Report on Statistical Properties of Kernel PCA

By Xuelin Zhu
Email: xuelin@umich.edu

August 25, 2023

Contents

1 Problem Setup and Results Overview

Suppose the data are iid observations from a population in $\mathbb{R}^p$, i.e., $x_1, \ldots, x_n \overset{\text{iid}}{\sim} X \in \mathbb{R}^p$. Prespecifying a number $d < p$, PCA finds the subspace of dimension d where the total variance of the projected data is maximized. However, PCA may behave poorly when the data exhibits nonlinearity, as its projection is limited in $\mathbb{R}^p$. In such cases, projecting the data to a higher-dimensional space can be advantageous, as demonstrated in the following example.

Example. Initially, the data resides in the $\mathbb{R}^2$ plane, forming a circular pattern. In this scenario, regardless of the chosen linear subspace, it is impossible to separate the two classes. However, by employing a *higher-dimensional projection* function defined as

$$\forall x \in \mathbb{R}^2 : \quad \phi(x) = \begin{pmatrix} y_1 = x_1 \\ y_2 = x_2 \\ y_3 = x_1^2 + x_2^2 \end{pmatrix},$$

it becomes feasible to achieve separation using a linear subspace in $\mathbb{R}^3$.

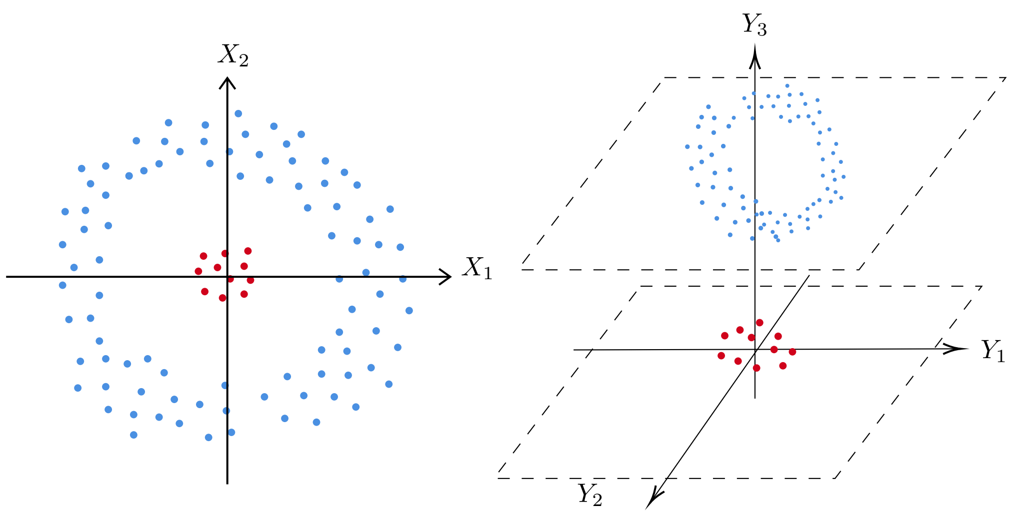

Figure 1: Higher dimension projection

This example serves to illustrate the concept of kernel PCA. By selecting a feature map $\phi : \mathbb{R}^p \mapsto \mathscr{H}$, where $\mathscr{H}$ represents a separable Hilbert space, we can transform the original data x_i, which is a random vector in $\mathbb{R}^d$, into a new set of variables $z_i := \phi(x_i) \overset{\text{iid}}{\sim} Z \in \mathscr{H}$. Since $\mathscr{H}$ is a Hilbert space, the inner product endowed on it provides the covariance structure, enabling us to perform PCA on the transformed variables z_i.

1.1 Setup and Assumptions

Here listed are the model assumptions and notations used in the paper. Subsequent theorems will introduce additional assumptions to further support the analysis.

> **Problem Setup.** Suppose $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} Z \in \mathscr{H}$ with $\mathbb{E}\|Z\|^4 < \infty$.

(a) For any integrable function $f : \mathscr{H} \mapsto \mathbb{R}$, denote the mean and sample mean by

$$Pf := \mathbb{E}[f(Z)] \text{ and } P_n f := \frac{1}{n}\sum_{i=1}^{n} f(Z_i).$$

(b) The collection of all Hilbert-Schmidt operators from $\mathscr{H}$ to $\mathscr{H}$ is denoted by $\text{HS}(\mathcal{H})$, and the rank one operator is defined as

$$\forall f, g \in \mathscr{H} \setminus \{0\}, \quad f \otimes g^*(h) := \langle g, h\rangle f.$$

For any $z \in \mathscr{H}$, denote $C_z := z \otimes z \in \text{HS}(\mathcal{H})$, and the non-centered covariance operators of Z and C_Z are denoted by

$$\begin{aligned} C_1 &:= \mathbb{E}(Z \otimes Z^*) = \mathbb{E}(C_Z), \\ C_2 &:= \mathbb{E}(C_Z \otimes C_Z^*), \end{aligned}$$

with the sampled version notation $C_{1,n} := \frac{1}{n}\sum_{i=1}^{n} Z_i \otimes Z_i^*$ and $C_{2,n} := \frac{1}{n}\sum_{i=1}^{n} C_{Z_i} \otimes C_{Z_i}^*$.

(c) An orthogonal projector in $\mathscr{H}$ is an operator U such that $U^2 = U = U^*$. And given a closed subspace $V \subset \mathscr{H}$, the unique orthogonal projector having range V and null space $V^\perp$ is denoted by Π_V. When V is of finite dimension, the operator $\Pi_{V^\perp}$ is not Hilbert-Schmidt. However with the abuse of notation, for a trace-class operator A, we denote

$$\langle \Pi_{V^\perp}, A\rangle := \text{tr}A - \langle \Pi_V, A\rangle \tag{1}$$

(d) Given a subspace V, the notations for the true and empirical reconstruction error of V are

$$R(V) := \mathbb{E}\left[\|Z - \Pi_V(Z)\|^2\right] = P\left\langle \Pi_{V^\perp}, C_Z\right\rangle = \left\langle \Pi_{V^\perp}, C_1\right\rangle \tag{2}$$

and

$$R_n(V) := \frac{1}{n}\sum_{i=1}^{n}\|Z_i - \Pi_V(Z_i)\|^2 = P_n\left\langle \Pi_{V^\perp}, C_Z\right\rangle = \left\langle \Pi_{V^\perp}, C_{1,n}\right\rangle, \tag{3}$$

where the second equality in both notations follows from (1) and

$$\|Z - \Pi_V(Z)\|^2 = \|Z\|^2 - \|\Pi_V(Z)\|^2 = \text{tr}C_Z - \langle \Pi_V, Z \otimes Z^*\rangle = \left\langle \Pi_{V^\perp}, Z \otimes Z\right\rangle.$$

Prespecifying d as the dimension of projected subspace, we denote by $\mathcal{V}_d$ the collection of all subspaces of dimension d, and

$$V_d := \arg\min_{V \in \mathcal{V}_d} R(V) \text{ and } \widehat{V}_d := \arg\min_{V \in \mathcal{V}_d} R_n(V).$$

Before delving into the derivations, it is beneficial to outline a few key facts that will provide assistance throughout the paper. And they can be easily showed by using a CONS considering the separability of $\mathscr{H}$.

For any $f, g \in \mathscr{H} \setminus \{0\}$, the rank one operator has the following properties:

$$\begin{aligned}\|f \otimes g^*\|_{\mathrm{HS}(\mathcal{H})} &= \|f\|_{\mathscr{H}} \|g\|_{\mathscr{H}},\\ \mathrm{tr} f \otimes g^* &= \langle f, g \rangle_{\mathscr{H}},\\ \langle f \otimes g^*, A \rangle_{\mathrm{HS}(\mathcal{H})} &= \langle Ag, f \rangle, \quad \forall A \in \mathrm{HS}(\mathcal{H}).\end{aligned}$$

For an orthogonal projector U, it has the properties:

$$\begin{aligned}\|U(h)\|_{\mathscr{H}}^2 &= \langle h, Uh \rangle_{\mathscr{H}} \leq \|h\|_{\mathscr{H}}^2,\\ \langle f \otimes g^*, U \rangle_{\mathrm{HS}(\mathcal{H})} &= \langle Uf, Ug \rangle_{\mathscr{H}}.\end{aligned}$$

An orthogonal projector U has rand $d < \infty$ if and only if $U \in \mathrm{HS}(\mathcal{H})$ with

$$\|U\|_{\mathrm{HS}(\mathcal{H})} = \sqrt{d} \text{ and } \mathrm{tr}(U) = d.$$

1.2 Results Overview

With sample size n, dimension d, and these notations, we can overview the results of the paper. Firstly, the authors recalled a "global bound" for the uncentered data by Shawe-Taylor et al(2005), taking the form

$$R(\widehat{V}_d) - R(V_d) \lesssim \sqrt{\frac{d}{n} \mathrm{tr}(C_2 - C_1 \otimes C_1^*)}, \tag{4}$$

where the symbol $\lesssim$ means some constants are forgetting for the overview purpose. Then by introducing a "local approach", the authors proved two better bounds, one taking the form

$$R(\widehat{V}_d) - R(V_d) \lesssim B_d \rho(d, n), \tag{5}$$

and the other taking the form

$$R(\widehat{V}_d) - R(V_d) \lesssim \sqrt{R(V_d)\rho(d, n)} + \rho(d, n), \tag{6}$$

where $\rho(d, n)$ is a quantity which is always smaller than the RHS of (4), and $B_d \lesssim [R(V_d) - R(V_{d-1})]^{-1}$. In practice, (5) behaves better than (4) when $n \to \infty$, and (6) behaves better than (4) when $d \to \infty$. An example of how these bounds behave is given in Figure 1 of the paper. Furthermore, the authors provided the bound in the form of (4) for the centered data.

Given that we will be directly dealing with random variables taking values in a Hilbert space, these results (bounds on reconstruction error) naturally apply to functional data.

1.3 Report Outline

The report consists of several sections covering different aspects of the topic. In Section 2, the global approach results are discussed, starting with the uncentered case in Section 2.1, followed by the centered case in Section 2.2.

Section 3 focuses on the local approach results. It begins with the introduction of a theorem in Section 3.1, which serves as a lemma for the subsequent sections. Then, in Section 3.2, the local approach is applied to get a bound when the sample size is large. Section 3.3 delves into the local approach for situations where the dimension is large. The proofs in Section 3 are independent of those in Section 2, so skipping Section 2 should not cause any issues for understanding Section 3.

In Section 4 of this report, we delve into the practical application of the theorems discussed in the previous sections, specifically in the context of kernel PCA. This section likely provides insights into how the obtained results and techniques can be effectively utilized in analyzing and understanding kernel PCA. The proofs presented in Section 4 solely depend on the results established in the previous sections. Therefore, skipping the proofs in Sections 2 and 3 will not hinder the comprehension of the material in Section 4.

Finally, in Section 5, the report concludes with a brief summary of the paper's key points and implications, as well as reflections on what I have learned and the areas I think I can focus on in the future.

2 Global Approach

In this section, we will present a reformulation of the global approach for uncentered data, as proposed by Shawe-Taylor et al. Furthermore, we will introduce the centered version developed by the author of this paper. It is important to note that the ordering of the theorems may not align exactly with the original paper, as I have grouped together the results that share similar outcomes.

2.1 Uncentered Case

To avoid any confusion, the theorems in this report are indexed consistently with those in the original paper.

Theorem 3.1. Assume $\|Z\|^2 \leq M$ a.s. and that $Z \otimes Z^*$ belongs a.s. to a set of $\mathrm{HS}(\mathscr{H})$ with bounded diameter L. Then for any $n \geq 2$, with probability at least $1 - 3e^{-\xi}$,

$$|R(\widehat{V}_d) - R_n(\widehat{V}_d)| \leq \sqrt{\frac{d}{n-1} \operatorname{tr} C'_{2,n}} + (M \wedge L)\sqrt{\frac{\xi}{2n}} + L\frac{\sqrt{d}\xi^{\frac{1}{4}}}{n^{\frac{3}{4}}}. \tag{7}$$

Also, with probability at least $1 - 2e^{-\xi}$,

$$0 \leq R(\widehat{V}_d) - R(V_d) \leq \sqrt{\frac{d}{n} \operatorname{tr} C'_2} + 2(M \wedge L)\sqrt{\frac{\xi}{2n}}, \tag{8}$$

where $C'_2 = C_2 - C_1 \otimes C_1^*$ and $C'_{2,n} = C_{2,n} - C_{1,n} \otimes C_{1,n}^*$.

Proof. We now focus on $|R(\widehat{V}_d) - R_n(\widehat{V}_d)|$. By (2) and (3), we can write

$$R(\widehat{V}_d) - R_n(\widehat{V}_d) = (P - P_n)\left\langle \Pi_{\widehat{V}_d^{\perp}}, C_Z \right\rangle \leq \sup_{V \in \mathcal{V}_d} (P - P_n)\left\langle \Pi_{V^{\perp}}, C_Z \right\rangle. \tag{9}$$

Firstly, we will establish an upper bound of $|\langle \Pi_{V^\perp}, C_z - C_{z'} \rangle| \leq L \wedge M$ for any $z, z' \in \operatorname{supp}(Z)$. This bound serves as a condition that allows us to apply McDiarmid's and Hoeffding's inequalities to get our desired result.

Let $z, z' \in \operatorname{supp}(Z)$. With V being of finite dimension d, by (1), we know

$$\langle \Pi_{V^\perp}, C_z \rangle = \operatorname{tr} C_z - \langle \Pi_V, z \otimes z^* \rangle = \|z\|^2 - \|\Pi_V(z)\|^2 = \|\Pi_{V^\perp}(z)\|^2 \leq \|z\|^2. \tag{10}$$

Further with the assumption $\|Z\|^2 \leq M$ a.s., (10) implies $0 \leq \langle \Pi_{V^\perp}, C_z \rangle \leq M$ a.s., and hence $|\langle \Pi_{V^\perp}, C_z - C_{z'} \rangle| \leq$

M. Another bound comes from the orthogonal decomposition

$$z \otimes z^* = z_V \otimes z_V^* + z_V \otimes z_{V^\perp}^* + z_{V^\perp} \otimes z_V^* + z_{V^\perp} \otimes z_{V^\perp}^*.$$

By contractivity property of an orthogonal projection, it follows that

$$\begin{aligned}
\|z \otimes z^* - z' \otimes z'^*\| &\geq \|z_{V^\perp} \otimes z_{V^\perp}^* - z'_{V^\perp} \otimes z'^*_{V^\perp}\| \\
&\geq |\|z_{V^\perp} \otimes z_{V^\perp}^*\| - \|z'_{V^\perp} \otimes z'^*_{V^\perp}\|| \\
&= \left| \|z_{V^\perp}\|^2 - \|z'_{V^\perp}\|^2 \right| \\
&= |\langle \Pi_{V^\perp}, z \otimes z^* - z' \otimes z'^* \rangle| .
\end{aligned}$$

Further with the assumption on diameter L of $\mathrm{HS}(\mathcal{H})$, we claim

$$|\langle \Pi_{V^\perp}, C_z - C_{z'} \rangle| \leq \|C_z - C_{z'}\| \leq L \quad \Rightarrow \quad |\langle \Pi_{V^\perp}, C_z - C_{z'} \rangle| \leq L \wedge M. \tag{11}$$

Letting $f(z_1, \ldots, z_n) := \sup_{V \in \mathcal{V}_d} (P_n - P) \langle \Pi_{V^\perp}, C_z \rangle$, then we have

$$\sup_{z_1,\ldots,z_n,z_i'} |f(z_1, \ldots, z_n) - f(z_1, \ldots, z_i', \ldots, z_n)| = \sup_{z's} \left| \sup_{V \in \mathcal{V}_d} P_n \langle \Pi_V, C_z \rangle - \sup_{V \in \mathcal{V}_d} P_n' \langle \Pi_V, C_z \rangle \right|,$$

where $P_n' \langle \Pi_V, C_z \rangle$ is the same as $P_n \langle \Pi_V, C_z \rangle$ except z_i is replaced by z_i',

$$\begin{aligned}
&\leq \sup_{z's} \left| \sup_{V \in \mathcal{V}_d} \frac{1}{n} \left\langle \Pi_{V^\perp}, C_{z_i} - C_{z_i'} \right\rangle \right| \\
&\leq \frac{1}{n} M \wedge L,
\end{aligned}$$

where the last equality holds since (11) holds for all choice of $V \in \mathcal{V}_d$ and all choice of $z's$ taking values in the range of Z. Applying McDiarmid's inequality to $f(Z_1, \ldots, Z_n)$ in Appendix B.1 by taking $c_i = \frac{1}{n} M \wedge L$ and ξ in the Appendix to be $(M \wedge L)\sqrt{\xi/(2n)}$, it follows that with probability at least $1 - e^{-\xi}$,

$$\sup_{V \in \mathcal{V}_d} (P_n - P) \langle \Pi_{V^\perp}, C_Z \rangle \leq \mathbb{E} \left[\sup_{V \in \mathcal{V}_d} (P_n - P) \langle \Pi_{V^\perp}, C_Z \rangle \right] + (M \wedge L) \sqrt{\frac{\xi}{2n}}. \tag{12}$$

A very similar derivation also leads to

$$\sup_{V \in \mathcal{V}_d} (P - P_n) \langle \Pi_{V^\perp}, C_Z \rangle \leq \mathbb{E} \left[\sup_{V \in \mathcal{V}_d} (P - P_n) \langle \Pi_{V^\perp}, C_Z \rangle \right] + (M \wedge L) \sqrt{\frac{\xi}{2n}}.$$

It remains to bound the expectation term, and then we will complete the $|R(\widehat{V}_d) - R_n(\widehat{V}_d)|$ part. By the definition of P_n and P, it can be checked that

$$(P_n - P) \langle \Pi_{V^\perp}, C_Z \rangle = \left\langle \Pi_V, \mathbb{E} C_Z - \frac{1}{n} \sum_{i=1}^n C_{Z_i} \right\rangle + \frac{1}{n} \sum_{i=1}^n \mathrm{tr}(C_{Z_i}) - \mathbb{E}[\mathrm{tr}(C_Z)].$$

With $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} Z$, taking expectation and sup yields

$$\begin{aligned}
\mathbb{E}\left[\sup_{V\in\mathcal{V}_d} (P_n - P)\left\langle \Pi_{V^\perp}, C_Z\right\rangle\right] &= \mathbb{E}\left[\sup_{V\in\mathcal{V}_d} \left\langle \Pi_V, \mathbb{E}C_Z - \frac{1}{n}\sum_{i=1}^n C_{Z_i}\right\rangle\right] \\
\text{(By CS inequality)} \quad &\leq \mathbb{E}\left\{\sup_{V\in\mathcal{V}_d} \|\Pi_V\| \cdot \left\|\mathbb{E}C_Z - \frac{1}{n}\sum_{i=1}^n C_{Z_i}\right\|\right\} \\
\left(\because \|\Pi_V\| = \sqrt{d}\right) \quad &= \sqrt{d}\,\mathbb{E}\left\{\left\|\mathbb{E}C_Z - \frac{1}{n}\sum_{i=1}^n C_{Z_i}\right\|\right\} \\
\text{(By Jensen's inequality)} \quad &\leq \sqrt{d}\sqrt{\mathbb{E}\left\{\left\|\mathbb{E}C_Z - \frac{1}{n}\sum_{i=1}^n C_{Z_i}\right\|^2\right\}} \\
&= \sqrt{d}\sqrt{\frac{1}{n^2}\mathbb{E}\left[\|\mathbb{E}C_Z - C_{Z_1} + \cdots + \mathbb{E}C_Z - C_{Z_n}\|^2\right]} \\
(\because Z_i \overset{\text{iid}}{\sim} Z) \quad &= \sqrt{d}\sqrt{\frac{1}{n^2}\cdot n\mathbb{E}\|\mathbb{E}C_Z - C_{Z_1}\|^2} \\
&= \sqrt{\frac{d}{n}}\sqrt{\mathbb{E}\left[\|\mathbb{E}C_Z - C_{Z_1}\|^2\right]}. \qquad (13)
\end{aligned}$$

It holds that $\mathbb{E}\left[\|\mathbb{E}\left[C_Z\right] - C_{Z_1}\|^2\right] = \frac{1}{2}\mathbb{E}\left[\|C_{Z_1} - C_{Z_2}\|^2\right]$. By applying Hoeffding's inequality in Appendix B.2 by taking $r = 2$, $U = \frac{1}{n(n-1)}\sum_{i\neq j}\|C_{Z_i} - C_{Z_j}\|^2$ and ξ in the Appendix to be $L^2\sqrt{\frac{\xi}{n}}$, it holds for at least probability $1 - e^{-\xi}$,

$$\mathbb{E}\left[\|C_Z - \mathbb{E}\left[C_{Z'}\right]\|^2\right] = \frac{1}{2}\mathbb{E}\left[\|C_{Z_1} - C_{Z_2}\|^2\right] \leq \frac{1}{2n(n-1)}\sum_{i\neq j}\left\|C_{Z_i} - C_{Z_j}\right\|^2 + L^2\sqrt{\frac{\xi}{n}}. \qquad (14)$$

Finally, claiming

$$\frac{1}{n^2}\sum_{i\neq j}\left\|C_{Z_i} - C_{Z_j}\right\|^2 = 2\,\mathrm{tr}\left(C_{2,n} - C_{1,n}\otimes C_{1,n}^*\right), \qquad (15)$$

by (9), (12), 13, (14) and (15) together, using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we have at least probability $1 - 3e^{-\xi}$,

$$|R(\widehat{V}_d) - R_n(\widehat{V}_d)| \leq \sqrt{\frac{d}{n-1}\,\mathrm{tr}\,C'_{2,n}} + (M\wedge L)\sqrt{\frac{\xi}{2n}} + L\frac{\sqrt{d}\xi^{\frac{1}{4}}}{n^{\frac{3}{4}}}.$$

Note that McDiarmid's inequality (12) is used twice to get a two sided bound. This is the reason for $3e^{-\xi}$.

We will check the claim (15) after this whole proof is done. Now let's focus on the $R(\widehat{V}_d) - R(V_d)$ part. By the definition of $\widehat{V}_d$, the following inequality naturally holds

$$0 \leq R(\widehat{V}_d) - R(V_d) \leq \left[R(\widehat{V}_d) - R(V_d)\right] + \left[R_n(V_d) - R_n(\widehat{V}_d)\right],$$

where, according to (9), (12) and (13), the first term is bounded, with at least probability $1 - e^{-\xi}$, by

$$\begin{aligned}
R(\widehat{V}_d) - R(V_d) &\leq \sqrt{\frac{d}{n}}\sqrt{\mathbb{E}\left[\|\mathbb{E}C_Z - C_{Z_1}\|^2\right]} + (M\wedge L)\sqrt{\frac{\xi}{2n}} \\
&= \sqrt{\frac{d}{n}}\sqrt{\mathrm{tr}(C_2 - C_1\otimes C_1^*)} + (M\wedge L)\sqrt{\frac{\xi}{2n}}
\end{aligned}$$

$$= \sqrt{\frac{d}{n}}\sqrt{\mathrm{tr}(C_2')} + (M \wedge L)\sqrt{\frac{\xi}{2n}}, \tag{16}$$

where the equality in the second row comes from the fact $\mathbb{E}\left[\|\mathbb{E}C_Z - C_{Z_1}\|^2\right] = \mathrm{tr}(C_2 - C_1 \otimes C_1^*)$. For the second term, we can compute by definition

$$R_n(V_d) - R_n(\widehat{V}_d) = \sup_{V \in \mathcal{V}_d} \frac{1}{n}\left\langle \Pi_{V^\perp}, \sum_{i=1}^n (C_{Z_i} - \mathbb{E}C_Z)\right\rangle =: U.$$

It is easy to verify that $\mathbb{E}U = 0$ and $U \leq M \wedge L$. By applying Hoeffding's inequality to U and taking $r = 1$ and ξ in the Appendix to be $(M \wedge L)\sqrt{\frac{\xi}{2n}}$, it follows that with probability at least $1 - e^{-\xi}$,

$$U \leq (M \wedge L)\sqrt{\frac{\xi}{2n}} \tag{17}$$

By (16) and (17), with probability at least $1 - 2e^{\xi}$,

$$0 \leq R(\widehat{V}_d) - R(V_d) \leq \sqrt{\frac{d}{n}\operatorname{tr} C_2'} + 2(M \wedge L)\sqrt{\frac{\xi}{2n}}.$$

Verification of Claim (15). Let's start from the LHS. We can compute

$$\begin{aligned}
\frac{1}{n^2}\sum_{i \neq j}\|C_{Z_i} - C_{Z_j}\|^2 &= \frac{1}{n^2}\sum_{i \neq j}\left(\|C_{Z_i}\|^4 + \|C_{Z_j}\|^4\right) - \frac{2}{n^2}\sum_{i \neq j}\langle Z_i, Z_j\rangle^2 \\
&= \frac{2(n-1)}{n^2}\sum_{i=1}^n \|Z_i\|^4 - \frac{2}{n^2}\sum_{i \neq j}\langle Z_i, Z_j\rangle^2.
\end{aligned}$$

And we can also compute the RHS, getting

$$\begin{aligned}
2\mathrm{tr}(C_{2,n} - C_{1,n} \otimes C_{1,n}^*) &= 2\mathrm{tr}(C_{2,n}) - 2\mathrm{tr}(C_{1,n} \otimes C_{1,n}^*) \\
&= \frac{2}{n}\sum_{i=1}^n \|Z_i\|^4 - 2\langle C_{1,n}, C_{1,n}\rangle \\
&= \frac{2}{n}\sum_{i=1}^n \|Z_i\|^4 - \frac{2}{n^2}\sum_{1 \leq i,j \leq n}\langle C_{Z_i}, C_{Z_j}\rangle \\
&= \frac{2}{n}\sum_{i=1}^n \|Z_i\|^4 - \frac{2}{n^2}\left\{\sum_{i \neq j}\langle Z_i, Z_j\rangle^2 + \sum_{i=1}^n \|Z_i\|^4\right\} \\
&= \frac{2n-2}{n^2}\sum_{i=1}^n \|Z_i\|^4 - \frac{2}{n^2}\sum_{i \neq j}\langle Z_i, Z_j\rangle^2.
\end{aligned}$$

Since their simplified results are the same, we complete the whole proof. □

Remark. The comments in the paper provided by the authors offer valuable insights to distinguish the essential differences between (7) and (8) for me. They explain that the computability of $R_n(\widehat{V}_d)$ in (7) allows it to serve as an empirical confidence interval for $R(\widehat{V}_d)$ when all the parameters are known. On the other hand, although the terms in (8) are not computable in practical settings, the upper bound establishes the convergence property of the estimated $\widehat{V}_d$ to the optimal V_d in terms of reconstruction error. Consequently,

(8) holds significant theoretical importance.

2.2 Centered Case

In practice, centering the data before PCA is a common procedure. In Section 3.5 of the paper, the authors showed a bound of the same order as Theorem 3.1 for the centered case. Before the theorem, several notations are needed:

$$\bar{Z} := Z - \mathbb{E}Z \in \mathscr{H}, \quad \bar{C}_Z := \bar{Z} \otimes \bar{Z}^* \in \mathrm{HS}(\mathcal{H}) \text{ and } \bar{C}_1 := \mathbb{E}(\bar{C}_Z) = C_1 - \mathbb{E}[Z] \otimes \mathbb{E}[Z]^*.$$

The corresponding sampled versions are denoted by hat:

$$\hat{Z} := Z - \frac{1}{n}\sum_{i=1}^{n} Z_i, \quad \bar{C}_{Z,n} := \hat{Z} \otimes \hat{Z}^*,$$

and

$$\bar{C}_{1,n} = \frac{1}{n-1}\sum_{i=1}^{n} \bar{C}_{Z_i,n} = C_{1,n} - \frac{1}{n(n-1)}\sum_{i \neq j} Z_i \otimes Z_j^* \text{ with } \mathbb{E}\left[\bar{C}_{1,n}\right] = \bar{C}_1.$$

All equalities above can be checked simply by the helpful facts in Section 1.1. Since data are centered, PCA is done by the followings:

$$\widehat{W}_d := \arg\min_{V \in \mathcal{V}_d} \frac{1}{n}\sum_{i=1}^{n} \|\widehat{Z}_i - \Pi_V(\widehat{Z}_i)\|^2 \text{ and } W_d := \arg\min_{V \in \mathcal{V}_d} \mathbb{E}\|\bar{Z} - \Pi_V(\bar{Z})\|^2.$$

And the notations for reconstruction errors are

$$\bar{R}_n(V) = \frac{1}{n-1}\sum_{j=1}^{n} \left\|\widehat{Z}_j - \Pi_V\left(\widehat{Z}_j\right)\right\|^2 = \left\langle \Pi_{V^\perp}, \bar{C}_{1,n} \right\rangle,$$
$$\bar{R}(V) = \mathbb{E}\left\|\bar{Z} - \Pi_V(\bar{Z})\right\|^2 = \left\langle \Pi_{V^\perp}, \bar{C}_1 \right\rangle.$$

The result presented bears a striking resemblance to Theorem 3.1, with both sharing a similar underlying idea in their proofs.

Theorem 3.5. Assume that $\|Z\|^2 \leq M$ a.s. Then for any $\xi > 1$ and $n \geq 10$, with probability greater than $1 - 5e^{-\xi}$, the following inequality holds:

$$|\bar{R}(\widehat{W}_d) - \bar{R}_n(\widehat{W}_d)| \leq \sqrt{\frac{d}{n}\operatorname{tr}(C_{2,n} - C_{1,n} \otimes C_{1,n}^*)} + 14M\sqrt{\frac{\xi}{2n}} + 2M\frac{\sqrt{d}\xi^{\frac{1}{4}}}{n^{\frac{3}{4}}},$$

also, with probability at least $1 - 3e^{-\xi}$,

$$0 \leq \bar{R}(\widehat{W}_d) - \bar{R}(W_d) \leq \sqrt{\frac{d}{n}\operatorname{tr}(C_2 - C_1 \otimes C_1^*)} + 17M\sqrt{\frac{\xi}{n}}.$$

Proof. First we focus on the $|\bar{R}(\widehat{W}_d) - \bar{R}_n(\widehat{W}_d)|$ part. By (2) and (3), we have

$$|\bar{R}(\widehat{W}_d) - \bar{R}_n(\widehat{W}_d)| = |\langle \widehat{W}_d, \bar{C}_1 - \bar{C}_{1,n} \rangle| \leq \sup_{V \in \mathcal{V}_d} |\langle \Pi_{V^\perp}, \bar{C}_1 - \bar{C}_{1,n} \rangle|.$$

Denoting $\mu = \mathbb{E}[Z]$, recall the following identities:

$$\bar{C}_1 = C_1 - \mu \otimes \mu^* \text{ and } \bar{C}_{1,n} = C_{1,n} - \frac{1}{n(n-1)} \sum_{i \neq j}^{n} Z_i \otimes Z_j^*$$

By triangle inequality and separating the sup, we get

$$\sup_{V \in \mathcal{V}_d} \left| \left\langle \Pi_{V^\perp}, \bar{C}_{1,n} - \bar{C}_1 \right\rangle \right| \leq \sup_{V \in \mathcal{V}_d} \left| \left\langle \Pi_{V^\perp}, C_{1,n} - C_1 \right\rangle \right| + \sup_{V \in \mathcal{V}_d} \left| \left\langle \Pi_{V^\perp}, \mu \otimes \mu^* - \frac{1}{n(n-1)} \sum_{i \neq j} Z_i \otimes Z_j^* \right\rangle \right| \tag{18}$$

Though we do not have the assumption for the diameter of $Z \otimes Z$ to be L as in Theorem 3.1, we can use $2M$ to be the diameter since it can be easily checked that $\|Z \otimes Z^* - Z' \otimes Z'^*\|^2 \leq 2M$. Then Theorem 3.1 guarantees that with probability at least $1 - 3e^{-\xi}$, the first term of (18) is bounded by

$$\sup_{V \in \mathcal{V}_d} \left| \left\langle \Pi_{V^\perp}, C_{1,n} - C_1 \right\rangle \right| \leq \sqrt{\frac{d}{n}} \sqrt{\operatorname{tr}\left(C_{2,n} - C_{1,n} \otimes C_{1,n}^*\right)} + M\sqrt{\frac{\xi}{2n}} + 2M\frac{\sqrt{d}\xi^{\frac{1}{4}}}{n^{\frac{3}{4}}}, \tag{19}$$

so it remains to get a bound for the second term in the RHS of (18).

Letting $G(z_1, \ldots, z_n) = \left\langle \Pi_{V^\perp}, \mu \otimes \mu^* - \frac{1}{n(n-1)} \sum_{i \neq j} z_i \otimes z_j^* \right\rangle$, we will use McDiarmid's inequality for the second term, which need the following bound as a condition. For any $z_i's$ it follows that

$$\begin{aligned}
&\left| G(z_1, \ldots, z_n) - G\left(z_1, \ldots, z_{i_0-1}, z'_{i_0}, z_{i_0+1}, \ldots, z_n\right) \right| \\
&\qquad = \left\langle \Pi_{V^\perp}, \mu \otimes \mu^* - \frac{1}{n(n-1)} \sum_{i \neq j}^{(z'_{i_0})} z_i \otimes z_j^* \right\rangle - \left\langle \Pi_{V^\perp}, \mu \otimes \mu^* - \frac{1}{n(n-1)} \sum_{i \neq j}^{(z_{i_0})} z_i \otimes z_j^* \right\rangle,
\end{aligned}$$

since the terms which do not contain i_0 cancel,

$$= \frac{1}{n(n-1)} \left\langle \Pi_{V^\perp}, \sum_{j \neq i_0} \left[\left(z_{i_0} - z'_{i_0}\right) \otimes z_j^* + z_j \otimes \left(z_{i_0}^* - z'^*_{i_0}\right) \right] \right\rangle,$$

then by the facts in Section 1.1, property of orthogonal projectors and triangle inequality,

$$\begin{aligned}
&= \frac{1}{n(n-1)} \left\| \Pi_{V^\perp} \left[\sum_{j \neq i_0} \left[\left(z_{i_0} - z'_{i_0}\right) \otimes z_j^* + z_j \otimes \left(z_{i_0}^* - z'^*_{i_0}\right) \right] \right] \right\|^2 \\
&\leq \frac{1}{n(n-1)} \left\| \sum_{j \neq i_0} \left[\left(z_{i_0} - z'_{i_0}\right) \otimes z_j^* + z_j \otimes \left(z_{i_0}^* - z'^*_{i_0}\right) \right] \right\| \\
&\leq \frac{2}{n(n-1)} \sum_{j \neq i_0} \left\| z'_{i_0} - z_{i_0} \right\| \|z_j\| \\
&\leq \frac{4M}{n}.
\end{aligned}$$

Therefore, by applying McDiarmid's inequality, taking $c_i = \frac{4M}{n}$ and ξ in the Appendix to be $\frac{n\xi^2}{8M^2}$, we get

$$\sup_{V \in \mathcal{V}_d} \left\langle \Pi_{V^\perp}, \mu \otimes \mu^* - \frac{1}{n(n-1)} \sum_{i \neq j} Z_i \otimes Z_j^* \right\rangle$$

$$\leq \mathbb{E}\left[\sup_{V\in\mathcal{V}_d}\left\langle \Pi_{V^\perp}, \mu\otimes\mu^* - \frac{1}{n(n-1)}\sum_{i\neq j} Z_i\otimes Z_j^*\right\rangle\right] + 4M\sqrt{\frac{\xi}{2n}}. \tag{20}$$

To deal with the expectation term in (20), Hoeffding's decomposition is considered. We define

$$S_d = \sup_{V\in\mathcal{V}_d} \frac{2}{n}\sum_{j=1}^{n}\left(\langle \Pi_{V^\perp}, \mu\otimes\mu^*\rangle - \langle \Pi_{V^\perp}(Z_j), \mu\rangle\right),$$
$$R_d = \sup_{V\in\mathcal{V}_d} -\frac{1}{n(n-1)}\sum_{i\neq j}\left(\left\langle \Pi_{V^\perp}, Z_i\otimes Z_j^*\right\rangle - \langle \Pi_{V^\perp}(Z_j), \mu\rangle - \langle \Pi_{V^\perp}(Z_i), \mu\rangle + \langle \Pi_{V^\perp}, \mu\otimes\mu^*\rangle\right),$$

and by separating the sup, it follows that

$$\sup_{V\in\mathcal{V}_d}\left\langle \Pi_{V^\perp}, \mu\otimes\mu^* - \frac{1}{n(n-1)}\sum_{i\neq j} Z_i\otimes Z_j^*\right\rangle \leq S_d + R_d,$$

yielding the expectation in (20) to be bounded by

$$\mathbb{E}\left[\sup_{V\in\mathcal{V}_d}\left\langle \Pi_{V^\perp}, \mu\otimes\mu^* - \frac{1}{n(n-1)}\sum_{i\neq j} Z_i\otimes Z_j^*\right\rangle\right] \leq \mathbb{E}S_d + \mathbb{E}R_d. \tag{21}$$

Now, by (18), (19), (20) and (21), collecting terms, we have at least probability $1-5e^{-\xi}$,

$$|\bar{R}(\widehat{W}_d) - \bar{R}_n(\widehat{W}_d)| \leq \sqrt{\frac{d}{n}}\sqrt{\operatorname{tr}\left(C_{2,n} - C_{1,n}\otimes C_{1,n}^*\right)} + 5M\sqrt{\frac{\xi}{2n}} + 2M\frac{\sqrt{d}\xi^{\frac{1}{4}}}{n^{\frac{3}{4}}} + \mathbb{E}\left[S_d\right] + \mathbb{E}\left[R_d\right].$$

Please note that the term $3e^{-\xi}$ originates from (19), while the term $2e^{-\xi}$ is derived from (20) which is applied twice in order to establish a two-sided bound. The remaining portion of this section relies on the utilization of Lemma 3.6 and Lemma 3.7 from the paper. They provide bounds for

$$\mathbb{E}\left[S_d\right] \leq 4\frac{\mathbb{E}\|Z\|^2}{\sqrt{n}} \leq 6M\sqrt{\frac{\xi}{2n}} \text{ and } \mathbb{E}\left[R_d\right] \leq \frac{6}{n-1}\mathbb{E}\|Z\|^2 \leq 3M\sqrt{\frac{\xi}{2n}}, \tag{22}$$

leading to the final result of the $|\bar{R}(\widehat{W}_d) - \bar{R}_n(\widehat{W}_d)|$ part.

As for the $\bar{R}(\widehat{W}_d) - \bar{R}(W_d)$ part, the definition yields a natural bound:

$$0 \leq \bar{R}(\widehat{W}_d) - \bar{R}\left(W_d\right) \leq \left(\bar{R}(\widehat{W}_d) - \bar{R}_n(\widehat{W}_d)\right) + \left(\bar{R}_n\left(W_d\right) - \bar{R}\left(W_d\right)\right). \tag{23}$$

Since we only need to consider the one sided bound, a little modification leads to the bound of the first term in (23):

$$\begin{aligned}\bar{R}(\widehat{W}_d) - \bar{R}_n(\widehat{W}_d) &\leq \sup_{V\in\mathcal{V}_d}\langle \Pi_{V^\perp}, \bar{C}_1 - \bar{C}_{1,n}\rangle \\ &\leq \sup_{V\in\mathcal{V}_d}\langle \Pi_{V^\perp}, C_{1,n} - C_1\rangle + \sup_{V\in\mathcal{V}_d}\left\langle \Pi_{V^\perp}, \mu\otimes\mu^* - \frac{1}{n(n-1)}\sum_{i\neq j} Z_i\otimes Z_j^*\right\rangle,\end{aligned}$$

where by (16), we have with at least probability $1-e^{-\xi}$,

$$\sup_{V\in\mathcal{V}_d}\langle \Pi_{V^\perp}, C_{1,n}-C_1\rangle \leq \sqrt{\frac{d}{n}}\sqrt{\operatorname{tr}\left(C_2-C_1\otimes C_1^*\right)}+M\sqrt{\frac{\xi}{2n}}$$

and by (20), (21) and (22), we have with at least probability $1-e^{-\xi}$,

$$\sup_{V\in\mathcal{V}_d}\left\langle \Pi_{V^\perp}, \mu\otimes\mu^*-\frac{1}{n(n-1)}\sum_{i\neq j} Z_i\otimes Z_j^*\right\rangle \leq 13M\sqrt{\frac{\xi}{2n}}.$$

Summing up, we get with at least probability $1-2e^{-\xi}$,

$$0\leq \bar{R}(\widehat{W}_d)-\bar{R}_n(\widehat{W}_d)\leq \sqrt{\frac{d}{n}\operatorname{tr}(C_2-C_1\otimes C_1^*)}+14M\sqrt{\frac{\xi}{2n}}. \tag{24}$$

It remains to consider the second term in the RHS of (23). By the definitions, we can write

$$\hat{R}_n(W_d)-\hat{R}(W_d)=\left\langle \Pi_{W_d^\perp}, \hat{C}_{1,n}\right\rangle-\left\langle \Pi_{V^\perp}, \hat{C}_1\right\rangle=\left\langle \Pi_{W_d^\perp}, \hat{C}_{1,n}\right\rangle-\mathbb{E}\left\langle \Pi_{V^\perp}, \hat{C}_{1,n}\right\rangle,$$

where the second equality is due to $\mathbb{E}\bar{C}_{1,n}=\bar{C}_1$. Now letting

$$g\left(z_1,z_2\right)=\frac{1}{2}\left\langle z_1-z_2, \Pi_{W_d^\perp}\left(z_1-z_2\right)\right\rangle,$$

simple algebra shows

$$\left\langle \Pi_{W_d^\perp}, \bar{C}_{1,n}\right\rangle=\frac{1}{n(n-1)}\sum_{i\neq j} g\left(Z_i,Z_j\right)=:U.$$

By the property of orthogonal projector, $g(Z_1,Z_2)\in[0,M]$. Hence, Hoeffding's inequality, taking $r=2$, $t=M\sqrt{\frac{\xi}{n}}$ and U denoted above, yields with probability at least $1-e^{-\xi}$,

$$0\leq \bar{R}_n\left(W_d\right)-\bar{R}\left(W_d\right)=U-\mathbb{E}U\leq M\sqrt{\frac{\xi}{n}}\leq 2M\sqrt{\frac{\xi}{2n}}. \tag{25}$$

Combining (23), (24) and (25), the final result follows: with at least probability $1-3e^{-\xi}$,

$$0\leq \bar{R}(\widehat{W}_d)-\bar{R}\left(W_d\right)\leq \sqrt{\frac{d}{n}\operatorname{tr}(C_2-C_1\otimes C_1^*)}+16M\sqrt{\frac{\xi}{2n}}$$

□

It is important to acknowledge that there is a discrepancy in the constant appearing in my reformulation, which is 16, compared to the value stated in the paper as 17. Similar issues with constant discrepancy arise in the subsequent theorems as well. Although I suspect that there may have been an error in collecting terms that I have yet to identify, this discrepancy does not hinder my understanding of the theorem and the overall proof outline. The crucial point to grasp is that regardless of this difference in constant value, the convergence property exhibits the same order and rate. Thus, whether empirical centering is applied before PCA or not, the fundamental behavior remains consistent.

3 Local Approach

In my perspective, the paper's notable contribution lies in the introduction of the *local* approach, which takes variance considerations into account and yields enhanced bounds. This approach represents a significant advancement when compared to the global approach. Furthermore, the authors' choice of applying the *localization* technique to two classes of functions allows for improvements in two directions, as we will explore in the following sections.

3.1 Prerequisites

The advancements presented in the paper do entail certain trade-offs. Specifically, these advancements necessitate the introduction of a complex quantity and an additional theorem. This section serves the purpose of establishing the prerequisite understanding needed for the subsequent sections. Sections 3.2 and 3.3 will delve into the main results of the paper.

For a compact, positive and self-adjoint operator L, we denote $\lambda(L) = (\lambda_1(L) \geq \lambda_2(L) \geq \dots)$ the sequence of its positive eigenvalues. And a key quantity that will be used frequently in this section is the following:

$$\rho(A, d, n) = \inf_{h \geq 0} \left\{ A\frac{h}{n} + \sqrt{\frac{d}{n} \sum_{j>h} \lambda_j\left(C_2'\right)} \right\}, \quad \text{where } C_2' = C_2 - C_1 \otimes C_1^*.$$

The subsequent results consistently outperform those in Theorem 3.1 due to

(a) the condition $\rho(A, d, n) < \sqrt{\frac{d}{n}\operatorname{tr}(C_2')}$ for all A,

(b) and the behavior of ρ as a power of n vary from $n^{-\frac{1}{2}}$ to n^{-1}, depending on the decay behavior of the eigenvalues of C_2'.

Remark. Although the subsequent results exhibit promising performance, the computation and comparison of the quantity ρ pose challenges. Though the theoretical way to show $\rho(A, d, n) < \sqrt{\frac{d}{n}\operatorname{tr}(C_2')}$ remains unclear to me, here below is an example under the case $n = 1000$, $p = 100$ and $d = 10$. And all the cases that have been tried show the same pattern.

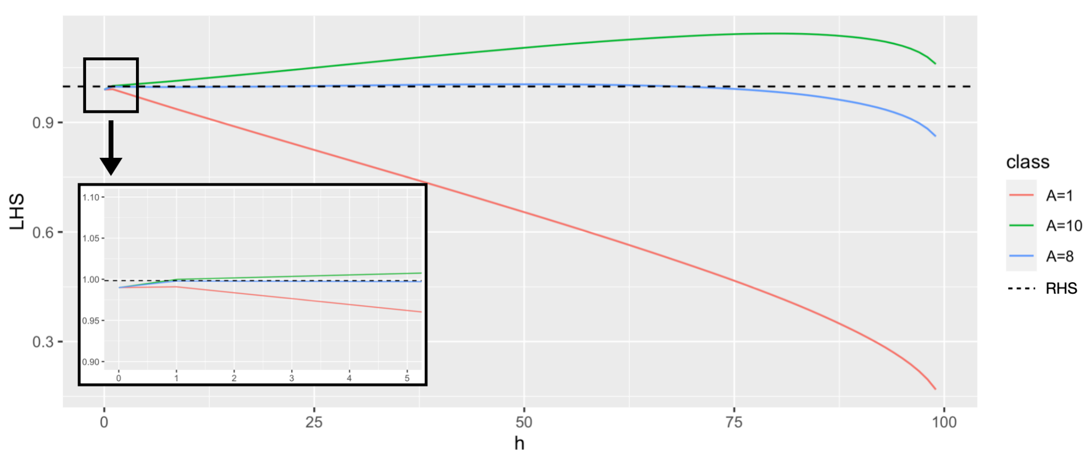

Figure 2: An Illustration of $\rho(A, d, n) < \sqrt{\frac{d}{n}\operatorname{tr}(C_2')}$

The core of our main results relies on a fundamental theorem by Bartlett et al, which serves as a pivotal component in our derivation. To state the theorem, we need the following notation: let $\mathcal{X}$ be a measurable space and $X_1, \ldots, X_n$ a n-uple of points in $\mathcal{X}$. For a class of functions $\mathcal{F}$ from $\mathcal{X}$ to $\mathbb{R}$, we denote

$$\mathcal{R}_n \mathcal{F} := \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f\left(X_i\right),$$

where $(\epsilon_i)_{i=1}^n$ are iid Rademacher variables. The *star-shaped hull* of a class of functions $\mathcal{F}$ is defined as

$$\operatorname{star}(\mathcal{F}) = \{\lambda f \mid f \in \mathcal{F}, \lambda \in [0,1]\}.$$

And a function $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is called *sub-root* if it is nonnegative, nondecreaing, and $\psi(r)/\sqrt{r}$ is nonincreasing. It can be shown that $\psi(r) = r$ has a unique positive solution (trivial case $\psi \equiv 0$ is excluded), and the solution is denoted by r^* and called the fixed point.

Theorem 3.3. Let $\mathcal{X}$ be a measurable space, P be a probability distribution on $\mathcal{X}$ and $X_1, \ldots, X_n$ an i.i.d. sample from P. Let $\mathcal{F}$ be a class of functions on X ranging in $[-1, 1]$ and assume that there exists some constant $B > 0$ such that for every $f \in \mathcal{F}, Pf^2 \leq BPf$. Let ψ be a sub-root function and r^* be the fixed point of ψ. If ψ satisfies

$$\psi(r) \geq B\mathbb{E}_{X,\varepsilon} \mathcal{R}_n \left\{f \in \operatorname{star}(\mathcal{F}) \mid Pf^2 \leq r\right\} \tag{26}$$

then for any $K > 1$ and $x > 0$, with probability at least $1 - e^{-x}$,

$$\forall f \in \mathcal{F}, \quad Pf \leq \frac{K}{K-1} P_n f + \frac{6K}{B} r^* + \frac{x(11 + 5BK)}{n} \tag{27}$$

also, with probability at least $1 - e^{-x}$

$$\forall f \in \mathcal{F}, \quad P_n f \leq \frac{K+1}{K} Pf + \frac{6K}{B} r^* + \frac{x(11 + 5BK)}{n} \tag{28}$$

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

$$\widehat{\psi}_n(r) \geq 2(10 \vee B)\mathbb{E}_{\varepsilon} \mathcal{R}_n \left\{f \in \operatorname{star}(\mathcal{F}) \mid P_n f^2 \leq 2r\right\} + \frac{(2(10 \vee B) + 11)x}{n} \tag{29}$$

then with probability $1 - 2e^{-x}$, it holds that $\widehat{r}^* \geq r^*$; as a consequence, with probability $1 - 3e^{-x}$, inequality (27) holds with r^* replaced by $\widehat{r}^*$; similarly for inequality (28).

As this theorem serves as a lemma for our main objective, the proof is not given in this paper.

3.2 Better Behavior as A Function of n

In this subsection, we will apply the localization technique to the excess loss class, leading to a result which behaves better than Theorem 3.1 as $n \to \infty$.

Theorem 3.2. Assume $\|Z\|^2 \leq M$ a.s. Let (λ_i) denote the ordered eigenvalues with multiplicity of

C_1, resp. (μ_i) the ordered distinct eigenvalues. Let $\tilde{d}$ be such that $\lambda_d = \mu_{\tilde{d}}$. Define

$$\gamma_d = \begin{cases} \mu_{\tilde{d}} - \mu_{\tilde{d}+1} & \text{if } \tilde{d} = 1 \text{ or } \lambda_d > \lambda_{d+1}, \\ \min\left(\mu_{\tilde{d}-1} - \mu_{\tilde{d}}, \mu_{\tilde{d}} - \mu_{\tilde{d}+1}\right) & \text{otherwise} \end{cases}$$

and $B_d = \left(\mathbb{E}\left\langle Z, Z'\right\rangle^4\right)^{\frac{1}{2}} \gamma_d^{-1}$ (where Z' is an independent copy of Z).

Then for all d, for all $\xi > 0$, with probability at least $1 - e^{-\xi}$ the following holds:

$$R(\widehat{V}_d) - R\left(V_d\right) \leq 24\rho\left(B_d, d, n\right) + \frac{\xi\left(11M + 7B_d\right)}{n}.$$

Before we start the proof, it is worth noting that when dealing with the multiplicity of eigenvalues, specifically when $\lambda_d(C_1)$ has a multiplicity greater than 1, the uniqueness of $V_d = \arg\min_{V \in \mathcal{V}_d} R(V)$ is not preserved. In the proofs presented in the previous section, this non-uniqueness does not affect the overall results. Nevertheless, for the current proof, the choice of V_d does matter. While the selection of V_d has no impact on the expectation Pf, it does influence the value of Pf^2, which is a part of the assumptions of Theorem 3.3. Fortunately, for each $V \in \mathcal{V}_d$, there exists an element $H_V \in \mathcal{V}_d$ such that

$$R\left(H_V\right) = \min_{H \in \mathcal{V}_d} R(H) = R\left(V_d\right),$$

and

$$\mathbb{E}\left[\left\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_Z \right\rangle^2\right] \leq 2B_d \mathbb{E}\left[\left\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_Z \right\rangle\right]. \tag{30}$$

This is guaranteed in Lemma A.1., and we will use H_V to avoid the Pf^2 issue.

Proof. Consider the class of functions

$$\mathcal{F}_d = \left\{ f_V : x \mapsto \left\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_x \right\rangle \mid V \in \mathcal{V}_d \right\},$$

where for each $V \in \mathcal{V}_d$, H_V is obtained by Lemma A.1. We will apply Theorem 3.3 to the class $M^{-1}\mathcal{F}_d$, and now we check the condition for it.

By the assumption $\|Z\|^2 \leq M$ a.s., suppose $z \in \mathscr{H}$ with $\|z\|^2 \leq M$. It follows that for any $V \in \mathcal{V}_d$,

$$f_V(z) = \left\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_z \right\rangle = \left\langle \Pi_{V^\perp}, C_z \right\rangle - \left\langle \Pi_{H_V^\perp}, C_z \right\rangle = \left\langle \Pi_{H_V}(z), z \right\rangle - \left\langle \Pi_V(z), z \right\rangle.$$

By the property of orthogonal projector, we have $\langle \Pi_{H_V}(z), z\rangle \leq \|z\|^2 \leq M$ and $\langle \Pi_V(z), z\rangle \leq \|z\|^2 \leq M$, yielding

$$-M \leq f_V(z) \leq M.$$

Since this bound holds for any choice of $V \in \mathcal{V}_d$, we claim $\forall g \in M^{-1}\mathcal{F}_d$, it holds that $g \in [-1, 1]$, verifying the first condition in Theorem 3.3. The second condition requires us to show $\forall g \in M^{-1}\mathcal{F}_d$, there exists a constant B such that $\mathbb{E}[g(Z)^2] \leq B\mathbb{E}[g(Z)]$. Since the function $g(\cdot)$ takes the form

$$g(Z) = \frac{1}{M}\left\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_Z \right\rangle,$$

we can write

$$\mathbb{E}\left[g(Z)^2\right] = \frac{1}{M^2}\mathbb{E}\left\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_Z \right\rangle^2 \leq \frac{2B_d}{M^2}\cdot \mathbb{E}\left\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_Z \right\rangle = \frac{2B_d}{M}\mathbb{E}[g(Z)]$$

where the second inequality comes from (30). And we can see that the constant $B = 2B_d/M$ in this proof.

The remaining task is to verify the last condition, which involves finding a sub-root function $\psi(r)$ and its fixed point r^* that satisfy the inequality:

$$\psi(r) \geq 2B_d\mathbb{E}_{Z,\varepsilon}\mathcal{R}_n\left\{f \in \operatorname{star}(\mathcal{F}) \mid Pf^2 \leq r\right\}, \tag{31}$$

Once we establish this, we can apply Theorem 3.3 to obtain the final result. In the next claim, I will present the desired form of $\psi(r)$ and utilize it to complete the main proof. The proof of the claim will be provided after the main proof.
Claim. The following function $\psi(r)$ satisfies (31):

$$\psi(r) = \frac{2B_d}{M}\left[\psi_0(r) + \sqrt{\frac{r}{n}}\right] \quad \text{where} \quad \psi_0(r) = \frac{1}{\sqrt{n}}\inf_{h\geq 0}\left(\sqrt{rh} + 2M^{-1}\sqrt{d\sum_{j\geq h+1}\lambda_j\left(C_2'\right)}\right),$$

with the fixed point

$$r^* \leq 8M^{-2}\left(\inf_{h\geq 0}\left\{\frac{B_d^2 h}{n} + B_d\sqrt{\frac{d}{n}\sum_{j\geq h+1}\lambda_j\left(C_2'\right)}\right\} + \frac{B_d^2}{n}\right) = 8M^{-2}\left[B_d\rho(B_d,d,n) + \frac{B_d^2}{n}\right]. \tag{32}$$

Let's continue the main proof. By applying Theorem 3.3 with $B = 2B_dM^{-1}$, we get $\forall K > 1$ and $\forall \xi > 0$,

$$\forall g \in M^{-1}\mathcal{F}_d: \quad Pg \leq \frac{K}{K-1}P_n g + \frac{3K}{B_dM^{-1}}r^* + \frac{x(11 + 10B_dM^{-1}K)}{n}$$

with at least probability $1 - e^{-\xi}$. By the definition of $\mathcal{F}_d$, it equivalently means that

$$\forall V \in \mathcal{V}_d: \quad P\left(\frac{f_V}{M}\right) \leq \frac{K}{K-1}P_n\left(\frac{f_V}{M}\right) + \frac{3K}{B_dM^{-1}}r^* + \frac{x(11 + 10B_dM^{-1}K)}{n}.$$

Some simplification and plugging (32), we get with at least probability $1 - e^{-\xi}$,

$$\forall V \in \mathcal{V}_d, \quad Pf_V \leq \frac{K}{K-1}P_nf_V + 24K\rho\left(B_d,d,n\right) + \frac{\xi\left(11M + 34B_dK\right)}{n}. \tag{33}$$

Now we choose $V = \widehat{V}_d$. Note that

$$P_nf_{\widehat{V}_d} = P_n\left\langle \Pi_{\widehat{V}_d^\perp} - \Pi_{H_{\widehat{V}_d^\perp}}, C_Z\right\rangle = P_n\left\langle \Pi_{\widehat{V}_d^\perp}, C_Z\right\rangle - P_n\left\langle \Pi_{H_{\widehat{V}_d^\perp}}, C_Z\right\rangle \leq 0, \tag{34}$$

where the last equality is from the definition: $\hat{V}_d = \arg\min R_n(V) = \arg\min\langle \Pi_{V^\perp}, C_{1,n}\rangle$. Also note that

$$Pf_{\widehat{V}_d} = \mathbb{E}\left\{\left\langle \Pi_{\widehat{V}_d^\perp} - \Pi_{H_{\widehat{V}_d^\perp}}, C_Z\right\rangle\right\} = \mathbb{E}\left\langle \Pi_{\widehat{V}_d^\perp}, C_Z\right\rangle - \mathbb{E}\left\langle \Pi_{H_{\widehat{V}_d^\perp}}, C_Z\right\rangle = R(\widehat{V}_d) - R(V_d), \tag{35}$$

where the last equality is from the facts in Section 1.1 and $R\left(H_V\right) = R\left(V_d\right)$. Combining the (33), (34) and

(35), we get for any $K > 1$ and $\xi > 0$, with at least probability $1 - e^{-\xi}$,

$$R(\widehat{V}_d) - R(V_d) \leq 24K\rho\left(B_d, d, n\right) + \frac{\xi\left(11M + 34B_dK\right)}{n}.$$

Letting $K \to 1$, our desired result follows. $\square$

Remark. Although the constant 34 used in this claim does not match the value of 7 stated in the paper, which may be due to an error on my part, the overall structure and approach remain consistent with the paper. The remaining task is to prove the claim used in the theorem, which is a non-trivial step.

Proof of Claim. Note that $\Pi_{V^\perp} - \Pi_{H_V^\perp} = \Pi_{H_V} - \Pi_V$ and

$$\left\|\Pi_V - \Pi_{H_V}\right\|^2 = \left\|\Pi_V\right\|^2 + \left\|\Pi_{H_V}\right\|^2 - 2\left\langle \Pi_V, \Pi_{H_V}\right\rangle \leq 4d,$$

where the last inequality comes from $V, H_V \in \mathcal{V}_d$. Therefore $\mathcal{F}_d$ is contained in a bigger collection of functions (by relaxing the condition):

$$\mathcal{F}_d = \left\{x \mapsto \left\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_x\right\rangle \mid V \in \mathcal{V}_d\right\} \subset \left\{x \mapsto \langle \Gamma, C_x\rangle \mid \Gamma \in \mathrm{HS}(\mathscr{H}), \|\Gamma\|^2 \leq 4d\right\}.$$

Since the latter set is convex and contains the origin, it therefore also contains $\mathrm{star}(\mathcal{F}_d)$, i.e.

$$\mathrm{star}(\mathcal{F}_d) \subset \left\{x \mapsto \langle \Gamma, C_x\rangle \mid \Gamma \in \mathrm{HS}(\mathscr{H}), \|\Gamma\|^2 \leq 4d\right\}. \tag{36}$$

Also note that for a function of the form $f(x) = \langle \Gamma, C_x\rangle$, we can compute

$$Pf^2 = \mathbb{E}\left[\langle \Gamma, C_Z\rangle^2\right] = \mathbb{E}\left[\langle \Gamma, \langle \Gamma, C_Z\rangle C_Z\rangle\right] = \mathbb{E}\left\langle \Gamma, C_Z \otimes C_Z^*(\Gamma)\right\rangle = \langle \Gamma, C_2\Gamma\rangle.$$

Therefore the condition $Pf^2 \leq r$ is equivalent to $\langle \Gamma, C_2\Gamma\rangle \leq r$, together with (36), leading to

$$\begin{aligned}\left\{g \in \mathrm{star}\left(M^{-1}\mathcal{F}_d\right) \mid Pg^2 \leq r\right\} &= M^{-1}\left\{g \in \mathrm{star}\left(\mathcal{F}_d\right) \mid Pg^2 \leq M^2 r\right\} \\ &\subset M^{-1}\left\{x \mapsto \langle \Gamma, C_x\rangle \mid \|\Gamma\|^2 \leq 4d, \langle \Gamma, C_2\Gamma\rangle \leq M^2 r\right\} := \mathcal{S}_r.\end{aligned} \tag{37}$$

Since our goal now is to upper bound $\mathbb{E}\mathbb{E}_\epsilon \mathcal{R}_n \mathcal{S}_r$, where we recall $\mathcal{R}_n\mathcal{F} = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i f\left(X_i\right)$, by splitting the sup and relaxing some conditions, we can get

$$\mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{S}_r \leq \mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{S}_{1,r} + \mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{S}_{2,r}, \tag{38}$$

where we define the set of constant functions

$$\mathcal{S}_{1,r} = M^{-1}\left\{x \mapsto \langle \Gamma, C_1\rangle \mid \langle \Gamma, C_2\Gamma\rangle \leq M^2 r\right\},$$

and the set of centered functions

$$\mathcal{S}_{2,r} = M^{-1}\left\{x \mapsto \langle \Gamma, C_x - C_1\rangle \mid \|\Gamma\|^2 \leq 4d, \left\langle \Gamma, \left(C_2 - C_1 \otimes C_1^*\right)\Gamma\right\rangle \leq M^2 r\right\}.$$

Note that in $\mathcal{S}_{2,r}$, the condition $\langle \Gamma, C_2\Gamma\rangle \leq M^2 r$ is replaced by $\left\langle \Gamma, \left(C_2 - C_1 \otimes C_1^*\right)\Gamma\right\rangle \leq M^2 r$. We can do it since $\langle \Gamma, C_1 \otimes C_1^*(\Gamma) \geq 0\rangle$.

By (36) and (38), it suffices to upper bound $\mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{S}_{1,r}$ and $\mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{S}_{2,r}$. The bound for the first class of functions is easy since it only contains constant functions. The following equation can be checked easily:

for a set of scalars $A \subset \mathbb{R}$,

$$\mathbb{E}\left[\sup_{a\in A}\left(a\sum_{i=1}^{n}\varepsilon_i\right)\right] = \frac{1}{2}(\sup A - \inf A)\mathbb{E}\left\{\left|\sum_{i=1}^{n}\varepsilon_i\right|\right\} \le \frac{1}{2}(\sup A - \inf A)\sqrt{n}, \tag{39}$$

leading to

$$\begin{aligned}
\mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{S}_{1,r} &\le \mathbb{E}\mathbb{E}_\epsilon \sup\left\{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i M^{-1}\left\langle \Gamma, C_1\right\rangle \mid \left\langle \Gamma, C_2\Gamma\right\rangle \le M^2 r\right\} \\
&= \mathbb{E}_\epsilon \sup\left\{\frac{1}{n}\sum_{i=1}^{n}\left\langle \Gamma, C_1\right\rangle M^{-1}\epsilon_i \mid \left\langle \Gamma, C_2\Gamma\right\rangle \le M^2 r\right\} \\
[\text{By}(39)] \quad &\le \frac{1}{2n}M\left(\sup\left\{\langle\Gamma, C_1\rangle \mid \langle\Gamma, C_2\Gamma\rangle \le M^2 r\right\} - \inf\left\{\langle\Gamma, C_1\rangle \mid \langle\Gamma, C_2\Gamma\rangle \le M^2 r\right\}\right)\sqrt{n} \\
&\le M\sqrt{n}\sup\left\{\langle\Gamma, C_1\rangle \mid \langle\Gamma, C_2\Gamma\rangle \le M^2 r\right\} \le \sqrt{\frac{r}{n}},
\end{aligned} \tag{40}$$

where the last inequality comes from

$$\langle\Gamma, C_2\Gamma\rangle = \mathbb{E}\left[\langle C_Z, \Gamma\rangle^2\right] \ge \mathbb{E}\left[\langle C_Z, \Gamma\rangle\right]^2 = \langle\Gamma, C_1\rangle^2 .$$

In the following, our goal is to upper bound $\mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{S}_{2,r}$, which is technical. By the above equation, we can tell $C_2' = C_2 - C_1 \otimes C_1^*$ is self-adjoint. Therefore, we can find an orthogonal basis (Φ_i) of eigenfunctions of the operator C_2' with eigenvalues (λ_i). By introducing this basis, the constraints in $\mathcal{S}_{2,r}$ are equivalent to

$$\|\Gamma\|^2 = \sum_i \langle\Gamma, \Phi_i\rangle^2 \le 4d, \quad \text{and} \quad \langle\Gamma, C_2'\Gamma\rangle = \sum_i \lambda_i\left(C_2'\right)\langle\Gamma, \Phi_i\rangle^2 \le M^2 r. \tag{41}$$

Then for any Γ satisfying the above constraints, and any integer $h \le \operatorname{rank}(C_2')$, we can upper bound

$$\begin{aligned}
&\sum_{j=1}^{n}\varepsilon_j\left\langle\Gamma, C_{Z_j} - C_1\right\rangle = \sum_{j=1}^{n}\varepsilon_j\left\langle\sum_{i=1}^{\infty}\langle\Gamma, \Phi_i\rangle\Phi_i, C_{Z_j} - C_1\right\rangle \\
&\quad = \sum_{i=1}^{h}\langle\Gamma, \Phi_i\rangle\left\langle\Phi_i, \sum_{j=1}^{n}\varepsilon_j\left(C_{Z_j} - C_1\right)\right\rangle + \sum_{i>h}\langle\Gamma, \Phi_i\rangle\left\langle\Phi_i, \sum_{j=1}^{n}\varepsilon_j\left(C_{Z_j} - C_1\right)\right\rangle \\
&\quad = \sum_{i=1}^{h}\sqrt{\lambda_i}\langle\Gamma, \Phi_i\rangle\frac{\left\langle\Phi_i, \sum_{j=1}^{n}\varepsilon_j\left(C_{Z_j} - C_1\right)\right\rangle}{\sqrt{\lambda_i}} + \sum_{i>h}\langle\Gamma, \Phi_i\rangle\left\langle\Phi_i, \sum_{j=1}^{n}\varepsilon_j\left(C_{Z_j} - C_1\right)\right\rangle,
\end{aligned}$$

by Cauchy-Schwartz inequality and using (41),

$$\begin{aligned}
&\le \sqrt{\sum_{i=1}^{h}\frac{\left\langle\Phi_i, \sum_{j=1}^{n}\varepsilon_j(C_{Z_j} - C_1)\right\rangle^2}{\lambda_i}} \cdot \sqrt{\sum_{i=1}^{h}\lambda_i\langle\Gamma, \Phi_i\rangle^2} + \sqrt{\sum_{i>h}\left\langle\Phi_i, \sum_{j=1}^{n}\varepsilon_j\left(C_{Z_j} - C_1\right)\right\rangle^2} \cdot \sqrt{\sum_{i>h}\langle\Gamma, \Phi_i\rangle^2} \\
&\le M\sqrt{r\sum_{i=1}^{h}\frac{\left\langle\Phi_i, \sum_{j=1}^{n}\varepsilon_j(C_{Z_j} - C_1)\right\rangle^2}{\lambda_i}} + 2\sqrt{d\sum_{i\ge h+1}\left\langle\Phi_i, \sum_{j=1}^{n}\varepsilon_j\left(C_{Z_j} - C_1\right)\right\rangle^2}.
\end{aligned} \tag{42}$$

Both terms in the last equation contain a common element $\left\langle \Phi_i, \sum_{j=1}^n \varepsilon_j \left(C_{Z_j} - C_1\right)\right\rangle^2$, and we can compute

$$\mathbb{E}\mathbb{E}_\varepsilon \left\langle \Phi_i, \sum_{j=1}^n \varepsilon_j \left(C_{Z_j} - C_1\right)\right\rangle^2 = \mathbb{E}\left\{\mathbb{E}_\varepsilon \left[\sum_{j=1}^n \left\langle \Phi_i, \left(C_{Z_j} - C_1\right)\right\rangle \varepsilon_j\right]^2\right\},$$

by Rademacher distribution of independent ε_i's,

$$= \mathbb{E}\left\{\sum_{j=1}^n \left\langle \Phi_i, \left(C_{Z_j} - C_1\right)\right\rangle^2\right\},$$

where the next equality is tedious but trivial, and it will be checked after the main proof,

$$\begin{aligned} &= n\mathbb{E}\left\langle \Phi_i, \left(C_{2,n} - C_1 \otimes C_1^*\right)\Phi_i\right\rangle \\ &= n\left\langle \Phi_i, C_2'\Phi_i\right\rangle = n\lambda_i\left(C_2'\right). \end{aligned} \tag{43}$$

Using the above knowledge (42) and (43) with Jensen's inequality (expectation will switch the order with square root), we can upper bound $\mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{S}_{2,r}$ by

$$\begin{aligned} \mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{S}_{2,r} &= \mathbb{E}\mathbb{E}_\varepsilon \sup_{\mathcal{S}_{2,r}} \frac{1}{n}\sum_{j=1}^n \frac{\varepsilon_j}{M}\left\langle \Gamma, C_{Z_j} - C_1\right\rangle \\ &= \frac{1}{nM}\mathbb{E}\mathbb{E}_\varepsilon \sup_{\mathcal{S}_{2,r}} \sum_{j=1}^n \varepsilon_j \left\langle \Gamma, C_{Z_j} - C_1\right\rangle \\ &\le \frac{1}{\sqrt{n}}\left(\sqrt{rh} + 2M^{-1}\sqrt{d\sum_{j\ge h+1}\lambda_j}\right). \end{aligned}$$

Since this inequality holds for all choice of $h \le \operatorname{rank}(C_2')$, we actually have

$$\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{S}_{2,r} \le \frac{1}{\sqrt{n}} \inf_{0\le h\le \operatorname{rank}(C_2')} \left(\sqrt{rh} + 2M^{-1}\sqrt{d\sum_{j\ge h+1}\lambda_j\left(C_2'\right)}\right).$$

Since C_2' is a covariance operator, its ordered eigenvalue sequence is either finite or decreasing to zero. So, we can generalize it and get

$$\mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{S}_{2,r} \le \frac{1}{\sqrt{n}} \inf_{h\ge 0} \left(\sqrt{rh} + 2M^{-1}\sqrt{d\sum_{j\ge h+1}\lambda_j\left(C_2'\right)}\right) := \psi_0(r). \tag{44}$$

Combining (37), (38), (40) and (44) yields

$$\mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \left\{g \in \operatorname{star}\left(M^{-1}\mathcal{F}_d\right) \mid Pg^2 \le r\right\} \le \mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{S}_r \le \mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{S}_{1,r} + \mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{S}_{2,r} \le \sqrt{\frac{r}{n}} + \psi_0(r).$$

Recall that the constant B in Theorem 3.2 is chosen to be $2B_d M^{-1}$. Therefore the choice of

$$\psi(r) = \frac{2B_d}{M}\left[\psi_0(r) + \sqrt{\frac{r}{n}}\right]$$

satisfies (31). To upper bound the fixed point r^* as (32), we only need to solve for all $h \geq 0$,

$$r^* \leq \frac{2M^{-1}B_d}{\sqrt{n}} \left\{ \left(h^{\frac{1}{2}}+1\right)\sqrt{r^*} + 2M^{-1}\sqrt{d\sum_{j\geq h+1}\lambda_j} \right\},$$

where the RHS is the ψ but excluding the infimum. Using the fact that $x \leq A\sqrt{x}+B$ implies $x \leq A^2+2B$, and the fact $(\sqrt{h}+1)^2 \leq 2h+2$ yields for all $h > 0$:

$$\begin{aligned}
r^* &\leq \frac{4M^{-2}B_d^2}{n}\left(\sqrt{h}+1\right)^2 + \frac{8M^{-2}B_d}{\sqrt{n}}\sqrt{d\sum_{j\geq h+1}\lambda_j} \\
&\leq \frac{4M^{-2}B_d^2}{n}(2h+2) + \frac{8M^{-2}B_d}{\sqrt{n}}\sqrt{d\sum_{j\geq h+1}\lambda_j} \\
&\leq 8M^{-2}\left\{ \frac{B_d^2 h}{n} + \frac{B_d^2}{n} + B_d\sqrt{d\sum_{j\geq h+1}\lambda_j} \right\}.
\end{aligned}$$

Taking infimum of all $h > 0$, we finally get

$$r^* \leq 8M^{-2}\left(\inf_{h\geq 0}\left\{ \frac{B_d^2 h}{n} + B_d\sqrt{\frac{d}{n}\sum_{j\geq h+1}\lambda_j\left(C_2'\right)} \right\} + \frac{B_d^2}{n} \right).$$

□

Verification of the the omitted equality. The remaining task is to show

$$\mathbb{E}\left\{ \sum_{j=1}^{n}\left\langle \Phi_i, \left(C_{Z_j}-C_1\right)\right\rangle^2 \right\} = n\mathbb{E}\left\langle \Phi_i, \left(C_{2,n}-C_1\otimes C_1^*\right)\Phi_i\right\rangle.$$

We start by computing from the LHS without the expectation:

$$\sum_{j=1}^{n}\left\langle \Phi_i, \left(C_{Z_j}-C_1\right)\right\rangle^2 = \sum_{j=1}^{n}\left\{ \left\langle C_{Z_j}, \Phi_i\right\rangle C_{Z_j} - \left\langle C_{Z_j}, \Phi\right\rangle C_1 - \left\langle C_1, \Phi\right\rangle C_{Z_j} + \left\langle C_1, \Phi\right\rangle C_1 \right\}.$$

Leaving the expectation, the RHS can be also computed as

$$n\left\langle \Phi_i, \left(C_{2,n}-C_1\otimes C_1^*\right)\Phi_i\right\rangle = \sum_{j=1}^{n}(C_{Z_j}\otimes C_{Z_j}^* - C_1\otimes C_1^*)(\Phi_i) = \sum_{j=1}^{n}\left\langle C_{Z_j}, \Phi_i\right\rangle C_{Z_j} - n\left\langle C_1, \Phi\right\rangle C_1.$$

Taking expectation on both LHS and RHS, we can easily tell they are equal. □

Remark. The crux of this paper lies in Theorem 3.3, which plays a crucial role in enhancing the bounds presented in both Theorem 3.2 and 3.5. However, it also makes the proof significantly more challenging to follow and think of for me.

3.3 Better Behavior as A Function of d

In this subsection, we employ the localization technique to examine the initial loss class, leading to an improved result compared to Theorem 3.1 when the dimension d is large (as shown in the example of Figure

1, where d exceeds 10). It is noteworthy that this result is obtained under a stronger condition, namely $\|Z\|^2 = M$ almost surely.

Theorem 3.4. Assume Z takes values on the sphere of radius $\sqrt{M}$, i.e. $\|Z\|^2 = M$ a.s. Then for all $d, n \geq 2, \xi > 0$, with probability at least $1 - 4e^{-\xi}$ the following holds:

$$|R(\widehat{V}_d) - R_n(\widehat{V}_d)| \leq c \left[\sqrt{R_n(\widehat{V}_d) \left(\rho_n(M,d,n) + M\frac{(\xi + \log n)}{n} \right)} + \rho_n(M,d,n) + \frac{M(\xi + \log n)}{n} \right],$$

where c is a universal constant $\left(c \leq 1.2 \times 10^5\right)$, and ρ_n is defined as the empirical counterpart of ρ:

$$\rho_n(A,d,n) = \inf_{h \geq 0} \left\{ A\frac{h}{n} + \sqrt{\frac{d}{n} \sum_{j>h} \lambda_j \left(C'_{2,n} \right)} \right\}, \quad \text{where} \;\; C'_{2,n} = C_{2,n} - C_{1,n} \otimes C^*_{1,n}.$$

Also, with probability at least $1 - 2e^{-\xi}$,

$$R(\widehat{V}_d) - R\left(V_d\right) \leq c \left[\sqrt{R\left(V_d\right) \left(\rho(M,d,n) + M\frac{\xi}{n} \right)} + \rho(M,d,n) + M\frac{\xi}{n} \right],$$

where c is a universal constant ($c \leq 80$).

Proof. We first focus on the $R(\widehat{V}_d) - R\left(V_d\right)$ part. To prove this result, we will apply Theorem 3.3 to the class of functions $M^{-1}\mathcal{G}_d$, where

$$\mathcal{G}_d = \left\{ f_V : x \mapsto \langle \Pi_{V^\perp}, C_x \rangle \mid V \in \mathcal{V}_d \right\}.$$

Now let's check the condition for Theorem 3.3. By the properties of orthogonal projector and $\mathcal{G}_d$, we have for any function $g \in M^{-1}\mathcal{G}_d$ and $z \in \mathscr{H}$ s.t. $\|z\|^2 = M$,

$$|g(z)| = M^{-1} \langle \Pi_{V^\perp}, C_z \rangle = M^{-1} \|\Pi_{V^\perp}(z)\|^2 \leq M^{-1}\|z\|^2 = 1,$$

and by $0 \leq g(Z) \leq 1$ a.s., we also tell $0 \leq g(Z)^2 \leq g(Z)$ a.s., and hence

$$Pg^2 = \mathbb{E} \langle \Pi_{V^\perp}, C_Z \rangle^2 \leq \mathbb{E} \langle \Pi_{V^\perp}, C_Z \rangle = Pg,$$

i.e. the first two conditions for Theorem 3.3 are satisfied. And we can see that the constant $B = 1$ in this theorem. The remaining task is to find the sub-root function satisfying

$$\psi(r) \geq \mathbb{E}_{Z,\varepsilon} \mathcal{R}_n \left\{ f \in \operatorname{star}(M^{-1}\mathcal{G}_d) \mid Pf^2 \leq r \right\} \tag{45}$$

and its fixed point. As in Theorem 3.2, I will present the desired ψ and r^* here to continue the main proof, and leave the their verification after the main proof.
Claim. The following function $\psi(r)$ satisfies (45):

$$\psi(r) = \sqrt{\frac{r}{n}} + \frac{1}{\sqrt{n}} \inf_{h \geq 0} \left(\sqrt{rh} + M^{-1} \sqrt{d \sum_{j \geq h+1} \lambda_j \left(C'_2 \right)} \right), \tag{46}$$

with the fixed point

$$r^* \leq 2M^{-1}\left[\rho(M,d,n) + \frac{M}{n}\right]. \tag{47}$$

Let's continue the main proof. By applying Theorem with $B = 1$ to $M^{-1}\mathcal{G}_d$, we get for all $K > 1$ and every $\xi > 0$, with probability at least $1 - e^{-\xi}$:

$$\forall g \in M^{-1}\mathcal{G}_d: \quad Pg \leq \frac{K}{K-1}P_n g + 6Kr^* + \frac{\xi\,(11+5K)}{n}.$$

By the definition of $\mathcal{G}_d$, it equivalently means that

$$\forall V \in \mathcal{V}_d: \quad P\left(\frac{f_V}{M}\right) \leq \frac{K}{K-1}P_n\left(\frac{f_V}{M}\right) + 6Kr^* + \frac{\xi(11+5K)}{n}.$$

Note that the functions take the form $f_V(z) = \langle \Pi_{V^\perp}, C_x\rangle$. Plugging the definition (2), (3) of $R(V)$ and $R_n(V)$ yields

$$\forall V \in \mathcal{V}_d: \quad R(V) \leq \frac{K}{K-1}R_n(V) + 6MKr^* + \frac{\xi M(11+5K)}{n}.$$

Further plugging the property (47) of r^* and applying to $\hat{V}_d$, we have

$$R(\hat{V}_d) \leq \frac{K}{K-1}R_n(\hat{V}_d) + 12K\rho(M,d,n) + \frac{\xi M(11+5K)+12MK}{n}. \tag{48}$$

And similarly (the other conclusion of Theorem 3.3) applying to V_d we have

$$R_n(V_d) \leq \frac{K+1}{K}R(V_d) + 12K\rho(M,d,n) + \frac{\xi M(11+5K)+12MK}{n}. \tag{49}$$

Focusing on (48), by the fact that $R_n(\hat{V}_d) \leq R_n(V_d)$, we further note that

$$R(\hat{V}_d) \leq \frac{K}{K-1}R_n(V_d) + 12K\rho(M,d,n) + \frac{\xi M(11+5K)+12MK}{n},$$

using (49) and $\frac{K}{K-1} \leq 2$,

$$\leq \frac{K+1}{K-1}R(V_d) + 36K\rho(M,d,n) + 3\frac{\xi M(11+5K)+12MK}{n},$$

using $\frac{K+1}{K-1} = 1 + \frac{1}{K}\cdot\frac{2K}{K-1}$ and $\frac{2K}{K-1} \leq 4$,

$$\begin{aligned}
&\leq R(V_d) + \frac{4}{K}R(V) + 36K\rho(M,d,n) + 3\cdot\frac{\xi M\cdot 10.5K + 12MK}{n} \\
&\leq R(V_d) + 36\left[\frac{1}{K}R(V) + K\left(\rho(M,d,n) + \frac{\xi M+1}{n}\right)\right].
\end{aligned} \tag{50}$$

The choice of

$$K = \max\left\{2, \frac{\sqrt{R(V_d)}}{\sqrt{\rho(M,d,n) + \frac{M\xi+1}{n}}}\right\}$$

leads to the final result. To explain why it is true, we need to discuss by cases. When

$$K = \frac{\sqrt{R(V_d)}}{\sqrt{\rho(M,d,n) + \frac{M\xi+1}{n}}},$$

plugging in K into (50) leads to

$$\begin{aligned} R(\widehat{V}_d) - R(V_d) &\le 72\sqrt{R(V_d)\left(\rho(M,d,n) + \frac{M\xi+1}{n}\right)} \\ &\le 72\left\{\sqrt{R(V_d)\left(\rho(M,d,n) + \frac{M\xi+1}{n}\right)} + \rho(M,d,n) + \frac{M\xi+1}{n}\right\}. \end{aligned}$$

When

$$K = 2 \iff \sqrt{R(V_d)} \le 2\sqrt{\left(\rho(M,d,n) + \frac{M\xi+1}{n}\right)},$$

we only plug one $\sqrt{R(V_d)}$ into RHS of (50) and leave one $\sqrt{R(V_d)}$ in the expression, yielding

$$\begin{aligned} R(\widehat{V}_d) - R(V_d) &\le 36\sqrt{R(V_d)\left(\rho(M,d,n) + \frac{M\xi+1}{n}\right)} + 72\left(\rho(M,d,n) + \frac{M\xi}{n}\right) \\ &\le 72\left\{\sqrt{R(V_d)\left(\rho(M,d,n) + \frac{M\xi+1}{n}\right)} + \rho(M,d,n) + M\frac{\xi}{n}\right\}, \end{aligned}$$

concluding the final result.

Proof of Claim. We now deal with the claim used previously. Denote

$$\begin{aligned} \mathcal{L}_r &:= \left\{g \in \operatorname{star}\left(M^{-1}\mathcal{G}_d\right) \mid Pg^2 \le r\right\} \\ &= \left\{g : x \mapsto \lambda M^{-1}\left(\|x\|^2 - \langle \Pi_V, C_x\rangle\right) \mid V \in \mathcal{V}_d, Pg^2 \le r, \lambda \in [0,1]\right\}. \end{aligned}$$

To upper bound $\mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{L}_r$, we can decompose every function in $\mathcal{L}_r$ as (also using $\|x\|^2 = M$)

$$M^{-1}\lambda\left(\|x\|^2 - \langle \Pi_V, C_x\rangle\right) = \lambda\left(1 - M^{-1}\langle \Pi_V, C_1\rangle\right) + M^{-1}\langle \lambda\Pi_V, C_1 - C_x\rangle.$$

For convenience, we denote the decomposition for each $g \in \mathcal{L}_r$ as $g = g_1 + g_2$ where

$$g_1(x) = \lambda\left(1 - M^{-1}\langle \Pi_V, C_1\rangle\right) \text{ and } g_2(x) = M^{-1}\langle \lambda\Pi_V, C_1 - C_x\rangle.$$

Now let's focus on the constraint in $\mathcal{L}_r$. Note that

$$\begin{aligned} r \ge Pg^2 &= M^{-2}\lambda^2 P\left(\|x\|^2 - \langle \Pi_V, C_x\rangle\right)^2 \\ &= \lambda^2\left(1 - 2M^{-1}\langle \Pi_V, C_1\rangle\right) + M^{-2}\langle \Pi_V, C_2\Pi_V\rangle \\ &= \left(\lambda\left(1 - M^{-1}\langle \Pi_V, C_1\rangle\right)\right)^2 + M^{-2}\langle \lambda\Pi_V, (C_2 - C_1 \otimes C_1^*)\lambda\Pi_V\rangle. \end{aligned}$$

Therefore, by relaxing conditions, we have the set relation (also note that g_1 is note related to x, thereby a constant function):

$$\left\{g_1 : x \mapsto \lambda\left(1 - M^{-1}\left\langle \Pi_V, C_1\right\rangle\right) \mid V \in \mathcal{V}_d, Pg^2 \le r, \lambda \in [0,1]\right\} \subset \left\{g_1 : x \mapsto c \mid 0 \le c \le \sqrt{r}\right\} =: \mathcal{L}_{1,r},$$

and

$$\begin{aligned}&\left\{g_2 : x \mapsto M^{-1}\left\langle \lambda\Pi_V, C_1 - C_x\right\rangle \mid V \in \mathcal{V}_d, Pg^2 \le r, \lambda \in [0,1]\right\} \\ &\quad \subset \left\{x \mapsto M^{-1}\left\langle \Gamma, C_1 - C_x\right\rangle \mid \|\Gamma\|^2 \le d, \left\langle \Gamma, \left(C_2 - C_1 \otimes C_1^*\right)\Gamma\right\rangle \le M^2 r\right\} =: \mathcal{L}_{2,r}\end{aligned}$$

Then by the definition of sup, we need to upper bound the RHS of:

$$\mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{L}_r \le \mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{L}_{1,r} + \mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{L}_{2,r}.$$

Since the form of functions and constraints on the functions are the same in pairs ($\mathcal{S}_{1,r}$, $\mathcal{R}_{1,r}$) and ($\mathcal{S}_{2,r}$, $\mathcal{R}_{2,r}$), using the same argument in Theorem 3.2, we conclude

$$\mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{L}_{1,r} \le \sqrt{\frac{r}{n}} \text{ and } \mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{L}_{2,r} \le \frac{1}{\sqrt{n}} \inf_{h \ge 0}\left(\sqrt{rh} + M^{-1}\sqrt{d \sum_{j \ge h+1} \lambda_j\left(C_2'\right)}\right),$$

and then we can upper bound

$$\begin{aligned}\mathbb{E}_{Z,\varepsilon} \mathcal{R}_n \left\{f \in \operatorname{star}\left(M^{-1}\mathcal{G}_d\right) \mid Pf^2 \le r\right\} &\le \mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{L}_{1,r} + \mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{L}_{2,r} \\ &\le \sqrt{\frac{r}{n}} + \frac{1}{\sqrt{n}} \inf_{h \ge 0}\left(\sqrt{rh} + M^{-1}\sqrt{d \sum_{j \ge h+1} \lambda_j\left(C_2'\right)}\right),\end{aligned}$$

where the last line is the expression of $\psi(r)$ defined in (46). Solving $r^* \le \psi(r^*)$ (using the same trick of Theorem 3.2) yields the property (47) of r^*, completing the claim.

The remaining task is to upper bound $|R(\widehat{V}_d) - R_n(\widehat{V}_d)|$. Most derivation will be familiar, but one more thing we will use is the empirical part (29) of Theorem 3.3. The class of functions is still chosen to be $M^{-1}\mathcal{G}_d$, so the first two conditions of Theorem 3.3 are verified in the previous part. It suffices to find a data-dependent sub-root function $\widehat{\psi}_n$ with fixed point $\hat{r}^*$ such that for a real constant $x \in \mathbb{R}$,

$$\widehat{\psi}_n(r) \ge 20\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{J}_r + \frac{31x}{n}, \text{ where } \mathcal{J}_r = \left\{j \in \operatorname{star}\left(M^{-1}\mathcal{G}_d\right) \mid P_n g^2 \le 2r\right\} \tag{51}$$

As in the previous claim, we decompose each $j \in \mathcal{J}_r$ as $j = j_1 + j_2$ where

$$g_1(x) = \lambda\left(1 - M^{-1}\left\langle \Pi_V, C_{1,n}\right\rangle\right) \text{ and } g_2(x) = M^{-1}\left\langle \lambda\Pi_V, C_{1,n} - C_x\right\rangle.$$

And the constraint on the functions can be transformed into

$$\begin{aligned}2r \ge P_n j^2 &= M^{-2}\lambda^2 P_n\left(\|x\|^2 - \left\langle \Pi_V, C_x\right\rangle\right)^2 \\ &= \lambda^2\left(1 - 2M^{-1}\left\langle \Pi_V, C_{1,n}\right\rangle\right) + M^{-2}\left\langle \Pi_V, C_{2,n}\Pi_V\right\rangle \\ &= \left(\lambda\left(1 - M^{-1}\left\langle \Pi_V, C_{1,n}\right\rangle\right)\right)^2 + M^{-2}\left\langle \lambda\Pi_V, \left(C_{2,n} - C_{1,n} \otimes C_{1,n}^*\right)\lambda\Pi_V\right\rangle.\end{aligned}$$

Therefore, by relaxing conditions, we have the set relation (note that j_1's are also constant functions):

$$\left\{j_1 : x \mapsto \lambda\left(1 - M^{-1}\left\langle \Pi_V, C_{1,n}\right\rangle\right) \mid V \in \mathcal{V}_d, P_n j^2 \le 2r, \lambda \in [0,1]\right\} \subset \left\{j_1 : x \mapsto c \mid 0 \le c \le \sqrt{2r}\right\} =: \mathcal{J}_{1,r},$$

and

$$\begin{aligned}
&\left\{j_2 : x \mapsto M^{-1}\left\langle \lambda\Pi_V, C_{1,n} - C_x\right\rangle \mid V \in \mathcal{V}_d, P_n j^2 \le 2r, \lambda \in [0,1]\right\} \\
&\quad \subset \left\{x \mapsto M^{-1}\left\langle \Gamma, C_{1,n} - C_x\right\rangle \mid \|\Gamma\|^2 \le d, \left\langle \Gamma, \left(C_{2,n} - C_{1,n} \otimes C_{1,n}^*\right)\Gamma\right\rangle \le 2M^2 r\right\} =: \mathcal{J}_{2,r},
\end{aligned}$$

with $\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{J}_r \le \mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{J}_{1,r} + \mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{J}_{2,r}$. Applying the same argument as in the previous claim leads to

$$\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{J}_{1,r} \le \sqrt{\frac{2r}{n}}.$$

Modifying the spectral decomposition from C_2 to $C_{2,n}$ yields (treating $C_{2,n}$ as observed data instead of random element)

$$\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{J}_{2,r} \le \frac{1}{\sqrt{n}} \inf_{h \ge 0}\left(\sqrt{rh} + M^{-1}\sqrt{d \sum_{j \ge h+1} \lambda_j\left(C'_{2,n}\right)}\right)$$

Thus we conclude

$$\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{J}_r \le \mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{J}_{1,r} + \mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{J}_{2,r}$$

Furthermore, we can choose

$$\widehat{\psi}_n(r) = 20\left\{\sqrt{\frac{2r}{n}} + \frac{1}{\sqrt{n}} \inf_{h \ge 0}\left(\sqrt{rh} + M^{-1}\sqrt{d \sum_{j \ge h+1} \lambda_j\left(C'_{2,n}\right)}\right)\right\} + \frac{31x}{n},$$

and it will satisfies (51) once we recall that

$$\psi(r) = \sqrt{\frac{2r}{n}} + \frac{1}{\sqrt{n}} \inf_{h \ge 0}\left(\sqrt{rh} + M^{-1}\sqrt{d \sum_{j \ge h+1} \lambda_j\left(C'_{2,n}\right)}\right)$$

is a sub-root function and so is $\widehat{\psi}_n(r) = 20\psi(r) + \frac{31x}{n}$. To proceed, we need a fact:

FACT: if ψ is a sub-root function with fixed point r^* and $\psi_1 = \alpha\psi + \beta$ for nonnegative α, β, then the fixed point r_1^* for ψ_1 satisfies $r_1^* \le 4(\alpha^2 r^* + \beta)$.

Using this fact, we can tell the fixed point $\hat{r}^*$ of $\widehat{\psi}_n$ should satisfy

$$\hat{r}^* \le 1600 r^* + \frac{124x}{n} \le 4800 M^{-1}\left[\rho_n(M, d, n) + \frac{M}{n}\right] + \frac{124x}{n},$$

where the second inequality comes from solving $\psi(r^*) \le r^*$. Now we can apply the empirical part of Theorem 3.2 to get for all $K > 2$, with probability at least $1 - 4e^{-\xi}$,

$$\forall V \in \mathcal{V}_d, \quad |R(V) - R_n(V)| \le c\left\{\frac{1}{K} R_n(V) + K\left[\rho_n(M, d, n) + \frac{\xi M}{n}\right]\right\},$$

where $c \leq 1.2 \times 10^5$ after a tedious calculation. Using the union bound, we can make this bound uniform over positive integer K in the range $[2, 3, \ldots, n]$ at the price of replacing ξ to $\xi + \log n$. Finally, applying this inequality to $\widehat{V}_d$ and choose

$$K = \max \left\{ 2, \left\lceil \left(\rho_n(M, d, n) + M \frac{\xi + \log n}{n} \right)^{-\frac{1}{2}} R(V_d)^{\frac{1}{2}} \right\rceil \right\},$$

concluding the final result. □

The highlighted statement is the direct copy from the original paper. I am guessing the word union bound refers to Boole's inequality, but I'm still working on thinking about how it works.

4 Kernel PCA

In the practice of kernel PCA, the distribution of the random element $Z \in \mathscr{H}$ is implicitly associated with the kernel function k. Specifically, given the kernel function k, the collection of functions $\mathscr{H}_k = \{k(x, \cdot) \mid x \in \mathcal{X}\}$ is a RKHS with

$$\forall f \in \mathscr{H}_k \quad \forall x \in \mathcal{X} \quad \langle f, k(x, .)\rangle = f(x),$$

and in particular

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

And the kernel PCA, in essence, is doing PCA for $Z = k(x, \cdot) \in \mathscr{H}_k$. This relation allows us to compute the covariance matrix of Z_i's by the kernel matrix (k_{ij}), which is computable from the data. The key relation between the covariance operator and kernel operator is the following.

Theorem 4.1. Let $(\mathcal{X}, P)$ be a probability space, $\mathscr{H}$ be a separable Hilbert space, X be a $\mathcal{X}$-valued random variable and Φ be a map from $\mathcal{X}$ to $\mathscr{H}$ such that for all $h \in \mathscr{H}$, $\langle h, \Phi(\cdot)\rangle$ is measurable and $\mathbb{E}\|\Phi(X)\|^2 < \infty$. Let C_Φ be the covariance operator associated to $\Phi(X)$ and $K_\Phi : L_2(P) \to L_2(P)$ be the integral operator defined as

$$\left(K_\Phi f\right)(t) = \mathbb{E}[f(X)\langle\Phi(X), \Phi(t)\rangle] = \int f(x)\langle\Phi(x), \Phi(t)\rangle dP(x)$$

Then K_Φ is a Hilbert-Schmidt, positive self-adjoint operator, and

$$\lambda\left(K_\Phi\right) = \lambda\left(C_\Phi\right) \tag{52}$$

In particular, K_Φ is a trace-class operator and $\operatorname{tr}\left(K_\Phi\right) = \mathbb{E}\|\Phi(X)\|^2 = \sum_{i \geq 1} \lambda_i\left(K_\Phi\right)$.

Proof. Since K_Φ is the integral operator defined by the symmetric and nonnegative kernel function

$$K(s, t) = \langle\Phi(s), \Phi(t)\rangle_{\mathscr{H}},$$

by Theorem 4.6.7 of your book, we claim K_Φ is HS, trace-class, positive self-adjoint operator since

$$\sum_i \lambda_i(K_\Phi) = \mathrm{tr}(K_\Phi) = \int K(s,s)dP(s) = \int \langle \Phi(s), \Phi(s)\rangle \, dP(s) = \mathbb{E}\|\Phi(X)\|^2 < \infty.$$

For convenience, let $\langle \cdot, \cdot \rangle$ be the inner product on $\mathscr{H}$ and $\langle \cdot, \cdot \rangle_{L_2}$ be the inner product on $L_2(P)$. Define the operator

$$\forall g \in \mathscr{H}: \quad \mathscr{T} g = \langle \Phi(\cdot), g \rangle \text{ where the RHS is a function from } \mathcal{X} \text{ to } \mathbb{R}.$$

Note that $\forall g \in \mathscr{H}$, the integral of the RHS square

$$\int \langle \Phi(x), g\rangle^2 \, dP(x) \le \int \|\Phi(x)\|^2 \|g\|^2 dP(x) = \|g\|^2 \mathbb{E}\|\Phi(X)\|^2 < \infty,$$

indicating $\mathscr{T}: \mathscr{H} \mapsto L_2(P)$. And for any $g_1, g_2 \in \mathscr{H}$, we have

$$\langle \mathscr{T} g_1, \mathscr{T} g_2 \rangle_{L_2} = \Big\langle \langle \Phi(\cdot), g_1\rangle, \langle \Phi(\cdot), g_2 \rangle \Big\rangle_{L_2} = \mathbb{E}\Big[\langle \Phi(X), g_1 \rangle \langle \Phi(X), g_2 \rangle \Big] = \langle g_1, C_\Phi g_2 \rangle, \tag{53}$$

which in particular shows that

$$\mathrm{Ker}(\mathscr{T}) = \mathrm{Ker}(\mathscr{T}^* \mathscr{T}) = \mathrm{Ker}(C_\Phi). \tag{54}$$

Let $\mathscr{T}(\mathscr{H}) := \{\mathscr{T} g : g \in \mathscr{H}\}$. We will prove the goal (52) in the following steps.

(i) $\mathscr{T}(\mathscr{H}) \subset L_2(P)$ since for $g \in \mathscr{H}$ by (53),

$$\int [\mathscr{T}(g)]^2 dP = \mathbb{E}\Big[\langle \Phi(X), g \rangle^2 \Big] \le \|g\|^2 \mathbb{E}\|\Phi(X)\|^2 < \infty.$$

(ii) $\mathrm{Ker}(K_\Phi)^\perp \subset \overline{\mathscr{T}(\mathscr{H})}$, the completed space of $\mathscr{T}(\mathscr{H})$ in $L_2(P)$.

Suppose $f \in \mathscr{T}(\mathscr{H})^\perp$, i.e. $\langle f, \mathscr{T} g \rangle_{L_2} = 0$ for all $g \in \mathscr{H}$. Since $\Phi: \mathcal{X} \mapsto \mathscr{H}$, in particular we have for all $t \in \mathcal{X}$,

$$0 = \langle f, \mathscr{T}\Phi(t) \rangle_{L_2} = \mathbb{E}\Big[f(X) \langle \Phi(X), \Phi(t) \rangle \Big] = (K_\Phi f)(t),$$

where the second and third equalities are from the definition of $\mathscr{T}$ and K_Φ. This means $K_\Phi f \equiv 0$ or equivalently $f \in \mathrm{Ker}(K_\Phi)$, and hence $\mathscr{T}(\mathscr{H})^\perp \subset \mathrm{Ker}(K_\Phi)$. Taking orthogonal complement on both sides, we reach the goal in step (ii):

$$\mathrm{Ker}(K_\Phi)^\perp \subset \mathscr{T}(\mathscr{H})^{\perp\perp} = \overline{\mathscr{T}(\mathscr{H})}.$$

(iii) $K_\Phi \mathscr{T} g = \mathscr{T} C_\Phi g$, as a function from $\mathcal{X}$ to $\mathbb{R}$.

By the definition of K_Φ and $\mathscr{T}$, it follows that for all $t \in \mathcal{X}$,

$$(K_\Phi \mathscr{T} g)(t) = \mathbb{E}\Big[\langle \Phi(X), g \rangle \langle \Phi(X), \Phi(t) \rangle \Big] = \langle C_\Phi \Phi(t), g \rangle = \langle \Phi(t), C_\Phi g \rangle = (\mathscr{T} C_\Phi)(t),$$

where the third equality is given by the property of the covariance operator C_Φ.

(iv) Let (λ_j, e_j), $j \geq 1$, be the nonzero eigenvalue/eigenfunction pairs of C_Φ. For convenience, assume λ_j are distinct and $\|e_j\|^2 = 1$. Then $\{\tilde{e}_j := \lambda_j^{-1/2} \mathscr{T} e_j\}$ is a CONS of $\overline{\mathscr{T}(\mathscr{H})}$.

First, the e_j are orthonormal since, by (53),

$$\langle \tilde{e}_i, \tilde{e}_j \rangle_{L_2} = \left\langle \lambda_i^{-\frac{1}{2}} \mathscr{T} e_i, \lambda_j^{-\frac{1}{2}} \mathscr{T} e_j \right\rangle_{L_2} = (\lambda_i \lambda_j)^{-\frac{1}{2}} \langle e_i, C_\Phi e_j \rangle = \delta_{ij}.$$

To show $\overline{\operatorname{span}(\tilde{e}_j)} = \overline{T(\mathscr{H})}$, since $\overline{\operatorname{span}(\tilde{e}_j)} \subset \overline{T(\mathscr{H})}$ is natural, it remains to show the other direction. Let $f \in \overline{\mathscr{T}(\mathscr{H})}$. By the definition of $\overline{\mathscr{T}(\mathscr{H})} \subset L_2(P)$, there exists a sequence $g_n \in \mathscr{H}$ s.t. $\mathscr{T} g_n \xrightarrow{L_2} f$ and $g_n \notin \operatorname{Ker}(\mathscr{T})$. By (54), it means $g_n \in \operatorname{Ker}(\mathscr{T})^\perp = (\operatorname{Ker} C_\Phi)^\perp = (\operatorname{Im} C_\Phi^*)^{\perp\perp} = \overline{\operatorname{Im} C_\Phi^*}$. Therefore, we can express $g_n \in \mathscr{H}$ in the Fourier expansion $g_n = \sum_j c_{n,j} e_j$ where $c_{nj} = \langle g_n, e_j \rangle$. Then

$$\mathscr{T} g_n = \sum_j c_{n,j} \mathscr{T} e_j = \sum_j c_{n,j} \lambda_j^{1/2} \tilde{e}_j \tag{55}$$

Since $\mathscr{T} g_n \xrightarrow{L_2} f$ by construction, it is necessary and sufficient that $f = \sum_k c_j \tilde{e}_j$ where $c_j = \lim_{n\to\infty} c_{nj} \lambda_j^{1/2}$. This concludes our goal: $\overline{\operatorname{span}(\tilde{e}_j)} \supset \overline{T(\mathscr{H})}$.

(v) Any eigenvalue of C_Φ is also an eigenvalue of K_Φ.

Suppose (λ, g) is an eigenvalue/eigenfunction pair of C_Φ in $\mathscr{H}$, i.e. $C_\Phi g = \lambda g$. Then by (iii), it follows that

$$K_\Phi \mathscr{T} g = \mathscr{T} C_\Phi g = \mathscr{T} \lambda g = \lambda \mathscr{T} g,$$

meaning that $(\lambda, \mathscr{T} g)$ is an eigenvalue/eigenfunction pair for K_Φ.

(vi) Any eigenvalue of K_Φ is also an eigenvalue of C_Φ.

Suppose (λ, f) is an eigenvalue/eigenfunction pair of K_Φ in $L_2(P)$ where $\lambda \neq 0$, i.e. $K_\Phi f = \lambda f$. By (ii), it follows that $f \in \operatorname{Ker}(K_\Phi)^\perp = \overline{\mathscr{T}(\mathscr{H})}$, meaning that there exists a sequence $g_n \in \mathscr{H}$ s.t. $\mathscr{T} g_n \xrightarrow{L_2} f$. And hence in $L_2(P)$,

$$\lambda \lim_{n\to\infty} \mathscr{T} g_n = \lambda f = K_\Phi f = \lim_{n\to\infty} K_\Phi \mathscr{T} g_n = \lim_{n\to\infty} \mathscr{T} C_\Phi g_n. \tag{56}$$

Following the argument used in (iv), we again use the Fourier expansion $g_n = \sum_j c_{n,j} e_j$ where $c_{nj} = \langle g_n, e_j \rangle$. With (55), the LHS of the above equality can be expressed as

$$\lambda \lim_{n\to\infty} \mathscr{T} g_n = \lambda \lim_{n\to\infty} \sum_j c_{nj} \lambda_j^{\frac{1}{2}} \tilde{e}_j = \lim_{n\to\infty} \sum_j c_{nj} \lambda \lambda_j^{\frac{1}{2}} \tilde{e}_j.$$

Using the Fourier expansion, the RHS of (55) can be expressed as

$$\lim_{n\to\infty} \mathscr{T} C_\Phi g_n = \lim_{n\to\infty} \sum_j \langle C_\Phi g_n, e_j \rangle \mathscr{T} e_j = \lim_{n\to\infty} \sum_j \lambda_j \langle g_n, e_j \rangle \mathscr{T} e_j = \lim_{n\to\infty} \sum_j c_{nj} \lambda_j^{\frac{3}{2}} \tilde{e}_j.$$

Therefore, the equality (56) implies the following identity holds in $L_2(P)$:

$$\lim_{n\to\infty} \sum_j c_{nj} \lambda \lambda_j^{\frac{1}{2}} \tilde{e}_j = \lim_{n\to\infty} \sum_j c_{nj} \lambda_j^{\frac{3}{2}} \tilde{e}_j$$

Since $\tilde{e}_j$ are orthonormal, we must have

$$\begin{cases} c_{nj}\lambda_j^{\frac{1}{2}} \to \text{constant } c_j \text{ for all } j, \\ \sum_j \left(c_{nj}\lambda\lambda_j^{\frac{1}{2}} - c_{nj}\lambda_j^{\frac{3}{2}}\right)^2 \to 0. \end{cases}$$

Or equivalently, we must have

$$\sum_j c_j^2(\lambda - \lambda_j)^2 = 0. \tag{57}$$

Since we assumed (λ, f) is a nonzero eigenvalue/eigenfunction pair, it means

$$0 \neq f = \lim_{n\to\infty} \mathscr{T} g_n = \lim_{n\to\infty} \sum_j c_{nj}\lambda_j^{\frac{1}{2}}\tilde{e}_j = \sum_j c_j \tilde{e}_j,$$

indicating that there exists some j s.t. $c_j \neq 0$. With the condition (56) together, it further means for this j,

$$\lambda_j = \lambda,$$

i.e. the eigenvalue of K_Φ, λ, equals to one of the eigenvalues of C_Φ.

□

Since in the kernel PCA setting, we are assuming $\Phi(X) \in \mathscr{H}$, where $\mathscr{H}$ is a RKHS with kernel function $k(\cdot,\cdot)$, further consequences of Theorem 4.1 are

$$\lambda\left(C_1\right) = \lambda\left(K_1\right), \quad \lambda\left(C_{1,n}\right) = \lambda\left(K_{1,n}\right), \quad \lambda\left(C_{2,n}\right) = \lambda\left(K_{2,n}\right), \quad \lambda\left(C_2'\right) = \lambda\left(K_2'\right),$$

and

$$\lambda\left(C_{2,n}'\right) := \lambda\left(C_{2,n} - C_{1,n} \otimes C_{1,n}^*\right) = \lambda\left(\left(I_n - \frac{1}{n}\mathbf{1}\right) K_{2,n} \left(I_n - \frac{1}{n}\mathbf{1}\right)\right) =: \lambda\left(K_{2,n}'\right),$$

where K_1 denotes the kernel integral operator with kernel k and the true probability distribution P, $K_{1,n}$ and $K_{2,n}$ are identified to the matrices $(k(X_i, X_j)/n)$ and $(k^2(X_i, X_j)/n)$ respectively. With Theorem 4.1, all results in previous sections of this paper can be applied immediately.

Theorem 4.2. Given a kernel function k, assume

(A1) $\mathscr{H}_k$ is separable;

(A2) For all $x \in \mathcal{X}$, $k(x,\cdot)$ is P-measurable;

Let $\mathcal{X}_0$ be the support of distribution P on $\mathcal{X}$; further assume

$$\sup_{x\in\mathcal{X}_0} k(x,x) \leq M \text{ and } \sup_{x,y\in\mathcal{X}_0} \left(k^2(x,x) + k^2(y,y) - 2k^2(x,y)\right) \leq L^2.$$

Denote $R(d) = \sum_{i>d} \lambda_d\left(K_1\right)$ and $R_n(d) = \sum_{i>d} \lambda_d\left(K_{1,n}\right)$. Then for any $n \geq 2$, either of the following inequalities holds with probability at least $1-e^{-\xi}$:

$$R(d)-R_n(d) \leq \sqrt{\frac{d}{n}\operatorname{tr} K_2'}+(M \wedge L)\sqrt{\frac{\xi}{2n}} \tag{58}$$

$$R(d)-R_n(d) \leq \sqrt{\frac{d}{n-1}\operatorname{tr} K_{2,n}'}+(M \wedge L)\sqrt{\frac{\xi}{2n}}+L\frac{\sqrt{d}\xi^{\frac{1}{4}}}{n^{\frac{3}{4}}} \tag{59}$$

$$R(d)-R_n(d) \geq -\sqrt{\frac{2\xi}{n}(M \wedge L)R(d)}-(M \wedge L)\frac{\xi}{3n} \tag{60}$$

$$R(d)-R_n(d) \geq -\sqrt{\frac{2\xi}{n}(M \wedge L)\left(R_n(d)-(M \wedge L)\frac{\xi}{3n}\right)_+}-(M \wedge L)\frac{\xi}{3n} \tag{61}$$

Under the stronger condition $k(x,x)=M$ for all $x \in \mathcal{X}_0$, either of the following inequalities holds with probability at least $1-e^{-\xi}$:

$$R(d)-R_n(d) \leq c\left(\sqrt{R(d)\left(\rho(M,d,n)+M\frac{\xi}{n}\right)}+\rho(M,d,n)+\frac{M\xi}{n}\right), \tag{62}$$

$$R(d)-R_n(d) \leq c\left(\sqrt{R_n(d)\left(\rho_n(M,d,n)+M\frac{(\xi+\log n)}{n}\right)}+\rho_n(d,n)+\frac{M(\xi+\log n)}{n}\right). \tag{63}$$

Proof. By the definition of $R(d), R_n(d)$ and the relation that $\lambda(C_1)=\lambda(K_1)$ and $\lambda(C_{1,n})=\lambda(K_{1,n})$, we have

$$R(d)=R(V_d) \text{ and } R_n(d)=R_n(V_d).$$

Then using the fact that $R(V_d) \leq R(\widehat{V}_d)$, applying Theorem 3.1 yields (58) and (59); applying Theorem 3.4 yields (62) and (63).

To show the relative bounds (60) and (61), note that

$$R(d)-R_n(d)=R\left(V_d\right)-R_n(\widehat{V}_d) \geq R(V_d)-R_n(V_d)=(P_n-P)\left\langle \Pi_{V_d}, C_Z\right\rangle. \tag{64}$$

Consider the function $f(z)=(P_n-P)\left\langle \Pi_{V_d}, C_z\right\rangle$. The assumptions

$$\sup_{x\in\mathcal{X}_0} k(x,x) \leq M \text{ and } \sup_{x,y\in\mathcal{X}_0}\left(k^2(x,x)+k^2(y,y)-2k^2(x,y)\right) \leq L^2,$$

mean (using same arguments as in the proof of Theorem 3.1) $f(Z) \in [a,b]$ with $a \geq 0$ and $|a-b| \leq M \wedge L$ a.s.. Applying Bernstein's inequality in Appendix B.3 to the function $(f-a) \in [0, M \wedge L]$ yields with probability at least $1-e^{-\xi}$,

$$(P_n-P)\left\langle \Pi_{V_d}, C_Z\right\rangle \geq -\sqrt{\frac{2\xi P(f-a)^2}{n}}-(M \wedge L)\frac{\xi}{3n}. \tag{65}$$

Further with the fact

$$P(f-a)^2 \leq (M \wedge L)(Pf-a) \leq (M \wedge L)Pf, \tag{66}$$

by (64), (65) and (66), we conclude with probability at least $1 - e^{-\xi}$,

$$R(d) - R_n(d) \geq -\sqrt{\frac{2\xi}{n}(M \wedge L)R(d)} - (M \wedge L)\frac{\xi}{3n},$$

concluding (60). To show (61), we first collect terms from the above equation, and say

$$R(d) + \sqrt{\frac{2\xi}{n}(M \wedge L)} \cdot \sqrt{R(d)} + (M \wedge L)\frac{\xi}{3n} - R_n(d) \geq 0,$$

Using the fact that

$$\begin{cases} x \geq 0 \\ x^2 + ax + b \geq 0 \end{cases} \quad \Rightarrow \quad x^2 \geq -b - a\sqrt{-(b \wedge 0)}$$

to $x = \sqrt{R(d)}$, we get (61). □

5 Application to the Horizontal FL

Horizontal Federated Learning (HFL) is a collaborative and privacy-preserving approach for multiple clients to collectively train a model. From a statistical perspective, we assume that independently and identically distributed (iid) observations are divided among different clients. Specifically, the i-th client stores data $\mathbb{X}_i \in \mathbb{R}^{n_i \times p}$, where n_i denotes the sample size and p represents the number of features. In the context of (kernel) PCA within HFL, the currently accepted algorithm [1] can be outlined as follows:

(1) Select a target lower dimension d (and a kernel function k for kernel PCA).

(2) Initialize a standard Gaussian matrix $\mathbb{V}^{(1)} \in \mathbb{R}^{p \times d}$.

(3) For each iteration $t = 1, \ldots, T$:

  (3.1) Broadcast $\mathbb{V}^{i(t)} \leftarrow \mathbb{V}^{(t)}$ for $i = 1, \ldots, m$.

  (3.2) Employ the power iteration algorithm to compute the d eigenvectors for $\mathbb{X}_i^\top \mathbb{X}_i$ (or kernel matrix):

$$\mathbb{V}^{i(t+1)} := \left(\mathbf{v}_1^{i(t+1)}, \ldots, \mathbf{v}_d^{i(t+1)}\right),$$

  where $\mathbf{v}_1^{i(t+1)}, \ldots, \mathbf{v}_d^{i(t+1)}$ represent the eigenvectors with the largest eigenvalues for the t-th iteration and i-th client.

  (3.3) Aggregate the vector list by $\mathbb{V}^{(t+1)} = \sum_{i=1}^m \frac{n_i}{n} \cdot \mathbb{V}^{i(t+1)}$.

(4) Return $\mathbb{V}^{(T+1)}$ whose column vectors form a spanning list of the estimated subspace.

Arbenz [2] demonstrates that the power iteration in (3.2) converges to the true eigenvectors. Moreover, experiments conducted by Li [1] and Cheung [3] indicate that the model performs well even for $T = 1$. Therefore, we aim to establish an upper bound on the reconstruction error for the HFL case with $T = 1$. And in fact, it is the only case that I can apply the results to, since once $T > 1$, the final estimate $\mathbb{V}^{(T+1)}$ is no longer a simple average.

5.1 Notations and Results

The results by Blanchard [4] can be directly applied to the HFL setting when $T = 1$ with a little modification. So, we follow his ideas. To adapt to the FL setting, we use the following notations.

Problem Setup. Suppose we have m clients, and the i-th client owns the data of n_i observations. In total, we have $n := n_1 + \cdots + n_m$ observations. Furthermore, let $Z_1^1, \dots, Z_{n_1}^1, \dots, Z_1^m, \dots, Z_{n_m}^m \overset{\text{iid}}{\sim} Z \in \mathcal{H}$ with $\mathbb{E}\|Z\|^4 < \infty$, where Z_j^i indicates the j-th observation stored in the i-th client.

(a) For any integrable function $f : \mathcal{H} \mapsto \mathbb{R}$, we denote the mean and sample mean as follows:

$$Pf := \mathbb{E}[f(Z)], \quad P_n f := \frac{1}{n}\sum_{i=1}^m \sum_{j=1}^{n_i} f(Z_j^i), \quad \text{and} \quad P_n^i f := \frac{1}{n_i}\sum_{j=1}^{n_i} f(Z_j^i).$$

(b) For any $z \in \mathcal{H}$, let $C_z := z \otimes z \in \mathcal{HS}$, and the non-centered covariance operators of Z and C_Z are denoted as:

$$C_1 := \mathbb{E}(Z \otimes Z) = \mathbb{E}(C_Z), \quad \text{and} \quad C_2 := \mathbb{E}(C_Z \otimes C_Z).$$

The sampled versions are denoted as:

$$C_{1,n} := \frac{1}{n}\sum_{i=1}^m \sum_{j=1}^{n_i} Z_j^i \otimes Z_j^i, \quad C_{1,n}^i := \frac{1}{n_i}\sum_{j=1}^{n_i} Z_j^i \otimes Z_j^i,$$

and

$$C_{2,n} := \frac{1}{n}\sum_{i=1}^m \sum_{j=1}^{n_i} C_{Z_j^i} \otimes C_{Z_j^i}, \quad C_{2,n}^i := \frac{1}{n_i}\sum_{j=1}^{n_i} C_{Z_j^i} \otimes C_{Z_j^i}.$$

(c) Given a subspace V, the notations for the true and empirical reconstruction error of V are:

$$\begin{aligned}
R(V) &:= \mathbb{E}[\|Z - \Pi_V(Z)\|^2] = P\langle \Pi_{V^\perp}, C_Z\rangle = \langle \Pi_{V^\perp}, C_1\rangle, \\
R_n(V) &:= \frac{1}{n}\sum_{i=1}^m \sum_{j=1}^{n_i} \|Z_j^i - \Pi_V(Z_j^i)\|^2 = P_n\langle \Pi_{V^\perp}, C_Z\rangle = \langle \Pi_{V^\perp}, C_{1,n}\rangle, \\
R_n^i(V) &:= \frac{1}{n_i}\sum_{j=1}^{n_i} \|Z_j^i - \Pi_V(Z_j^i)\|^2 = P_n^i\langle \Pi_{V^\perp}, C_Z\rangle = \langle \Pi_{V^\perp}, C_{1,n}^i\rangle.
\end{aligned}$$

We can easily verify that $R_n(V) = \sum_{i=1}^m \frac{n_i}{n} R_n^i(V)$. Prespecifying d as the dimension of the projected subspace, we denote by $\mathcal{V}_d$ the collection of all subspaces of dimension d, and we define:

$$V_d := \arg\min_{V \in \mathcal{V}_d} R(V), \quad \widehat{V}_d := \arg\min_{V \in \mathcal{V}_d} R_n(V) \text{ and } \widehat{V}_d^i := \arg\min_{V \in \mathcal{V}_d} R_n^i(V).$$

Following the HFL procedure, we can observe that $\widehat{V}_d^i = \text{col}(\mathbb{V}^i)$, where $\text{col}(\cdot)$ denotes the column space. If we define:

$$\mathbb{V}^i := \left(\mathbf{v}_1^i, \dots, \mathbf{v}_d^i\right) \quad \text{for } i = 1, \dots, m,$$

the aggregation step (3.3) indicates that the final estimated low-dimensional subspace is given by:

$$\widehat{V} := \operatorname{span}\left(\sum_{i=1}^{m} \frac{n_i}{n}\mathbf{v}_1^i, \ldots, \sum_{i=1}^{m} \frac{n_i}{n}\mathbf{v}_d^i\right) \tag{67}$$

when $T = 1$. Fortunately, we can directly apply the relevant Theorem 3.1 from Blanchard's work [4] to each client in the HFL setting. By doing so, we can obtain bounds on the total reconstruction error in HFL. Taking Theorem 3.1 as an example, we can derive the corresponding result.

Result. *Assume $\|Z\|^2 \le M$ a.s. and that $Z \otimes Z$ belongs a.s. to a set of* $\mathrm{HS}(\mathcal{H})$ *with bounded diameter L. Then for any $n_i \ge 2$, with probability at least $1 - 3me^{-\xi}$,*

$$|R(\widehat{V}) - R_n(\widehat{V})| \le \sum_{i=1}^{m} \frac{n_i}{n}\left\{\sqrt{\frac{d}{n_i - 1}\operatorname{tr}(C_{2,n}^i)'} + (M \wedge L)\sqrt{\frac{\xi}{2n_i}} + L\frac{\sqrt{d}\xi^{\frac{1}{4}}}{n_i^{\frac{3}{4}}}\right\}. \tag{68}$$

Also, with probability at least $1 - 2me^{-\xi}$,

$$0 \le R(\widehat{V}) - R(V_d) \le \sum_{i=1}^{m} \frac{n_i}{n}\left\{\sqrt{\frac{d}{n_i}\operatorname{tr}(C_2^i)'} + 2(M \wedge L)\sqrt{\frac{\xi}{2n_i}}\right\}, \tag{69}$$

where $C_2' = C_2 - C_1 \otimes C_1^*$ and $(C_{2,n}^i)' = C_{2,n}^i - C_{1,n}^i \otimes C_{1,n}^i$.

Remark. Though the subspace $\widehat{V}$ may be of less than dimension d by possible linear dependence of the spanning list, this result also holds.

Proof. W.L.O.G., assume $\widehat{V}$ is of dimension d. By noting that for any $i = 1, \ldots, m$

$$R(\widehat{V}) - R_n^i(\widehat{V}) = \left(P - P_n^i\right)\left\langle \Pi_{\widehat{V}^\perp}, C_Z\right\rangle \le \sup_{V \in \mathcal{V}_d}\left(P - P_n^i\right)\left\langle \Pi_{V^\perp}, C_Z\right\rangle,$$

we can apply the same argument in the first part of proof of Theorem 3.1 in [4], which leads to

$$\left|R(\widehat{V}) - R_n^i(\widehat{V})\right| \le \left|\sup_{V \in \mathcal{V}_d}\left(P - P_n^i\right)\left\langle \Pi_{V^\perp}, C_Z\right\rangle\right| \le \sqrt{\frac{d}{n_i - 1}\operatorname{tr}(C_{2,n}^i)'} + (M \wedge L)\sqrt{\frac{\xi}{2n_i}} + L\frac{\sqrt{d}\xi^{\frac{1}{4}}}{n_i^{\frac{3}{4}}}$$

to hold with probability at least $1 - 3e^{-\xi}$. Further by the fact $R_n(V) = \sum_{i=1}^{m} \frac{n_i}{n} R_n^i(V)$, we know

$$\left|R(\widehat{V}) - R_n(\widehat{V})\right| = \sum_{i=1}^{m} \frac{n_i}{n}\left|R(\widehat{V}) - R_n^i(\widehat{V})\right| \le \sum_{i=1}^{m} \frac{n_i}{n}\left\{\sqrt{\frac{d}{n_i - 1}\operatorname{tr}(C_{2,n}^i)'} + (M \wedge L)\sqrt{\frac{\xi}{2n_i}} + L\frac{\sqrt{d}\xi^{\frac{1}{4}}}{n_i^{\frac{3}{4}}}\right\}.$$

with probability at least $1 - 3me^{-\xi}$, completing the first part.

To prove the second part, we recall the fact

$$R(V) := \mathbb{E}[\|Z - \Pi_V(Z)\|^2] = P\langle \Pi_{V^\perp}, C_Z\rangle,$$

and plug $\widehat{V}$ and $\widehat{V}_d^i$ in to get

$$R(\widehat{V}) = \left\langle \mathbb{E}\left[\Pi_{\widehat{V}^\perp}\right], C_1\right\rangle \quad \text{and} \quad R(\widehat{V}_d^i) = \left\langle \mathbb{E}\left[\Pi_{(\widehat{V}_d^i)^\perp}\right], C_1\right\rangle.$$

By the construction (67) of $\widehat{V}$, I think (but not confirm) that $R(\widehat{V}) = R(\widehat{V}_d^i)$ for all $i = 1, \ldots, m$. If this is true, then it follows that

$$R(\widehat{V}) - R(V_d) = \sum_{i=1}^{m} \frac{n_i}{n} \left[R(\widehat{V}_d^i) - R(V_d) \right] \le \sum_{i=1}^{m} \frac{n_i}{n} \left\{ \sqrt{\frac{d}{n_i} \operatorname{tr}(C_2^i)'} + 2(M \wedge L) \sqrt{\frac{\xi}{2n_i}} \right\},$$

with probability at least $1 - 2me^{-\xi}$, where the second inequality comes from the second part of Theorem 3.1. $\square$

Using the same idea and fact that $R(\widehat{V}) = R(\widehat{V}_d^i)$, the following variant of Theorem 3.2 also holds.

Result. *Assume $\|Z\|^2 \le M$ a.s. Let (λ_i) denote the ordered eigenvalues with multiplicity of C_1, resp. (μ_i) the ordered distinct eigenvalues. Let $\tilde{d}$ be such that $\lambda_d = \mu_{\tilde{d}}$. Define*

$$\gamma_d = \begin{cases} \mu_{\tilde{d}} - \mu_{\tilde{d}+1} & \text{if } \tilde{d} = 1 \text{ or } \lambda_d > \lambda_{d+1}, \\ \min\left(\mu_{\tilde{d}-1} - \mu_{\tilde{d}}, \mu_{\tilde{d}} - \mu_{\tilde{d}+1}\right) & \text{otherwise} \end{cases}$$

and $B_d = \left(\mathbb{E}\left\langle Z, Z'\right\rangle^4\right)^{\frac{1}{2}} \gamma_d^{-1}$ (where Z' is an independent copy of Z).
Then for all d, for all $\xi > 0$, with probability at least $1 - me^{-\xi}$ the following holds:

$$R(\widehat{V}) - R(V_d) \le \sum_{i=1}^{m} \frac{n_i}{n} \left\{ 24\rho\left(B_d, d, n_i\right) + \frac{\xi\left(11M + 7B_d\right)}{n_i} \right\}, \tag{70}$$

where $n = n_1 + \cdots + n_m$.

Moreover, the bounds for the centered case in Theorem 3.5 also hold.

5.2 Summary and Open Problems

Based on the results (68), (69), and (70), we can draw some conclusions for HFL in terms of reconstruction error bounds in the kernel PCA setting:

(1) When the number of samples is evenly distributed, i.e., $n_1 = n_2 = \cdots = n_m$, using the averaged subspace $\widehat{V}$ or the estimated subspace $\widehat{V}_d^i$ by the i-th client alone yields no difference in these bounds.

(2) When the number of samples is not evenly distributed, and assuming that n_1 is the smallest among all n_i's, the averaged subspace $\widehat{V}$ provides a lower bound compared to $\widehat{V}_d^1$. This implies that HFL benefits clients with less data by achieving better reconstruction error bounds.

(3) Combining all the data into a single client and performing PCA can significantly decrease the reconstruction error bounds. However, this approach poses potential privacy risks as all the data is concentrated in one location. In contrast, HFL mitigates these risks by allowing model training to be performed collaboratively without the need for centralized data storage.

There are other interesting problems to explore, such as deriving bounds for the case when $T > 1$ or when dealing with vertically partitioned data. However, I do not have ideas for how to start these problems.

All in all, the analytical methods in KPCA and FL papers seem different. The KPCA focuses in a mathematical view and holds statistical model assumptions. Whereas, the latter one starts from a practical view. It states clearly about their data, algorithms and results. Though the two topics are related, I find myself somewhat constrained in coming up with a clear problem that is worth solving and can be solved. I suspect the underlying cause might be my insufficient number of readings. Is it a good idea to read more until I actually have an idea of problem?

References

[1] Xiang Li, Shusen Wang, Kun Chen, and Zhihua Zhang. Communication-efficient distributed svd via local power iterations. In *International Conference on Machine Learning*, pages 6504–6514. PMLR, 2021.

[2] Peter Arbenz, Daniel Kressner, and DME Zürich. Lecture notes on solving large scale eigenvalue problems. *D-MATH, EHT Zurich*, 2(3):377, 2012.

[3] Yiu-ming Cheung, Juyong Jiang, Feng Yu, and Jian Lou. Vertical federated principal component analysis and its kernel extension on feature-wise distributed data. *arXiv preprint arXiv:2203.01752*, 2022.

[4] Gilles Blanchard, Olivier Bousquet, and Laurent Zwald. Statistical properties of kernel principal component analysis. *Machine Learning*, 66:259–294, 2007.