The Appendix is divided into the following six sections,

A. *Preliminaries and Structural Results*
B. *Equivalence of Optimization problem in $\mathcal{H}$ and $L^2(\mathcal{X}, \rho)$*
C. *Proof of the main Theorem*
D. *Examples of ESL*
E. *Experiments*
F. *Auxiliary Results*

A Preliminaries and Structural Results

We begin this section by giving some definitions and structural theorems that are required for the proofs. We first introduce a few additional notation. Consider two separable Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$. For an operator $\mathcal{D} : \mathcal{H}_1 \to \mathcal{H}_2$, we use $\|\mathcal{D}\|_{\mathcal{L}^p(\mathcal{H}_1, \mathcal{H}_2)}$ to denote its p^{th} schatten norm, assuming that it is finite. We omit the p when we talk about about Hilbert-Schmidt norm i.e. $p = 2$.

A.1 Covariance operators C and C_m

The covariance operators in the RKHS and the random feature space are defined as follows:

Definition A.1. $\mathrm{C} : \mathcal{H} \to \mathcal{H}$ is the co-variance operator of the random variables $k(\mathrm{x}, \cdot)$ with measure ρ, defined as:

$$\mathrm{C}f := \int_{\mathcal{X}} k(\mathrm{x}, \cdot) f(\mathrm{x}, t) d\rho(\mathrm{x}, t)$$

C is compact and self-adjoint, which implies C has a spectral decomposition as follows:

$$\mathrm{C} = \sum_{i=1}^{\infty} \bar{\lambda}_i \bar{\phi}_i \otimes_{\mathcal{H}} \bar{\phi}_i$$

where $\bar{\lambda_i}, \bar{\phi}_i$'s are the eigenvalues and eigenfunctions of C. Also, $\bar{\phi_i}$ are a unitary basis for $\mathcal{H}$.

Definition A.2. $\mathrm{C}_m : \mathcal{F} \to \mathcal{F}$ is the covariance operator in the random feature space, defined as

$$\mathrm{C}_m := \mathbb{E}_\rho \left[\mathrm{z}(\mathrm{x}, t) \otimes_{\mathcal{F}} \mathrm{z}(\mathrm{x}, t) \right]$$

Equivalently, for any $\beta \in \mathcal{F}$, $\mathrm{C}_m \beta = \int_{\mathcal{X}} \langle \mathrm{z}(\mathrm{x}, t), \beta \rangle \, \mathrm{z}(\mathrm{x}, t) d\rho(\mathrm{x}, t)$.

C_m is compact and self-adjoint which implies that C_m has a spectral decomposition as follows:

$$\mathrm{C}_m = \sum_{i=1}^{m} \lambda_i \phi_i \otimes_{\mathcal{F}} \phi_i$$

The kernel integral operators and its approximation based on random features are defined as follows:

Definition A.3. The kernel integral operator $\mathrm{L} : L^2(\mathcal{X}, \rho) \to L^2(\mathcal{X}, \rho)$ is defined as follows:

$$\mathrm{L}g = \int_{\mathcal{X}} k(\mathrm{x}, \cdot) g(\mathrm{x}) d\rho(\mathrm{x}) \; \forall \; g \in L^2(\mathcal{X}, \rho)$$

Definition A.4. $\mathrm{L}_m : L^2(\mathcal{X}, \rho) \to L^2(\mathcal{X}, \rho)$ is the (approximated) kernel integral operator, defined as:

$$(\mathrm{L}_m g)(\cdot) = \int_{\mathcal{X}} k_m(\mathrm{x}, \cdot) g(\mathrm{x}) d\rho(\mathrm{x})$$

We state the classical Mercer's and Bochner's theorems for completeness.

Theorem A.5 (Mercer's Theorem)**.** *For every positive definition kernel $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exits a set Ω with measure π, and functions $\varpi(\cdot) : \mathcal{X} \times \Omega \to \mathbb{R}$ such that the kernel has an integral representation of the following form,*

$$k(\mathrm{x}, \mathrm{y}) = \int_{\Omega} z(\mathrm{x}, \omega) z(\mathrm{y}, \omega) d\pi(\omega) \; \forall \; \mathrm{x}, \mathrm{y} \in \mathcal{X}$$

In particular, for shift-invariant kernels, we have

Theorem A.6 (Bochner's Theorem Rudin [2017])**.** *A continuous, real-valued, symmetric shift-invariant kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive-definite kernel if and only if there exits a non-negative measure $\pi(\omega)$ such that $k(x-y) = \int_{\mathcal{X}} \mathrm{e}^{i\omega^\top (x-y)} d\pi(\omega)$ i.e the inverse Fourier transform of $k(x-y)$*

For a comprehensive list of kernels with their Fourier transform see Table 1 of Xie et al. [2015].

We now define operators to lift functions and operators from and into different spaces. These are crucially used in the analysis of the algorithm. See Figure 3 in for a schematic of the lifting operators.

A.2 Inclusions operators $\mathrm{I}, \mathfrak{I}$

We first recall the definitions of Inclusion operators $\mathrm{I}, \mathfrak{I}$.

Definition A.7. [Inclusion Operators I and $\mathfrak{I}$] The inclusion operator is defined

$$\mathrm{I} : \mathcal{H} \to L^2(\mathcal{X}, \rho), \ (\mathrm{I}f) = f, \ \text{where } f \in \mathcal{H}$$

Also, for an operator $\mathrm{D} \in HS(\mathcal{H})$ with spectral decomposition $\mathrm{D} = \sum_{i \in I \subset \mathbb{R}} \mu_i \psi_i \otimes \psi_i$,

$$\mathfrak{I} : HS(\mathcal{H}) \to HS(\rho), \ \mathfrak{I}\mathrm{D} := \sum_{i \in I \subset \mathbb{R}} \mu_i \frac{\mathrm{I}\psi_i}{\sqrt{\langle \mathrm{C}\psi_i, \psi_i \rangle_{\mathcal{H}}}} \otimes \frac{\mathrm{I}\psi_i}{\sqrt{\langle \mathrm{C}\psi_i, \psi_i \rangle_{\mathcal{H}}}}$$

In Proposition A.8, we show that the adjoint of the Inclusion operator I is

$$\mathrm{I}^* : L^2(\mathcal{X}, \rho) \to \mathcal{H}, (\mathrm{I}^* g)(\cdot) = \int k(\mathrm{x}, \cdot) g(\mathrm{x}) d\rho(\mathrm{x}).$$

Moreover, In Proposition A.9 we show that the covariance operator and the kernel integral operator can be expressed in terms of I and I$*$ as $\mathrm{C} = \mathrm{I}^*\mathrm{I}$ and $\mathrm{L} = \mathrm{I}\mathrm{I}^*$.

Proposition A.8. *The following holds with regard to the inclusion operator,*

(a). The adjoint of the Inclusion operator I *is given by* $(\mathrm{I}^* g)(\cdot) = \int_{\mathcal{X}} k(\mathrm{x}, \cdot) g(\mathrm{x}) d\rho(\mathrm{x})$.

(b). I *and* I^* *are Hilbert-Schmidt.*

Proof of Proposition A.8. (a). We first show that the adjoint of the Inclusion operator I is given by $(\mathrm{I}^* g)(\cdot) = \int_{\mathcal{X}} k(\mathrm{x}, \cdot) g(\mathrm{x}) d\rho(\mathrm{x})$. For $f \in \mathcal{H}$ and $g \in L^2(\mathcal{X}, \rho)$, we have that

$$\begin{aligned}
\langle \mathrm{I}f, g \rangle_\rho &= \langle f, g \rangle_\rho && \text{(Definition of I)} \\
&= \int_{\mathcal{X}} f(\mathrm{x}) g(\mathrm{x}) d\rho(\mathrm{x}) \\
&= \int_{\mathcal{X}} \langle k(\mathrm{x}, \cdot), f \rangle_{\mathcal{H}} \, g(\mathrm{x}) d\rho(\mathrm{x}) && \text{(Reproducing property)} \\
&= \int_{\mathcal{X}} \langle k(\mathrm{x}, \cdot) g(\mathrm{x}), f \rangle_{\mathcal{H}} \, d\rho(\mathrm{x}) && \text{(Linearity of inner product)} \\
&= \left\langle \int_{\mathcal{X}} k(\mathrm{x}, \cdot) g(\mathrm{x}) d\rho(\mathrm{x}), f \right\rangle_{\mathcal{H}} && \text{(Fubini's Theorem)} \\
&= \langle \mathrm{I}^* g, f \rangle_{\mathcal{H}}
\end{aligned}$$

(b). Let $\{\bar{\mathrm{e}}_i\}_{i=1}^{\infty}$ be an orthonormal basis for $\mathcal{H}$. We have,

$$\begin{aligned}
\|\mathrm{I}\|^2_{\mathcal{L}(\mathcal{H},\rho)} &= \sum_{i=1}^{\infty} \|\mathrm{I}\bar{\mathrm{e}}_i\|^2_{\rho} && \text{(Pythagoras Theorem)} \\
&= \sum_{i=1}^{\infty} \|\bar{\mathrm{e}}_i\|^2_{\rho} && \text{(Definition 3.5)} \\
&= \sum_{i=1}^{\infty} \int_{\mathcal{X}} \langle k(\mathrm{x},\cdot), \bar{\mathrm{e}}_i\rangle^2_{\mathcal{H}}\, d\rho(\mathrm{x}) && \text{(Reproducing Property)} \\
&= \int_{\mathcal{X}} \sum_{i=1}^{\infty} \langle k(\mathrm{x},\cdot), \bar{\mathrm{e}}_i\rangle^2_{\mathcal{H}}\, d\rho(\mathrm{x}) && \text{(Fubini's Theorem)} \\
&= \int_{\mathcal{X}} k(x,x) d\rho(\mathrm{x}) \leq \tau^2 < \infty && \text{(Assumption 3.1)}
\end{aligned}$$

For the adjoint I^*, we have $\|\mathrm{I}^*\|_{\mathcal{L}(\rho,\mathcal{H})} = \|\mathrm{I}\|_{\mathcal{L}(\mathcal{H},\rho)} < \infty$ □

Proposition A.9. *The following properties hold,*

(a). The covariance operator and the kernel integral operator satisfy $\mathrm{C} = \mathrm{I}^*\mathrm{I}$ *and* $\mathrm{L} = \mathrm{I}\mathrm{I}^*$ *respectively.*

(b). C *and* L *are trace-class*

Proof of Proposition A.9. (a). We first show that $\mathrm{C} = \mathrm{I}^*\mathrm{I}$. For any $f \in L^2(\mathcal{X},\rho)$, we have

$$\begin{aligned}
\mathrm{I}^*\mathrm{I}f &= \mathrm{I}^* f && \text{(Definition A.7)} \\
&= \int_{\mathcal{X}} k(\mathrm{x},\cdot) f(\mathrm{x}) d\rho(\mathrm{x}) && \text{(Proposition A.8)} \\
&= \mathrm{C}f && \text{(Definition A.1)}
\end{aligned}$$

We now show that $\mathrm{L} = \mathrm{I}\mathrm{I}^*$. For any $g \in L^2(\mathcal{X},\rho)$, we have

$$\begin{aligned}
\mathrm{I}\mathrm{I}^* g &= \mathrm{I}\left(\int_{\mathcal{X}} k(\mathrm{x},\cdot) g(\mathrm{x}) d\rho(\mathrm{x})\right) && \text{(Proposition A.8)} \\
&= \int_{\mathcal{X}} k(\mathrm{x},\cdot) g(\mathrm{x}) d\rho(\mathrm{x}) && \text{(Definition A.7)} \\
&= \mathrm{L}g
\end{aligned}$$

(b). Now we show that C and L are trace-class.

$$\begin{aligned}
\|\mathrm{C}\|_{\mathcal{L}^1(\mathcal{H})} &= \|\mathrm{I}^*\mathrm{I}\|_{\mathcal{L}^1(\mathcal{H})} && \text{(Proposition A.9)} \\
&= \|\mathrm{I}\|^2_{\mathcal{L}^2(\mathcal{H})} < \infty && \text{(Proposition A.8)}
\end{aligned}$$

Similarly,

$$\begin{aligned}
\|\mathrm{L}\|_{\mathcal{L}^1(\rho)} &= \|\mathrm{I}\mathrm{I}^*\|_{\mathcal{L}^1(\rho)} && \text{(Proposition A.9)} \\
&= \|\mathrm{I}\|^2_{\mathcal{L}^2(\mathcal{H})} < \infty && \text{(Proposition A.8)}
\end{aligned}$$

□

A.3 Approximation operators A and $\mathfrak{A}$

We first recall the definitions of approximation operators A and $\mathfrak{A}$.

Definition A.10. [Approximation Operators A and $\mathfrak{A}$] The Approximation operator A is defined as

$$\mathrm{A} : \mathcal{F} \to L^2(\mathrm{X},\rho), (\mathrm{A}\mathrm{v})(\cdot) = \langle \mathrm{z}(\cdot), \mathrm{v}\rangle\,, \text{ where } \mathrm{v} \in \mathcal{F}$$

For an operator $\mathrm{D} \in HS(\mathcal{F})$ with rank k with spectral decomposition $\mathrm{D} = \sum_{i=1}^{\infty} \mu_i \psi_i \otimes \psi_i$, let Ψ be the matrix with eigenvectors ψ_i as columns and Φ be the matrix with eigenvectors of C_m as columns (see Definition 3.4). Define

$$\mathrm{R}^* = \underset{\mathrm{R}^\top \mathrm{R} = \mathrm{R}\mathrm{R}^\top = \mathrm{I}}{\arg\min} \|\Psi \mathrm{R} - \Phi\|_{\mathcal{F}}^2, \ \tilde{\Psi} := \Psi \mathrm{R}^*$$

Let $\tilde{\psi}_i$ be the columns of $\tilde{\Psi}$, define

$$\mathfrak{A} : HS(\mathcal{F}) \to HS(\rho), \ \mathfrak{A}\mathrm{D} := \sum_{i=1}^{k} \mu_i \frac{\mathrm{A}\tilde{\psi}_i}{\sqrt{\langle \mathrm{C}_m \tilde{\psi}_i, \tilde{\psi}_i \rangle_{\mathcal{F}}}} \otimes_{\rho} \frac{\mathrm{A}\tilde{\psi}_i}{\sqrt{\langle \mathrm{C}_m \tilde{\psi}_i, \tilde{\psi}_i \rangle_{\mathcal{F}}}}$$

Note that the definition of the approximation operator $\mathfrak{A}$ requires knowledge of the co-variance matrix C_m to find the optimal rotation matrix R^*, but this is solely for the purpose of analysis and is not used in the algorithm in any form.

In Proposition A.11, we show that the adjoint of the Approximation Operator is

$$\mathrm{A}^* : L^2(\mathrm{X}, \rho) \to \mathcal{F}, (\mathrm{A}^* f)_i = \int_{\mathcal{X}} f(\mathrm{x}) z_{\omega_i}(\mathrm{x}) d\rho(\mathrm{x}).$$

Moreover, in Proposition A.12 we show that that the approximate covariance operator and the approximate kernel integral operator can be expressed in terms of the Approximation operator A as $\mathrm{C}_m = \mathrm{A}^*\mathrm{A}$ and $\mathrm{L}_m = \mathrm{A}\mathrm{A}^*$.

Proposition A.11. *The approximation operator satisfies the following properties,*

(a). The adjoint of A *is* $(\mathrm{A}^* f)_i = \frac{1}{\sqrt{m}} \int_{\mathcal{X}} f(\mathrm{x}) z_{\omega_i}(\mathrm{x}) d\rho(\mathrm{x})$

(b). A *and* A^* *are Hilbert-Schmidt.*

Proof of Proposition A.11. (a). First we show that the adjoint of A is $(\mathrm{A}^* f)_i = \frac{1}{\sqrt{m}} \int_{\mathcal{X}} f(\mathrm{x}) z_{\omega_i}(\mathrm{x}) d\rho(\mathrm{x})$. For $\mathrm{v} \in \mathcal{F}, f \in L^2(\mathcal{X}, \rho)$, we have,

$$\begin{aligned}
\langle \mathrm{Av}, f \rangle_{\rho} &= \int_{\mathcal{X}} (\mathrm{Av})(\mathrm{x}) f(\mathrm{x}) d\rho(\mathrm{x}) \\
&= \int_{\mathcal{X}} \langle \mathrm{z}(\mathrm{x}), \mathrm{v} \rangle_{\mathcal{F}} f(\mathrm{x}) d\rho(\mathrm{x}) && \text{(Definition A.10)} \\
&= \left\langle \int_{\mathcal{X}} \mathrm{z}(\mathrm{x}) f(\mathrm{x}) d\rho(\mathrm{x}), \mathrm{v} \right\rangle_{\mathcal{F}} && \text{(Fubini's Theorem)} \\
&= \langle \mathrm{A}^* f, \mathrm{v} \rangle_{\mathcal{F}}
\end{aligned}$$

(b). Let $\{\mathrm{e}_i\}$ be an orthonormal basis for $\mathcal{F}$.

$$\begin{aligned}
\|\mathrm{A}\|_{\mathcal{L}(\mathcal{F},\rho)}^2 &= \sum_{i=1}^{m} \|\mathrm{Ae}_i\|_{\rho}^2 && \text{(Pythagoras Theorem)} \\
&= \sum_{i=1}^{m} \|\langle \mathrm{z}(\cdot), \mathrm{e}_i \rangle_{\mathcal{F}}\|_{\rho}^2 && \text{(Definition 3.6)} \\
&= \sum_{i=1}^{m} \int_{\mathcal{X}} (\langle \mathrm{z}(\mathrm{x}), \mathrm{e}_i \rangle_{\mathcal{F}})^2 d\rho(\mathrm{x}) \\
&\leq \sum_{i=1}^{m} \int_{\mathcal{X}} \|\mathrm{z}(\mathrm{x})\|_{\mathcal{F}}^2 d\rho(\mathrm{x}) < m\tau^2 < \infty
\end{aligned}$$

where third last and second inequality follows from Cauchy Schwartz inequality and Assumption 3.1 respectively.

Similarly, to show A^* is Hilbert-Schmidt, we note that $\|\mathrm{A}^*\|_{\mathcal{L}^2(\rho,\mathcal{F})} = \|\mathrm{A}\|_{\mathcal{L}^2(\mathcal{F},\rho)}^2 < \infty$ □

In the following proposition, we show how the Covariance operator C_m and kernel Integral operator L_m are related.

Proposition A.12. *The following properties hold,*

(a). C_m *and* L_m *satisfy that* $\mathrm{C}_m = \mathrm{A}^*\mathrm{A}, \mathrm{L}_m = \mathrm{A}\mathrm{A}^*$

(b). C_m *and* L_m *are trace-class.*

Proof of Proposition A.12. (a). We first show the first part of the Proposition. For any $\mathrm{v} \in \mathcal{F}$, we have,

$$\begin{aligned}
\mathrm{A}^*\mathrm{A}\mathrm{v} &= \int_{\mathcal{X}} \langle \mathrm{z}(\mathrm{x}), \mathrm{v} \rangle_{\mathcal{F}} \, \mathrm{z}(\mathrm{x}) d\rho(\mathrm{x}) && \text{(Definition A.10 and Proposition A.11)} \\
&= \mathbb{E}_\rho \left[\mathrm{z}(\mathrm{x}) \otimes_{\mathcal{F}} \mathrm{z}(\mathrm{x}) \right] \mathrm{v} \\
&= \mathrm{C}_m \mathrm{v} && \text{(Definition A.2)}
\end{aligned}$$

For any $g \in L^2(\mathcal{X}, \rho)$,

$$\begin{aligned}
\mathrm{A}\mathrm{A}^* g &= \frac{1}{m} \sum_{i=1}^{m} z_{\omega_i}(\cdot) \int_{\mathcal{X}} z_{\omega_i}(\mathrm{x}) g(\mathrm{x}) d\rho(\mathrm{x}) && \text{(Definition A.10 and Proposition A.11)} \\
&= \int_{\mathcal{X}} \sum_{i=1}^{m} \frac{1}{\sqrt{m}} z_{\omega_i}(\cdot) \frac{1}{\sqrt{m}} z_{\omega_i}(\mathrm{x}) g(\mathrm{x}) d\rho(\mathrm{x}) && \text{(Fubini's Theorem)} \\
&= \int_{\mathcal{X}} \langle \mathrm{z}(\mathrm{x}), \mathrm{z}(\cdot) \rangle_{\mathcal{F}} \, g(\mathrm{x}) d\rho(\mathrm{x}) \\
&= \int_{\mathcal{X}} k_m(\mathrm{x}, \cdot) g(\mathrm{x}) d\rho(\mathrm{x}, t) && \text{(Definition of the approximate kernel mapping)} \\
&= \mathrm{L}_m g && \text{(Definition A.4)}
\end{aligned}$$

(b). Now we show that C_m and L_m are trace-class.

$$\begin{aligned}
\|\mathrm{C}_m\|_{\mathcal{L}^1(\mathcal{F})} &= \|\mathrm{A}^*\mathrm{A}\|_{\mathcal{L}^1(\mathcal{F})} && \text{(Proposition A.12)} \\
&= \|\mathrm{A}\|^2_{\mathcal{L}^2(\mathcal{F})} < \infty && \text{(Proposition A.11)}
\end{aligned}$$

Similarly,

$$\begin{aligned}
\|\mathrm{L}_m\|_{\mathcal{L}^1(\rho)} &= \|\mathrm{A}\mathrm{A}^*\|_{\mathcal{L}^1(\rho)} && \text{(Proposition A.12)} \\
&= \|\mathrm{A}^*\|^2_{\mathcal{L}^2(\rho)} < \infty && \text{(Proposition A.11)}
\end{aligned}$$

□

A.4 Kernel integral operator L and its approximation L_m

We first recall the definition of Kernel integral operator L and its approximation L_m.

Definition A.13. The kernel integral operator $\mathrm{L} : L^2(\mathcal{X}, \rho) \to L^2(\mathcal{X}, \rho)$ is defined as follows:

$$\mathrm{L} g = \int_{\mathcal{X}} k(\mathrm{x}, \cdot) g(\mathrm{x}) d\rho(\mathrm{x}) \; \forall \; g \in L^2(\mathcal{X}, \rho)$$

Definition A.14. $\mathrm{L}_m : L^2(\mathcal{X}, \rho) \to L^2(\mathcal{X}, \rho)$ is the (approximated) kernel integral operator, defined as:

$$(\mathrm{L}_m g)(\cdot) = \int_{\mathcal{X}} k_m(\mathrm{x}, \cdot) g(\mathrm{x}) d\rho(\mathrm{x})$$

We now show in Proposition A.15 that spectral decomposition of the kernel integral operator L can be given in terms of the eigenfunctions and the eigenvalues of the covariance operator C.

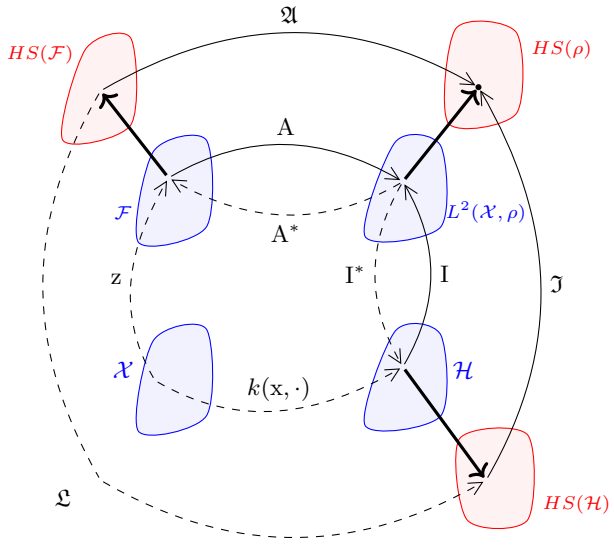

Figure 3: Maps between the data domain ($\mathcal{X}$), space of square integrable functions on $\mathcal{X}$ ($L^2(\mathcal{X},\rho)$), the RKHS of kernel $k(\cdot,\cdot)$, and RKHS of the approximate feature map, as well as maps between Hilbert-Schmidt operators on these spaces.

Proposition A.15. *The spectral decomposition of* L *is:*

$$\mathrm{L} = \sum_{i=1}^{\infty} \bar{\lambda}_i \frac{\mathrm{I}\bar{\phi}_i}{\sqrt{\bar{\lambda}_i}} \otimes_\rho \frac{\mathrm{I}\bar{\phi}_i}{\sqrt{\bar{\lambda}_i}}$$

where $\frac{\mathrm{I}\bar{\phi}_i}{\sqrt{\bar{\lambda}_i}}$ *are the (unit norm) eigenfunctions of* L *with eigenvalues* $\bar{\lambda}_i$

Proof of Proposition A.15. First we show that the operators L and $\sum_{i=1}^{\infty} \bar{\lambda}_i \frac{\mathrm{I}\bar{\phi}_i}{\sqrt{\bar{\lambda}_i}} \otimes_\rho \frac{\mathrm{I}\bar{\phi}_i}{\sqrt{\bar{\lambda}_i}}$ agree on any function $g \in L^2(\mathcal{X},\rho)$.

$$\begin{aligned}
\sum_{i=1}^{\infty} \bar{\lambda}_i \frac{\mathrm{I}\bar{\phi}_i}{\sqrt{\bar{\lambda}_i}} \otimes_\rho \frac{\mathrm{I}\bar{\phi}_i}{\sqrt{\bar{\lambda}_i}} g &= \sum_{i=1}^{\infty} \mathrm{I}\bar{\phi}_i \left\langle \mathrm{I}\bar{\phi}_i, g \right\rangle_\rho && \text{(Definition of outer product)} \\
&= \sum_{i=1}^{\infty} \bar{\phi}_i \left\langle \bar{\phi}_i, g \right\rangle_\rho && \text{(Definition A.7)} \\
&= \sum_{i=1}^{\infty} \bar{\phi}_i \int_{\mathcal{X}} \left\langle \bar{\phi}_i, k(\mathrm{x},\cdot) \right\rangle_{\mathcal{H}} g(\mathrm{x},t) d\rho(\mathrm{x},t) && \text{(Reproducing property)} \\
&= \sum_{i=1}^{\infty} \bar{\phi}_i \left\langle \bar{\phi}_i, \int_{\mathcal{X}} k(\mathrm{x},\cdot) g(\mathrm{x},t) d\rho(\mathrm{x},t) \right\rangle_{\mathcal{H}} && \text{(Fubini's Theorem)} \\
&= \int_{\mathcal{X}} k(\mathrm{x},\cdot) g(\mathrm{x},t) d\rho(\mathrm{x},t) && \text{(Pythagoras theorem)} \\
&= \mathrm{L}g
\end{aligned}$$

where the second to last equality holds by Pythagoras theorem, since $\bar{\phi}_i$'s are basis for $\mathcal{H}$. Now we show that $\frac{\mathrm{I}\bar{\phi}_i}{\sqrt{\bar{\lambda}_i}}$ are eigenfunctions of L with eigenvalues $\bar{\lambda}_i$.

$$
\begin{aligned}
\mathrm{L}\frac{\mathrm{I}\bar{\phi}_i}{\sqrt{\bar{\lambda}_i}} &= \mathrm{I}\mathrm{I}^*\frac{\mathrm{I}\bar{\phi}_i}{\sqrt{\bar{\lambda}_i}} && \text{(Proposition A.9)}\\
&= \mathrm{I}\mathrm{C}\frac{\bar{\phi}_i}{\sqrt{\bar{\lambda}_i}} && \text{(Proposition A.9)}\\
&= \bar{\lambda}_i\frac{\mathrm{I}\bar{\phi}_i}{\sqrt{\bar{\lambda}_i}} && (\bar{\phi}_i \text{ is an eigenfunction of C})
\end{aligned}
$$

Moreover, they have unit norms and are mutually orthogonal.

$$
\begin{aligned}
\left\langle \frac{\mathrm{I}\bar{\phi}_i}{\sqrt{\bar{\lambda}_i}}, \frac{\mathrm{I}\bar{\phi}_j}{\sqrt{\bar{\lambda}_j}} \right\rangle_\rho &= \frac{1}{\sqrt{\bar{\lambda}_i\bar{\lambda}_j}}\left\langle \mathrm{I}^*\mathrm{I}\bar{\phi}_i, \bar{\phi}_j\right\rangle_{\mathcal{H}} && \text{(Proposition A.8)}\\
&= \frac{1}{\sqrt{\bar{\lambda}_i\bar{\lambda}_j}}\left\langle \mathrm{C}\bar{\phi}_i, \bar{\phi}_j\right\rangle_{\mathcal{H}} && \text{(Proposition A.9)}\\
&= \frac{1}{\sqrt{\bar{\lambda}_i\bar{\lambda}_j}}\lambda_i\delta_{ij} = \delta_{ij}
\end{aligned}
$$

□

Similarly, we characterize the relationship between the spectral decomposition of the approximate kernel integral operator L_m and the approximate covariance operator C_m in Proposition A.16.

Proposition A.16. *The spectral decomposition of* L_m *is:*

$$
\mathrm{L}_m = \sum_{i=1}^{m} \lambda_i \frac{\mathrm{A}\phi_i}{\sqrt{\lambda_i}} \otimes_\rho \frac{\mathrm{A}\phi_i}{\sqrt{\lambda_i}}
$$

where $\frac{\mathrm{A}\phi_i}{\sqrt{\lambda_i}}$ *are the (unit norm) eigenfunctions of* L_m *with eigenvalues* λ_i

Proof of Proposition A.16. First we prove that L_m and $\sum_{i=1}^m \lambda_i \frac{\mathrm{A}\phi_i}{\sqrt{\lambda_i}} \otimes_\rho \frac{\mathrm{A}\phi_i}{\sqrt{\lambda_i}}$ agree on any function g in $L^2(\mathcal{X}, \rho)$:

$$
\begin{aligned}
\sum_{i=1}^{m} \lambda_i \frac{\mathrm{A}\phi_i}{\sqrt{\lambda_i}} \otimes_\rho \frac{\mathrm{A}\phi_i}{\sqrt{\lambda_i}} g &= \sum_{i=1}^{m} \left\langle A\phi_i, g\right\rangle_\rho A\phi_i && \text{(Definition of outer product)}\\
&= \sum_{i=1}^{m} \int_{\mathcal{X}} \left\langle \mathrm{z}(\mathrm{x},t), \phi_i\right\rangle_{\mathcal{F}} g(\mathrm{x},t) d\rho(\mathrm{x},t) \left\langle \mathrm{z}(\cdot), \phi_i\right\rangle_{\mathcal{F}} && \text{(Definition A.10)}\\
&= \int_{\mathcal{X}} \sum_{i=1}^{m} \left\langle \mathrm{z}(\mathrm{x},t), \phi_i\right\rangle_{\mathcal{F}} \left\langle \mathrm{z}(\cdot), \phi_i\right\rangle_{\mathcal{F}} g(\mathrm{x},t) d\rho(\mathrm{x},t) && \text{(Fubini's Theorem)}\\
&= \int_{\mathcal{X}} \sum_{i=1}^{m} \frac{1}{\sqrt{m}} z_{\omega_i}(\mathrm{x},t) \frac{1}{\sqrt{m}} z_{\omega_i}(\cdot) g(\mathrm{x},t) d\rho(\mathrm{x},t) && \text{(Change of basis)}\\
&= \int_{\mathcal{X}} \left\langle \mathrm{z}(\mathrm{x},t), \mathrm{z}(\cdot)\right\rangle_{\mathcal{F}} g(\mathrm{x},t) d\rho(\mathrm{x},t)\\
&= \int_{\mathcal{X}} k_m(\mathrm{x},\cdot) g(\mathrm{x},t) d\rho(\mathrm{x},t)\\
&= \mathrm{L}_m g
\end{aligned}
$$

The fourth equality follows from the fact that inner products are invariant under orthogonal change of basis. Now we show that $\frac{\mathrm{A}\phi_i}{\sqrt{\lambda_i}}$ are eigenfunctions of L_m with the corresponding eigenvalue λ_i

$$\begin{aligned}
\mathrm{L}_m \frac{\mathrm{A}\phi_i}{\lambda_i} &= \mathrm{A}\mathrm{A}^* \frac{\mathrm{A}\phi_i}{\sqrt{\lambda_i}} && \text{(Proposition A.12)} \\
&= \frac{\mathrm{A}\mathrm{C}_m\phi_i}{\sqrt{\lambda_i}} && \text{(Proposition A.12)} \\
&= \lambda_i \frac{\mathrm{A}\phi_i}{\sqrt{\lambda_i}} && (\phi_i \text{ is an eigenvector of } \mathrm{C}_m)
\end{aligned}$$

Moreover, they have unit norms and are mutually orthogonal.

$$\begin{aligned}
\left\langle \frac{\mathrm{A}\phi_i}{\sqrt{\lambda_i}}, \frac{\mathrm{A}\phi_j}{\sqrt{\lambda_j}} \right\rangle_\rho &= \frac{1}{\sqrt{\lambda_i\lambda_j}} \left\langle \mathrm{A}^*\mathrm{A}\phi_i, \phi_j \right\rangle_{\mathcal{F}} && \text{(Proposition A.11)} \\
&= \frac{1}{\sqrt{\lambda_i\lambda_j}} \left\langle \mathrm{C}_m\phi_i, \phi_j \right\rangle_{\mathcal{F}} && \text{(Proposition A.12)} \\
&= \frac{1}{\sqrt{\lambda_i\lambda_j}} \lambda_i\delta_{ij} = \delta_{ij}
\end{aligned}$$

which completes the proof. □

The following is an important lemma which shows that the kernel integral operator and its approximation can be seen as true and empirical covariance operators in $HS(\rho)$ associated with the random variable z_ω. This allows us to use concentration of measure tools to bound the approximation error in $H(\rho)$.

Lemma A.17. $\mathrm{L} = \mathbb{E}_\omega \left[z_\omega \otimes_\rho z_\omega \right], \mathrm{L}_m = \frac{1}{m}\sum_{i=1}^m z_{\omega_i} \otimes_\rho z_{\omega_i}$

Proof of Lemma A.17. For any $f, g \in L^2(\mathcal{X}, \rho)$ it holds that

$$\begin{aligned}
\langle \mathrm{L}f, g \rangle_\rho &= \left\langle \int_{\mathcal{X}} k(\mathrm{x}, \cdot) f(\mathrm{x}, t) d\rho(\mathrm{x}, t), g \right\rangle_\rho && \text{(Definition A.3)} \\
&= \left\langle \int_{\mathcal{X}} \int_\Omega z_\omega(\mathrm{x}, t) z_\omega(\cdot) f(\mathrm{x}, t) d\pi(\omega) d\rho(\mathrm{x}, t), g \right\rangle_\rho && \text{(Theorem A.6)} \\
&= \left\langle \int_\Omega \int_{\mathcal{X}} z_\omega(\mathrm{x}, t) f(\mathrm{x}, t) d\rho(\mathrm{x}, t) z_\omega(\cdot) d\pi(\omega), g \right\rangle_\rho && \text{(Fubini's theorem)} \\
&= \left\langle \int_\Omega \langle z_\omega(\cdot), f \rangle_\rho z_\omega(\cdot) d\pi(\omega), g \right\rangle_\rho \\
&= \left\langle \int_\Omega z_\omega(\cdot) \otimes_\rho z_\omega(\cdot) d\pi(\omega) f, g \right\rangle_\rho && \text{(Definition of outer product)} \\
&= \langle \mathbb{E}_\omega \left[z_\omega \otimes_\rho z_\omega \right] f, g \rangle_\rho
\end{aligned}$$

Similarly, for any $f, g \in L^2(\mathcal{X}, \rho)$ we have that

$$
\begin{aligned}
\langle \mathrm{L}_m f, g\rangle_\rho &= \left\langle \int_{\mathcal{X}} k_m(\mathrm{x},\cdot) f(\mathrm{x},t) d\rho(\mathrm{x},t), g \right\rangle_\rho && \text{(Definition A.4)} \\
&= \left\langle \int_{\mathcal{X}} \langle \mathrm{z}(\mathrm{x},t), \mathrm{z}(\cdot)\rangle_{\mathcal{F}} f(\mathrm{x},t) d\rho(\mathrm{x},t), g \right\rangle_\rho && \text{(Definition of } k_m\text{)} \\
&= \left\langle \int_{\mathcal{X}} \frac{1}{m}\sum_{i=1}^m z_{\omega_i}(\mathrm{x},t) z_{\omega_i}(\cdot) f(\mathrm{x},t) d\rho(\mathrm{x},t), g \right\rangle_\rho \\
&= \left\langle \frac{1}{m}\sum_{i=1}^m \int_{\mathcal{X}} z_{\omega_i}(\mathrm{x},t) f(\mathrm{x},t) d\rho(\mathrm{x},t) z_{\omega_i}(\cdot), g \right\rangle_\rho \\
&= \left\langle \frac{1}{m}\sum_{i=1}^m \langle z_{\omega_i}(\cdot), f\rangle_\rho z_{\omega_i}(\cdot), g \right\rangle_\rho \\
&= \left\langle \frac{1}{m}\sum_{i=1}^m z_{\omega_i}(\cdot) \otimes_\rho z_{\omega_i}(\cdot) f, g \right\rangle_\rho && \text{(Definition of outer product)} \\
&= \langle \mathrm{L}_m f, g\rangle_\rho && \text{(Definition A.4)}
\end{aligned}
$$

which completes the proof. □

The following Proposition shows the relation between the outer products in two separable Hilbert spaces; this is useful in the proof of the main theorem.

Proposition A.18. *For any Hilbert-Schmidt Operator* $\mathrm{B} : \mathcal{H}_1 \to \mathcal{H}_2$*, it holds that* $\mathrm{B}u \otimes_{\mathcal{H}_2} \mathrm{B}v = \mathrm{B}(u \otimes_{\mathcal{H}_1} v)\mathrm{B}^*$*, where* $\mathrm{u}, \mathrm{v} \in \mathcal{H}_1$.

Proof of Proposition A.18. For any $f \in \mathcal{H}_2$, the following equalities hold:

$$
\begin{aligned}
(\mathrm{B}u \otimes_{\mathcal{H}_2} \mathrm{B}v) f &= \mathrm{B}u \langle \mathrm{B}v, f\rangle_{\mathcal{H}_2} && \text{(Definition of outer product)} \\
&= \mathrm{B}u \langle v, \mathrm{B}^* f\rangle_{\mathcal{H}_1} && \text{(Proposition A.11)} \\
&= \mathrm{B}(u \langle v, \mathrm{B}^* f\rangle_{\mathcal{H}_1}) \\
&= \mathrm{B}(u \otimes_{\mathcal{H}_1} v)\mathrm{B}^* f && \text{(Definition of outer product)}
\end{aligned}
$$

□

Finally, we state the assumptions that we make throughout the paper.

Assumption A.19. The kernel function k is a Mercer's kernel(see Theorem A.5) and has the following integral representation, $k(\mathrm{x},\mathrm{y}) = \int_\Omega z(\mathrm{x},\omega) z(\mathrm{y},\omega) d\pi(\omega)\ \forall \mathrm{x},\mathrm{y} \in \mathcal{X}$ where $(\mathcal{H}, k)$ is a separable RKHS of real-valued functions on $\mathcal{X}$ with a bounded positive definite kernel k. We also assume that there exists $\tau > 1$ such that $|z(\mathrm{x},\omega)| \leq \tau$ for all $x \in \mathcal{X}, \omega \in \Omega$. Furthermore, we assume that the operator $\mathrm{L}^{\frac{1}{2}}$ exists.

B Equivalence of optimization problems in $\mathcal{H}$ and $L^2(\mathcal{X}, \rho)$

We now show that Kernel PCA in the RKHS $\mathcal{H}$ and $L^2(\mathcal{X}, \rho)$ are equivalent under some assumptions which we show are naturally satisfied in the case of Kernel PCA with random features. Ths allows us to transfer our generalization bounds established in $L^2(\mathcal{X}, \rho)$ to $\mathcal{H}$.

The Kernel PCA problem essentially reduces to solving the following optimization problem:

$$\max_{\mathrm{P} \in \mathcal{P}^k_{\mathcal{H}}} \langle \mathrm{P}, \mathrm{C} \rangle_{HS(\mathcal{H})} \tag{OPT-1}$$

For any $\mathrm{P} \in \mathcal{P}^k_{HS(\mathcal{H})}$, by spectral decomposition, P has an eigendecompostion given by $\mathrm{P} = \sum_{i=1}^k \mathrm{u}_i \otimes_\rho \mathrm{u}_i$ where $\mathrm{u}_i \in \mathcal{H}, i \in [k]$ are a set of orthonormal functions. We define an operator $\mathrm{U} : \mathbb{R}^k \to \mathcal{H}$ such that $\mathrm{Ub} = \sum_{i=1}^k \mathrm{b}_i \mathrm{u}_i$, where $\mathrm{b} \in \mathbb{R}^k$.

Proposition B.1. U *satisfies the following properties.*

(a) U *is Hilbert-Schmidt.*

(b) The adjoint of U *is* $\mathrm{U}^* : \mathcal{H} \to \mathbb{R}^k$ *such that* $(\mathrm{U}^* f)_i = \langle \mathrm{u}_i, f \rangle_{\mathcal{H}}$ *where* $f \in \mathcal{H}$.

(c) $\mathrm{P} = \mathrm{U}\mathrm{U}^*$ *and* $\mathrm{U}^*\mathrm{U} = \mathrm{I}_k$

Proof. (a) First we show that the operator U is Hilbert-Schmidt. Let $\{\mathrm{e}_i\}_{i=1}^k$ be the canonical basis of $\mathbb{R}^k$.

$$\begin{aligned} \|\mathrm{U}\|^2_{\mathcal{L}^2(\mathbb{R}^k, \mathcal{H})} &= \sum_{i=1}^k \|\mathrm{U}\mathrm{e}_i\|^2_{\mathcal{H}} && \text{(Pythagoras Theorem)} \\ &= \sum_{i=1}^k \|\mathrm{u}_i\|^2_{\mathcal{H}} = k \end{aligned}$$

(b) Let U^* be the adjoint of U. We now show that $(\mathrm{U}^* f)_i = \langle \mathrm{u}_i, f \rangle_{\mathcal{H}}$. For any $\mathrm{b} \in \mathbb{R}^k, f \in \mathcal{H}$,

$$\begin{aligned} \langle \mathrm{U}^* f, \mathrm{b} \rangle &= \langle f, \mathrm{Ub} \rangle_{\mathcal{H}} \\ &= \left\langle f, \sum_{i=1}^k \mathrm{b}_i \mathrm{u}_i \right\rangle_{\mathcal{H}} \\ &= \sum_{i=1}^k \langle f, \mathrm{u}_i \rangle_{\mathcal{H}} \, \mathrm{b}_i \\ &= \langle \mathrm{d}, \mathrm{b} \rangle \end{aligned}$$

where $\mathrm{d} \in \mathbb{R}^k, \mathrm{d}_i = \langle f, \mathrm{u}_i \rangle_{\mathcal{H}}$

(c) For the first part, for any $f \in \mathcal{H}$, we have,

$$\begin{aligned} \mathrm{P} f &= \sum_{i=1}^k (\mathrm{u}_i \otimes_{\mathcal{H}} \mathrm{u}_i) f \\ &= \sum_{i=1}^k \langle \mathrm{u}_i, f \rangle_{\mathcal{H}} \, \mathrm{u}_i \\ &= \sum_{i=1}^k (\mathrm{U}^* f)_i \mathrm{u}_i \\ &= \mathrm{U}\mathrm{U}^* f \end{aligned}$$

Now we show that the constraint $\mathrm{P} \in \mathcal{P}^k_{\mathcal{H}}$ reduces to $\mathrm{U}^*\mathrm{U} = \mathrm{I}_k$.

For any $\mathrm{b} \in \mathbb{R}^k$,

$$\begin{aligned}\mathrm{U}^*\mathrm{U}\mathrm{b} &= \sum_{i=1}^{k} \mathrm{b}_i \mathrm{U}^* \mathrm{u}_i \\ &= \sum_{i=1}^{k} \mathrm{b}_i \mathrm{d}_i\end{aligned}$$

where $\mathrm{d}_i \in \mathbb{R}^k, (d_i)_j = \langle \mathrm{u}_j, \mathrm{u}_i \rangle_{\mathcal{H}}$. Note that since u_i's are orthonormal functions, therefore, $\mathrm{d}_i = \mathrm{e}_i$, where e_i's is the canonical basis of $\mathbb{R}^k$. Therefore,

$$\mathrm{U}^*\mathrm{U}\mathrm{b} = \sum_{i=1}^{k} \mathrm{b}_i \mathrm{e}_i = \mathrm{b}$$

□

We now write the optimization problem OPT-1 in terms of U as,

$$\max_{\mathrm{U}^*\mathrm{U}=\mathrm{I}_k} \langle \mathrm{U}\mathrm{U}^*, \mathrm{C} \rangle_{HS(\mathcal{H})} \tag{OPT-2}$$

Now consider $\mathrm{v}_i \in L^2(\mathcal{X}, \rho), i \in [k]$ such that $\mathrm{u}_i = \mathrm{I}^*\mathrm{v}_i$. Note that the existence of v_i is guaranteed from the construction of RKHS from eigenfunctions of L (for details, see [Sejdinovic and Gretton, 2012, Theorem 51]). We now define an operator $\mathrm{V} : \mathbb{R}^k \to L^2(\mathcal{X}, \rho)$ such that $\mathrm{V}\mathrm{b} = \sum_{i=1}^{k} \mathrm{b}_i \mathrm{v}_i$, where $\mathrm{b} \in \mathbb{R}^k$. We have the following proposition about V.

Proposition B.2. V *satisfies the following properties,*

(a) V *is Hilbert-Schmidt.*

(b) The adjoint of V *is* $\mathrm{V}^* : L^2(\mathcal{X}, \rho) \to \mathcal{H}$*, defined as* $(V^* f)_i = \langle v_i, f \rangle_\rho$

(c) $\langle \mathrm{V}\mathrm{V}^*, \mathrm{L}^2 \rangle_{HS(\rho)} = \langle \mathrm{U}\mathrm{U}^*, \mathrm{C} \rangle_{HS(\mathcal{H})}$ *and* $\mathrm{V}^*\mathrm{L}\mathrm{V} = \mathrm{U}^*\mathrm{U} = \mathrm{I}_k$

Proof. The proofs of *(a)* and *(b)* are similar to that of Proposition B.1.
(c). We start with the first part. The objective in terms of V is,

$$\begin{aligned}
\langle \mathrm{P}, \mathrm{C} \rangle_{HS(\mathcal{H})} &= \left\langle \sum_{i=1}^{k} \mathrm{u}_i \otimes_{\mathcal{H}} \mathrm{u}_i, \mathrm{C} \right\rangle_{HS(\mathcal{H})} \\
&= \left\langle \sum_{i=1}^{k} \mathrm{I}^*\mathrm{v}_i \otimes_{\mathcal{H}} \mathrm{I}^*\mathrm{v}_i, \mathrm{C} \right\rangle_{HS(\mathcal{H})} && (\mathrm{u}_i = \mathrm{I}^*\mathrm{v}_i) \\
&= \left\langle \sum_{i=1}^{k} \mathrm{I}(\mathrm{v}_i \otimes_\rho \mathrm{v}_i)\mathrm{I}^*, \mathrm{C} \right\rangle_{HS(\mathcal{H})} && \text{(Proposition A.18)} \\
&= \left\langle \sum_{i=1}^{k} \mathrm{v}_i \otimes_\rho \mathrm{v}_i, \mathrm{I}\mathrm{C}\mathrm{I}^* \right\rangle_{HS(\rho)} && \text{(Definition of adjoint)} \\
&= \left\langle \sum_{i=1}^{k} \mathrm{v}_i \otimes_\rho \mathrm{v}_i, \mathrm{I}\mathrm{I}^*\mathrm{I}\mathrm{I}^* \right\rangle_{HS(\rho)} && \text{(Proposition A.9)} \\
&= \left\langle \sum_{i=1}^{k} \mathrm{v}_i \otimes_\rho \mathrm{v}_i, \mathrm{L}^2 \right\rangle_{HS(\rho)} && \text{(Proposition A.9)} \\
&= \langle \mathrm{V}\mathrm{V}^*, \mathrm{L}^2 \rangle_{HS(\rho)}
\end{aligned}$$

For the second part, for any $\mathrm{b} \in \mathbb{R}^k$, we have,

$$\begin{aligned}
\mathrm{U}^*\mathrm{U}\mathrm{b} &= \sum_{i=1}^{k} \mathrm{b}_i \mathrm{U}^*\mathrm{u}_i \\
&= \sum_{i=1}^{k} \mathrm{b}_i \mathrm{U}^*\mathrm{I}^*\mathrm{v}_i && (\mathrm{u}_i = \mathrm{I}^*\mathrm{v}_i) \\
&= \sum_{i=1}^{k} \mathrm{b}_i \mathrm{d}_i
\end{aligned}$$

where $\mathrm{d}_i \in \mathbb{R}^k, (\mathrm{d}_i)_j = \langle \mathrm{u}_j, \mathrm{I}^*\mathrm{v}_i \rangle_{\mathcal{H}} = \langle \mathrm{I}^*\mathrm{v}_j, \mathrm{I}^*\mathrm{v}_i \rangle_{\mathcal{H}} = \langle \mathrm{v}_j, \mathrm{II}^*\mathrm{v}_i \rangle_{\rho} = \langle \mathrm{v}_j, \mathrm{L}\mathrm{v}_i \rangle_{\rho}$, where the third equality follows from the property of adjoints, and the last equality because $\mathrm{L} = \mathrm{II}^*$.
Since $\mathrm{U}^*\mathrm{U}\mathrm{b} = \mathrm{b}$, so $\mathrm{d}_i = \mathrm{e}_i$. Therefore, we get $\langle \mathrm{v}_j, \mathrm{L}\mathrm{v}_i \rangle = \delta_{ij}$

Let us now look at the j^{th} element of $\mathrm{V}^*\mathrm{L}\mathrm{V}\mathrm{b}$,

$$\begin{aligned}
(\mathrm{V}^*\mathrm{L}\mathrm{V}\mathrm{b})_j &= (\mathrm{V}^*\mathrm{L} \sum_{i=1}^{k} \mathrm{b}_i \mathrm{v}_i)_j && \text{(Definition of V)} \\
&= \sum_{i=1}^{k} (\mathrm{b}_i \mathrm{V}^*\mathrm{L}\mathrm{v}_i)_j \\
&= \sum_{i=1}^{k} \mathrm{b}_i \langle \mathrm{v}_j, \mathrm{L}\mathrm{v}_i \rangle_{\rho} \\
&= \sum_{i=1}^{k} \mathrm{b}_i \delta_{ij} = \mathrm{b}_j
\end{aligned}$$

□

We can now restate the optimization problem in terms of V.

$$\max_{\mathrm{V}^*\mathrm{L}\mathrm{V} = \mathrm{I}_k} \left\langle \mathrm{V}\mathrm{V}^*, \mathrm{L}^2 \right\rangle_{HS(\rho)} \tag{OPT-3}$$

Now, let $\mathrm{w}_i = \mathrm{L}^{1/2}\mathrm{v}_i$. Note that w_i is well-defined since we assume that $\mathrm{L}^{1/2}$ exists (See Assumption 3.1). Define $\mathrm{W} : \mathbb{R}^k \to L^2(\mathcal{X}, \rho)$, such that $\mathrm{W}\mathrm{b} = \sum_{i=1}^{k} \mathrm{b}_i \mathrm{w}_i$.

Proposition B.3. W *satisfies the following properties,*

(a) W *is Hilbert-Schmidt.*

(b) The adjoint of W *is* $\mathrm{W}^* : L^2(\mathcal{X}, \rho) \to \mathbb{R}^k$ $(\mathrm{W}^* f)_i = \langle \mathrm{w}_i, f \rangle_{\rho}$.

(c) $\mathrm{W} = \mathrm{L}^{1/2}\mathrm{V}$, $\left\langle \mathrm{V}\mathrm{V}^*, \mathrm{L}^2 \right\rangle_{HS(\rho)} = \langle \mathrm{W}\mathrm{W}^*, \mathrm{L} \rangle_{HS(\rho)}$ *and* $\mathrm{W}^*\mathrm{W} = \mathrm{V}^*\mathrm{L}\mathrm{V} = \mathrm{I}_k$

Proof. The proofs of *(a)* and *(b)* are similar to that of Proposition B.1.
(c) For the first part, for any $\mathrm{b} \in \mathbb{R}^k$, we have

$$\begin{aligned}
\mathrm{W}\mathrm{b} &= \sum_{i=1}^{k} \mathrm{b}_i \mathrm{w}_i \\
&= \sum_{i=1}^{k} \mathrm{b}_i \mathrm{L}^{1/2}\mathrm{v}_i \\
&= \mathrm{L}^{1/2} \sum_{i=1}^{k} \mathrm{b}_i \mathrm{v}_i \\
&= \mathrm{L}^{1/2}\mathrm{V}\mathrm{b}
\end{aligned}$$

Note that since L is self-ajoint, $\mathrm{L}^{1/2}$ is self-adjoint too. The objective in terms of W is

$$\begin{aligned}\left\langle \mathrm{VV}^*, \mathrm{L}^2 \right\rangle_{HS(\rho)} &= \left\langle \mathrm{L}^{1/2}\mathrm{VV}^*\mathrm{L}^{1/2}, \mathrm{L} \right\rangle_{HS(\rho)} && \text{(Definition of adjoint)} \\ &= \left\langle \mathrm{WW}^*, \mathrm{L} \right\rangle_{HS(\rho)}\end{aligned}$$

Equivalently, we can restate the constraint in terms of W as,

$$\begin{aligned}\mathrm{V}^*\mathrm{LV} &= \mathrm{V}^*\mathrm{L}^{1/2}\mathrm{L}^{1/2}\mathrm{V} \\ &= (\mathrm{L}^{1/2}\mathrm{V})^*(\mathrm{L}^{1/2}\mathrm{V}) \\ &= \mathrm{W}^*\mathrm{W} = \mathrm{I}_k\end{aligned}$$

□

We now restate the optimization problem in terms of W.

$$\max_{\mathrm{W}^*\mathrm{W}=\mathrm{I}_k} \left\langle \mathrm{WW}^*, \mathrm{L} \right\rangle_{HS(\rho)} \tag{OPT-4}$$

We now state this equivalence of objective in the following Lemma.

Lemma B.4 (Equivalence of Objective)**.**

$$\left\langle \mathrm{P}, \mathrm{C} \right\rangle_{HS(\mathcal{H})} = \left\langle \mathrm{WW}^*, \mathrm{L} \right\rangle_{HS(\rho)}$$

where the relation between P *and* W *is presented via Propositions B.1,B.2 and B.3.*

Proof. One direction of implication simply simply follows from the construction in Propositions B.1, B.2 and B.3. In particular, from Propositions B.1, B.2 and B.3, we conclude that OPT-1 $\implies$ OPT-2 $\implies$ OPT-3 $\implies$ OPT-4. It is easy to see that OPT-3 $\implies$ OPT-2 $\implies$ OPT-1 where the first implication simply follows from the construction of u_i's and the second from Proposition B.1. However, showing that OPT-4 $\implies$ OPT-3 is conditioned on w_i's lying in the range of $\mathrm{L}^{1/2}$ because otherwise there might not exist v_i's such that $\mathrm{w}_i = \mathrm{L}^{1/2}\mathrm{v}_i$. In Lemma B.5, we show that when using random features, with the approximation operator defined in Definition 3.6, the functions obtained via random feature approximation lies in the range of $\mathrm{L}^{1/2}$ with probability 1 on the support of π. This establishes the equivalence claimed. □

We now formally show that vectors from $\mathcal{F}$ lifted to $L^2(\mathcal{X}, \rho)$ via the approximation operator A lie in the range of $\mathrm{L}^{1/2}$ almost surely with respect to measure π. The proof of the following lemma closely follows [Rudi and Rosasco, 2017, Lemma 2].

Lemma B.5. *For every* $\mathrm{v} \in \mathcal{F}$, $\mathrm{Av} \in L^2(\mathcal{X}, \rho)$ *lies in the range of* $\mathrm{L}^{1/2}$ *almost surely on the support of* π.

Proof. Let $\Pi \in HS(\rho)$ denote the projection operator projecting to the range of $\mathrm{L}^{1/2}$. Then $(\mathrm{I}_\rho - \Pi)\mathrm{L}^{1/2} f = 0 \; \forall \; f \in L^2(\mathcal{X}, \rho)$ as $(\mathrm{I}_\rho - \Pi)$ is the projection to the orthogonal complement to the range of $\mathrm{L}^{1/2}$. From this, we have, $\mathrm{Tr}\left((\mathrm{I}_\rho - \Pi)\mathrm{L}^{1/2}\mathrm{L}^{1/2}(\mathrm{I}_\rho - \Pi)\right) = \mathrm{Tr}\left((\mathrm{I}_\rho - \Pi)\mathrm{L}(\mathrm{I}_\rho - \Pi)\right) = 0$.

$$\begin{aligned}\mathrm{Tr}\left((\mathrm{I}_\rho - \Pi)\mathrm{L}(\mathrm{I}_\rho - \Pi)\right) &= \mathrm{Tr}\left((\mathrm{I}_\rho - \Pi)\int_\Omega z_\omega \otimes_\rho z_\omega d\pi(\omega)(\mathrm{I}_\rho - \Pi)\right) \\ &= \int_\Omega \mathrm{Tr}\left((\mathrm{I}_\rho - \Pi)(z_\omega \otimes_\rho z_\omega)(\mathrm{I}_\rho - \Pi)\right) d\pi(\omega) \\ &= \int_\Omega \mathrm{Tr}\left((\mathrm{I}_\rho - \Pi)z_\omega \otimes_\rho (\mathrm{I}_\rho - \Pi)z_\omega\right) d\pi(\omega) \\ &= \int_\Omega \|(\mathrm{I}_\rho - \Pi)z_\omega\|_\rho^2 \, d\pi(\omega) = 0\end{aligned}$$

From the above equation, we see that $\|(\mathrm{I}_\rho - \Pi)z_\omega\|_\rho = 0$ almost surely on the support of π. This implies that $(\mathrm{I}_\rho - \Pi)z_\omega = 0$ a.s. on the support of π.

Now we show that all functions of interest i.e. anything lifted from $\mathcal{F}$ to $L^2(\mathcal{X}, \rho)$ lie in the range of $\mathrm{L}^{1/2}$. Let $\mathrm{v} \in \mathcal{F}, f \in L^2(\mathcal{X}, \rho)$.

$$\begin{aligned}\langle(\mathrm{I}_\rho - \Pi)f, \mathrm{A}\mathrm{v}\rangle_\rho &= \langle \mathrm{A}^*(\mathrm{I}_\rho - \Pi)f, \mathrm{v}\rangle_{\mathcal{F}} \\ &= \sum_{i=1}^{m}((\mathrm{I}_\rho - \Pi)z_{\omega_i} f)\mathrm{v}_i = 0\end{aligned}$$

This is because ω_i's are drawn from π, and we already argued that $(\mathrm{I}_\rho - \Pi)z_\omega = 0$ a.s. on the support of π. Since this holds for any $\mathrm{v} \in \mathcal{F}$ and $f \in L^2(\mathcal{X}, \rho)$, this implies that $\mathrm{A}\mathrm{v}$ lies in the range of $\mathrm{L}^{1/2}$ for all $\mathrm{v} \in \mathcal{F}$. □

Moreover, note that since $\mathrm{L}_m = \mathfrak{A}\mathrm{C}_m$ (See Proposition A.16), the eigenfunctions of L_m are the lifted eigenvectors of C_m from $\mathcal{F}$ to $L^2(\mathcal{X}, \rho)$. This equivalence entails that any candidate solution of OPT-2 has an equivalent candidate solution for OPT-4, and they would both have the same objective.

As already hinted, the solution of Kernel PCA with random features might not lie in the constraint set of rank k projection operators over $L^2(\mathcal{X}, \rho)$. By the equivalence of the optimization problems OPT-1 and OPT-4, we violate the constraint in $\mathcal{H}$ as well. We, however, remarked that we counter this problem by showing a fast $O(1/\sqrt{n})$ speed of convergence to the constraint set in $L^2(\mathcal{X}, \rho)$. A natural question to ask is does the speed of convergence to the constraint set preserved too? We answer affirmatively as shown below.

We now use this equivalence to give a reduction from a candidate solution OPT-4 to OPT-2. Let $\tilde{\mathrm{P}} = \sum_{i=1}^{k} \tilde{\mathrm{p}}_i \otimes_\rho \tilde{\mathrm{p}}_i$ be the output of some algorithm for Kernel PCA with random features, lifted through the approximation operator $\mathfrak{A}$. We have show in Theorem C.1 that $\mathfrak{A}\mathrm{P}^k_{\mathrm{C}_m} = \mathrm{P}^k_{\mathrm{L}_m}$ is a rank k projection over $L^2(\mathcal{X}, \rho)$. Let $\mathrm{P}_{\mathrm{L}_m} = \sum_{i=1}^{k} \tilde{\mathrm{q}}_i \otimes \tilde{\mathrm{q}}_i$. Since $\tilde{\mathrm{p}}_i$ and $\tilde{\mathrm{q}}_i$ lie in the range of $\mathrm{L}^{1/2}$, $\forall\, i \in [k]$ (See Lemma B.5), there exists p_i's and q_i's such that $\tilde{\mathrm{p}}_i = \mathrm{L}^{1/2}\mathrm{p}_i$ and $\tilde{\mathrm{q}}_i = \mathrm{L}^{1/2}\mathrm{q}_i$, $i \in [k]$. Define $\mathrm{P} := \sum_{i=1}^{k} \mathrm{I}^*\mathrm{p}_i \otimes_{\mathcal{H}} \mathrm{I}^*\mathrm{p}_i$, and $\mathrm{Q} := \sum_{i=1}^{k} \mathrm{I}^*\mathrm{q}_i \otimes_{\mathcal{H}} \mathrm{I}^*\mathrm{q}_i$.

First we show that Q is a projection operator in $HS(\mathcal{H})$.

$$\begin{aligned}\langle \mathrm{I}^*\mathrm{q}_i, \mathrm{I}^*\mathrm{q}_j\rangle_{\mathcal{H}} &= \langle \mathrm{q}_i, \Pi\mathrm{I}^*\mathrm{q}_j\rangle_\rho && \text{(Definition of adjoints)} \\ &= \langle \mathrm{q}_i, \mathrm{L}\mathrm{q}_j\rangle_\rho && \text{(Proposition A.9)} \\ &= \left\langle \mathrm{L}^{1/2}\mathrm{q}_i, \mathrm{L}^{1/2}\mathrm{q}_j\right\rangle_\rho && \text{(Definition of adjoints)} \\ &= \langle \tilde{\mathrm{q}}_i, \tilde{\mathrm{q}}_j\rangle_\rho = \delta_{ij}\end{aligned}$$

Now, let us look at the rate of convergence P to $\mathcal{P}^k_{HS(\mathcal{H})}$.

Lemma B.6 (Equivalence of convergence to the constraint set)**.**

$$d(\bar{\mathrm{P}}, \mathcal{P}^k_{HS(\mathcal{H})}) \leq \left\|\tilde{\mathrm{P}} - \mathfrak{A}\mathrm{C}_m\right\|_{HS(\rho)}$$

Proof.

$$
\begin{aligned}
d(\bar{\mathrm{P}}, \mathcal{P}^k_{HS(\mathcal{H})}) &\leq \|\mathrm{P}-\mathrm{Q}\|_{HS(\mathcal{H})} \\
&= \left\| \sum_{i=1}^{k} \mathrm{I}^* \mathrm{p}_i \otimes_{\mathcal{H}} \mathrm{I}^* \mathrm{p}_i - \mathrm{I}^* \mathrm{q}_i \otimes_{\mathcal{H}} \mathrm{I}^* \mathrm{q}_i \right\|_{HS(\mathcal{H})} \\
&= \left\| \mathrm{I}(\sum_{i=1}^{k} \mathrm{p}_i \otimes_{\rho} \mathrm{p}_i - \mathrm{q}_i \otimes_{\rho} \mathrm{q}_i)\mathrm{I}^* \right\|_{HS(\mathcal{H})} && \text{(Proposition A.18)} \\
&= \left\| \mathrm{L}(\sum_{i=1}^{k} \mathrm{p}_i \otimes_{\rho} \mathrm{p}_i - \mathrm{q}_i \otimes_{\rho} \mathrm{q}_i) \right\|_{HS(\rho)} \\
&= \left\| \mathrm{L}^{1/2}(\sum_{i=1}^{k} \mathrm{p}_i \otimes_{\rho} \mathrm{p}_i - \mathrm{q}_i \otimes_{\rho} \mathrm{q}_i)\mathrm{L}^{1/2} \right\|_{HS(\rho)} && \text{(Cyclic property)} \\
&= \left\| \sum_{i=1}^{k} \mathrm{L}^{1/2}\mathrm{p}_i \otimes_{\rho} \mathrm{L}^{1/2}\mathrm{p}_i - \mathrm{L}^{1/2}\mathrm{q}_i \otimes_{\rho} \mathrm{L}^{1/2}\mathrm{q}_i \right\|_{HS(\rho)} && \text{(Proposition A.18)} \\
&= \left\| \sum_{i=1}^{k} \tilde{\mathrm{p}}_i \otimes_{\rho} \tilde{\mathrm{p}}_i - \tilde{\mathrm{q}}_i \otimes_{\rho} \tilde{\mathrm{q}}_i \right\|_{HS(\rho)} \\
&= \left\| \tilde{\mathrm{P}} - \mathfrak{A}\mathrm{C}_m \right\|_{HS(\rho)}
\end{aligned}
$$

□

In Lemma C.5, we will bound $\left\| \tilde{\mathrm{P}} - \mathfrak{A}\mathrm{C}_m \right\|_{HS(\rho)}$ which implies the bound given in the main theorem 4.2.

We now combine the above relations into a definition to lift operators from $HS(\mathcal{F})$ to $HS(\mathcal{H})$ and then discuss that the operator is well-defined.

Definition B.7 (Operator $\mathfrak{L}$)**.** Let $\tilde{\mathrm{P}} \in HS(\mathcal{F})$ and $\mathfrak{A}\tilde{\mathrm{P}} = \sum_{i=1}^{k} \tilde{\mathrm{p}}_i \otimes_{\rho} \tilde{\mathrm{p}}_i$ be $\tilde{\mathrm{P}}$ lifted to $L^2(\mathcal{X}, \rho)$. Consider the equivalence relation $\mathrm{p}_i \sim \mathrm{p}_j$ if $\mathrm{L}^{1/2}\mathrm{p}_i = \mathrm{L}^{1/2}\mathrm{p}_j$. Let $[\mathrm{p}_i]$ be the equivalence class such that $\mathrm{L}^{1/2}\mathrm{p}_i = \tilde{\mathrm{p}}_i$. The operator $\mathfrak{L} : HS(\mathcal{F}) \to HS(\mathcal{H})$ is defined as,

$$
\mathfrak{L}\widehat{P} = \sum_{i=1}^{k} \mathrm{I}^* \mathrm{p}_i \otimes_{\mathcal{H}} \mathrm{I}^* \mathrm{p}_i
$$

Here I^* is the restriction of the operator I^* to the quotient space $L^2(\mathcal{X}, \rho)/\sim$.

We now discuss that the operator $\mathfrak{L}$ is indeed a well defined operator. We guarantee by Lemma B.5 that there is at least one element in $[\mathrm{p}_i]$ such that $\tilde{\mathrm{p}}_i = \mathrm{L}^{1/2}\mathrm{p}_i$. It remains to argue that all the elements in the equivalence class $[\mathrm{p}_i]$ are being mapped to the same element in $\mathcal{H}$ through I^*. Let p_i and p_j be two elements of $[\mathrm{p}_i]$. Since $\mathrm{L}^{1/2}\mathrm{p}_i = \mathrm{L}^{1/2}\mathrm{p}_j$, therefore $\mathrm{L}^{1/2}(\mathrm{p}_i - \mathrm{p}_j) = 0$. This implies that $\mathrm{p}_j = \mathrm{p}_i + \mathrm{Ker}(\mathrm{L}^{1/2})$. Note that any $\mathrm{r}_i \in \mathrm{Ker}(\mathrm{L}^{1/2})$ will be mapped by I^* to 0, i.e. $\mathrm{I}^*\mathrm{r}_i = 0$. It follows from linearity of I^* that $\mathrm{I}^*\tilde{\mathrm{p}}_j = \mathrm{I}^*\mathrm{p}_i$. Thus this maps an equivalence class to a single element in $\mathcal{H}$.

C Proof of the main Theorem

From we have already established the problems in $HS(\rho)$ and $HS(\mathcal{H})$, we focus on error decomposition and bounding the corresponding error terms in $L^2(\mathcal{X}, \rho)$. Solving KPCA by using a kernel approximation, one needs to consider two different sources of error. First, the error coming from approximating the true kernel operator by random features. Second, the statistical error due to estimating the covariance operator using iid samples from the unknown distribution. Thus, we distinguish between and base our proof around these two sources of error, namely *approximation error* and *estimation error*. In particular we decompose our objective as:

$$\left\langle \mathfrak{I}\mathrm{P}^k_{\mathrm{C}}, \mathfrak{I}\mathrm{C} \right\rangle_{HS(\rho)} - \left\langle \mathfrak{A}\widehat{\mathrm{P}}, \mathfrak{I}\mathrm{C} \right\rangle_{HS(\rho)} = \underbrace{\langle \mathfrak{I}\mathrm{P}^k_{\mathrm{C}}, \mathfrak{I}\mathrm{C} \rangle_{HS(\rho)} - \langle \mathfrak{A}\mathrm{P}^k_{\mathrm{C}_m}, \mathfrak{I}\mathrm{C} \rangle_{HS(\rho)}}_{\epsilon_a\text{: Approximation Error}}$$
$$+ \underbrace{\langle \mathfrak{A}\mathrm{P}^k_{\mathrm{C}_m}, \mathfrak{I}\mathrm{C} \rangle_{HS(\rho)} - \langle \mathfrak{A}\widehat{\mathrm{P}}, \mathfrak{I}\mathrm{C} \rangle_{HS(\rho)}}_{\epsilon_e\text{: Estimation Error}}.$$

The first term in the decomposition is interpreted as approximation error because it essentially captures the error incurred by approximating the kernel function with random features. The second term in the decomposition is interpreted as estimation error as it is the error incurred in the original statistical estimation problem. In what follows, we give a bound on each of the error terms and provide a detailed analysis. Throughout this section, we use the following Lemma that shows the relation between different projection operators.

Lemma C.1. *$\mathfrak{I}\mathrm{P}^k_{\mathrm{C}}$ and $\mathfrak{A}\mathrm{P}^k_{\mathrm{L}}$ are rank k projection operators in $L^2(\mathcal{X}, \rho)$. Furthermore, it holds that $\mathfrak{I}\mathrm{P}^k_{\mathrm{C}} = \mathrm{P}^k_{\mathrm{L}}$ and $\mathfrak{A}\mathrm{P}^k_{\mathrm{C}_m} = \mathrm{P}^k_{\mathrm{L}_m}$.*

Proof of Lemma C.1. We have

$$\mathfrak{I}\mathrm{P}^k_{\mathrm{C}} = \sum_{i=1}^{k} \frac{\mathrm{I}\bar{\phi}_i}{\sqrt{\bar{\lambda}_i}} \otimes_\rho \frac{\mathrm{I}\bar{\phi}_i}{\sqrt{\bar{\lambda}_i}} = \mathrm{P}^k_{\mathrm{L}}$$

where the second inequality follows from Lemma A.9. Similarly,

$$\mathfrak{A}\mathrm{P}^k_{\mathrm{C}_m} = \sum_{i=1}^{k} \frac{\mathrm{A}\phi_i}{\sqrt{\lambda_i}} \otimes_\rho \frac{\mathrm{A}\phi_i}{\sqrt{\lambda_i}} = \mathrm{P}^k_{\mathrm{L}_m}$$

where the second inequality follows from Lemma A.12. and $\mathfrak{A}\mathrm{P}^k_{\mathrm{C}}$ □

C.1 Approximation Error

The main idea behind controlling the approximation error is to use the local Rademacher complexity of the kernel class Massart [2000], Bartlett et al. [2002]. More precisely, we use the following result in [Blanchard et al., 2007], which allows us to get rates depending both on the number of features used and the decay of the spectrum of the operator C_2.

Theorem C.2 (Blanchard et al. [2007])**.** *Assume $\|\zeta\|^2 \leq M$ almost surely, and let (λ_i) denote the ordered eigenvalues of $\mathrm{C} := \mathrm{E}[\zeta\zeta^\top]$, and further assume that (λ_i) are distinct. Let $B_k := \frac{\sqrt{\mathrm{E}[\langle \zeta, \zeta' \rangle^4]}}{\lambda_k - \lambda_{k+1}}$, where ζ' is and iid copy of ζ. Then for all δ, with overwhelming probability of at least $1 - e^{-\delta}$ it holds that*

$$\langle P^k_{\widehat{\mathrm{C}}^\perp}, \mathrm{C} \rangle - \langle P^k_{\mathrm{C}^\perp}, \mathrm{C} \rangle \leq 24\kappa(B_k, k, n) + \frac{11\delta(M + B_k)}{n}$$

where κ is defined as follows:

$$\kappa(B_k, k, n) = \inf_{h \geq 0} \left\{ \frac{B_k h}{n} + \sqrt{\frac{k}{n} \sum_{j>h} \lambda_i(\mathrm{C}')} \right\}$$

Lemma C.3 (Approximation Error)**.** *With probability at least $1 - \frac{\delta}{2}$, we have*

$$\langle \mathrm{P}^k_{\mathrm{L}^\perp_m}, \mathrm{L} \rangle - \langle \mathrm{P}^k_{\mathrm{L}^\perp}, \mathrm{L} \rangle \leq 24\kappa(B_k, k, m) + \frac{11 \log(\delta/2)\, \tau^2 + 7B_k}{m}$$

Proof. We first note that $\mathfrak{A}P^k_{\mathrm{C}_m} = \sum_{i=1}^k \frac{\mathrm{A}\phi_i}{\sqrt{\lambda_i}} \otimes \frac{\mathrm{A}\phi_i}{\sqrt{\lambda_i}}$ (see definition of the approximation operator in A.10). The following holds for the approximation error:

$$\begin{aligned}
\epsilon_a &= \left\langle \mathfrak{I}P^k_{\mathrm{C}}, \mathfrak{I}\mathrm{C} \right\rangle_{HS(\rho)} - \left\langle \mathfrak{A}P^k_{\mathrm{C}_m}, \mathfrak{I}\mathrm{C} \right\rangle_{HS(\rho)} && (5)\\
&= \left\langle \sum_{i=1}^k \frac{\mathrm{I}\bar{\phi}_i}{\sqrt{\bar{\lambda}_i}} \otimes_\rho \frac{\mathrm{I}\bar{\phi}_i}{\sqrt{\bar{\lambda}_i}}, \sum_{i\in I\subset\mathbb{R}} \bar{\lambda}_i \left(\frac{\mathrm{I}\bar{\phi}_i}{\sqrt{\bar{\lambda}_i}} \otimes_\rho \frac{\mathrm{I}\bar{\phi}_i}{\sqrt{\bar{\lambda}_i}} \right) \right\rangle_{HS(\rho)} && \text{(definition A.7)}\\
&- \left\langle \sum_{i=1}^k \frac{\mathrm{A}\phi_i}{\sqrt{\lambda_i}} \otimes_\rho \frac{\mathrm{A}\phi_i}{\sqrt{\lambda_i}}, \sum_{i\in I\subset\mathbb{R}} \bar{\lambda}_i \left(\frac{\mathrm{I}\bar{\phi}_i}{\sqrt{\bar{\lambda}_i}} \otimes_\rho \frac{\mathrm{I}\bar{\phi}_i}{\sqrt{\bar{\lambda}_i}} \right) \right\rangle_{HS(\rho)} && \text{(definition A.7, A.10)}\\
&= \langle P^k_{\mathrm{L}}, \mathrm{L} \rangle_{HS(\rho)} - \langle P^k_{\mathrm{L}_m}, \mathrm{L} \rangle_{HS(\rho)} && \text{(Lemma A.15 and Lemma C.1)}\\
&= \langle P^k_{\mathrm{L}^\perp_m}, \mathrm{L} \rangle_{HS(\rho)} - \langle P^k_{\mathrm{L}^\perp}, \mathrm{L} \rangle_{HS(\rho)} && \text{(properties of the orthogonal subspace)}
\end{aligned}$$

We have already showed that L and L_m in the right hand side of the equation (5) are true and empirical covariance operators respectively (see Lemma A.17). As required by Theorem C.2, we need to show that norm of the random variables z_ω are bounded. We have

$$\begin{aligned}
\|z_\omega\|^2 &= \langle z_\omega, z_\omega \rangle_\rho \\
&= \int_{\mathcal{X}} z_\omega(\mathrm{x},t)^2 d\rho(\mathrm{x},t) \\
&\leq \tau^2
\end{aligned}$$

where the last inequality follows Assumption 3.1.

Invoking Theorem C.2, we have with probability at least $1-\delta$,

$$\langle P^k_{\mathrm{L}^\perp_m}, \mathrm{L} \rangle - \langle P^k_{\mathrm{L}^\perp}, \mathrm{L} \rangle \leq 24\kappa(B_k, k, m) + \frac{11\log(\delta)\tau^2 + 7B_k}{m}$$

where $\kappa(B_k,k,m) = \inf_{h\geq 0}\left\{ \frac{B_k h}{m} + \sqrt{\frac{k\sum_{j>h}\lambda_j(C'_2)}{m}} \right\}$. □

Lemma C.4 (Approximation Error - Good decay)**.** *When the spectrum of operator C'_2 has an exponential decay, i.e. $\lambda_j(\mathrm{C}'_2) = \alpha^j$ for some $\alpha < 1$, then with probability at least $1-\delta$, we have*

$$\langle \mathrm{P}^k_{\mathrm{L}^\perp_m}, \mathrm{L} \rangle - \langle \mathrm{P}^k_{\mathrm{L}^\perp}, \mathrm{L} \rangle \leq \frac{24B_k\log(m)}{\log(1/\alpha)\, m} + \frac{k + (1-\alpha)(11\log(\delta)\tau^2 + 7B_k)}{(1-\alpha)m}$$

Proof. When $\lambda_i(\mathrm{C}'_2)$ have an exponential decay, i.e $\lambda_j(\mathrm{C}'_2) = \alpha^j$ for some $\alpha < 1$, we have

$$\sum_{j>h} \lambda_j(\mathrm{C}'_2) = \frac{\alpha^{h+1}}{1-\alpha}$$

Set $h = \lceil -\log_\alpha(m) \rceil - 1$, we get

$$\sum_{j>h} \lambda_j(\mathrm{C}'_2) \leq \frac{1}{(1-\alpha)m}$$

Now,

$$\begin{aligned}
\kappa(B_k,k,m) &= \inf_{h\geq 0}\left\{ \frac{B_k h}{m} + \sqrt{\frac{k\sum_{j>h}\lambda_j(C'_2)}{m}} \right\} \\
&\leq \frac{-B_k\log_\alpha m}{m} + \frac{k}{(1-\alpha)m} \\
&= \frac{B_k\log(m)}{\log(1/\alpha)\, m} + \frac{k}{(1-\alpha)m}
\end{aligned}$$

where the last equality follows from the identity $\log_b a = \frac{\log_d(a)}{\log_d(b)}$

So essentially, $\kappa(B_k, k, m) = O\left(\frac{\log(m)}{m}\right)$. Therefore, we get

$$\begin{aligned}\epsilon_a = \left\langle \mathfrak{I}P^k_{\mathrm{C}}, \mathfrak{I}\mathrm{C}\right\rangle_{HS(\rho)} - \left\langle \mathfrak{A}P^k_{\mathrm{C}_m}, \mathfrak{I}\mathrm{C}\right\rangle_{HS(\rho)} &\leq \frac{24B_k \log(m)}{\log(1/\alpha)\, m} + \frac{k}{(1-\alpha)m} + \frac{11\log(\delta)\,\tau^2 + 7B_k}{m} \\ &= \frac{24B_k \log(m)}{\log(1/\alpha)\, m} + \frac{k + (1-\alpha)(11\log(\delta)\,\tau^2 + 7B_k)}{(1-\alpha)m}\end{aligned}$$

which completes the proof. □

C.2 Estimation Error

We first remind the reader that $\mathfrak{A}P^k_{C_m}$ is a projection operator in $L^2(\mathcal{X}, \rho)$ (See Lemma C.1). However, the problem we face is that $\mathfrak{A}\widehat{P}$ might not be a projection operator in $L^2(\mathcal{X}, \rho)$. This is because the lifting is accomplished by lifting a particular set of eigenvectors of $\widehat{\mathrm{P}}_{\mathcal{A}}$ through A, and we remark that A doesn't necessarily preserve norms and angles between elements. To get around this predicament, we show that lifted operator converges to a projection operator, i.e the lifted set of eigenvectors go to an orthogonal set of functions in $L^2(\mathcal{X}, \rho)$. Moreover, from Lemma B.6, we have that this convergence in $HS(\rho)$ is equivalent to convergence in $HS(\mathcal{H})$.

Lemma C.5. *When the number of samples* $n \geq \frac{2\lambda_1^2 q_{\mathcal{A}}(1/\delta, \log(m), \log(n))^2}{\lambda_k^2(\sqrt{2}-1)}$*, with probability at least* $1 - \frac{\delta}{2}$*, we have*

$$\begin{aligned}d(\mathfrak{A}\mathrm{C}_m, \mathcal{P}^k_{HS(\rho)}) \leq \left\|\mathfrak{A}P^k_{\mathrm{C}_m} - \mathfrak{A}\widehat{P}_{\mathcal{A}}\right\|_{HS(\rho)} &\leq \left\|\mathfrak{A}P^k_{\mathrm{C}_m} - \mathfrak{A}\widehat{P}_{\mathcal{A}}\right\|_{\mathcal{L}^1(\rho)} \\ &\leq \frac{\lambda_1}{(\sqrt{2}-1)}\sqrt{\sum_{i=1}^{k}\left(\frac{2\lambda_i + 4\lambda_1}{\lambda_i^2}\right)^2}\frac{q_{\mathcal{A}}(1/\delta, \log(m), \log(n))}{n}\end{aligned}$$

Proof of Lemma C.5. Since $\mathfrak{A}\mathrm{C}_m$ is a rank k projection operator in $HS(\rho)$ (from Lemma C.1), the first inequality follows trivially. The second inequality is just from the property of norms that schatten norms $\|\mathrm{D}\|_{\mathcal{L}^p(\rho)}$ decreases with increasing p. We focus on proving the third inequality below.

Let $\widehat{\mathrm{P}}_{\mathcal{A}} = \tilde{\Phi}\tilde{\Phi}^\top$ be an eigendecomposition of the output $\widehat{\mathrm{P}}_{\mathcal{A}}$. Let

$$\mathrm{R}^* = \underset{\mathrm{R}^\top\mathrm{R} = \mathrm{R}\mathrm{R}^\top = \mathrm{I}}{\arg\min} \left\|\tilde{\Phi}\mathrm{R} - \Phi_k\right\|_F^2$$

where Φ_k is the matrix corresponding top top k eigenvectors of C_m. Define $\widehat{\Phi} := \tilde{\Phi}\mathrm{R}^*$. This means that we rotate the eigenvectors of our output to a basis such that it is closest to the truth (in element-wise metric sense). An important point on why we can do this is that this rotation (or any other rotation for that matter) doesn't change the output, i.e. $\widehat{\Phi}\widehat{\Phi}^\top = \tilde{\Phi}\mathrm{R}^*\mathrm{R}^{*\top}\tilde{\Phi}^\top = \tilde{\Phi}\tilde{\Phi}^\top = \widehat{\mathrm{P}}_{\mathcal{A}}$. We now lifting the output by lifting this rotated set of eigenvectors. We have, $\mathfrak{A}\widehat{\mathrm{P}} = \sum_{i=1}^k \frac{\mathrm{A}\widehat{\phi}_i}{\sqrt{\widehat{\lambda}_i}} \otimes_\rho \frac{\mathrm{A}\widehat{\phi}_i}{\sqrt{\widehat{\lambda}_i}}$, where $\widehat{\lambda}_i := \left\langle \widehat{\phi}_i, \mathrm{C}_m\widehat{\phi}_i\right\rangle_{\mathcal{F}}$.

$$
\begin{aligned}
\left\|\mathfrak{A}P^k_{C_m} - \mathfrak{A}\widehat{P}_{\mathcal{A}}\right\|_{\mathcal{L}^1(\rho)} &= \left\|\sum_{i=1}^k \frac{\mathrm{A}\phi_i}{\sqrt{\lambda_i}} \otimes \frac{\mathrm{A}\phi_i}{\sqrt{\lambda_i}} - \sum_{i=1}^k \frac{\mathrm{A}\widehat{\phi}_i}{\sqrt{\widehat{\lambda}_i}} \otimes \frac{\mathrm{A}\widehat{\phi}_i}{\sqrt{\widehat{\lambda}_i}}\right\|_{\mathcal{L}^1(\rho)} \\
&= \left\|\mathrm{A}\sum_{i=1}^k \left(\frac{\phi_i}{\sqrt{\lambda_i}} \otimes \frac{\phi_i}{\sqrt{\lambda_i}} - \frac{\widehat{\phi}_i}{\sqrt{\widehat{\lambda}_i}} \otimes \frac{\widehat{\phi}_i}{\sqrt{\widehat{\lambda}_i}}\right)\mathrm{A}^*\right\|_{\mathcal{L}^1(\rho)} \\
&\leq \|\mathrm{A}\| \left\|\sum_{i=1}^k \left(\frac{1}{\lambda_i}\phi_i \otimes \phi_i - \frac{1}{\widehat{\lambda}_i}\widehat{\phi}_i \otimes \widehat{\phi}_i\right)\mathrm{A}^*\right\|_{\mathcal{L}^1(\mathcal{F},\rho)} \\
&\leq \|\mathrm{A}\| \, \|\mathrm{A}^*\| \left\|\sum_{i=1}^k \left(\frac{1}{\lambda_i}\phi_i \otimes \phi_i - \frac{1}{\widehat{\lambda}_i}\widehat{\phi}_i \otimes \widehat{\phi}_i\right)\right\|_{\mathcal{L}^1(\mathcal{F})} \\
&\leq \lambda_1 \left\|\sum_{i=1}^k \left(\frac{1}{\lambda_i}\phi_i \otimes \phi_i - \frac{1}{\widehat{\lambda}_i}\widehat{\phi}_i \otimes \widehat{\phi}_i\right)\right\|_{\mathcal{L}^1(\mathcal{F})}
\end{aligned}
$$

Where third and fourth inequalities follows from the fact that for trace-class operators $\|AB\|_{\mathcal{L}^1} \leq \|A\|_2 \, \|B\|_{\mathcal{L}^1}$. See [Reed and Simon, 1972, Exercise 28, Page 218].

Adding and subtracting $\frac{1}{\lambda_i}\widehat{\phi}_i \otimes \widehat{\phi}_i$ inside the summation to get

$$
\begin{aligned}
&\leq \lambda_1 \left\|\sum_{i=1}^k \frac{1}{\lambda_i}\phi_i \otimes \phi_i - \frac{1}{\lambda_i}\widehat{\phi}_i \otimes \widehat{\phi}_i + \frac{1}{\lambda_i}\widehat{\phi}_i \otimes \widehat{\phi}_i - \frac{1}{\widehat{\lambda}_i}\widehat{\phi}_i \otimes \widehat{\phi}_i\right\|_{\mathcal{L}^1(\mathcal{F})} \\
&\leq \lambda_1 \left\|\sum_{i=1}^k \left(\frac{1}{\lambda_i}\left(\phi_i \otimes \phi_i - \widehat{\phi}_i \otimes \widehat{\phi}_i\right) + \left(\frac{1}{\lambda_i} - \frac{1}{\widehat{\lambda}_i}\right)\widehat{\phi}_i \otimes \widehat{\phi}_i\right)\right\|_{\mathcal{L}^1(\mathcal{F})} \\
&\leq \lambda_1 \sum_{i=1}^k \frac{1}{\lambda_i}\left\|\phi_i \otimes \phi_i - \widehat{\phi}_i \otimes \widehat{\phi}_i\right\|_{\mathcal{L}^1(\mathcal{F})} + \left|\frac{1}{\lambda_i} - \frac{1}{\widehat{\lambda}_i}\right| \\
&\leq \lambda_1 \sum_{i=1}^k \frac{1}{\lambda_i}\left\|\phi_i \otimes \phi_i - \widehat{\phi}_i \otimes \widehat{\phi}_i\right\|_{\mathcal{L}^1(\mathcal{F})} + \left|\frac{\lambda_i - \widehat{\lambda}_i}{\lambda_i\widehat{\lambda}_i}\right| \\
&\leq \lambda_1 \sum_{i=1}^k \frac{2}{\lambda_i}\left\|\phi_i - \widehat{\phi}_i\right\|_2 + \frac{4\lambda_1}{\lambda_i^2}\left\|\phi_i - \widehat{\phi}_i\right\|_2 \\
&\leq \lambda_1 \sum_{i=1}^k \left(\frac{2\lambda_i + 4\lambda_1}{\lambda_i^2}\right)\left\|\phi_i - \widehat{\phi}_i\right\|_2 \\
&\leq \lambda_1 \sqrt{\sum_{i=1}^k \left(\frac{2\lambda_i + 4\lambda_1}{\lambda_i^2}\right)^2}\left\|\Phi_k - \widehat{\Phi}\right\|_{\mathcal{F}} \\
&\leq \frac{\lambda_1}{2(\sqrt{2}-1)}\sqrt{\sum_{i=1}^k \left(\frac{2\lambda_i + 4\lambda_1}{\lambda_i^2}\right)^2}\left\|\mathrm{P}^k_{\mathrm{C}_m} - \widehat{\mathrm{P}}\right\|_F^2 \\
&\leq \frac{\lambda_1}{(\sqrt{2}-1)}\sqrt{\sum_{i=1}^k \left(\frac{2\lambda_i + 4\lambda_1}{\lambda_i^2}\right)^2}\frac{q_{\mathcal{A}}(1/\delta, \log(m), \log(n))}{n}
\end{aligned}
$$

The second to last inequality follows from Lemma C.6 and Lemma C.7. □

Lemma C.6. $\|\phi_i \otimes \phi_i - \widehat{\phi}_i \otimes \widehat{\phi}_i\|_{\mathcal{L}^1(\mathcal{F})} \leq 2\|\phi_i - \widehat{\phi}_i\|_2 \; \forall \; i \in [k]$

Proof.

$$\begin{aligned}\|\phi_i \otimes \phi_i - \widehat{\phi}_i \otimes \widehat{\phi}_i\|_{\mathcal{L}^1(\mathcal{F})} &= \|\phi_i \otimes \phi_i - \widehat{\phi}_i \otimes \phi_i + \widehat{\phi}_i \otimes \phi_i - \widehat{\phi}_i \otimes \widehat{\phi}_i\|_{\mathcal{L}^1(\mathcal{F})} \\ &\leq \|(\phi_i - \widehat{\phi}_i) \otimes \phi_i\|_{\mathcal{L}^1(\mathcal{F})} + \|\widehat{\phi}_i \otimes (\phi_i - \widehat{\phi}_i)\|_{\mathcal{L}^1(\mathcal{F})} \\ &= 2\|\phi_i - \widehat{\phi}_i\|_2\end{aligned}$$

□

Lemma C.7. *When the number of samples* $n \geq \frac{2q_{\mathcal{A}}(1/\delta, \log(m), \log(n))^2 \lambda_1^2}{\lambda_i^2(\sqrt{2}-1)}$*, with probability at least* $1-\delta$*,* $\forall\, i \in [k]$ *we have,*

$$\left|\frac{\lambda_i - \widehat{\lambda}_i}{\lambda_i \widehat{\lambda}_i}\right| \leq \frac{4\lambda_1}{\lambda_i^2} \left\|\phi_i - \widehat{\phi}_i\right\|_2$$

where $C_{\mathcal{A}}$ *is a constant specific to the algorithm* $\mathcal{A}$*.*

The numerator is bounded as follows

Proof.

$$\begin{aligned}|\lambda_i - \widehat{\lambda}_i| &= |\phi_i^\top \mathrm{C}_m \phi_i - \widehat{\phi}_i^\top \mathrm{C}_m \widehat{\phi}_i| \\ &= |\phi_i^\top \mathrm{C}_m \phi_i - \phi_i^\top \mathrm{C}_m \widehat{\phi}_i + \phi_i^\top \mathrm{C}_m \widehat{\phi}_i - \widehat{\phi}_i^\top \mathrm{C}_m \widehat{\phi}_i| \\ &= |\phi_i^\top \mathrm{C}_m (\phi_i - \widehat{\phi}_i) + (\phi_i - \widehat{\phi}_i)^\top \mathrm{C}_m \widehat{\phi}_i| \\ &\leq \|\mathrm{C}_m \phi_i\|_2 \|\phi_i - \widehat{\phi}_i\|_2 + \|\phi_i - \widehat{\phi}_i\|_2 \|\mathrm{C}_m \widehat{\phi}_i\|_2 \\ &= (\lambda_i + \widehat{\lambda}_i)\|\phi_i - \widehat{\phi}_i\|_2 \\ &\leq (\lambda_i + \lambda_1)\|\phi_i - \widehat{\phi}_i\|_2 \\ &\leq 2\lambda_1 \|\phi_i - \widehat{\phi}_i\|\end{aligned}$$

where the second inequality holds since $\widehat{\lambda}_i < \lambda_1$ by definition of $\widehat{\lambda}_i$, and the last inequality follows because $\widehat{\lambda}_i \leq \lambda_1$.
The denominator is lower bounded similarly as

$$\begin{aligned}\lambda_i \widehat{\lambda}_i &\geq \lambda_i(\lambda_i - 2\lambda_1 \|\phi_i - \widehat{\phi}_i\|) \\ &\geq \frac{\lambda_i^2}{2}\end{aligned}$$

where the first inequality follows from the bound on the numerator and the last inequality follows when $2\lambda_1 \left\|\phi_i - \widehat{\phi}_i\right\| \leq \frac{\lambda_i}{2}$. From Lemma C.10, we know that $\left\|\phi_i - \widehat{\phi}_i\right\|_2^2 \leq \frac{1}{2(\sqrt{2}-1)} \left(\frac{q_{\mathcal{A}}(1/\delta, \log(m), \log(n))}{n}\right)$ with probability at least $1-\delta$. Combining, we get, with probability at least $1-\delta$,

$$\left\|\phi_i - \widehat{\phi}_i\right\|_2 \leq \sqrt{\frac{1}{2(\sqrt{2}-1)} \left(\frac{q_{\mathcal{A}}(1/\delta, \log(m), \log(n))}{n}\right)} \leq \frac{\lambda_i}{4\lambda_1}$$

The above holds when the number of samples $n \geq \frac{2\lambda_1^2 q_{\mathcal{A}}(1/\delta, \log(m), \log(n))^2}{\lambda_i^2(\sqrt{2}-1)}$. Combining, we get

$$\left|\frac{\lambda_i - \widehat{\lambda}_i}{\lambda_i \widehat{\lambda}_i}\right| \leq \frac{4\lambda_1}{\lambda_i^2} \left\|\phi_i - \widehat{\phi}_i\right\|_2$$

with probability at least $1-\delta$ and when $n \geq \dfrac{2\lambda_1^2 q_{\mathcal{A}}(1/\delta, \log(m), \log(n))^2}{\lambda_i^2(\sqrt{2}-1)}$. □

Note that in particular since Oja's algorithm has a warm-up phase, the lower bound on the denominator

Lemma C.8. *For rank k orthogonal matrices $\mathrm{U} \in \mathbb{R}^{m\times k}$ and $\mathrm{V} \in \mathbb{R}^{m\times k}$, i.e. $\mathrm{U}^\top\mathrm{U} = \mathrm{V}^\top\mathrm{V} = \mathrm{I}_k$, the following holds,*

$$\left\|\mathrm{U} - \mathrm{V}\widehat{\mathrm{R}}\right\|_F^2 \leq \frac{1}{2(\sqrt{2}-1)}\left\|\mathrm{U}\mathrm{U}^\top - \mathrm{V}\mathrm{V}^\top\right\|_F^2,$$

where

$$\widehat{\mathrm{R}} = \underset{\mathrm{R}^\top\mathrm{R}=\mathrm{R}\mathrm{R}^\top=\mathrm{I}_k}{\arg\min} \|\mathrm{U} - \mathrm{V}\mathrm{R}\|_F^2$$

Proof. Proof in [Ge et al., 2017, Lemma 6]. □

Since Φ_k and $\widehat{\Phi}$ are rank k orthogonal matrices, from Lemma C.8, we have

$$\left\|\Phi_k - \widehat{\Phi}\right\|_F^2 \leq \frac{1}{2(\sqrt{2}-1)}\left\|\mathrm{P}_{\mathrm{C}_m}^k - \widehat{\mathrm{P}}\right\|_F^2$$

Lemma C.9. *For any efficient subspace learner $\mathcal{A}$, we have $\left\|\mathrm{P}_{\mathrm{C}_m}^k - \widehat{\mathrm{P}}\right\|_F^2 \leq \frac{2q_{\mathcal{A}}(1/\delta,\log(m),\log(n))}{n}$ with probability at least $1 - \frac{\delta}{2}$.*

Proof.

$$\begin{aligned}
\left\|\mathrm{P}_{\mathrm{C}_m}^k - \widehat{\mathrm{P}}_{\mathcal{A}}\right\|_F^2 &= \left\|\mathrm{P}_{\mathrm{C}_m}^k\right\|_F^2 + \left\|\widehat{\mathrm{P}}\right\|_F^2 - 2\left\langle\widehat{\mathrm{P}}, \mathrm{P}_{\mathrm{C}_m}^k\right\rangle \\
&= 2\left(k - \left\langle\widehat{\mathrm{P}}, \mathrm{P}_{\mathrm{C}_m}^k\right\rangle\right) \\
&= 2\left(\left\langle\mathrm{I} - \mathrm{P}_{\mathrm{C}_m}^k, \widehat{\mathrm{P}}\right\rangle\right) \\
&= 2\left\|\left(\Phi_k^\perp\right)^\top\widehat{\Phi}\right\|_F^2 \\
&\leq \frac{2q_{\mathcal{A}}(1/\delta, \log(m), \log(n))}{n}
\end{aligned}$$

where the last inequality follows from the definition of efficient subspace learner. □

Lemma C.10. *With probability at least $1 - \delta$,*

$$\left\|\phi_i - \widehat{\phi}_i\right\|_2 \leq \frac{1}{2(\sqrt{2}-1)}\left(\frac{q_{\mathcal{A}}(1/\delta, \log(m), \log(n))}{n}\right)$$

where $q_{\mathcal{A}}(1/\delta, \log(m), \log(n))$ is specific to the algorithm $\mathcal{A}$.

Proof.

$$\begin{aligned}
\left\|\phi_i - \widehat{\phi}_i\right\|_2^2 &\leq \sum_{i=1}^k \left\|\phi_i - \widehat{\phi}_i\right\|_2^2 \\
&= \left\|\Phi_k - \widehat{\Phi}\right\|_F^2 \\
&\leq \frac{1}{2(\sqrt{2}-1)}\left\|\mathrm{P}_{\mathrm{C}_m}^k - \widehat{\mathrm{P}}_{\mathcal{A}}\right\|_F^2 \\
&\leq \frac{1}{2(\sqrt{2}-1)}\left(\frac{q_{\mathcal{A}}(1/\delta, \log(m), \log(n))}{n}\right)
\end{aligned}$$

□

where the second inequality holds from Lemma C.8 and the definition of $\widehat{\mathrm{P}}$, and the last inequality holds from Lemma C.9

Lemma C.11 (Estimation Error). *When the number of samples* $n \geq \dfrac{2\lambda_1^2 q_{\mathcal{A}}(1/\delta, \log(m), \log(n))^2}{\lambda_k^2(\sqrt{2}-1)}$, *then with probability at least* $1-\delta$, *we have*

$$\epsilon_e \leq \frac{\lambda_1^2}{(\sqrt{2}-1)} \sqrt{\sum_{i=1}^{k} \left(\frac{2\lambda_i + 4\lambda_1}{\lambda_i^2}\right)^2} \frac{q_{\mathcal{A}}(1/\delta, \log(m), \log(n))^2}{n}$$

Proof.

$$\begin{aligned}
\epsilon_e &= \langle \mathfrak{A} P^k_{\mathrm{C}_m}, \mathfrak{IC} \rangle_{HS(\rho)} - \langle \mathfrak{A}\widehat{P}_{\mathcal{A}}, \mathfrak{IC} \rangle_{HS(\rho)} \\
&= \left\langle \mathfrak{A} P^k_{\mathrm{C}_m} - \mathfrak{A}\widehat{P}_{\mathcal{A}}, \mathfrak{IC} \right\rangle_{HS(\rho)} \\
&\leq \left\| \mathfrak{A} P^k_{\mathrm{C}_m} - \mathfrak{A}\widehat{P}_{\mathcal{A}} \right\|_{\mathcal{L}^1(\rho)} \|\mathfrak{IC}\|_2 \\
&\leq \lambda_1 \left\| \mathfrak{A} P^k_{\mathrm{C}_m} - \mathfrak{A}\widehat{P}_{\mathcal{A}} \right\|_{\mathcal{L}^1(\rho)} \\
&\leq \frac{\lambda_1^2}{(\sqrt{2}-1)} \sqrt{\sum_{i=1}^{k} \left(\frac{2\lambda_i + 4\lambda_1}{\lambda_i^2}\right)^2} \frac{q_{\mathcal{A}}(1/\delta, \log(m), \log(n))}{n}
\end{aligned}$$

The last inequality follows from Lemma C.5. □

We now invoke the approximation and the estimation error bounds i.e. Lemma C.3 and Lemma C.11 with failure probabilities $\delta/2$ each. We then apply a union bound over them and get that with probability at least $1-\delta$,

$$\langle \mathfrak{I}\mathrm{P}^k_{\mathrm{C}}, \mathfrak{IC} \rangle_\rho - \langle \mathfrak{A}\widehat{\mathrm{P}}_{\mathcal{A}}, \mathfrak{IC} \rangle_\rho \leq \frac{cB_k}{\sqrt{n}} + \frac{c'(k + \log(\delta/2) + 7B_k)}{\sqrt{n}\log(n)} + \sqrt{\frac{q_{\mathcal{A}}(2/\delta, \log(m), \log(n))}{n}},$$

This concludes the proof of the main theorem.

Also note that since $d(\mathfrak{A}\widehat{\mathrm{P}}_{\mathcal{A}}, \mathcal{P}_{HS(\rho)})$ decays as $O(1/\sqrt{n})$, we can bound the suboptimality of $\mathfrak{A}\widehat{\mathrm{P}}_{\mathcal{A}}$ projected onto the set of projection operators $\mathcal{P}^k_{HS(\rho)}$. It is now easy to give a bound on the objective with respect to the projection $\tilde{\mathrm{P}}_{\mathcal{A}} \in \mathcal{P}_{HS(\rho)}$ of $\mathfrak{A}\widehat{\mathrm{P}}_{\mathcal{A}}$ onto the set of projection operators:

Corollary C.12. *Let* $\tilde{\mathrm{P}}_{\mathcal{A}}$ *be the projection of* $\mathfrak{A}\widehat{\mathrm{P}}_{\mathcal{A}}$ *onto the set* $\mathcal{P}_{HS(\rho)}$. *Under the same conditions as in theorem 4.2, we have*

$$\langle \mathfrak{I}\mathrm{P}^k_{\mathrm{C}}, \mathfrak{IC} \rangle_\rho - \langle \tilde{\mathrm{P}}_{\mathcal{A}}, \mathfrak{IC} \rangle_\rho \leq 24\kappa(B_k, k, m) + \frac{\log(\delta/2) + 7B_k}{m} + 2\sqrt{\frac{q_{\mathcal{A}}(2/\delta, \log(m), \log(n))}{n}}$$

Proof.

$$\begin{aligned}
\langle \mathfrak{I}\mathrm{P}^k_{\mathrm{C}}, \mathfrak{IC} \rangle_\rho - \langle \tilde{\mathrm{P}}_{\mathcal{A}}, \mathfrak{IC} \rangle_\rho &= \langle \mathfrak{I}\mathrm{P}^k_{\mathrm{C}}, \mathfrak{IC} \rangle_\rho - \langle \mathfrak{A}\widehat{\mathrm{P}}_{\mathcal{A}}, \mathfrak{IC} \rangle_\rho + \langle \mathfrak{A}\widehat{\mathrm{P}}_{\mathcal{A}} - \tilde{\mathrm{P}}_{\mathcal{A}}, \mathfrak{IC} \rangle \\
&\leq 24\kappa(B_k, k, m) + \frac{\log(\delta/2) + 7B_k}{m} + \sqrt{\frac{q_{\mathcal{A}}(2/\delta, \log(m), \log(n))}{n}} \\
&\quad + d\left(\mathfrak{A}\widehat{\mathrm{P}}_{\mathcal{A}}, \mathcal{P}_{HS(\rho)}\right) \|\mathfrak{IC}\|_{HS(\rho)} \\
&\leq 24\kappa(B_k, k, m) + \frac{\log(\delta/2) + 7B_k}{m} + 2\sqrt{\frac{q_{\mathcal{A}}(2/\delta, \log(m), \log(n))}{n}},
\end{aligned}$$

where the second to last inequality follows from Cauchy-Schwartz in $HS(\rho)$. □

We now give the proof of Corollary 4.3.

Proof of Corollary 4.3.

$$\begin{aligned}\langle \mathfrak{I}\mathrm{P}^k_{\mathrm{C}}, \mathfrak{I}\mathrm{C}\rangle_\rho - \langle \mathfrak{A}\widehat{\mathrm{P}}_{\mathcal{A}}, \mathfrak{I}\mathrm{C}\rangle_\rho &= \langle \mathfrak{I}\mathrm{P}^k_{\mathrm{C}}, \mathfrak{I}\mathrm{C}\rangle_{HS(\rho)} - \langle \mathfrak{A}\mathrm{P}^k_{\mathrm{C}_m}, \mathfrak{I}\mathrm{C}\rangle_{HS(\rho)} \\ &\quad + \langle \mathfrak{A}\mathrm{P}^k_{\mathrm{C}_m}, \mathfrak{I}\mathrm{C}\rangle_{HS(\rho)} - \langle \mathfrak{A}\widehat{\mathrm{P}}, \mathfrak{I}\mathrm{C}\rangle_{HS(\rho)} \\ &\leq \frac{24B_k \log(m)}{\log(1/\alpha)\, m} + \frac{k + (1-\alpha)(11\log(\delta/2)\, M + 7B_k)}{(1-\alpha)m} \\ &\quad + \lambda_1 \sqrt{\frac{q_{\mathcal{A}}(2/\delta, \log(m), \log(n))}{n}},\end{aligned}$$

with probability at least $1-\delta$. The last inequality follows from Lemma C.4 and Lemma C.11 with a union bound over them. □

D Examples of ESL

In this section, we instantiate our framework with two popular learning algorithms, Empirical Risk Minimization (ERM) and Oja's Algorithm, and show that they satisfy the requirements of ESL.

D.1 Empirical Risk Minimizer

A natural candidate for an efficient subspace learner is the Empirical Risk Minimizer, which we call as $\mathcal{A}_{ERM}$. We first show that $\mathcal{A}_{ERM}$ satisfies the sufficient condition of Definition 4.1 and then show that $\mathcal{A}_{ERM}$ is an efficient subspace learner. We then discuss its computational aspects. Let $\{\mathrm{x}_i\}_{i=1}^n$ be n data samples and $\{\mathrm{z}(\mathrm{x})\}_{i=1}^n$ be the corresponding representations in $\mathcal{F}$. The empirical covariance matrix in $\mathcal{F}$ is defined as

$$\widehat{\mathrm{C}}_m = \frac{1}{n}\sum_{i=1}^{n} \mathrm{z}(\mathrm{x}_i)\mathrm{z}(\mathrm{x}_i)^\top$$

The algorithm $\mathcal{A}_{ERM}$ computes the top k eigenvectors of $\widehat{\mathrm{C}}_m$, and returns a rank k orthogonal matrix say $\widehat{\Phi}$. Let the corresponding projection matrix be $\widehat{\mathrm{P}}_{ERM}$. We first state the bound on covariance matrices C_m and $\widehat{\mathrm{C}}_m$.

Lemma D.1 (Covariance Estimation)**.** *With probability at least* $1-\delta$*,*

$$\left\|\widehat{\mathrm{C}}_m - \mathrm{C}_m\right\|_2 \le \frac{\kappa}{3n}\log\left(\frac{\delta}{2m}\right) + \sqrt{\frac{\kappa}{3n}\log\left(\frac{\delta}{2m}\right)^2 + \log\left(\frac{\delta}{2m}\right)\frac{\kappa\lambda_1}{n}}$$

Proof.

$$\begin{aligned}
\left\|\widehat{\mathrm{C}}_m - \mathrm{C}_m\right\|_2 &= \left\|\frac{1}{n}\sum_{i=1}^{n}\mathrm{z}(\mathrm{x}_i)\mathrm{z}(\mathrm{x}_i)^\top - \mathrm{C}_m\right\|_2 \\
&= \left\|\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{z}(\mathrm{x}_i)\mathrm{z}(\mathrm{x}_i)^\top - \mathrm{C}_m\right)\right\|_2 \\
&= \left\|\sum_{i=1}^{n}\Xi_i\right\|_2
\end{aligned}$$

where $\Xi_i = \frac{1}{n}\left(\mathrm{z}(\mathrm{x}_i)\mathrm{z}(\mathrm{x}_i)^\top - \mathrm{C}_m\right)$. Ξ_i's are 0 mean random matrices, i.e. $\mathbb{E}\left[\Xi_i\right] = 0\ \forall\ i \in [n]$. Note that $\|\mathrm{z}(\mathrm{x})\|_2^2 = \int_{\mathcal{X}} \mathrm{z}_\omega(\mathrm{x})^2 \mathrm{d}\rho(\mathrm{x}) \le \tau^2$, since $\mathrm{z}_\omega(\mathrm{x}) \le \tau\ \forall\ \omega \in \Omega, \mathrm{x} \in \mathcal{X}$ by Assumption 3.1. We have,

$$\begin{aligned}
\|\Xi_i\|_2 &\le \frac{1}{n}\left(\left\|\mathrm{z}(\mathrm{x}_i)\mathrm{z}(\mathrm{x}_i)^\top\right\|_2 + \|\mathrm{C}_m\|_2\right) \\
&= \frac{1}{n}\left(\mathrm{Tr}\left(\mathrm{z}(\mathrm{x}_i)\mathrm{z}(\mathrm{x}_i)^\top\right) + \left\|\mathbb{E}_{\mathrm{x}}\left[\mathrm{z}(\mathrm{x})\mathrm{z}(\mathrm{x})^\top\right]\right\|_2\right) \\
&\le \frac{1}{n}\left(\|\mathrm{z}(\mathrm{x}_i)\|^2 + \mathbb{E}_{\mathrm{x}}\left[\left\|\mathrm{z}(\mathrm{x})\mathrm{z}(\mathrm{x})^\top\right\|\right]_2\right) \\
&\le \frac{1}{n}\left(\|\mathrm{z}(\mathrm{x}_i)\|^2 + \mathbb{E}_{\mathrm{x}}\left[\|\mathrm{z}(\mathrm{x})\|_2^2\right]\right) \\
&\le \frac{2\tau^2}{n}
\end{aligned}$$

so that $L(\Xi) := \max_i\{\|\Xi_i\|_2\} \leq \frac{2\tau^2}{n}$. where in the second inequality, we apply Jensen's inequality. Define $v(\Xi) := \left\|\sum_{i=1}^n \mathbb{E}\left[\Xi_i\Xi_i^\top\right]\right\|_2$. We have,

$$
\begin{aligned}
v(\Xi) &= \left\|\sum_{i=1}^n \mathbb{E}\left[\Xi_i\Xi_i^\top\right]\right\|_2 \\
&= \left\|\sum_{i=1}^n \mathbb{E}\left[\frac{1}{n^2}\left(\mathrm{z}(\mathrm{x}_i)\mathrm{z}(\mathrm{x}_i)^\top - \mathrm{C}_m\right)\left(\mathrm{z}(\mathrm{x}_i)\mathrm{z}(\mathrm{x}_i)^\top - \mathrm{C}_m\right)^\top\right]\right\|_2 \\
&= \frac{1}{n^2}\left\|\sum_{i=1}^n \mathbb{E}\left[\|\mathrm{z}(\mathrm{x}_i)\|^2\,\mathrm{z}(\mathrm{x}_i)^\top\mathrm{z}(\mathrm{x}_i) - \mathrm{z}(\mathrm{x}_i)^\top\mathrm{z}(\mathrm{x}_i)\mathrm{C}_m - \mathrm{C}_m\mathrm{z}(\mathrm{x}_i)^\top\mathrm{z}(\mathrm{x}_i) + \mathrm{C}_m^2\right]\right\|_2 \\
&\leq \frac{1}{n^2}\left\|\sum_{i=1}^n \mathbb{E}\left[\tau^2\mathrm{z}(\mathrm{x}_i)^\top\mathrm{z}(\mathrm{x}_i) - \mathrm{z}(\mathrm{x}_i)^\top\mathrm{z}(\mathrm{x}_i)\mathrm{C}_m - \mathrm{C}_m\mathrm{z}(\mathrm{x}_i)^\top\mathrm{z}(\mathrm{x}_i) + \mathrm{C}_m^2\right]\right\|_2 \\
&= \frac{1}{n^2}\left\|\sum_{i=1}^n \tau^2\mathrm{C}_m - \mathrm{C}_m^2 - \mathrm{C}_m^2 + \mathrm{C}_m^2\right\|_2 \\
&= \frac{1}{n}\left\|\tau^2\mathrm{C}_m - \mathrm{C}_m^2\right\|_2 \\
&\leq \frac{\tau^2}{n}\left\|\mathrm{C}_m\right\|_2 \\
&\leq \frac{\tau^2\lambda_1}{n}
\end{aligned}
$$

where the second last inequality holds because C_m is a positive semi-definite matrix. From matrix Bernstein concentration (Lemma F.2, restated from [Tropp et al., 2015]), we have, with probability at least $1-\delta$

$$
\begin{aligned}
\left\|\widehat{\mathrm{C}}_m - \mathrm{C}_m\right\|_2 = \left\|\sum_1^m \Xi_i\right\|_2 &\leq \frac{L(\Xi)}{6}\log\left(\frac{\delta}{2m}\right) + \sqrt{\frac{L(\Xi)^2}{12}\log\left(\frac{\delta}{2m}\right)^2 + \log\left(\frac{\delta}{2m}\right)v(\Xi)} \\
&\leq \frac{\tau^2}{3n}\log\left(\frac{\delta}{2m}\right) + \sqrt{\frac{\tau^2}{3n}\log\left(\frac{\delta}{2m}\right)^2 + \log\left(\frac{\delta}{2m}\right)\frac{\tau^2\lambda_1}{n}}
\end{aligned}
$$

□

In the following lemma, we show that $\mathcal{A}_{ERM}$ is an efficient subspace learner.

Lemma D.2. *$\mathcal{A}_{ERM}$ is an effcient subspace learner.*

Proof. We invoke Theorem F.1 with the sub-multiplicative norm being the spectral norm. With $\mathrm{A} = \mathrm{C}_m, \mathrm{B} = \widehat{\mathrm{C}}_m, \mathrm{U} = \widehat{\Phi}, \mathrm{V} = \Phi_k^\perp$. Let $\epsilon = \left\|\mathrm{C}_m - \widehat{\mathrm{C}}_m\right\|_2$. .From Weyl's inequality, we have $\lambda_k(\widehat{\mathrm{C}}_m) \geq \lambda_k - \epsilon = \lambda_{k+1} + \mathrm{gap} - \epsilon \geq \lambda_{k+1}$, if $\epsilon < \mathrm{gap}$. Therefore, setting $\mu = \lambda_{k+1}$, and

$\alpha = \text{gap} = \lambda_k - \lambda_{k+1}$, then with probability $1-\delta$, we get

$$\begin{aligned}
\left\|(\Phi_k^{\perp})^\top \widehat{\Phi}\right\|_F^2 &\leq k \left\|(\Phi_k^{\perp})^\top \widehat{\Phi}\right\|_2^2 \leq \frac{k\epsilon}{\alpha^2} \\
&\leq \frac{k}{\alpha^2}\left(\frac{\tau}{3n}\log\left(\frac{\delta}{2m}\right) + \sqrt{\frac{\tau}{3n}\log\left(\frac{\delta}{2m}\right)^2 + \log\left(\frac{\delta}{2m}\right)\frac{\tau\lambda_1}{n}}\right)^2 \\
&\leq \frac{k}{\alpha^2}\left(\frac{\tau}{3n}\log\left(\frac{\delta}{2m}\right) + \sqrt{\frac{2\lambda_1\tau}{n}\log\left(\frac{\delta}{2m}\right)}\right)^2 \\
&\leq \frac{k}{\alpha^2}\left(\frac{\lambda_1\tau^2}{n}\log\left(\frac{\delta}{2m}\right)^2\right)
\end{aligned}$$

Setting $q_{ERM}(1/\delta, \log(m), \log(n)) = \frac{\lambda_1\tau^2}{\alpha^2}\log\left(\frac{\delta}{2m}\right)^2 = \frac{k\lambda_1\tau^2}{(\lambda_k-\lambda_{k+1})^2}\log\left(\frac{\delta}{2m}\right)^2$, we get,

$$\left\|(\Phi_k^{\perp})^\top \widehat{\Phi}\right\|_F^2 \leq \frac{q_{ERM}(1/\delta, \log(m), \log(n))}{n}$$

□

Space and Computational Complexity of ERM: ERM requires computing and storing the empirical covariance matrix $\widehat{\mathrm{C}}_m$, which takes $O(m^2)$ memory. A rank k SVD on $\widehat{\mathrm{C}}_m$, generally, takes $O(m^2k)$ computations. We note that there are methods to scale this up but it is out of the scope of this work.

D.2 Oja's Algorithm

Having shown that ERM achieves optimal statistical rates, we now discuss a (relatively) more efficient algorithm in terms of space and computational complexity. We leverage the recent analysis of the classical Oja's algorithm and show how the algorithmic parameters affect the main result. We first restate the theorem statement from the analysis of Oja in Allen-Zhu and Li [2016a].

Theorem D.3. *Let* $\text{gap} := \lambda_k - \lambda_{k+1} \in \left(0, \frac{1}{k}\right]$ *and* $\Lambda := \sum_{i=1}^k \lambda_i \in (0,1]$*, for every* $\epsilon, \delta \in (0,1)$ *define learning rates*

$$T_0 = \Theta\left(\frac{4k\Lambda}{\text{gap}^2\,\delta^2}\right), T_1 = \Theta\left(\frac{\Lambda}{\text{gap}^2}\right), \eta_t = \begin{cases} \Theta\left(\frac{1}{\text{gap}\,T_o}\right) & 1 \leq t \leq T_0 \\ \Theta\left(\frac{1}{\text{gap}^2\,T_1}\right) & T_0 < t \leq T_0 + T_1 \\ \Theta\left(\frac{1}{\text{gap}(t-T_0)}\right) & t > T_0 + T_1 \end{cases}$$

Let Z *be the column orthonormal matrix consisting of all eigenvectors of* C_m *with values no more than* λ_{k+1}*. Then the output* Q_T *of the algorithm satisfies with at least* $1 - \frac{\delta}{2}$*,*

$$\textit{for every } T = T_0 + T_1 + \Theta\left(\frac{T_1}{\epsilon}\right), \textit{ it satisfies } \left\|\mathrm{Z}^\top \mathrm{Q}_T\right\|_F^2 \leq \epsilon$$

The above theorem gives guarantees of the form required by the definition of efficient subspace learner. Therefore, implicitly, Oja is an efficient subspace learner. This is formally stated in the following lemma.

Lemma D.4. $\mathcal{A}_{oja}$ *is an Efficient Subspace Learner.*

Proof. From Theorem D.3, we have

$$\begin{aligned}\left\|\mathrm{Z}^\top \mathrm{Q}_n\right\|_F^2 &\leq \epsilon \\ &= \tilde{\Theta}\left(\frac{T_1}{n - T_0 - T_1}\right) \\ &\leq \tilde{\Theta}\left(\frac{2\Lambda}{\mathrm{gap}^2\, n}\right) && \text{(for large } n\text{)}\end{aligned}$$

Setting $q_{oja}(1/\delta, \log(m), \log(n)) = \tilde{\Theta}\left(\frac{\Lambda}{\mathrm{gap}^2}\right)$, we get,

$$\left\|\mathrm{Z}^\top \mathrm{Q}_n\right\|_F^2 \leq \frac{q_{oja}(1/\delta, \log(m), \log(n))}{n}$$

□

Moreover the requirement of an initial constant number of samples as stated in Theorem 4.2 also appears in Theorem D.3 as warm-up phase. Therefore, the requirement of initial samples can be absorbed in the warm-up phase of Oja.

Space and Computational Complexity of Oja's Algorithm: Oja's algorithm takes $O(mk)$ memory. The per iteration computational cost is $O(mk)$. Therefore, for an ϵ-suboptimal solution, the total computational cost is $O\left(\frac{mk}{\epsilon^2}\right)$.

E Experiments

We now need some lemmas which gives us analytical forms which would be used to calculate the objective with respect to empirical measure in the experiments.

Let $\widehat{\mathrm{P}}_{\mathcal{A}}$ be the output of an efficient subspace learner $\mathcal{A}$. Let $\widehat{\mathrm{P}}_{\mathcal{A}} = \tilde{\Phi}\tilde{\Phi}^\top$ be its eigendecomposition. We define $\widehat{\Phi} = \tilde{\Phi}\mathrm{R}^*$, where let

$$\mathrm{R}^* = \underset{\mathrm{R}^\top\mathrm{R}=\mathrm{R}\mathrm{R}^\top=\mathrm{I}}{\arg\min} \left\| \tilde{\Phi}\mathrm{R} - \Phi_k \right\|_{\mathcal{F}}^2$$

The following gives gives an explit form for R^*.

Lemma E.1. *For any orthogonal matrix $\tilde{\Phi} \in \mathbb{R}^{m\times k}$ and $\Phi \in \mathbb{R}^{m\times k}$, the solution of the optimization problem*

$$\underset{\mathrm{R}^\top\mathrm{R}=\mathrm{R}\mathrm{R}^\top=\mathrm{I}}{\arg\min} \left\| \tilde{\Phi}\mathrm{R} - \Phi \right\|_{\mathcal{F}}^2$$

is $\mathrm{R}^* = \tilde{\Phi}^\top\Phi$

Proof.

$$\underset{\mathrm{R}^\top\mathrm{R}=\mathrm{R}\mathrm{R}^\top=\mathrm{I}}{\arg\min} \left\| \tilde{\Phi}\mathrm{R} - \Phi_k \right\|_{\mathcal{F}}^2 = \underset{\mathrm{R}^\top\mathrm{R}=\mathrm{R}\mathrm{R}^\top=\mathrm{I}}{\arg\min} -\mathrm{Tr}\left(\mathrm{R}^\top\tilde{\Phi}^\top\Phi_k\right)$$

$$\max_{\mathrm{R}^\top\mathrm{R}=\mathrm{R}\mathrm{R}^\top=\mathrm{I}} \mathrm{Tr}\left(\mathrm{R}^\top\tilde{\Phi}^\top\Phi_k\right) \leq \|\mathrm{R}\|_F \left\|\tilde{\Phi}^\top\Phi_k\right\|_F = k$$

Note that $\mathrm{Tr}\left(\mathrm{R}^*\tilde{\Phi}^\top\Phi_k\right) = k$. So, the maximum is achieved at $\mathrm{R} = \mathrm{R}^* = \tilde{\Phi}^\top\Phi_k$. □

We have $\widehat{\Phi} = \tilde{\Phi}\mathrm{R}^*$. We now use this and apply Lemma E.2 to evaluate the objective.

Lemma E.2. *For a projection matrix $\mathrm{P} = \mathrm{U}\mathrm{U}^\top = \sum_{i=1}^k u_i \otimes_{\mathcal{F}} u_i$, $\langle \mathfrak{A}\mathrm{P}, \mathfrak{I}C\rangle_\rho = \frac{1}{n}\mathrm{Tr}\left(\mathrm{V}^\top\mathrm{K}\mathrm{V}\right)$, where $\mathrm{V} = \Phi^\top\mathrm{U}\mathrm{S}^{-\frac{1}{2}}$ and $\mathrm{S} = \mathrm{diag}\left(\lambda_1, \lambda_2, \ldots, \lambda_k\right), \lambda_i = \langle \mathrm{C}_m u_i, u_i\rangle_{\mathcal{F}}$*

Proof of Lemma E.2.

$$\begin{aligned}
\langle \mathfrak{A}\mathrm{P}, \mathfrak{I}\mathrm{C}\rangle_{HS(\rho)} &= \left\langle \sum_{i=1}^k \frac{\mathrm{A}u_i}{\sqrt{\lambda_i}} \otimes_\rho \frac{\mathrm{A}u_i}{\sqrt{\lambda_i}}, \sum_{j=1}^n \bar{\phi_j} \otimes_\rho \bar{\phi_j} \right\rangle_{HS(\rho)} \\
&= \sum_{i,j=1}^{k,n} \frac{1}{\lambda_i} \left\langle \mathrm{A}u_i, \bar{\phi_j} \right\rangle_\rho^2 \\
&= \sum_{i,j=1}^{k,n} \frac{1}{\lambda_i} \left\langle u_i, \mathrm{A}^*\bar{\phi_j} \right\rangle_{\mathcal{F}}^2
\end{aligned}$$

where the second equality follows from bi-linearity of inner products, third from the definition of adjoints.

$$\begin{aligned}
\left\langle u_i, \mathrm{A}^*\bar{\phi_j} \right\rangle_{\mathcal{F}} &= \sum_{l=1}^m (u_i)_l \left(\mathrm{A}^*\bar{\phi_j}\right)_l \\
&= \sum_{l=1}^m (u_i)_l \frac{1}{n} \sum_{q=1}^n \bar{\phi_j}(x_q)\mathrm{z}(x_q)_l = \frac{1}{n} u_i^\top \Phi\bar{\Phi}_j
\end{aligned}$$

where $\bar{\Phi_j} \in \mathbb{R}^n$ and $\left(\bar{\Phi_j}\right)_q = \bar{\phi_j}(x_q)$.

Note that

$$\begin{aligned}\bar{\lambda_j} &= \left\langle \mathrm{C}\bar{\phi}_j, \bar{\phi}_j \right\rangle_{\mathcal{H}} = \left\langle \mathrm{I}^*\mathrm{I}\bar{\phi}_j, \bar{\phi}_j \right\rangle_{\mathcal{H}} \\ &= \left\langle \mathrm{I}\bar{\phi}_j, \mathrm{I}\bar{\phi}_j \right\rangle_{\rho} = \left\langle \bar{\phi}_j, \bar{\phi}_j \right\rangle_{\rho} \\ &= \frac{1}{n}\sum_{q=1}^{n} \bar{\phi}_j(x_q)^2 = \frac{1}{n}\left\|\bar{\Phi}_j\right\|_2^2\end{aligned}$$

where the third equality follows from the property of adjoints.
Moreover, $\left(\mathrm{V}_j^*\right)^\top \mathrm{K}\mathrm{V}_j^* = \bar{\lambda_j}$. Therefore $\bar{\Phi_j}^\top \bar{\Phi}_j = n\bar{\lambda_j} = n\left(\mathrm{V}_j^*\right)^\top \mathrm{K}\mathrm{V}_j^*$.
So we have $\bar{\Phi}_j = \sqrt{n}\mathrm{K}^{1/2}\mathrm{V}_j^*$. Hence,

$$\begin{aligned}\langle \mathfrak{A}\mathrm{P}, \mathfrak{J}\mathrm{C} \rangle_{HS(\rho)} &= \sum_{i,j=1}^{k,n} \frac{1}{\lambda_i}\left(\frac{1}{n} u_i^\top \Phi\sqrt{n}\mathrm{K}^{1/2}\mathrm{V}_j^*\right)^2 \\ &= \frac{1}{n}\sum_{i,j=1}^{k,n} \frac{1}{\lambda_i}\left(u_i^\top \Phi\mathrm{K}^{1/2}\mathrm{V}_j^*\right)^2 \\ &= \frac{1}{n}\sum_{i,j=1}^{k,n} \left(\frac{1}{\sqrt{\lambda_i}} u_i^\top \Phi\mathrm{K}^{1/2}\mathrm{V}_j^*\right)^2 \\ &= \frac{1}{n}\sum_{i,j=1}^{k,n} \left(\mathrm{V}_i^\top \mathrm{K}^{1/2}\mathrm{V}_j^*\right)^2\end{aligned}$$

where $\mathrm{V} = \Phi^\top \mathrm{U}\mathrm{S}^{-\frac{1}{2}}$. Therefore, we have,

$$\begin{aligned}\langle \mathfrak{A}\mathrm{P}, \mathfrak{J}\mathrm{C} \rangle_{HS(\rho)} &= \frac{1}{n}\sum_{i,j=1}^{k,n} \mathrm{Tr}\left(\mathrm{V}_i^\top \mathrm{K}^{1/2}\mathrm{V}_j^*\right)^2 \\ &= \frac{1}{n}\sum_{i,j=1}^{k,n} \mathrm{Tr}\left(\mathrm{V}_i^\top \mathrm{K}^{1/2}\mathrm{V}_j^*(\mathrm{V}_j^*)^\top \mathrm{K}^{1/2}\mathrm{V}_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{k} \mathrm{Tr}\left(\mathrm{V}_i^\top \mathrm{K}^{1/2}\sum_{j=1}^{n}\mathrm{V}_j^*(\mathrm{V}_j^*)^\top \mathrm{K}^{1/2}\mathrm{V}_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{k} \mathrm{Tr}\left(\mathrm{V}_i^\top \mathrm{K}^{1/2}\mathrm{V}^*(\mathrm{V}^*)^\top \mathrm{K}^{1/2}\mathrm{V}_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{k} \mathrm{Tr}\left(\mathrm{V}_i^\top \mathrm{K}\mathrm{V}_i\right) \\ &= \frac{1}{n}\mathrm{Tr}\left(\mathrm{V}^T\mathrm{K}\mathrm{V}\right)\end{aligned}$$

□

F Auxillary Results

Here we state some Auxillary results used in the proofs.

Theorem F.1 (Generalized Gap free Wedin Theorem)**.** *For $\epsilon > 0$, let* A *and* B *be two PSD matrices. For every $\mu > 0, \alpha > 0$, let* U *be column orthonormal matrix consisting of eigenvectors of* A *with eigenvalue $\leq \mu$, let* V *be column orthonormal matrix consisting of eigenvectors of* B *with eigenvalue $\geq \mu + \alpha$, then we have*

$$\left\|\mathrm{U}^\top \mathrm{V}\right\| \leq \frac{\|A - B\|}{\alpha}$$

where the norm $\|\cdot\|$ is any sub-multiplicative norm.

Proof. The above theorem is stated in [Allen-Zhu and Li, 2016b, Lemma B.3] in the sense of spectral norm. For the sake of completeness, we present the proof and show that it can easily be generalized to any sub-multiplicative norm.

Let the SVD of A and B be $\mathrm{A} = \mathrm{U}\Sigma U^\top + \mathrm{U}'\Sigma'\mathrm{U}^\top, \mathrm{B} = \mathrm{V}\tilde{\Sigma}\mathrm{V}^\top + \mathrm{V}'\tilde{\Sigma}'\mathrm{V}'^\top$, where Σ is a diagonal matrix which contains all eigenvalues of A which are $\leq \mu$. Similarly, $\tilde{\Sigma}$ contains all eigenvalues $\geq \mu + \alpha$. Let $\mathrm{E} := \mathrm{A} - \mathrm{B}$.

$$\Sigma\mathrm{U}^\top = \mathrm{U}^\top \mathrm{A} = \mathrm{U}^\top(\mathrm{B} + \mathrm{E})$$

where the first equality follows because U is orthogonal to U′. Multiply by V on the right on both sides, we get,

$$\Sigma\mathrm{U}^\top \mathrm{V} = \mathrm{U}^\top \mathrm{B}\mathrm{V} + \mathrm{U}^\top \mathrm{E}\mathrm{V} = \mathrm{U}^\top \mathrm{V}\tilde{\Sigma} + \mathrm{U}^\top \mathrm{E}\mathrm{V}$$

where the second equality follows because V is orthogonal to V′. Multiplying by $\tilde{\Sigma}^{-1}$ on the right on both sides, we get,

$$\Sigma\mathrm{U}^\top \mathrm{V}\tilde{\Sigma}^{-1} = \mathrm{U}^\top \mathrm{V} + \mathrm{U}^\top \mathrm{E}\mathrm{V}\tilde{\Sigma}^{-1}$$

Taking any sub-multiplicative norm on the left hand side, we obtain an upper bound on it as follows,

$$\begin{aligned}\left\|\Sigma\mathrm{U}^\top \mathrm{V}\tilde{\Sigma}^{-1}\right\| &\leq \|\Sigma\|_2 \left\|\tilde{\Sigma}^{-1}\right\|_2 \left\|U^\top \mathrm{V}\right\| \\ &\leq \frac{\mu}{\mu + \alpha} \left\|\mathrm{U}^\top \mathrm{V}\right\|\end{aligned}$$

where the first inequality follows from the property of sub-multiplicative norms, and the second from the definition of Σ and $\tilde{\Sigma}$.
Similarly, taking any sub-multiplicative norm on the right hand side, we get a lower bound on it as follows,

$$\begin{aligned}\left\|\mathrm{U}^\top \mathrm{V} + \mathrm{U}^\top \mathrm{E}\mathrm{V}\tilde{\Sigma}^{-1}\right\| &\geq \left\|\mathrm{U}^\top \mathrm{V}\right\| - \left\|\mathrm{U}^\top \mathrm{E}\mathrm{V}\tilde{\Sigma}^{-1}\right\| \\ &\geq \left\|\mathrm{U}^\top \mathrm{V}\right\| - \left\|U^\top\right\|_2 \|\mathrm{E}\| \|\mathrm{V}\|_2 \left\|\tilde{\Sigma}^{-1}\right\|_2 \\ &\geq \left\|\mathrm{U}^\top \mathrm{V}\right\| - \frac{\|\mathrm{E}\|}{\mu + \alpha}\end{aligned}$$

where the first inequality follows from (reverse) triangle inequality, the second from property of sub-multiplicative norms and third because U and V are orthonormal matrices and by definition of $\tilde{\Sigma}$. Combining both the bounds, we get,

$$\begin{aligned}&\left\|\mathrm{U}^\top \mathrm{V}\right\| \left(1 - \frac{\mu}{\mu + \alpha}\right) \leq \frac{\|\mathrm{E}\|}{\mu + \alpha} \\ \implies &\left\|\mathrm{U}^\top \mathrm{V}\right\| \leq \frac{\|\mathrm{E}\|}{\alpha}\end{aligned}$$

□

Theorem F.2 (Matrix Bernstein [Tropp et al., 2015]). *Let* $\mathrm{S}_1, \mathrm{S}_2, \ldots \mathrm{S}_n$ *be* n *i.i.d* $d_1 \times d_2$ *random matrices such that* $\mathbb{E}\mathrm{S}_i = 0, \|\mathrm{S}_i\| \leq L \; \forall i \in [n]$. *Let* $\mathrm{Z} = \sum_{i=1}^{n} \mathrm{S}_i$. *Let* $v(\mathrm{Z})$ *denote the matrix variance statistic of the sum defined as,*

$$v(\mathrm{Z}) = \max\{\mathbb{E}\mathrm{Z}\mathrm{Z}^\top, \mathbb{E}\mathrm{Z}^\top\mathrm{Z}\}$$

Then, with probability at least $1 - \delta$, *we have,*

$$\mathbb{P}\left\{\|\mathrm{Z}\| \geq t\right\} \leq (d_1 + d_2) \exp\left(\frac{-t^2/2}{v(\mathrm{Z}) + Lt/3}\right) \; \forall t \; \geq 0$$

Theorem F.3 (Local Rademacher Complexity [Bartlett et al., 2002]). *Let* $\mathcal{X}$ *be a measurable space. Let* $\mathcal{P}$ *be a probability distribution on* $\mathcal{X}$ *and let* $\mathrm{x}_1, \mathrm{x}_2 \ldots \mathrm{x}_n$ *be i.i.d. samples drawn from* $\mathcal{P}$. *Let* $\mathcal{P}_n$ *denote the empirical measure. Let* $\mathcal{F}$ *be a class of functions on* $\mathcal{X}$ *ranging from* $[-1, 1]$ *and assume that there exists some constant* B *such that for every* $f \in \mathcal{F}, \mathcal{P}^2 f \leq B\mathcal{P}f$. *Let* ψ *be a sub-root function and let* r^* *be the fixed point of* ψ. *If* ψ *satisfies*

$$\psi(r) \geq B\mathbb{E}_{X,\sigma}\left[\mathcal{R}_n\{f \in star(\mathcal{F}) | \mathcal{P}f^2 \leq r\}\right]$$

where $star(\mathcal{F}) = \{\lambda f | f \in \mathcal{F}, \lambda \in [0, 1]\}$ *is the star shaped hull of* $\mathcal{F}$ *and* $\mathcal{R}_n\mathcal{F} = \sup_{\mathrm{f}} \in \mathcal{F}\frac{1}{n}\sum_{i=1}^{n} \sigma_i f(\mathrm{x}_i)$ *is the empirical Rademacher complexity of* $\mathcal{F}$ *given data points* $\{\mathrm{x}_i\}_{i=1}^n$; *then for every* $K > 0$ *and* $\mathrm{x} > 0$, *with probability at least* $1 - e^{-\delta}$

$$\forall \, f \in \mathcal{F}, \mathcal{P}f \leq \frac{K}{K-1}\mathcal{P}_n f + \frac{6K}{B}r^* + \frac{\delta(11 + 5BK)}{n} \tag{6}$$

Also, with probability at least $1 - e^{-\delta}$

$$\forall f \in \mathcal{F}, \mathcal{P}_n f \leq \frac{K}{K+1}\mathcal{P}f + \frac{6K}{B}r^* + \frac{\delta(11 + 5BK)}{n} \tag{7}$$

Furthermore, if $\widehat{\psi}_n$ *is a data-dependent sub-root function with fixed point* $\widehat{r}^*$ *such that*

$$\psi^*(r) > 2(10 \vee B)\mathbb{E}_\sigma\left[\mathcal{R}_n\{f \in star(\mathcal{F}) | \mathcal{P}^n f^2 \leq 2r\}\right] + \frac{2(10 \vee B + 11)\delta}{n}$$

then with probability at least $1 - 2e^{\delta}$, *it holds that* $\widehat{r}^* \geq r^*$; *as a consequence, equations 6 and 7 holds with* r^* *replaced by* $\widehat{r}^*$