I}\|^2_{\mathcal{L}^2(\mathcal{H}, L^2(\mathbb{P}))} = \|\mathfrak{I}^*\mathfrak{I}\|_{\mathcal{L}^1(\mathcal{H})}$, Σ is trace-class.

(iv) For any $f \in L^2(\mathbb{P})$,

$$\begin{aligned}
(\mathfrak{I}\mathfrak{I}^*) f = \mathfrak{I}(\mathfrak{I}^* f) &= \mathfrak{I}\left(\int_{\mathcal{X}} k(\cdot, x) f(x) \, d\mathbb{P}(x) - m_{\mathbb{P}} f_{\mathbb{P}}\right) \\
&= \int_{\mathcal{X}} k(\cdot, x) f(x) \, d\mathbb{P}(x) - m_{\mathbb{P}} f_{\mathbb{P}} - \int_{\mathcal{X}} \int_{\mathcal{X}} k(y, x) f(x) \, d\mathbb{P}(x) \, d\mathbb{P}(y) \\
&\qquad + f_{\mathbb{P}} \int_{\mathcal{X}} \int_{\mathcal{X}} k(y, x) \, d\mathbb{P}(x) \, d\mathbb{P}(y) \\
&= \Upsilon f - \Upsilon 1 \langle 1, f\rangle_{L^2(\mathbb{P})} - 1 \langle \Upsilon 1, f\rangle_{L^2(\mathbb{P})} + 1 \langle 1, \Upsilon 1\rangle_{L^2(\mathbb{P})} \langle 1, f\rangle_{L^2(\mathbb{P})} \\
&= \Upsilon f - \Upsilon(1 \otimes_{L^2(\mathbb{P})} 1) f - (1 \otimes_{L^2(\mathbb{P})} 1) \Upsilon f \\
&\qquad + (1 \otimes_{L^2(\mathbb{P})} 1) \Upsilon (1 \otimes_{L^2(\mathbb{P})} 1) f
\end{aligned}$$

and the result follows, where in the last line we use the fact that Υ is self-adjoint, which follows from (Steinwart and Christmann, 2008, Theorem 4.27). Since $\|\mathfrak{I}^*\|^2_{\mathcal{L}^2(L^2(\mathbb{P}), \mathcal{H})} = \|\mathfrak{I}\mathfrak{I}^*\|_{\mathcal{L}^1(L^2(\mathbb{P}))}$, it follows that $\mathfrak{I}\mathfrak{I}^*$ is trace-class. □

The following result presents a representation for Υ if k satisfies (A_4).

Lemma D.1. *Suppose* (A_4) *holds. Then* $\Upsilon = \int_\Theta \varphi(\cdot,\theta) \otimes_{L^2(\mathbb{P})} \varphi(\cdot,\theta)\, d\Lambda(\theta)$.

Proof. Since $k(x,y) = \int_\Theta \varphi(x,\theta)\varphi(y,\theta)\, d\Lambda(\theta)$, for any $f \in L^2(\mathbb{P})$,

$$\begin{aligned}
\Upsilon f &= \int_{\mathcal{X}} k(\cdot,x) f(x)\, d\mathbb{P}(x) = \int_{\mathcal{X}} \int_\Theta \varphi(\cdot,\theta)\varphi(x,\theta)\, d\Lambda(\theta) f(x)\, d\mathbb{P}(x) \\
&\overset{(*)}{=} \int_\Theta \varphi(\cdot,\theta) \left(\int_{\mathcal{X}} \varphi(x,\theta) f(x)\, d\mathbb{P}(x) \right) d\Lambda(\theta) \\
&= \int_\Theta \varphi(\cdot,\theta) \langle \varphi(\cdot,\theta), f \rangle_{L^2(\mathbb{P})}\, d\Lambda(\theta) \\
&= \int_\Theta \left(\varphi(\cdot,\theta) \otimes_{L^2(\mathbb{P})} \varphi(\cdot,\theta) \right) f\, d\Lambda(\theta) = \left(\int_\Theta \varphi(\cdot,\theta) \otimes_{L^2(\mathbb{P})} \varphi(\cdot,\theta)\, d\Lambda(\theta) \right) f,
\end{aligned}$$

where Fubini's theorem is applied in $(*)$. □

D.3 Properties of the Approximation Operator

The following result presents the properties of the approximation operator, $\mathfrak{A}$.

Proposition D.3. *Define*

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

where $\varphi_{i,\mathbb{P}} := \int_{\mathcal{X}} \varphi_i(x)\, d\mathbb{P}(x)$ *and* $\sup_{x\in\mathcal{X}} |\varphi_i(x)| \le \sqrt{\frac{\kappa}{m}}$ *for all* $i \in [m]$ *with* $\kappa < \infty$. *Then the following hold:*

(i) $\mathfrak{A}^* : L^2(\mathbb{P}) \to \mathcal{H}_m,\ f \mapsto \sum_{i=1}^m \left(\langle f, \varphi_i \rangle_{L^2(\mathbb{P})} - f_{\mathbb{P}} \varphi_{i,\mathbb{P}} \right) \varphi_i$.

(ii) $\mathfrak{A}$ *and* $\mathfrak{A}^*$ *are Hilbert-Schmidt.*

(iii) $\Sigma_m = \mathfrak{A}^*\mathfrak{A}$ *is trace-class.*

(iv) $\mathfrak{A}\mathfrak{A}^* = \Pi - (1 \otimes_{L^2(\mathbb{P})} 1)\Pi - \Pi(1 \otimes_{L^2(\mathbb{P})} 1) + (1 \otimes_{L^2(\mathbb{P})} 1)\Pi(1 \otimes_{L^2(\mathbb{P})} 1)$ *is trace-class where* $\Pi := \sum_{i=1}^m \varphi_i \otimes_{L^2(\mathbb{P})} \varphi_i : L^2(\mathbb{P}) \to L^2(\mathbb{P})$.

Proof. The proof is similar to that of Proposition D.2.

(i) For any $g = \sum_{i=1}^m \beta_i \varphi_i \in \mathcal{H}_m$ and $f \in L^2(\mathbb{P})$,

$$\begin{aligned}\langle \mathfrak{A}^* f, g\rangle_{\mathcal{H}_m} = \langle f, \mathfrak{A} g\rangle_{L^2(\mathbb{P})} &= \int_{\mathcal{X}} \left(\sum_{i=1}^m \beta_i(\varphi_i(x) - \varphi_{i,\mathbb{P}})\right) f(x) d\mathbb{P}(x) \\ &= \sum_{i=1}^m \beta_i(\langle f, \varphi_i\rangle_{L^2(\mathbb{P})} - f_{\mathbb{P}}\varphi_{i,\mathbb{P}}),\end{aligned}$$

and the result follows from the definition of $\langle \cdot, \cdot \rangle_{\mathcal{H}_m}$.

(ii) For any orthonormal basis $(e_j)_j$ in $L^2(\mathbb{P})$,

$$\begin{aligned}\|\mathfrak{A}^*\|^2_{\mathcal{L}^2(L^2(\mathbb{P}),\mathcal{H}_m)} &= \sum_j \|\mathfrak{A}^* e_j\|^2_{\mathcal{H}_m} = \sum_j \sum_{i=1}^m \left(\langle e_j, \varphi_i\rangle_{L^2(\mathbb{P})} - e_{j,\mathbb{P}}\varphi_{i,\mathbb{P}}\right)^2 \\ &= \sum_j \sum_{i=1}^m \langle e_j, \varphi_i\rangle^2_{L^2(\mathbb{P})} + e^2_{j,\mathbb{P}}\varphi^2_{i,\mathbb{P}} - 2e_{j,\mathbb{P}}\varphi_{i,\mathbb{P}}\langle e_j, \varphi_i\rangle_{L^2(\mathbb{P})} \\ &= \sum_{i=1}^m \|\varphi_i\|^2_{L^2(\mathbb{P})} + \sum_{i=1}^m \varphi^2_{i,\mathbb{P}} \sum_j \langle e_j, 1\rangle^2_{L^2(\mathbb{P})} \\ &\qquad - \sum_{i=1}^m \varphi_{i,\mathbb{P}} \sum_j \langle e_j, (\varphi_i \otimes_{L^2(\mathbb{P})} 1 + 1 \otimes_{L^2(\mathbb{P})} \varphi_i) e_j\rangle_{L^2(\mathbb{P})} \\ &= \sum_{i=1}^m \|\varphi_i\|^2_{L^2(\mathbb{P})} + \sum_{i=1}^m \varphi^2_{i,\mathbb{P}} - 2\sum_{i=1}^m \varphi_{i,\mathbb{P}}\langle \varphi_i, 1\rangle_{L^2(\mathbb{P})} \\ &\leq \sum_{i=1}^m \|\varphi_i\|^2_{L^2(\mathbb{P})} \leq \kappa < \infty,\end{aligned}$$

and so $\mathfrak{A}$ and $\mathfrak{A}^*$ are Hilbert-Schmidt.

(iii) For any $f = \sum_{i=1}^m \beta_i \varphi_i \in \mathcal{H}_m$,

$$\begin{aligned}\mathfrak{A}^*\mathfrak{A} f &= \mathfrak{A}^* \left(\sum_{i=1}^m \beta_i(\varphi_i - \varphi_{i,\mathbb{P}})\right) = \sum_{i=1}^m \beta_i \mathfrak{A}^*(\varphi_i - \varphi_{i,\mathbb{P}}) \\ &= \sum_{i=1}^m \beta_i \sum_{j=1}^m (\langle \varphi_i, \varphi_j\rangle_{L^2(\mathbb{P})} - \varphi_{i,\mathbb{P}}\varphi_{j,\mathbb{P}})\varphi_j \\ &= \sum_{j=1}^m \left\langle \sum_{i=1}^m \beta_i \varphi_i, \varphi_j \right\rangle_{L^2(\mathbb{P})} \varphi_j \\ &\qquad - \left(\int_{\mathcal{X}} \sum_{i=1}^m \beta_i \varphi_i(x) d\mathbb{P}(x)\right) \left(\int_{\mathcal{X}} \sum_{j=1}^m \varphi_j(x)\varphi_j d\mathbb{P}(x)\right) \\ &= \int_{\mathcal{X}} \left(\sum_{i=1}^m \beta_i \varphi_i(x)\right) \left(\sum_{j=1}^m \varphi_j(x)\varphi_j\right) d\mathbb{P}(x)\end{aligned}$$

$$
\begin{aligned}
&\quad - \left(\int_{\mathcal{X}} f(x)\, d\mathbb{P}(x)\right)\left(\int_{\mathcal{X}} k_m(\cdot, x)\, d\mathbb{P}(x)\right) \\
&= \int_{\mathcal{X}} f(x) k_m(\cdot, x)\, d\mathbb{P}(x) - \left(\int_{\mathcal{X}} f(x)\, d\mathbb{P}(x)\right)\left(\int_{\mathcal{X}} k_m(\cdot, x)\, d\mathbb{P}(x)\right) \\
&= \Sigma_m f.
\end{aligned}
$$

That Σ_m is trace class is implied by *(ii)*.

(iv) For any $f \in L^2(\mathbb{P})$,

$$
\begin{aligned}
\mathfrak{A}\mathfrak{A}^* f &= \sum_{i=1}^{m} (\langle f, \varphi_i \rangle_{L^2(\mathbb{P})} - f_{\mathbb{P}} \varphi_{i,\mathbb{P}})(\varphi_i - \varphi_{i,\mathbb{P}}) \\
&= \sum_{i=1}^{m} (\langle f, \varphi_i \rangle_{L^2(\mathbb{P})} - \langle f, 1 \rangle_{L^2(\mathbb{P})} \langle \varphi_i, 1 \rangle_{L^2(\mathbb{P})})(\varphi_i - \langle \varphi_i, 1 \rangle_{L^2(\mathbb{P})}) \\
&= \Pi f - \langle \Pi 1, f \rangle_{L^2(\mathbb{P})} - \Pi(1 \otimes_{L^2(\mathbb{P})} 1) f + \langle (1 \otimes_{L^2(\mathbb{P})} 1) \Pi 1, f \rangle_{L^2(\mathbb{P})} \\
&= \Pi f - (1 \otimes_{L^2(\mathbb{P})} 1) \Pi f - \Pi(1 \otimes_{L^2(\mathbb{P})} 1) f + (1 \otimes_{L^2(\mathbb{P})} 1) \Pi (1 \otimes_{L^2(\mathbb{P})} 1) f
\end{aligned}
$$

and the result follows. $\mathfrak{A}\mathfrak{A}^*$ is trace-class since $\mathfrak{A}^*$ is Hilbert-Schmidt. □

Appendix E

Probabilistic Inequalities

In this appendix, we collect Bernstein's inequality for Hilbert-valued random elements (quoted from Yurinsky, 1995) and Tropp's inequality for operator-valued random elements (quoted from Rudi et al., 2013, Theorem A.1), that are used to prove the results of this paper. Based on these two results, Theorem E.3 presents a Bernstein-type inequality for the operator and Hilbert-Schmidt norms of a operator-valued U-statistics.

Theorem E.1 (Bernstein's inequality in separable Hilbert spaces)**.** *Let $(\Omega, \mathcal{A}, P)$ be a probability space, H be a separable Hilbert space, $B > 0$ and $\theta > 0$. Furthermore, let $\xi_1, \ldots, \xi_n : \Omega \to H$ be zero mean i.i.d. random variables satisfying*

$$\mathbb{E}\|\xi_1\|_H^r \leq \frac{r!}{2}\theta^2 B^{r-2}, \ \forall\ r > 2.$$

Then for any $0 < \delta < 1$,

$$P^n\left\{(\xi_i)_{i=1}^n : \left\|\frac{1}{n}\sum_{i=1}^n \xi_i\right\|_H \geq \frac{2B\log\frac{2}{\delta}}{n} + \sqrt{\frac{2\theta^2\log\frac{2}{\delta}}{n}}\right\} \leq \delta.$$

Theorem E.2 (Tropp's inequality for operators)**.** *Let $(Z_i)_{i=1}^n$ be independent copies of the random variable Z with law P taking values in the space of bounded self-adjoint operators for a separable Hilbert space H. Suppose there exists $S \in \mathcal{L}^2(H)$ such that $\mathbb{E}[(Z - \mathbb{E}[Z])^2] \preceq S$ and $0 < M < \infty$ such that $\|Z\|_{\mathcal{L}^\infty(H)} \leq M$ almost everywhere. Let*

$d := \frac{\|S\|_{\mathcal{L}^1(H)}}{\|S\|_{\mathcal{L}^\infty(H)}}$ *and* $\sigma^2 := \|S\|_{\mathcal{L}^\infty(H)}$. *Then for* $0 < \delta \leq d$,

$$P^n \left\{ (Z_i)_{i=1}^n : \left\| \frac{1}{n} \sum_{i=1}^n Z_i - \mathbb{E}[Z] \right\|_{\mathcal{L}^\infty(H)} \geq \frac{\beta M}{n} + \sqrt{\frac{3\beta\sigma^2}{n}} \right\} \leq \delta,$$

where $\beta := \frac{2}{3} \log \frac{4d}{\delta}$.

Theorem E.3. *Let* $(\mathcal{X}, P)$ *be a measurable space and* $Z : \mathcal{X} \times \mathcal{X} \to \mathcal{L}^2(H)$ *with* $Z(x, y) = Z(y, x)$ *for all* $x, y \in H$, *where* H *is a separable Hilbert space. Let*

$$D = \int \int Z(x, y)\, dP(x)\, dP(y)$$

with

$$\widehat{D} = \frac{1}{n(n-1)} \sum_{i \neq j}^n Z(X_i, X_j)$$

being its U-statistic estimator, where $(X_i)_{i=1}^n \overset{i.i.d.}{\sim} P$ *and* $n \geq 2$. *Define* $\psi(x) = \mathbb{E}_Y [Z(x, Y)]$ *and let* $\sup_{x,y \in \mathcal{X}} \|Z(x, y)\|_{\mathcal{L}^2(H)} \leq M$. *Then the following hold:*

(i) Suppose $Z : \mathcal{X} \times \mathcal{X} \to \mathcal{S}(H)$, *where* $\mathcal{S}(H)$ *is the space of self-adjoint Hilbert-Schmidt operators on* H, $\mathbb{E}[(\psi(X) - D)^2] \preceq S$, $\sigma^2 := \|S\|_{\mathcal{L}^\infty(H)}$, *and* $\sup_{x \in \mathcal{X}} \|\psi(x)\|_{\mathcal{L}^\infty(H)} \leq R$. *Then for* $0 < \delta \leq d$,

$$P^n \left\{ (X_i)_{i=1}^n : \left\| \widehat{D} - D \right\|_{\mathcal{L}^\infty(H)} \leq \frac{2\beta R}{n} + \sqrt{\frac{12\beta\sigma^2}{n}} + \frac{8M \log \frac{3}{\delta}}{n} \right\} \geq 1 - 2\delta,$$

where $\beta := \frac{2}{3} \log \frac{4d}{\delta}$ *and* $d := \frac{\|S\|_{\mathcal{L}^1(H)}}{\|S\|_{\mathcal{L}^\infty(H)}}$.

(ii) Suppose $\mathbb{E} \|\psi(X) - D\|_{\mathcal{L}^2(H)}^2 \leq \sigma_1^2$. *Then*

$$P^n \left\{ (X_i)_{i=1}^n : \left\| \widehat{D} - D \right\|_{\mathcal{L}^2(H)} \leq \frac{16M \log \frac{3}{\delta}}{n} + \sqrt{\frac{8\sigma_1^2 \log \frac{2}{\delta}}{n}} \right\} \geq 1 - 2\delta.$$

Proof. (i) The Hoeffding decomposition of $\widehat{D}$ yields

$$\widehat{D} - D = 2\left[\frac{1}{n}\sum_{i=1}^{n}\psi(X_i) - D\right] + \left[\frac{1}{n(n-1)}\sum_{i\neq j}^{n} Z(X_i, X_j) - \psi(X_i) - \psi(X_j) + D\right],$$

which implies

$$\left\|\widehat{D} - D\right\|_{\mathcal{L}^\infty(H)} \leq \left\|\frac{1}{n(n-1)}\sum_{i\neq j}^{n} Z(X_i, X_j) - \psi(X_i) - \psi(X_j) + D\right\|_{\mathcal{L}^\infty(H)} + 2\left\|\frac{1}{n}\sum_{i=1}^{n}\psi(X_i) - D\right\|_{\mathcal{L}^\infty(H)}. \tag{E.1}$$

The first term can be bounded by applying Theorem E.2 since $\mathbb{E}[\psi(X)] = D$. We now bound the second term as follows. Define $h(X_i, X_j) := Z(X_i, X_j) - \psi(X_i) - \psi(X_j) + D$. Applying Markov's inequality to the second term, we obtain that for any $\epsilon > 0$ and $t > 0$,

$$P^n\left\{(X_i)_{i=1}^n : \left\|\frac{1}{n(n-1)}\sum_{i\neq j}^{n} h(X_i, X_j)\right\|_{\mathcal{L}^\infty(H)} \geq \epsilon\right\} \leq e^{-t\epsilon}\mathbb{E}\exp\left\|t'\sum_{i\neq j}^{n} h(X_i, X_j)\right\|_{\mathcal{L}^\infty(H)}, \tag{E.2}$$

where $t' := \frac{t}{n(n-1)}$. Consider

$$\begin{aligned}
&\mathbb{E}\exp\left\|t'\sum_{i\neq j}^{n} h(X_i, X_j)\right\|_{\mathcal{L}^\infty(H)} \\
&= \mathbb{E}\exp\left\|t'\sum_{i\neq j}^{n}\Big[Z(X_i, X_j) - \mathbb{E}_{X_j'}Z(X_i, X_j') \right. \\
&\qquad\qquad \left. -\mathbb{E}_{X_i'}Z(X_j, X_i') + \mathbb{E}_{X_i', X_j'}Z(X_i', X_j')\Big]\right\|_{\mathcal{L}^\infty(H)} \\
&= \mathbb{E}\exp\left\|t'\mathbb{E}_{(X_i')_{i=1}^n|(X_i)_{i=1}^n}\sum_{i\neq j}^{n}\Big[Z(X_i, X_j) - Z(X_i, X_j') \right. \\
&\qquad\qquad \left. -Z(X_j, X_i') + Z(X_i', X_j')\Big]\right\|_{\mathcal{L}^\infty(H)}
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{E}\exp\left[\mathbb{E}_{(X_i')_{i=1}^n|(X_i)_{i=1}^n}\left\|t'\sum_{i\neq j}^n\Big[Z(X_i,X_j)-Z(X_i,X_j')\right.\right.\\
&\qquad\qquad\left.\left.-Z(X_i',X_j)+Z(X_i',X_j')\Big]\right\|_{\mathcal{L}^\infty(H)}\right]\\
&\overset{(*)}{=}\mathbb{E}\exp\left[\mathbb{E}_{(X_i')_{i=1}^n|(X_i)_{i=1}^n}\left\|t'\sum_{i\neq j}^n\left(\delta_{X_i}-\delta_{X_i'}\right)\left(\delta_{X_j}-\delta_{X_j'}\right)Z\right\|_{\mathcal{L}^\infty(H)}\right]\\
&=\mathbb{E}\exp\left[\mathbb{E}_{\epsilon^{(1)}}\mathbb{E}_{\epsilon^{(2)}}\mathbb{E}_{(X_i')_{i=1}^n|(X_i)_{i=1}^n}\left\|t'\sum_{i\neq j}^n\epsilon_i^{(1)}\left(\delta_{X_i}-\delta_{X_i'}\right)\right.\right.\\
&\qquad\qquad\left.\left.\times\epsilon_j^{(2)}\left(\delta_{X_j}-\delta_{X_j'}\right)Z\right\|_{\mathcal{L}^\infty(H)}\right]
\end{aligned}
$$

$$
\overset{(\dagger)}{\leq}\mathbb{E}\exp\left[\mathbb{E}_{\epsilon^{(1)}}\mathbb{E}_{\epsilon^{(2)}}\left\|t'\sum_{i\neq j}^n\epsilon_i^{(1)}\epsilon_j^{(2)}\left(\delta_{X_i}-\delta_{X_i'}\right)\left(\delta_{X_j}-\delta_{X_j'}\right)Z\right\|_{\mathcal{L}^\infty(H)}\right],
$$

where $(\epsilon_i^{(1)})_i$ and $(\epsilon_i^{(2)})_i$ are independent Rademacher random variables. In $(*)$, δ_x denotes a Dirac measure supported on x and we use the notation $Qf := \int f(x)\,dQ(x)$ with Q being a Dirac measure. In $(\dagger)$, the expectation is jointly over $(X_i, X_i')_{i=1}^n$ which is obtained through an application of Jensen's inequality. Therefore,

$$
\begin{aligned}
&\mathbb{E}\exp\left\|t'\sum_{i\neq j}^n h(X_i,X_j)\right\|_{\mathcal{L}^\infty(H)}\\
&\leq\mathbb{E}\exp\left[\mathbb{E}_{\epsilon^{(1)}}\mathbb{E}_{\epsilon^{(2)}}\left\|t'\sum_{i\neq j}^n\epsilon_i^{(1)}\epsilon_j^{(2)}Z(X_i,X_j)\right\|_{\mathcal{L}^\infty(H)}\right.\\
&\quad+\mathbb{E}_{\epsilon^{(1)}}\mathbb{E}_{\epsilon^{(2)}}\left\|t'\sum_{i\neq j}^n\epsilon_i^{(1)}\epsilon_j^{(2)}Z(X_i,X_j')\right\|_{\mathcal{L}^\infty(H)}\\
&\quad+\mathbb{E}_{\epsilon^{(1)}}\mathbb{E}_{\epsilon^{(2)}}\left\|t'\sum_{i\neq j}^n\epsilon_i^{(1)}\epsilon_j^{(2)}Z(X_i',X_j)\right\|_{\mathcal{L}^\infty(H)}\\
&\quad\left.+\mathbb{E}_{\epsilon^{(1)}}\mathbb{E}_{\epsilon^{(2)}}\left\|t'\sum_{i\neq j}^n\epsilon_i^{(1)}\epsilon_j^{(2)}Z(X_i',X_j')\right\|_{\mathcal{L}^\infty(H)}\right].
\end{aligned}
$$

Since $(X_i)_{i=1}^n$ and $(X_i')_{i=1}^n$ are i.i.d., we have

$$\begin{aligned}
&\mathbb{E}\exp\left\|t'\sum_{i\neq j}^{n}h(X_i,X_j)\right\|_{\mathcal{L}^\infty(H)} \\
&\leq \mathbb{E}\exp\left[4t'\mathbb{E}_{\epsilon^{(1)}}\mathbb{E}_{\epsilon^{(2)}}\left\|\sum_{i\neq j}^{n}\epsilon_i^{(1)}\epsilon_j^{(2)}Z(X_i,X_j)\right\|_{\mathcal{L}^\infty(H)}\right] \qquad \text{(E.3)} \\
&\leq \mathbb{E}\exp\left[4t'\mathbb{E}_{\epsilon^{(1)}}\mathbb{E}_{\epsilon^{(2)}}\left\|\sum_{i\neq j}^{n}\epsilon_i^{(1)}\epsilon_j^{(2)}Z(X_i,X_j)\right\|_{\mathcal{L}^2(H)}\right] \\
&\leq \mathbb{E}\exp\left[4t'\sqrt{\mathbb{E}_{\epsilon^{(1)}}\mathbb{E}_{\epsilon^{(2)}}\left\|\sum_{i\neq j}^{n}\epsilon_i^{(1)}\epsilon_j^{(2)}Z(X_i,X_j)\right\|_{\mathcal{L}^2(H)}^2}\right],
\end{aligned}$$

where the last inequality follows from Jensen's inequality. We will now bound

$$\begin{aligned}
&\mathbb{E}_{\epsilon^{(1)}}\mathbb{E}_{\epsilon^{(2)}}\left\|\sum_{i\neq j}^{n}\epsilon_i^{(1)}\epsilon_j^{(2)}Z(X_i,X_j)\right\|_{\mathcal{L}^2(H)}^2 \\
&\quad=\mathbb{E}_{\epsilon^{(1)}}\mathbb{E}_{\epsilon^{(2)}}\sum_{i\neq j}^{n}\sum_{k\neq l}^{n}\epsilon_i^{(1)}\epsilon_j^{(2)}\epsilon_k^{(1)}\epsilon_l^{(2)}\langle Z(X_i,X_j),Z(X_k,X_l)\rangle_{\mathcal{L}^2(H)}.
\end{aligned}$$

We consider the following cases.

Case 1: $i=k, j=l$

$$\begin{aligned}
\mathbb{E}_{\epsilon^{(1)}}\mathbb{E}_{\epsilon^{(2)}}\left\|\sum_{i\neq j}^{n}\epsilon_i^{(1)}\epsilon_j^{(2)}Z(X_i,X_j)\right\|_{\mathcal{L}^2(H)}^2 &= \sum_{i\neq j}^{n}\|Z(X_i,X_j)\|_{\mathcal{L}^2(H)}^2 \\
&\leq n(n-1)M^2.
\end{aligned}$$

Case 2: $i = k, j \neq l$

$$\mathbb{E}_{\epsilon^{(1)}}\mathbb{E}_{\epsilon^{(2)}} \left\| \sum_{i\neq j}^{n} \epsilon_i^{(1)}\epsilon_j^{(2)} Z(X_i, X_j) \right\|^2_{\mathcal{L}^2(H)}$$
$$= \mathbb{E}_{\epsilon^{(2)}} \sum_{i\neq j\neq l}^{n} \epsilon_j^{(2)}\epsilon_l^{(2)} \langle Z(X_i, X_j), Z(X_i, X_l)\rangle_{\mathcal{L}^2(H)} = 0.$$

Case 3: $i \neq k, j = l$

$$\mathbb{E}_{\epsilon^{(1)}}\mathbb{E}_{\epsilon^{(2)}} \left\| \sum_{i\neq j}^{n} \epsilon_i^{(1)}\epsilon_j^{(2)} Z(X_i, X_j) \right\|^2_{\mathcal{L}^2(H)}$$
$$= \mathbb{E}_{\epsilon^{(1)}} \sum_{i\neq j\neq k}^{n} \epsilon_i^{(1)}\epsilon_k^{(1)} \langle Z(X_i, X_j), Z(X_k, X_j)\rangle_{\mathcal{L}^2(H)} = 0.$$

Case 4: $i \neq k, j \neq l$

$$\mathbb{E}_{\epsilon^{(1)}}\mathbb{E}_{\epsilon^{(2)}} \left\| \sum_{i\neq j}^{n} \epsilon_i^{(1)}\epsilon_j^{(2)} Z(X_i, X_j) \right\|^2_{\mathcal{L}^2(H)} = 0.$$

Therefore,

$$\mathbb{E}\exp \left\| t' \sum_{i\neq j}^{n} h(X_i, X_j) \right\|_{\mathcal{L}^\infty(H)} \leq \exp\left[\frac{4tM}{\sqrt{n(n-1)}}\right] \leq \exp\left[\frac{8Mt}{n}\right]$$

for $n \geq 2$ as $n - 1 \geq \frac{n}{4}$. Using this in (E.2) and choosing $t = \frac{n}{8M}$, we obtain

$$P^n \left\{ (X_i)_{i=1}^n : \left\| \frac{1}{n(n-1)} \sum_{i\neq j}^{n} h(X_i, X_j) \right\|_{\mathcal{L}^\infty(H)} \geq \epsilon \right\} \leq 3\exp\left(-\frac{n\epsilon}{8M}\right),$$

which is equivalent to

$$P^n \left\{ (X_i)_{i=1}^n : \left\| \frac{1}{n(n-1)} \sum_{i\neq j}^{n} h(X_i, X_j) \right\|_{\mathcal{L}^\infty(H)} \geq \frac{8M}{n} \log\frac{3}{\delta} \right\} \leq \delta. \tag{E.4}$$

Combining (E.4) with the bound on the first term in (E.1) yields the result.

(ii) As in (i), we first write

$$\left\|\widehat{D} - D\right\|_{\mathcal{L}^2(H)} \leq \left\| \frac{1}{n(n-1)} \sum_{i \neq j} Z(X_i, X_j) - \psi(X_i) - \psi(X_j) + D \right\|_{\mathcal{L}^2(H)} + 2 \left\| \frac{1}{n} \sum_{i=1}^{n} \psi(X_i) - D \right\|_{\mathcal{L}^2(H)} . \tag{E.5}$$

The first term in (E.5) is bounded through an application of Theorem E.1. For the second term, we replicate the analysis between (E.2) and (E.3) with the operator norm being replaced by the Hilbert-Schmidt norm. Since the analysis between (E.3) and (E.4) anyway relies only on the Hilbert-Schmidt norm, the result follows by employing that $\log \frac{2}{\delta} < \log \frac{3}{\delta}$. □

Bibliography

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

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

Bach, F. R., Thibaux, R., and Jordan, M. I. (2005). Computing regularization paths for learning multiple kernels. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17*, Cambridge, MA. MIT Press.

Balasubramanian, K., Li, T., and Yuan, M. (2021). On the optimality of kernel-embedding based goodness-of-fit tests. *Journal of Machine Learning Research*, 22(1):1–45.

Bhatia, R. (1997). *Matrix Analysis*. Springer-Verlag, New York.

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

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

Cohen, M. B., lee, Y. T., Musco, C., Musco, C., Peng, R., and Sidford, A. (2015). Uniform sampling for matrix approximation. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 181–190. ACM.

Cohn, D. L. (2013). *Measure Theory*. Birkhäuser Basel.

Cortes, C., Mohri, M., and Talwalkar, A. (2010). On the impact of kernel approximation on learning accuracy. In Teh, Y. W. and Titterington, M., editors, *Proc. of the 13th International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 113–120. PMLR.

De Vito, E., Rosasco, L., and Toigo, A. (2014). Learning sets with separating kernels. *Applied and Computational Harmonic Analysis*, 37(2):185–217.

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

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

Drineas, P., Magdon-Ismail, M., Mahoney, M. W., and Woodruff, D. P. (2012). Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13:3475–3506.

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

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

Gao, C., Ma, Z., Ren, Z., and Zhou, H. H. (2015). Minimax estimation in sparse canonical correlation analysis. *The Annals of Statistics*, 43(5).

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

Gohberg, I. and Goldberg, S. (2003). *Basic Classes of Linear Operators*. Birkhaäuser Basel.

Hsieh, C., Si, S., and Dhillon, I. S. (2014). A divide-and-conquer solver for kernel support vector machines. In *International Conference on Machine Learning (ICML)*.

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

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

Kato, T. (1987). Variation of discrete spectra. *Communications in Mathematical Physics*, 111:501–504.

Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, 41(2):495–502.

Kimeldorf, G. S. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33:82–95.

Kumar, S., Mohri, M., and Talwalkar, A. (2012). Sampling methods for the Nyström method. *Journal of Machine Learning Research*, 13:981–1006.

Li, T. and Yuan, M. (2019). On the optimality of Gaussian kernel based nonparametric tests against smooth alternatives. arXiv preprint arXiv:1909.03302.

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

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

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

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

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

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

Schölkopf, B., Herbrich, R., and Smola, A. (2001). A generalized representer theorem. In *Proc. of the 14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory*, pages 416–426, London, UK. Springer-Verlag.

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

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels.* MIT Press, Cambridge, MA.

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

Smale, S. and Zhou, D.-X. (2007). Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26:153–172.

Sriperumbudur, B. and Sterge, N. (2020). Approximate kernel PCA using random features: Computational vs. statistical trade-off. arXiv preprint arXiv:1706.06296v3.

Sriperumbudur, B. K. and Szabó, Z. (2015). Optimal rates for random Fourier features. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 1144–1152. Curran Associates, Inc.

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

Sterge, N., Sriperumbudur, B., Lorenzo, L. R., and Rudi, A. (2020). Gain with no pain: Efficiency of kernel-PCA by Nyström sampling. In Chiappa, S. and Calandra, R., 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.

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

Wang, S., Gittens, A., and Mahoney, M. W. (2019). Scalable kernel k-means clustering with nyström approximation: Relative-error bounds. *Journal of Machine Learning Research*, 20(12):1–49.

Wendland, H. (2005). *Scattered Data Approximation*. Cambridge University Press, Cambridge, UK.

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

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

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

Yurinsky, V. (1995). *Sums and Gaussian Vectors*, volume 1617 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin.

Zhang, Y., Duchi, J., and Wainwright, M. (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16(102):3299–3340.

Vita

Nicholas J. Sterge

Education:

- Doctor of Philosophy, Statistics, Pennsylvania State University, 2021
- Bachelor of Science, Mathematics, Wake Forest University, 2015

Publications:

- **Gain With No Pain: Efficiency of Kernel-PCA by Nyström Sampling**. **N. Sterge**, B. K. Sriperumbudur, L. Rosasco and A. Rudi
  International Conference on Artificial Intelligence and Statistics, 2020.
- **Approximate Kernel PCA Using Random Features: Computational vs. Statistical Trade-Off**. B. K. Sriperumbudur and **N. Sterge**
  Under Revision, Annals of Statistics
- **Statistical Optimality and Computational Efficiency of Nyström Kernel PCA**. **N. Sterge** and B. K. Sriperumbudur
  Under Review, Journal of Machine Learning Research