On the Error-Propagation of Inexact Hotelling's Deflation for Principal Component Analysis

Fangshuo Liao [1] **Junhyung Lyle Kim** [1] **Cruz Barnum** [2] **Anastasios Kyrillidis** [1]

Abstract

Principal Component Analysis (PCA) aims to find subspaces spanned by the so-called *principal components* that best represent the variance in the dataset. The deflation method is a popular meta-algorithm that sequentially finds individual principal components, starting from the most important ones and working towards the less important ones. However, as deflation proceeds, numerical errors from the imprecise estimation of principal components propagate due to its sequential nature. This paper mathematically characterizes the error propagation of the inexact Hotelling's deflation method. We consider two scenarios: i) when the sub-routine for finding the leading eigenvector is abstract and can represent various algorithms; and ii) when power iteration is used as the sub-routine. In the latter case, the additional directional information from power iteration allows us to obtain a tighter error bound than the sub-routine agnostic case. For both scenarios, we explicitly characterize how the errors progress and affect subsequent principal component estimations.

1. Introduction

Principal Component Analysis (PCA) (Pearson, 1901; Hotelling, 1933) is a fundamental tool for data analysis with applications that range from statistics to machine learning and can be used for dimensionality reduction, classification, and clustering, to name a few (Majumdar, 2009; Wang et al., 2013; d'Aspremont et al., 2007; Jiang et al., 2011; Zou et al., 2006). Since its proposition Pearson (1901), many algorithms for PCA have been devised, even recently (Allen-Zhu & Li, 2017), to efficiently find approximations

---
*Equal contribution [1]Department of Computer Science, Rice University, Houston, U.S.A. [2]Department of Computer Science, UIUC. Correspondence to: Fangshuo Liao <Fangshuo.Liao@rice.edu>.
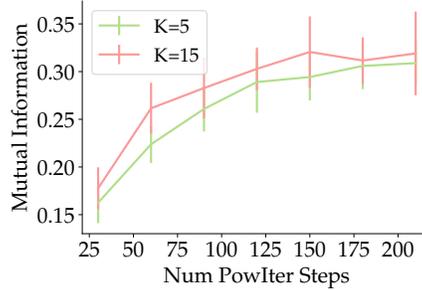
Figure 1. *Spectral clustering of MNIST dataset using inexact deflation method in Algorithm 1.* As the number of power iteration steps increases (x-axis), clustering performance, measured by the mutual information metric, also improves. A similar pattern is observed for recovering different numbers of eigenvectors.

of the first principal component, which is a vector that captures the most variance in the data. Mathematically, single-component PCA can be described as:

$$\mathbf{u}_1^* = \text{PCA}(\boldsymbol{\Sigma}) = \underset{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\|_2 = 1}{\arg\max} \ \mathbf{v}^\top \boldsymbol{\Sigma} \mathbf{v}, \quad (1)$$

where $\boldsymbol{\Sigma} = \frac{1}{n} \cdot \mathbf{Y}\mathbf{Y}^\top \in \mathbb{R}^{d \times d}$ is the empirical covariance matrix given a set of n centered d-dimensional datapoints $\mathbf{Y} \in \mathbb{R}^{d \times n}$.

It is sometimes helpful to know not only the first but also the top-K principal components, especially in the high-dimensional regime. This is becoming more relevant in the modern machine learning era, where the number of features reaches as large as billions (Fedus et al., 2022; Touvron et al., 2023). Mathematically, the multi-component PCA ($K \geq 1$) can be formulated as the following constrained optimization problem:

$$\mathbf{U}^* = \underset{\mathbf{V} \in \{\mathbf{Q}_{:,:K} : \mathbf{Q} \in \text{SO}(d)\}}{\arg\max} \langle \boldsymbol{\Sigma}\mathbf{V}, \mathbf{V} \rangle, \quad (2)$$

where $\text{SO}(d)$ denotes the group of rotations about a fixed point in d-dimensional Euclidean space. Since $\mathbf{U}^*$ contains the first K columns of an orthogonal matrix, each principal component $\mathbf{U}_{:,k}$ must be a unit vector, and the pairwise principal components, say $\mathbf{U}_{:,k_1}^*, \mathbf{U}_{:,k_2}^*$ for $k_1 \neq k_2$, are orthogonal. A popular method for solving (2) is the *deflation method* (Hotelling, 1933) *via sequentially solving* (1), which is the focus of our work.

Algorithm 1 Deflation with inexact PCA

Require: $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$; # of eigenvectors K; sub-routine for top eigenvector PCA($\cdot, \cdot$); # of iterations t.
Ensure: Approximate eigenvectors $\mathbf{V} = \{\mathbf{v}_k\}_{k=1}^{K}$.

1: $\mathbf{V} \leftarrow \emptyset$
2: $\boldsymbol{\Sigma}_1 \leftarrow \boldsymbol{\Sigma}$
3: **for** $k = 1, \ldots, K$ **do**
4:      $\mathbf{v}_k \leftarrow$ PCA $(\boldsymbol{\Sigma}_k, t)$
5:      $\boldsymbol{\Sigma}_{k+1} \leftarrow \boldsymbol{\Sigma}_k - \mathbf{v}_k \mathbf{v}_k^\top \boldsymbol{\Sigma}_k \mathbf{v}_k \mathbf{v}_k^\top$
6:      $\mathbf{V} \leftarrow \mathbf{V} \cup \{\mathbf{v}_k\}$
7: **end for**
8: **return** $\mathbf{V}$

Deflation proceeds as follows. After approximating the top component $\mathbf{u}_1^*$, the matrix $\boldsymbol{\Sigma}$ is further processed to exist on the subspace that is orthogonal to the subspace which is spanned by the first component. This process is repeated by applying single-component PCA in (1) on the deflated matrix, which leads to an approximation of the second component $\mathbf{u}_2^*$, and so on, as below:

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}; \quad \mathbf{v}_k = \text{PCA}\left(\boldsymbol{\Sigma}_k, t\right);$$
$$\boldsymbol{\Sigma}_{k+1} = \boldsymbol{\Sigma}_k - \mathbf{v}_k \mathbf{v}_k^\top \boldsymbol{\Sigma}_k \mathbf{v}_k \mathbf{v}_k^\top, \tag{3}$$

where $\text{PCA}\left(\boldsymbol{\Sigma}_k, \cdot\right)$ returns a normalized approximation of the top eigenvector of the deflated matrix $\boldsymbol{\Sigma}_k$. One estimates the subsequent principal components by running the same single-component PCA algorithm repetitively.

In this work, we focus on the more realistic scenario where the sub-routine, $\text{PCA}\left(\boldsymbol{\Sigma}_k, \cdot\right)$, incurs numerical errors due to limited computation budget and finite precision: even solving (1) for a single component, our estimate is only an approximation to the true principal component. The overall procedure with inexact $\text{PCA}\left(\boldsymbol{\Sigma}_k, \cdot\right)$ sub-routine is described in Algorithm 1. Line 4 performs the sub-routine algorithm for t iterations, denoted by $\text{PCA}(\boldsymbol{\Sigma}_k, t)$, which approximately computes the top eigenvector of $\boldsymbol{\Sigma}_k$. In line 5, $\boldsymbol{\Sigma}_k$ is deflated using $\mathbf{v}_k$; hence, the numerical error from approximately solving $\mathbf{v}_k$ in line 4 affects the quality of the deflated matrix $\boldsymbol{\Sigma}_{k+1}$ in the next iteration.

To motivate why studying the accumulation of errors is critical, we apply spectral clustering (Pothen et al., 1990) on the MNIST dataset and measure its performance using the mutual information metric[1]. Spectral clustering involves computing the top-K eigenvectors from a similarity matrix and uses the corresponding entries in the K eigenvectors as features for the following clustering procedure. In our experiments, eigenvectors are computed using the deflation in Algorithm 1 with different precision levels of the PCA

---

[1]Higher mutual information indicates more accurate clustering. For more details about the experiment, please see Appendix D.

sub-routine, indicated by different power iteration steps. We can observe from Figure 1 that, as the number of power iteration steps increases, the mutual information increases, showing a higher clustering accuracy. This shows that the errors incurred from inexactly solving each top eigenvector in the deflation process indeed have a non-negligible impact on the downstream machine learning tasks.

This interdependent accumulation of errors coming from approximately solving the sub-routine –which results in the less accurate deflation matrix, and further affects the quality of the approximation of the top eigenvector in the next iteration– is precisely what we characterize in this work. To the best of our knowledge, *this is the first work that analyzes the inexact setting, although this procedure is vastly used in practice.* Our contributions can be summarized as follows:

- In Theorem 3.1, we characterize how the errors from approximately solving the sub-routine (PCA) propagate into the overall error of the deflation procedure. Informally, the error between the approximate principal component and the actual one needs to be exponentially small to control the accuracy for the top-K eigenvectors for an increasing K, where the base of the exponential growth depends on the eigengap of the target matrix.

- In Theorem 4.1, we consider the case where PCA is fixed to the power iteration (Müntz et al., 1913). By leveraging the directional information of the numerical error vector produced by each power iteration, we improve the error bound in Theorem 3.1 (which is agnostic to the PCA sub-routine) to exponential growth with constant base.

- Along with the proof, we also present empirical results to justify the core ideas in our theoretical analysis.

Related work. Most works focus on the convergence proof of numerically solving the top eigenvector of matrices (Lanczos, 1950; Jain et al., 2016; Xu et al., 2018b). However, these results only concern the case of the first eigenvector, and their result does not readily apply to numerical algorithms, such as deflation, that solve for multiple top eigenvectors.

Iterative methods for multi-component PCA include *Schur complement deflation* (Zhang, 2006), *orthogonalized deflation* (Sriperumbudur et al., 2007), and *projection methods* (Saad, 1988). Among them, Hotelling's deflation, our focus, was reported to have the highest accuracy in practice (Danisman et al., 2014). For the comparison of these methods, we refer the readers to Mackey (2008, Section 2), which extends some of the above deflation methods to the sparse PCA setting (Zou et al., 2006; Johnstone & Lu, 2009), which is out of scope of our work. Importantly, all the works mentioned above assume the PCA step in (3) is solved exactly, i.e., $t = \infty$. Works that analyze the resulting error when solving multiple top eigenvectors include Allen-Zhu & Li (2017);

Gemp et al. (2021). However, the theoretical analysis in Allen-Zhu & Li (2017) and Gemp et al. (2021) only applies to specific algorithms such as EigenGame and Oja++. At the same time, our work focuses on deflation, a widely used algorithmic framework for solving multiple eigenvectors.

Subspace methods such as orthogonal iterations (Golub & Van Loan, 2013) are also widely used for discovering multiple eigenvectors. However, the deflation serves a different purpose than subspace methods. In particular, the sequential discovery of the eigenvectors in the deflation methods applies to scenarios where users have a dynamic necessity of the eigenvectors and require flexibility, especially in modern ML settings (Fawzi et al., 2022; Coulaud et al., 2013). For instance, compared with orthogonal iteration, which outputs all k eigenvectors at the same time, the deflation method will be more suitable for scenarios where the top principal component is known but more eigenvectors need to be computed depending on the need, and scenarios where eigenvectors are output in a consecutive way to facilitate downstream tasks.

In terms of the analysis technique, our work utilizes Weyl's inequality (Weyl, 1912) and Davis-Kahan $\sin \Theta$ Theorem (Davis & Kahan, 1970) to characterize the differences between eigenvalues of two matrices. In our fine-grained analysis in Section 4, we also utilize the Neumann expansion introduced by Eldridge et al. (2017) and characterize the subspace that the top eigenvector of a perturbed matrix lies in with a similar technique to Chen et al. (2020). However, the problem they study is entirely orthogonal to ours.

2. Problem Setup and Notations

Notation. For a vector $\mathbf{a}$, we use $\|\mathbf{a}\|_2$ to denote its ℓ_2-norm. For a matrix $\mathbf{A}$, $\|\mathbf{A}\|_2$ to denote its spectral norm and $\|\mathbf{A}\|_F$ its Frobenius norm. $\lambda_j(\mathbf{A})$ denotes its j-th eigenvalue.

Problem setup. Given a symmetric matrix $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$ with eigenvectors $\mathbf{u}_1^*, \ldots \mathbf{u}_d^*$ and eigenvalues $\lambda_1^*, \ldots, \lambda_d^*$, sorted in descending order (c.f., Assumption 2.1), we want to estimate the top $K \leq d$ eigenvectors using the deflation method. Formally, let PCA be a sub-routine that returns an approximation of the top eigenvector of a given matrix. The deflation method repeatedly computes approximation $\mathbf{v}_k$ of the principal eigenvector of the current deflation matrix $\mathbf{\Sigma}_k$. It forms a new deflation matrix $\mathbf{\Sigma}_{k+1}$ by subtracting the component of $\mathbf{v}_k$ from $\mathbf{\Sigma}_k$. This process is made formal in Algorithm 1 and is mathematically defined in (3). If the sub-routine PCA is solved precisely, then the output satisfies $\mathbf{v}_k = \mathbf{u}_k^*$. In this ideal scenario, the sequence of deflation matrices has the following form:

$$\mathbf{\Sigma}_{k+1}^* = \mathbf{\Sigma}_k^* - \mathbf{u}_k^* \mathbf{u}_k^{*\top} \mathbf{\Sigma}_k^* \mathbf{u}_k^* \mathbf{u}_k^{*\top}; \quad \mathbf{\Sigma}_1^* = \mathbf{\Sigma}. \quad (4)$$

We term the scenario in (4) as *ideal deflation* and the matrices $\{\mathbf{\Sigma}_k^*\}_{k=1}^d$ as *"ground-truth" deflation matrices*. The

ideal deflation has the nice property that $\mathbf{u}_k^*$ is the top eigenvector of $\mathbf{\Sigma}_k^*$. Therefore, solving for the top-K eigenvectors can be exactly reduced to repeatedly solving for the top-1 eigenvector of the "ground-truth" deflation matrices.

Yet, solving for the top eigenvector exactly is challenging, which makes the ideal deflation almost impossible in practice. Thus, we are interested in a more general case, where PCA computes the top-eigenvector *inexactly*, and results in numerical error $\boldsymbol{\delta}_k$, as defined below:

$$\boldsymbol{\delta}_k := \text{PCA}(\mathbf{\Sigma}_k, t) - \mathbf{u}_k = \mathbf{v}_k - \mathbf{u}_k; \quad \|\boldsymbol{\delta}_k\|_2 > 0. \quad (5)$$

Here, $\mathbf{u}_k$ is the top eigenvector of $\mathbf{\Sigma}_k$; recall that $\mathbf{\Sigma}_k$ in (3) is constructed recursively using $\mathbf{v}_k$. When $\mathbf{v}_k$ is solved inexactly, we cannot guarantee that $\mathbf{\Sigma}_k = \mathbf{\Sigma}_k^*$. Consequently, it is almost always the case that the top eigenvector of $\mathbf{\Sigma}_k$ does not equal to $\mathbf{u}_k^*$, namely $\|\mathbf{u}_k - \mathbf{u}_k^*\|_2 > 0$. However, since PCA only knows $\mathbf{\Sigma}_k$, its output $\mathbf{v}_k$ converges to $\mathbf{u}_k$ instead of $\mathbf{u}_k^*$ when t is large. This difference between the top eigenvector of the matrices $\mathbf{\Sigma}_k$ and $\mathbf{\Sigma}_k^*$ further complicates the analysis of the deflation process in (3), since $\mathbf{v}_k$ returned by PCA$(\mathbf{\Sigma}_k, t)$ approximates $\mathbf{u}_k$ instead of $\mathbf{u}_k^*$, which further builds up the difference between $\mathbf{v}_k$ and $\mathbf{u}_k^*$ and thus the subsequent $\|\mathbf{\Sigma}_{k+1} - \mathbf{\Sigma}_{k+1}^*\|_F$.

Our work aims to characterize this complicated error propagation through the deflation steps. To build up our analysis, we make the following assumption on the matrix $\mathbf{\Sigma}$.

Assumption 2.1. The matrix $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$ is a real symmetric matrix with eigenvalues and eigenvectors $\{\lambda_i^*, \mathbf{u}_i^*\}_{i=1}^d$, satisfying $1 = \lambda_1^* > \lambda_2^* > \cdots > \lambda_d^* > 0$.

Assumption 2.1 states the real symmetric matrix $\mathbf{\Sigma}$ has strictly decaying eigenvalues. Observing such phenomenon is quite common in practice; see Papailiopoulos et al. (2014, Figure 1), Papyan (2020, Figure 1), and Goujaud et al. (2022, Figure 1) for instance. Further, $\lambda_1^* = 1$ is not strictly required but assumed without loss of generality.

Assumption 2.1 guarantees that $\mathbf{\Sigma}$ is positive definite. By construction, all $\mathbf{\Sigma}_k^*$'s are guaranteed to be positive semidefinite. To enforce a similar condition on $\mathbf{\Sigma}_k$'s, we need to characterize the difference between the eigenvalues of $\mathbf{\Sigma}_k^*$ and $\mathbf{\Sigma}_k$: for λ_k being the top eigenvalue of $\mathbf{\Sigma}_k$, we need to control $|\lambda_k^* - \lambda_k|$. This is possible by Weyl's inequality:

Lemma 2.2 (Weyl's Inequality (Weyl, 1912)). *Let $\mathbf{M}, \mathbf{M}^* \in \mathbb{R}^{d \times d}$ be real symmetric matrices. Let σ_j, σ_j^* be the j-th eigenvalue of $\mathbf{M}$ and $\mathbf{M}^*$, respectively. Then:*

$$|\sigma_j^* - \sigma_j| \leq \|\mathbf{M}^* - \mathbf{M}\|_2.$$

A direct consequence of Weyl's inequality to our scenario is that $|\lambda_k^* - \lambda_k| \leq \|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\|_2$.

Another tool that we will utilize in the following sections of our paper is the Davis-Kahan $\sin \Theta$ Theorem:

Lemma 2.3 (sin Θ Theorem (Davis & Kahan, 1970)). *Let* $\mathbf{M}^* \in \mathbb{R}^{d \times d}$ *and let* $\mathbf{M} = \mathbf{M}^* + \mathbf{H}$. *Let* $\mathbf{a}_1^*$ *and* $\mathbf{a}_1$ *be the top eigenvectors of* $\mathbf{M}^*$ *and* $\mathbf{M}$, *respectively. Then we have:*

$$\sin \angle \{\mathbf{a}_1^*, \mathbf{a}_1\} \leq \frac{\|\mathbf{H}\|_2}{\min_{j \neq k} |\sigma_k^* - \sigma_j|}.$$

3. Sub-routine Agnostic Error Propagation

In this section, we aim to provide a characterization of the error propagation in the deflation methods in (3) that is agnostic to the detail of the sub-routine PCA. In particular, the only information known about PCA is the magnitude of the error of PCA, namely $\|\boldsymbol{\delta}_k\|_2 = \|\mathbf{v}_k - \mathbf{u}_k\|_2 = \|\text{PCA}(\boldsymbol{\Sigma}_k, t) - \mathbf{u}_k\|_2$.

Theorem 3.1. *Consider the scenario of looking for the top-K eigenvectors of* $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$, *and suppose* $\boldsymbol{\Sigma}$ *satisfies Assumption 2.1. Let* $\mathcal{T}_j = \lambda_j^* - \lambda_{j+1}^*$, *for* $j \in [d-1]$, *and* $\mathcal{T}_d = \lambda_d^*$. *Also, let* $\mathcal{T}_{K,\min} = \min_{k \in [K]} \mathcal{T}_k$, *and let* $\boldsymbol{\delta}_k$ *and* $\mathbf{v}_k$ *be defined in (5). If* $\|\boldsymbol{\delta}_k\|_2$*'s are small enough such that*

$$\sum_{k'=1}^{k-1} \lambda_{k'}^* \|\boldsymbol{\delta}_{k'}\|_2 \prod_{j=k'+1}^{k-1} \left(3 + \frac{2\lambda_j^*}{\mathcal{T}_j}\right) \leq \frac{1}{20} \mathcal{T}_{K,\min}. \quad (6)$$

then the output of Algorithm 1 satisfies for all $k \in [K]$:

$$\|\mathbf{v}_k - \mathbf{u}_k^*\|_2 \leq \|\mathbf{u}_k - \mathbf{u}_k^*\|_2 + \|\boldsymbol{\delta}_k\|_2$$
$$\leq 5 \sum_{k'=1}^{k} \frac{\lambda_{k'}^*}{\lambda_k^*} \|\boldsymbol{\delta}_{k'}\|_2 \prod_{j=k'+1}^{k} \left(3 + \frac{2\lambda_j^*}{\mathcal{T}_j}\right). \quad (7)$$

Remark 1. Theorem 3.1 characterizes how the errors from approximately solving the sub-routine (PCA) propagate into the overall error of the deflation procedure. The condition in (6) is to make sure that the difference between the "ground-truth" deflated matrices in (4) and the empirical ones in (3) is controlled and is sufficiently small, i.e., $\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_F \leq \frac{1}{4}\mathcal{T}_{k,\min}$ for all $k \in [K]$. As a simplification, this condition can be guaranteed as long as

$$\|\boldsymbol{\delta}_k\|_2 \leq \frac{\mathcal{T}_{K,\min}}{20K} \prod_{j=k+1}^{K-1} \left(3 + \frac{2\lambda_j^*}{\mathcal{T}_j}\right)^{-1}.$$

Remark 2. Notice that the error upper bound in (7) takes the form of a summation over components that depend on the error of the sub-routine $\{\|\boldsymbol{\delta}_{k'}\|_2\}_{k'=1}^k$. In particular, not only do the number of summands grow as k grows, but each summand also has a multiplicative factor of $\prod_{j=k'+1}^{k} \left(3 + \frac{2\lambda_j^*}{\mathcal{T}_j}\right)$ which grows as k becomes larger. When the eigen spectrum of $\boldsymbol{\Sigma}^*$ decays slowly, this multiplicative factor can grow near factorially. For instance, in the case of a power-law decay spectrum where $\lambda_j = \frac{1}{j}$, the multiplicative factor becomes $\prod_{j=k'+1}^{k} (3 + 2j)$, since

$\mathcal{T}_j = \frac{1}{j(j+1)}$. In this case, the errors $\{\|\boldsymbol{\delta}_{k'}\|_2\}_{k'=1}^k$ need to be nearly factorially small as we attempt to solve for more eigenvectors. To be more specific, based on Theorem 3.1, to guarantee that $\|\mathbf{v}_k - \mathbf{u}_k^*\|_2 \leq \varepsilon$, it is *necessary* to have:

$$\|\boldsymbol{\delta}_{k'}\| \leq \frac{\varepsilon \lambda_k^*}{5 \lambda_{k'}^*} \prod_{j=k'+1}^{k} \left(3 + \frac{2\lambda_j^*}{\mathcal{T}_j}\right)^{-1}.$$

The corollary below states a *sufficient* condition to guarantee that $\|\mathbf{v}_k - \mathbf{u}_k^*\|_2 \leq \varepsilon$ for all $k \in [K]$.

Corollary 3.2. *Consider the scenario of solving the top-K eigenvectors of* $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$, *and suppose* $\boldsymbol{\Sigma}$ *satisfies Assumption 2.1. Let* $\mathcal{T}_j = \lambda_j^* - \lambda_{j+1}^*$ *for* $j \in [d-1]$, *and* $\mathcal{T}_d = \lambda_d^*$. *Also let* $\mathcal{T}_{K,\min} = \min_{k \in [K]} \mathcal{T}_k$. *Finally, let* $\boldsymbol{\delta}_i$ *and* $\mathbf{v}_i$ *be defined as in (5). If for all* $k \in [K]$ *it holds that:*

$$\|\boldsymbol{\delta}_k\|_2 \leq \frac{\min\{\varepsilon \lambda_K^*, \mathcal{T}_{K,\min}\}}{20K} \prod_{j=k+1}^{K} \left(3 + \frac{2\lambda_j^*}{\mathcal{T}_j}\right)^{-1}, \quad (8)$$

then the output of Algorithm 1 satisfies $\|\mathbf{v}_k - \mathbf{u}_k^*\|_2 \leq \varepsilon$ *for all* $k \in [K]$.

Remark 3. Corollary 3.2 characterizes how accurately the sub-routine (PCA) has to be solved to achieve $\|\mathbf{v}_k - \mathbf{u}_k^*\|_2 \leq \varepsilon, \forall k \in [K]$, i.e., the desired accuracy of the overall deflation procedure.

Below, we provide lower bounds on the number of sub-routine iterations to satisfy (8), assuming a linearly-converging sub-routine with convergence rate α_k exists. In Section 4, we improve this rate by specifically setting the sub-routine to be the *power iteration* (Müntz et al., 1913).

Corollary 3.3. *Under the same assumptions as Corollary 3.2, we further assume a sub-routine exists for computing the top eigenvalue with linear convergence rate* $\alpha_k \in (0, 1)$. *If* $\|\boldsymbol{\delta}_k\|_2 \leq c_0 \cdot \alpha_k^t$ *for some constant* $c_0 > 0$ *for all* $k \in [K]$, *where* t *is the number of steps in the sub-routine satisfying:*

$$t \geq \Omega\left(\frac{\log \frac{c_0 K}{\min\{\varepsilon \lambda_k^*, \mathcal{T}_{K,\min}\}} + \sum_{j=k+1}^{K} \log\left(\frac{\lambda_j^*}{\mathcal{T}_j} + 1\right)}{\log \alpha_k^{-1}}\right), \quad (9)$$

then the output of Algorithm 1 satisfies $\|\mathbf{v}_k - \mathbf{u}_k^*\|_2 \leq \varepsilon$ *for all* $k \in [K]$.

Remark 4. Corollary 3.3 is constructed by further assuming that $\|\boldsymbol{\delta}_k\|_2 \leq c_0 \cdot \alpha_k^t$ based on Corollary 3.2 and derive the lower bound on t. The linear convergence property of solving the top eigenvector is shared among multiple existing algorithms (Xu et al., 2018a; Golub & Van Loan, 1996). In particular, when PCA is chosen to be the power iteration, we can show that in our setting $\|\boldsymbol{\delta}_k\|_2 \leq c_0 \left(\frac{\lambda_2(\boldsymbol{\Sigma}_k)}{\lambda_k}\right)^t$. In

the case of a power-law decay spectrum, when we treat c_0 as constants, the bound in (9) can be simplified to

$$t \geq \Omega \left(\log \frac{K}{\varepsilon} + \sum_{j=k+1}^{K} \log (j+1) \right).$$

Applying a slightly loose bound over the summation, we can see that based on the analysis of Theorem 3.1, one needs $t \geq \Omega \left(\log \frac{K}{\varepsilon} + K \log K \right)$ to achieve ε-accuracy. Section 4 compares this rate with the lower bound we obtain when PCA is fixed to power iteration.

3.1. Proof Overview

The proof involves bounding $\|\mathbf{u}_k - \mathbf{u}_k^*\|_2$ and $\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\|_F$. The former quantifies the difference between the true k-th eigenvector of $\mathbf{\Sigma}$ and the leading eigenvector of the k-th deflation step $\mathbf{\Sigma}_k$. The latter quantifies the difference between the "ground-truth" deflation matrices defined in (4) and the ones in practice. Recall that $\mathbf{u}_k^*$ is the top eigenvector of $\mathbf{\Sigma}_k^*$ (c.f., Lemma C.1). Let $\mathrm{nev}_1 (\mathbf{\Sigma}_k)$ denote a normalized top eigenvector of $\mathbf{\Sigma}_k$. Since both $\mathrm{nev}_1 (\mathbf{\Sigma}_k)$ and $-\mathrm{nev}_1 (\mathbf{\Sigma}_k)$ are unit-norm eigenvectors of $\mathbf{\Sigma}_k$, we will choose $\mathbf{u}_k$ to be the one such that $\mathbf{u}_k^\top \mathbf{u}_k^* \geq 0$:

$$\mathbf{u}_k := \underset{s \in \{\pm 1\}}{\arg \min} \| s \cdot \mathrm{nev}_1 (\mathbf{\Sigma}_k) - \mathbf{u}_k^* \|_2 \cdot \mathrm{nev}_1 (\mathbf{\Sigma}_k). \tag{10}$$

The difference between $\mathbf{v}_k$, the k-th approximate eigenvector returned by Algorithm 1, and $\mathbf{u}_k^*$, the k-th ground-truth eigenvector, consists of two components: $i)$ $\|\mathbf{u}_k - \mathbf{u}_k^*\|_2$, the difference between the top eigenvectors of the empirical deflated matrix $\mathbf{\Sigma}_k$ and the "ground-truth" deflated matrix $\mathbf{\Sigma}_k^*$; and $ii)$ $\|\mathbf{v}_k - \mathbf{u}_k\|_2$, the numerical error from not solving the sub-routine $\mathrm{PCA}(\cdot)$ exactly. We can decompose the difference $\|\mathbf{v}_k - \mathbf{u}_k^*\|_2$ in the following sense:

$$\|\mathbf{v}_k - \mathbf{u}_k^*\|_2 \leq \underbrace{\|\mathbf{v}_k - \mathbf{u}_k\|_2}_{:=\|\boldsymbol{\delta}_k\|_2} + \|\mathbf{u}_k - \mathbf{u}_k^*\|_2. \tag{11}$$

where the last equality follows from the definition in (5). The norm of $\|\boldsymbol{\delta}_k\|_2$ depends largely on the sub-routine PCA and can be controlled as long as t, the number of sub-routine iterations, is large enough. Therefore, our analysis will mainly focus on upper-bounding $\|\mathbf{u}_k - \mathbf{u}_k^*\|_2$. Intuitively, $\|\mathbf{u}_k - \mathbf{u}_k^*\|_2$ depends on the difference between $\mathbf{\Sigma}_k$ and $\mathbf{\Sigma}_k^*$. As a building block of our proof, we will first provide a characterization of $\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\|_F$.

Lemma 3.4. *Suppose that* $\|\boldsymbol{\delta}_k\|_2 \leq \frac{1}{6}$ *for all* $k \in [K-1]$, *then we have that for all* $k \in [K-1]$,

$$\begin{aligned} \left\| \mathbf{\Sigma}_{k+1} - \mathbf{\Sigma}_{k+1}^* \right\|_F &\leq 3 \left\| \mathbf{\Sigma}_k - \mathbf{\Sigma}_k^* \right\|_F + 5\lambda_k^* \left\| \boldsymbol{\delta}_k \right\|_2 \\ &\quad + 2\lambda_k^* \left\| \mathbf{u}_k - \mathbf{u}_k^* \right\|_2. \end{aligned} \tag{12}$$

Lemma 3.4 upper-bounds $\left\| \mathbf{\Sigma}_{k+1} - \mathbf{\Sigma}_{k+1}^* \right\|_F$ in terms of $\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\|_F$, $\|\mathbf{u}_k - \mathbf{u}_k^*\|_2$, and $\|\boldsymbol{\delta}_k\|_2$. To establish an recursive characterization of $\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\|_F$, we need to obtain an upper bound for $\|\mathbf{u}_k - \mathbf{u}_k^*\|_2$. Indeed, this can be easily obtained by applying the Davis-Kahan $\sin \Theta$ theorem (Davis & Kahan, 1970; Eldridge et al., 2017).

Lemma 3.5. *Let* $\mathcal{T}_k := \min_{j \neq k} |\lambda_k^* - \lambda_j^*|$. *If* $\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\|_F \leq \frac{1}{4} \mathcal{T}_k$, *then we shall have that*

$$\|\mathbf{u}_k - \mathbf{u}_k^*\|_2 \leq \frac{2}{\mathcal{T}_k} \|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\|_F. \tag{13}$$

Plugging (13) into (12) gives a recurrence that depends purely on $\|\boldsymbol{\delta}_k\|_2$ and the spectrum of $\mathbf{\Sigma}$:

$$\begin{aligned} \left\| \mathbf{\Sigma}_{k+1} - \mathbf{\Sigma}_{k+1}^* \right\|_F &\leq \left(3 + \frac{2}{\mathcal{T}_k} \right) \|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\|_F \\ &\quad + 5\lambda_k^* \|\boldsymbol{\delta}_k\|_2. \end{aligned}$$

Unrolling this recurrence gives a closed-form upper bound for $\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\|_F$. Combining this upper bound with (13) and plugging the result into (11) gives us an upper bound for $\|\mathbf{v}_k - \mathbf{u}_k^*\|_2$.

4. Error Propagation When Using Power Iteration

In this section, we turn our focus to a specific sub-routine algorithm (PCA in line 5 of Algorithm 1) for finding the top eigenvectors: the power iteration method (Müntz et al., 1913). Given a symmetric matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$ with rank r and an initialization vector $\mathbf{x}_0 \in \mathbb{R}^d$, the power iteration executes:

$$\hat{\mathbf{x}}_{t+1} = \mathbf{M} \mathbf{x}_t; \quad \mathbf{x}_{t+1} = \frac{\hat{\mathbf{x}}_{t+1}}{\|\hat{\mathbf{x}}_{t+1}\|_2}. \tag{14}$$

Assuming that PCA is the power iteration not only gives us the convergence rate of $\|\boldsymbol{\delta}_k\|_2$ in terms of the number of steps in each power iteration, but also provides the directional information $\boldsymbol{\delta}_k$. In particular, one could see that $\left| \boldsymbol{\delta}_k^\top \mathbf{u}_{k+j-1} \right| \leq c_0 \left(\frac{\lambda_j(\Sigma)_k}{\lambda_k^*} \right)^t$ for some constant c_0[2]. Namely, the component of $\boldsymbol{\delta}_k$ along $\mathbf{u}_j$ has an exponentially smaller magnitude if j is larger, depending on the eigen spectrum of $\mathbf{\Sigma}^*$. By leveraging this additional information, our theorem below shows that we can improve the bound on $\|\mathbf{v}_k - \mathbf{u}_k^*\|_2$, when PCA is specified to the power iteration, compared to the sub-routine agnostic result in Theorem 3.1.

Theorem 4.1. *Consider the problem of looking for the top-K eigenvectors of $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$ under Assumption 2.1. Let $\mathcal{T}_i = \lambda_i^* - \lambda_{i+1}^*$ for $i \in [d-1]$ and $\mathcal{T}_d = \lambda_d^*$, and let $\mathcal{T}_{k,\min} = \min_{i \in [k]} \mathcal{T}_i$. Assume that there exists a constant $c_0 \in (0, \infty)$ such that the initialized vector $\mathbf{x}_{0,k}$ in the k-th*

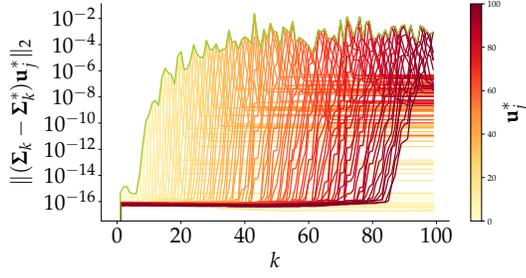[2] For the detail of proving this statement please see Lemma C.6

Figure 2. Dynamics of $\left\| (\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*) \mathbf{u}_j^* \right\|_2$ with respect to the change of k. Each $\mathbf{u}_j^*$ is represented by a different color, with light color for small j and dark color for large j. Experiments done for $\boldsymbol{\Sigma} \in \mathbb{R}^{100 \times 100}$, $\lambda_k^* = \frac{1}{k}$, $\{\mathbf{u}_k^*\}_{k=1}^d$ being randomly generated orthogonal basis, and $t = 200$. The orthogonal basis $\{\mathbf{u}_k^*\}_{k=1}^d$ is generated by randomly sampling a matrix with I.I.D. Gaussian entries, and computing its left singular vectors.

power iteration procedure satisfies $\left| \mathbf{x}_{0,k}^\top \mathbf{u}_k \right| > c_0^{-1}$ *for all* $k \in [K]$.[3] *Let* $\mathcal{G} = \max_{k \in [K]} \left(1 + \frac{c_0 \lambda_k^* \lambda_{k+1}^*}{\lambda_k^* - \lambda_{k+1}^*} \right)$. *If we use* t *steps of power iteration to solve for the top eigenvectors, where* t *satisfies:*

$$
t \geq \begin{cases} \log 2\mathcal{G}_k \left(\log \frac{\lambda_{k'+1}^* + 7\lambda_{k'}^*}{\lambda_{k'}^* + 7\lambda_{k'+1}^*} \right)^{-1}, & \forall k \leq K \\ \frac{1}{\log \lambda_k^* - \log \lambda_{k+1}^*}, & \forall k \leq d \end{cases} \quad (15)
$$

Moreover, if t *is also large enough to guarantee that*

$$
\sum_{k'=1}^{K-1} 8^{K-k'} \frac{\lambda_{k'}^*}{\mathcal{T}_{k'}} \left(\frac{7\lambda_{k+1}^* + \lambda_k^*}{7\lambda_k^* + \lambda_{k+1}^*} \right)^t \leq \frac{\mathcal{T}_{K,\min}}{140 c_0}, \quad (16)
$$

Then we can guarantee that for all $k \in [K]$

$$
\begin{aligned}
&\left\| \mathbf{v}_k - \mathbf{u}_k^* \right\|_2 \\
&\leq 3 \sum_{k'=1}^{k} 8^{k-k'} \frac{\lambda_{k'}^*}{\lambda_k^*} \left(5 \left\| \boldsymbol{\delta}_{k'} \right\|_2 + \frac{7c_0}{\mathcal{T}_{k'}} \left(\frac{\lambda_{k'+1}^*}{\lambda_{k'}^*} \right)^t \right).
\end{aligned} \quad (17)
$$

Remark 5. Similar to Theorem 3.1, the requirement of t in (16) is to guarantee that $\left\| \boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^* \right\|_F \leq \frac{1}{8} \mathcal{T}_{K,\min}$. In particular, (16) combined with (15) can be satisfied by the following simplified condition on t

$$
t \geq \begin{cases} \left(\frac{\max\left\{ \log 2\mathcal{G}, 3(K-k) + \log \frac{140 c_0 K}{\mathcal{T}_{K,\min}^2} \right\}}{\log\left(7\lambda_k^* + \lambda_{k+1}^*\right) - \log\left(7\lambda_{k+1}^* + \lambda_k^*\right)} \right), & \forall k \leq K \\ \frac{1}{\log \lambda_k^* - \log \lambda_{k+1}^*}, & \forall k \leq d \end{cases} \quad (18)
$$

---

[3]This is a standard assumption to guarantee the convergence of the power iteration, as in Xu et al. (2018a); Golub & Van Loan (1996)

The upper bound on $\left\| \mathbf{v}_k - \mathbf{u}_k^* \right\|_2$ in (17) also takes a similar form to (7). Recall that when PCA is chosen to be the power iteration, we have $\left\| \boldsymbol{\delta}_k \right\|_2 \leq c_0 \left(\frac{\lambda_2(\boldsymbol{\Sigma}_k)}{\lambda_k} \right)^t \approx c_0 \left(\frac{\lambda_{k+1}^*}{\lambda_k^*} \right)^t$. Therefore, one can treat $\left(5 \left\| \boldsymbol{\delta}_{k'} \right\|_2 + \frac{7c_0}{\mathcal{T}_{k'}} \left(\frac{\lambda_{k'+1}^*}{\lambda_{k'}^*} \right)^t \right)$ in (17) as roughly $\frac{C}{\mathcal{T}_{k'}} \left\| \boldsymbol{\delta}_{k'} \right\|_2$ for some constant C. Thus, the $\prod_{j=k+1}^{K} \left(3 + \frac{2\lambda_j^*}{\mathcal{T}_j} \right)$ scaling of each summand in (7) that depends on $\mathcal{T}_j$ is improved to the exponential scaling of 8^{K-k} in (17). This improvement is further made clear in the corollary below.

Corollary 4.2. *Under the same assumptions as in Theorem 4.1, if we use* t *steps of power iteration to solve for the top eigenvectors such that:*

$$
t \geq \begin{cases} \Omega\left(\frac{\max\left\{ \log \mathcal{G}, K-k + \log \frac{c_0 K}{\epsilon \mathcal{T}_{K,\min}} \right\}}{\log\left(7\lambda_k^* + \lambda_{k+1}^*\right) - \log\left(7\lambda_{k+1}^* + \lambda_k^*\right)} \right), & \forall k \leq K, \; and \\ \Omega\left(\frac{1}{\log \lambda_k^* - \log \lambda_{k+1}^*} \right), & \forall k \leq d, \end{cases} \quad (19)
$$

then we can guarantee $\left\| \mathbf{v}_k - \mathbf{u}_k \right\|_2 \leq \varepsilon$ *for all* $k \in [K]$.

Remark 6. Corollary 4.2 is obtained by setting $\left\| \boldsymbol{\delta}_k \right\|_2 \leq c_0 \left(\frac{\lambda_2(\boldsymbol{\Sigma}_k)}{\lambda_k} \right)^t$ in power iteration and notice that $\frac{\lambda_2(\boldsymbol{\Sigma}_k)}{\lambda_k} \leq \frac{7\lambda_{k+1}^* + \lambda_k^*}{7\lambda_k^* + \lambda_{k+1}^*}$ when $\left\| \boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^* \right\|_2 \leq \frac{1}{8} \mathcal{T}_k$. When we treat c_0 and $\{\lambda_k^*\}_{k=1}^d$ as constants, $\mathcal{G}$ also becomes a constant. In this case, the lower bound on t in (19) becomes $t \geq \Omega\left(\log \frac{K}{\epsilon} + K \right)$. Compared with the $t \geq \Omega\left(\log \frac{K}{\epsilon} + K \log K \right)$ requirement we derived in the discussion of Corollary 3.3. Thus, our lower bound in Corollary 4.2 saves a factor of $\log K$.

Remark 7. In the discussion about Theorem 4.1 and Corollary 4.2 above, we deliberately ignored the effect of c_0 and $\mathcal{G}$. For $\mathcal{G}$, we observe that $\mathcal{G}$ can also be written as $\mathcal{G} = \max_{k \in [K]} \left(1 + c_0 \left(\frac{1}{\lambda_{k+1}^*} - \frac{1}{\lambda_k^*} \right)^{-1} \right)$. If the eigenvalues of $\boldsymbol{\Sigma}$ follow a power law decay as $\lambda_k^* = \frac{1}{k^\gamma}$ for some $\gamma \geq 1$, then we have $\frac{1}{\lambda_{k+1}^*} - \frac{1}{\lambda_k^*} \geq 1$ and thus $\mathcal{G} \leq c_0 + 1$. Next, we focus on c_0. If in each power iteration sub-routine, the initialized vector $\mathbf{x}_{0,k}$ is generated randomly, then $\left| \mathbf{x}_{0,k}^\top \mathbf{u}_k \right|$ is independent of $\mathbf{u}_k$, and $\mathbb{P}\left(\left| \mathbf{x}_{0,k}^\top \mathbf{u}_k \right| > 0 \right) = 1$. This idea can be made concrete in the following lemma

Lemma 4.3. *Let* $\mathbf{x}_{0,k}$ *be sampled uniformly from a unit sphere in* $\mathbb{R}^d$. *Then with probability at least* $1 - \frac{2K}{3d}$ *we have that* $\left| \mathbf{x}_{0,k}^\top \mathbf{u}_k \right| \geq \frac{1}{\sqrt{1+2d^3}}$ *for all* $k \in [K]$.

The proof of Lemma 4.3 follows the idea of (Wang et al., 2020), and is deferred to Appendix B.1. Next, we will first explore the property of power iteration crucial to our proof
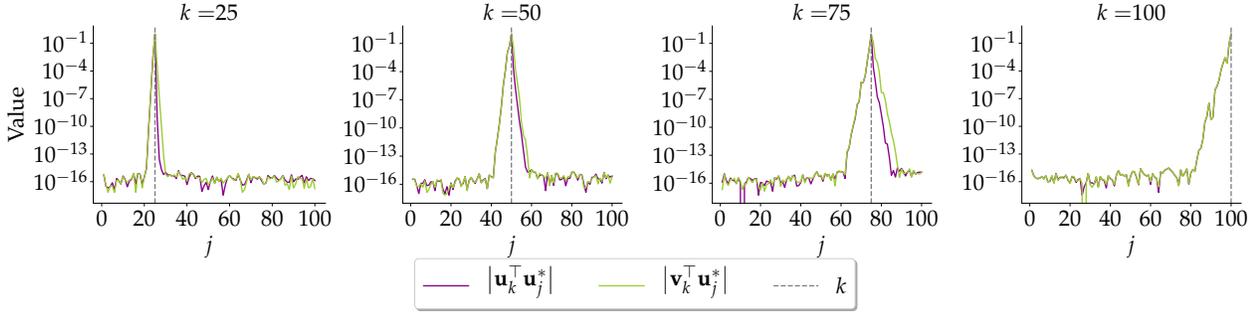
Figure 3. The comparison among the dynamics of $\mathbf{u}_k^\top \mathbf{u}_j^*$ and $\mathbf{v}_k^\top \mathbf{u}_j^*$ with respect to the change of k for $j \in \{25, 50, 75, 100\}$. Experiments are performed for $\boldsymbol{\Sigma} \in \mathbb{R}^{100 \times 100}$, $\lambda_k^* = \frac{1}{k}$, $\{\mathbf{u}_k^*\}_{k=1}^d$ being randomly generated orthogonal basis, with $t = 200$. The results show that both $\left|\mathbf{u}_k^\top \mathbf{u}_j^*\right|$ and $\left|\mathbf{v}_k^\top \mathbf{u}_j^*\right|$ are small only when k is near j. The orthogonal basis $\{\mathbf{u}_k^*\}_{k=1}^d$ is generated by randomly sampling a matrix with I.I.D. Gaussian entries, and computing its left singular vectors.

of Theorem 4.1 in Section 4.1, and then provide an overview of the proof in Section 4.2.

4.1. What Does Power Iteration Tell Us?

Recall our previous observation that the error vector returned by power iteration has different magnitudes along different eigenvector directions of the matrix of interest. In this section, we perform the same analysis, assuming that the matrix passed into the power iteration sub-routine is a perturbation of some "ground-truth" matrix. Let $\mathbf{a}_j^*$ be the j-th eigenvector of $\mathbf{M}^*$ and $\mathbf{M} = \mathbf{M}^* + \mathbf{H}$. Given the definition of $\mathbf{x}_t$ as the result of a t-step power iteration on $\mathbf{M}$, defined in (14), we will show that $\mathbf{x}_t^\top \mathbf{a}_j^*$'s take different forms depending on the index of the eigenvector j.

Lemma 4.4. *Let $\mathbf{x}_t$ be the result of running power iteration starting from $\mathbf{x}_0$ for t iterations, as defined in (14). Let $\mathbf{M} = \mathbf{M}^* + \mathbf{H}$. Moreover, let $\sigma_j^*, \mathbf{a}_j^*$ be the j-th eigenvalue and eigenvector of $\mathbf{M}^*$, and let σ_j be the j-th eigenvalue of $\mathbf{M}$. Assume that there exists $c_0 \in (0, \infty)$ such that $\left|\mathbf{x}_0^\top \mathbf{a}_1\right| \geq c_0^{-1}$. Then, for all $j = 2, \ldots, r$, we have:*

$$\left|\mathbf{x}_t^\top \mathbf{a}_j^*\right| \leq c_0 \left(\left(\frac{\sigma_j^*}{\sigma_1}\right)^t + \frac{\sigma_j^*}{\sigma_1 - \sigma_j^*} \left\|\mathbf{H}\mathbf{a}_j^*\right\|_2 \right). \quad (20)$$

Lemma 4.4 upper bounds the alignment of the vector in the tth step of the power iteration with any incorrect eigenvector $\mathbf{a}_j^*$. The magnitude of $\mathbf{x}_t^\top \mathbf{a}_j^*$ depends both on $\left(\frac{\sigma_j^*}{\sigma_1}\right)^t$ and how the perturbation matrix aligns with the j-th eigenvector. When $\mathbf{M}^*$ has a decaying spectrum, the first term in the upper bound of $\left|\mathbf{x}_t^\top \mathbf{a}_j^*\right|$ in (20) decreases as j becomes larger. Overloading Lemma 4.4 in the notations of our deflation procedure, we have

$$\left|\mathbf{v}_k^\top \mathbf{u}_j^*\right| \leq c_0 \left(\left(\frac{\lambda_j^*}{\lambda_1}\right)^t + \frac{\lambda_j^*}{\lambda_1 - \lambda_j^*} \left\|\left(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\right)\mathbf{u}_j^*\right\|_2 \right)$$

In short, the component of $\mathbf{v}_k$ along an incorrect direction, $\left|\mathbf{v}_k^\top \mathbf{u}_j^*\right|$, is determined by both the convergence along the jth eigenvector, and by the difference between the two deflated matrix along the direction of $\mathbf{u}_j^*$. This calls for analysis on $\left\|\left(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\right)\mathbf{u}_j^*\right\|_2$. Indeed, since the matrix residue $\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*$ is resulted from $\mathbf{v}_k - \mathbf{u}_k$'s, it is natural to believe that $\left\|\left(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\right)\mathbf{u}_j^*\right\|_2$ can also be upper-bounded by quantities depending on j. Our next lemma characterizes this behavior

Lemma 4.5. *Let $\mathbf{v}_k$ be the output of the k-th power iteration. For all $k \in [d]$ and $j \geq k$, we have:*

$$\left\|\left(\boldsymbol{\Sigma}_k^* - \boldsymbol{\Sigma}_k\right)\mathbf{u}_j^*\right\|_2 \leq \sum_{k'=1}^{k-1} \lambda_{k'} \left|\mathbf{v}_{k'}^\top \mathbf{u}_j^*\right|. \quad (21)$$

The proof of Lemma 4.5 is deferred to Appendix B.3. In particular, (21) in Lemma 4.5 upper-bounds $\left\|\left(\boldsymbol{\Sigma}_k^* - \boldsymbol{\Sigma}_k\right)\mathbf{u}_j^*\right\|_2$ using a linear combination of $\left\{\left|\mathbf{v}_{k'}^\top \mathbf{u}_j^*\right|\right\}_{k'=1}^{k-1}$. Indeed, recall that $\mathbf{v}_k$ is the output of the k-th power iteration procedure. This means that we can invoke Lemma 4.4 with $\mathbf{M} = \boldsymbol{\Sigma}_k, \mathbf{M}^* = \boldsymbol{\Sigma}_k^*$, and correspondingly $\mathbf{H} = \boldsymbol{\Sigma}_k^* - \boldsymbol{\Sigma}_k$. Plugging the bound of $\left\|\left(\boldsymbol{\Sigma}_k^* - \boldsymbol{\Sigma}_k\right)\mathbf{u}_j^*\right\|_2$ from (21) into then gives the recursively defined $\left|\mathbf{v}_k^\top \mathbf{u}_j^*\right|$:

$$\left|\mathbf{v}_k^\top \mathbf{u}_j^*\right| \leq c_0 \left(\left(\frac{\lambda_j^*}{\lambda_k}\right)^t + \frac{\lambda_j^*}{\lambda_k - \lambda_j^*} \sum_{k'=1}^{k-1} \lambda_{k'} \left|\mathbf{v}_{k'}^\top \mathbf{u}_j^*\right| \right). \quad (22)$$

(22) shows the independence of each $\left\{\left|\mathbf{v}_k^\top \mathbf{u}_j^*\right|\right\}_{k=1}^d$ for different j, which allows each $\left\{\left|\mathbf{v}_k^\top \mathbf{u}_j^*\right|\right\}_{k=1}^d$ to grow at different speeds. To see whether this characterization of independent growth is crucial, observe that in Figure 3, each $\left|\mathbf{v}_k^\top \mathbf{u}_j^*\right|$ only goes through a fast growth when j is near k. This behavior results in the phenomena that at each step k, there are only a few j's where $\left\|\left(\boldsymbol{\Sigma}_k^* - \boldsymbol{\Sigma}_k\right)\mathbf{u}_j^*\right\|_2$ is large.

7
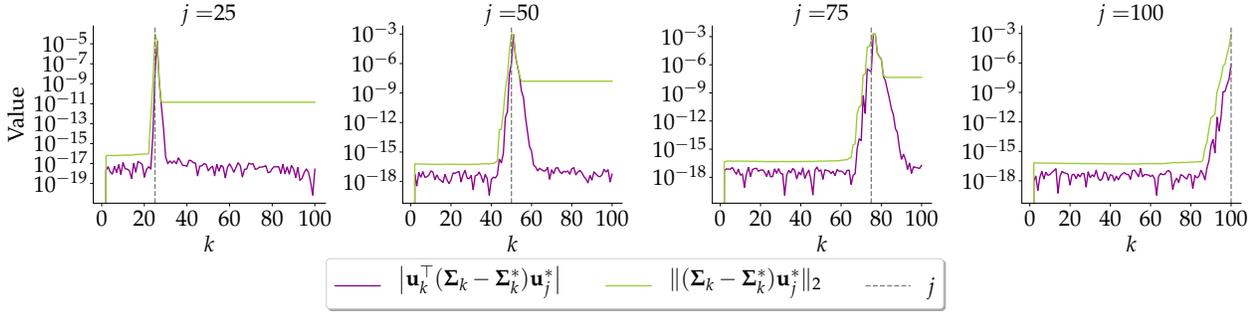
Figure 4. The comparison of between the dynamics of $\left|\mathbf{u}_k^\top \left(\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\right) \mathbf{u}_k^*\right|$ and $\left\|\left(\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\right) \mathbf{u}_j^*\right\|_2$ with respect to the change of k for $j \in \{25, 50, 75, 100\}$. Experiments are performed for $\mathbf{\Sigma} \in \mathbb{R}^{100 \times 100}$, $\lambda_k^* = \frac{1}{k}$, $\{\mathbf{u}_k^*\}_{k=1}^d$ as randomly generated orthogonal basis, with $t = 200$. The results show that $\left\|\left(\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\right) \mathbf{u}_j^*\right\|_2$ is a good approximation of $\left|\mathbf{u}_k^\top \left(\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\right) \mathbf{u}_k^*\right|$ when the latter is large. The orthogonal basis $\{\mathbf{u}_k^*\}_{k=1}^d$ is generated by randomly sampling a matrix with I.I.D. Gaussian entries, and computing its left singular vectors.

We can see this as we unroll this recursive form and obtain

$$\left\|\left(\mathbf{\Sigma}_k^* - \mathbf{\Sigma}_k\right) \mathbf{u}_j^*\right\|_2 \leq \sum_{k'=1}^{k-1} \lambda_{k'} \left|\mathbf{v}_{k'}^\top \mathbf{u}_j^*\right|$$
$$\leq c_0 \sum_{k'=1}^{k-1} \mathcal{G}_k^{k-k'-1} \left(\frac{\lambda_j^*}{\lambda_k}\right)^{t-1} \quad (23)$$

where $\mathcal{G}_k = 1 + \frac{c_0 \lambda_k^* \lambda_{k+1}^*}{\lambda_k^* - \lambda_{k+1}^*}$. (23) upper bounds $\left\|\left(\mathbf{\Sigma}_k^* - \mathbf{\Sigma}_k\right) \mathbf{u}_j^*\right\|_2$ with a summation of the linear decay terms resulted from the power iteration. For the detailed computation of unrolling the recursive definition, we refer the reader to Appendix B.7 and Appendix C (see Lemma C.3). Noticeably, (23) provides a closed-form upper bound of $\left\|\left(\mathbf{\Sigma}_k^* - \mathbf{\Sigma}_k\right) \mathbf{u}_j^*\right\|_2$, which further demonstrates the independent growth of $\left\|\left(\mathbf{\Sigma}_k^* - \mathbf{\Sigma}_k\right) \mathbf{u}_j^*\right\|_2$ as k increases for different j. This behavior corresponds to what we observed in experiments in Figure 2, where each $\left\|\left(\mathbf{\Sigma}_k^* - \mathbf{\Sigma}_k\right) \mathbf{u}_j^*\right\|_2$ grows sequentially.

4.2. Proof Overview of Theorem 4.1

In this section, we will demonstrate how the properties of power iteration, in particular (23) is used to construct a proof for Theorem 4.1. We start with the following lemma.

Lemma 4.6. *Let $\mathbf{M}^* \in \mathbb{R}^{d \times d}$ be a rank-r matrix and let $\mathbf{M} = \mathbf{M}^* + \mathbf{H}$ for some $\mathbf{H} \in \mathbb{R}^{d \times d}$. Let $\sigma_i^*, \mathbf{a}_i^*$ be the i-th eigenvalue and eigenvector of $\mathbf{M}^*$, and $\sigma_i, \mathbf{a}_i$ be the i-th eigenvalue and eigenvector of $\mathbf{M}$. Then, for all $i, j \in [r]$ with $\sigma_i \neq \sigma_j^*$, we have:*

$$\mathbf{a}_i^\top \mathbf{a}_j^* = \frac{\mathbf{a}_i^\top \mathbf{H} \mathbf{a}_j^*}{\sigma_i - \sigma_j^*}.$$

Unfortunately, we cannot directly characterize $\mathbf{u}_k^\top \mathbf{u}_k^*$ using Lemma 4.6 by plugging in $i = j = 1$ and $\mathbf{M} = \mathbf{\Sigma}_k, \mathbf{M}^* = \mathbf{\Sigma}_k^*$, since, as far as we know, it is impossible to show that

$\sigma_i = \lambda_k \neq \lambda_k^* = \sigma_j^*$. However, Lemma 4.6 will be useful to upper-bound $\left(\mathbf{u}_k^\top \mathbf{u}_j^*\right)^2$ when $j > k$, since in this case, we can guarantee that $\lambda_k \geq \lambda_k^* - \left\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\right\|_2 > \lambda_j^*$ as long as $\left\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\right\|_2 < \mathcal{T}_k$, where $\mathcal{T}_k$ is the k-th eigengap of $\mathbf{\Sigma}$. Under such conditions, we have:

$$\left|\mathbf{u}_k^\top \mathbf{u}_j^*\right| \leq \frac{\left\|\left(\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\right) \mathbf{u}_j^*\right\|_2}{\lambda_k - \lambda_j^*}. \quad (24)$$

Indeed, going from Lemma 4.6 to (24) loosens the bound by ignoring the directional information of $\mathbf{u}_k$. However, through experiments, we observe that (24) is a good enough upper bound since $\left\|\left(\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\right) \mathbf{u}_j^*\right\|_2$ already stays at a much smaller magnitude than $\left\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\right\|_2$. This is reflected in Figure 4, where we can see that the curve $\left\|\left(\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\right) \mathbf{u}_j^*\right\|_2$ enjoys the same decrease as $\left|\mathbf{u}_k^\top \left(\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\right) \mathbf{u}_j^*\right|$ when j first surpasses k, and stops decreasing after reaching a small magnitude (lower than 10^{-5}). (24) successfully applied the property of the power iteration in (23) to bound each $\left|\mathbf{u}_k^\top \mathbf{u}_j^*\right|$. Our next lemma connects the upper bound of each $\left|\mathbf{u}_k^\top \mathbf{u}_j^*\right|$ to a lower bound of $\left|\mathbf{u}_k^\top \mathbf{u}_k^*\right|$.

Lemma 4.7. *Suppose that $\left\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\right\|_F \leq \frac{1}{8}\lambda_k^*$, then we have*

$$\left(\mathbf{u}_k^\top \mathbf{u}_k^*\right)^2 \geq 1 - \frac{2.4}{\lambda_k^{*2}} \left\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\right\|_F^2 - \sum_{j=k+1}^d \left(\mathbf{u}_k^\top \mathbf{u}_j^*\right)^2$$
$$(25)$$

Our proof of Lemma 4.7 is based on Neumann expansion (Eldridge et al., 2017; Chen et al., 2020), a powerful technique for analyses that involve the product between perturbation matrix and eigenvectors. Indeed, the summation of the last term of (25) may seem to introduce an additional factor of d. However, since the sequence $\left\{\left|\mathbf{u}_k^\top \mathbf{u}_j^*\right|\right\}_{j=k+1}^d$ decays fast, as observed in Figure 3, the summation will eventually be independent of d. Combining (23), (24), and (25) gives the following lemma that upper bounds $\left\|\mathbf{u}_k - \mathbf{u}_k^*\right\|_2$.

Lemma 4.8. *Assume that there exists a constant* $c_0 \in (0, \infty)$ *such that the initialized vector* $\mathbf{x}_{0,k'}$ *in the* k'-*th power iteration procedure satisfies* $\left| \mathbf{x}_{0,k'}^\top \mathbf{u}_{k'} \right| > c_0^{-1}$ *for all* $k' \in [k]$. *Let* $\mathcal{T}_k = \lambda_k^* - \lambda_{k+1}^*$, $\mathcal{T}_{k,\min} = \min_{k' \in [k]} \mathcal{T}_{k'}$ *and* $\mathcal{G}_k = 1 + \frac{c_0 \lambda_k^* \lambda_{k+1}^*}{\lambda_k^* - \lambda_{k+1}^*}$. *Suppose* $\|\mathbf{\Sigma}_{k'} - \mathbf{\Sigma}_{k'}^*\|_F \leq \frac{1}{8} \mathcal{T}_{k,\min}$, *and* $t \geq \log 2\mathcal{G}_k \left(\log \frac{\lambda_{k'+1}^* + 7\lambda_{k'}^*}{\lambda_{k'}^* + 7\lambda_{k'+1}^*} \right)^{-1}$ *for all* $k' \in [k-1]$. *Then, we have that:*

$$\|\mathbf{u}_k - \mathbf{u}_k^*\|_2^2 \leq \frac{4.8}{\lambda_k^*} \|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\|_2^2 + \frac{11 c_0^2}{2\mathcal{T}_k^2} \sum_{j=k+1}^d \left(\frac{\lambda_j^*}{\lambda_k^*} \right)^{2t}.$$

Lemma 4.8 shows a similar result to Lemma 3.5. In particular, the first term in the upper bound of Lemma 4.8 also scales with $\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\|_2$, and the second term involves a summation over $\left(\frac{\lambda_j^*}{\lambda_k^*} \right)^{2t}$ for $j > k$, which can be made small as long as t is large enough. The upper bound in Lemma 4.8 shows a significant improvement in that the scaling factor in front of $\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\|_2$ depends only on λ_k^* compared with the $O\left(\frac{\sqrt{d-k}}{\mathcal{T}_k} \right)$ scaling in Lemma 3.5. This allows us to improve over the nearly factorial bound in Theorem 3.1. After obtaining Lemma 4.8, the rest of the proof for Theorem 4.1 becomes similar to what we have discussed in Section 3.1. We can plug the upper bound in Lemma 4.8 into Lemma 3.4 and unroll the recurrence to obtain a closed-form upper bound for $\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\|_F$. After that, we can finish the proof by plugging the upper bound for $\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\|_F$ back into Lemma 4.8 and utilize (11) to arrive at the desired upper bound for $\|\mathbf{v}_k - \mathbf{u}_k^*\|_2$. The detailed proof of Theorem 4.1 is deferred to Appendix B.7.

5. Conclusion and Future Work

Numerical algorithms for solving the top eigenvector of a given matrix often produce non-negligible errors, which will propagate and accumulate through multiple deflation steps. In this paper, we mathematically analyze this error propagation for a scenario agnostic to the sub-routine for solving the top eigenvectors and the case where this sub-routine is fixed to power iteration. In particular, for the sub-routine agnostic case, our analysis gives an exponential growth of the errors as one aims to solve more eigenvectors, where the base of the exponential growth depends on the eigengap. This upper bound on the errors is improved to exponential growth with a constant base by utilizing the directional information of the error vector produced by power iteration. Our result implies a lower bound on the number of power iteration steps required to solve for multiple leading eigenvectors using deflation up to a certain accuracy.

This paper focuses on the case where $\mathbf{\Sigma}$ is positive definite and has distinct eigenvalues. However, the analysis can also be extended to accommodate repeated and zero eigenvalues. When eigenvalues are repeated, the corresponding eigenvector and the method for measuring the error of the approximate eigenvector need to be redefined. Instead of considering a single eigenvector for each eigenvalue, we can generalize the approach by considering the subspace spanned by the eigenvectors corresponding to each repeated eigenvalue. The core ideas of our analysis can be applied to account for the error propagation between subspaces, but one must account for additional error propagation within subspaces. To extend the analysis to positive semi-definite matrices, $\mathbf{\Sigma}$ is of rank r and we are interested in finding top-K eigenvectors with $K \leq r$. In this case, our Theorem 3.1 still holds, as it does not depend on eigenvalues beyond the Kth. Theorem 4.1 may require slight modification, but we hypothesize that the case where $\mathbf{\Sigma}$ is low rank will relax the condition on t in (16) to $k \leq r$ instead of $k \leq d$.

While this paper focuses on the sub-routine agnostic scenario and the power iteration, future work can consider extending the analysis to the case where other algorithms, such as Lanczos method (Golub & Van Loan, 1996) and power iteration with momentum (Xu et al., 2018a), are used to approximate the top eigenvectors in the deflation process.

Acknowledgements

Impact Statement

This paper aims to deepen the theoretical understanding of principal component analysis, a long-standing method in machine learning. The theoretical results presented herein may change the computation power used for principal component analysis. While the direct impact of this work may be limited to computational efficiency gains, the broader implications extend beyond this, potentially catalyzing advancements in various domains where principal component analysis finds application. However, no specific societal consequences can be directly attributed to the work itself.

References

Allen-Zhu, Z. and Li, Y. First efficient convergence for streaming k-pca: a global, gap-free, and near-optimal rate, 2017.

Belkin, M. and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003. doi: 10.1162/089976603321780317.

Chen, Y., Cheng, C., and Fan, J. Asymmetry helps: Eigenvalue and eigenvector analyses of asymmetrically perturbed low-rank matrices, 2020.

Coulaud, O., Giraud, L., Ramet, P., and Vasseur, X. Deflation and augmentation techniques in krylov subspace methods for the solution of linear systems, 2013.

Danisman, Y., Yilmaz, M., Ozkaya, A., and Comlekciler, I. A comparison of eigenvalue methods for principal component analysis. *Appl. and Comput. Math*, 13:316–331, 2014.

d'Aspremont, A., El Ghaoui, L., Jordan, M., and Lanckriet, G. A direct formulation for sparse pca using semidefinite programming. *SIAM review*, 49(3):434–448, 2007.

Davis, C. and Kahan, W. M. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. ISSN 00361429. URL http://www.jstor.org/stable/2949580.

Deng, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Eldridge, J., Belkin, M., and Wang, Y. Unperturbed: spectral analysis beyond davis-kahan, 2017.

Fawzi, A., Balog, M., Huang, A., and et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610:47–53, 2022. doi: 10.1038/s41586-022-05172-4. URL https://doi.org/10.1038/s41586-022-05172-4.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270, 2022.

Gemp, I., McWilliams, B., Vernade, C., and Graepel, T. Eigengame: {PCA} as a nash equilibrium. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=NzTU59SYbNq.

Golub, G. H. and Van Loan, C. F. *Matrix Computations*. The Johns Hopkins University Press, third edition, 1996.

Golub, G. H. and Van Loan, C. F. *Matrix computations*. JHU press, 2013.

Goujaud, B., Scieur, D., Dieuleveut, A., Taylor, A. B., and Pedregosa, F. Super-acceleration with cyclical step-sizes. In *International Conference on Artificial Intelligence and Statistics*, pp. 3028–3065. PMLR, 2022.

Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Hubert, L. J. and Arabie, P. Comparing partitions. *Journal of Classification*, 2:193–218, 1985. URL https://api.semanticscholar.org/CorpusID:189915041.

Jain, P., Jin, C., Kakade, S. M., Netrapalli, P., and Sidford, A. Streaming pca: Matching matrix bernstein and near-optimal finite sample guarantees for oja's algorithm, 2016.

Jiang, R., Fei, H., and Huan, J. Anomaly localization for network data streams with graph joint sparse pca. In *Proceedings of the 17th ACM SIGKDD*, pp. 886–894. ACM, 2011.

Johnstone, I. M. and Lu, A. Y. Sparse principal components analysis. *arXiv preprint arXiv:0901.4392*, 2009.

Kuczynski, J. and Woźniakowski, H. Estimating the largest eigenvalue by the power and lanczos algorithms with a random start. *SIAM Journal on Matrix Analysis and Applications*, 13(4):1094–1122, 1992.

Lanczos, C. An iterative method for the solution of the eigenvalue problem of linear differential and integral. 1950. URL https://api.semanticscholar.org/CorpusID:478182.

Mackey, L. Deflation methods for sparse pca. *Advances in neural information processing systems*, 21, 2008.

Majumdar, A. Image compression by sparse pca coding in curvelet domain. *Signal, image and video processing*, 3 (1):27–34, 2009.

Müntz, H. et al. Direct solution of the ésecular equation and of some transcendent analogous problems. *CR Acad. Sci. Paris*, 156:43–46, 1913.

Papailiopoulos, D., Kyrillidis, A., and Boutsidis, C. Provable deterministic leverage score sampling. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 997–1006, 2014.

Papyan, V. Traces of class/cross-class structure pervade deep learning spectra. *The Journal of Machine Learning Research*, 21(1):10197–10260, 2020.

Pearson, K. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

Pothen, A., Simon, H. D., and Liou, K.-P. Partitioning sparse matrices with eigenvectors of graphs. *SIAM journal on matrix analysis and applications*, 11(3):430–452, 1990.

Saad, Y. Projection and deflation method for partial pole assignment in linear state feedback. *IEEE Transactions on Automatic Control*, 33(3):290–297, 1988.

Sriperumbudur, B. K., Torres, D. A., and Lanckriet, G. R. Sparse eigen methods by DC programming. In *Proceedings of the 24th international conference on Machine learning*, pp. 831–838, 2007.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Wang, P., Zhou, Z., and So, A. M.-C. A nearly-linear time algorithm for exact community recovery in stochastic block model. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10126–10135. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/wang20ac.html.

Wang, Z., Han, F., and Liu, H. Sparse principal component analysis for high dimensional multivariate time series. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pp. 48–56, 2013.

Weyl, H. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

Xu, P., He, B., De Sa, C., Mitliagkas, I., and Re, C. Accelerated stochastic power iteration. In *International Conference on Artificial Intelligence and Statistics*, pp. 58–67. PMLR, 2018a.

Xu, P., He, B., De Sa, C., Mitliagkas, I., and Re, C. Accelerated stochastic power iteration. In Storkey, A. and Perez-Cruz, F. (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 58–67. PMLR, 09–11 Apr 2018b. URL https://proceedings.mlr.press/v84/xu18a.html.

Zhang, F. *The Schur complement and its applications*, volume 4. Springer Science & Business Media, 2006.

Zou, H., Hastie, T., and Tibshirani, R. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

A. Missing Proofs from Section 3

A.1. Proof of Lemma 3.4

Recall that $\mathbf{\Sigma}_{k+1} = \mathbf{\Sigma}_k - \mathbf{v}_k \mathbf{v}_k^\top \mathbf{\Sigma}_k \mathbf{v}_k \mathbf{v}_k^\top$. Since $\mathbf{v}_k = \mathbf{u}_k + \boldsymbol{\delta}_k$, we can write $\mathbf{\Sigma}_{k+1}$ as

$$
\begin{aligned}
\mathbf{\Sigma}_{k+1} &= \mathbf{\Sigma}_k - (\mathbf{u}_k + \boldsymbol{\delta}_k)(\mathbf{u}_k + \boldsymbol{\delta}_k)^\top \mathbf{\Sigma}_k (\mathbf{u}_k + \boldsymbol{\delta}_k)(\mathbf{u}_k + \boldsymbol{\delta}_k)^\top \\
&= \mathbf{\Sigma}_k - (\mathbf{u}_k + \boldsymbol{\delta}_k)\left(\mathbf{u}_k^\top \mathbf{\Sigma}_k \mathbf{u}_k + \boldsymbol{\delta}_k^\top \mathbf{\Sigma}_k \mathbf{u}_k + \mathbf{u}_k^\top \mathbf{\Sigma}_k \boldsymbol{\delta}_k + \boldsymbol{\delta}_k^\top \mathbf{\Sigma}_k \boldsymbol{\delta}_k\right)(\mathbf{u}_k + \boldsymbol{\delta}_k)^\top \\
&= \mathbf{\Sigma}_k - \left(\lambda_k + 2\lambda_k \mathbf{u}_k^\top \boldsymbol{\delta}_k + \boldsymbol{\delta}_k^\top \mathbf{\Sigma}_k \boldsymbol{\delta}_k\right)(\mathbf{u}_k + \boldsymbol{\delta}_k)(\mathbf{u}_k + \boldsymbol{\delta}_k)^\top \\
&= \underbrace{\mathbf{\Sigma}_k - \lambda_k \mathbf{u}_k \mathbf{u}_k^\top}_{\tilde{\mathbf{\Sigma}}_{k+1}} - \underbrace{\lambda_k \left(\boldsymbol{\delta}_k \mathbf{u}_k^\top + \mathbf{u}_k \boldsymbol{\delta}_k^\top + \boldsymbol{\delta}_k \boldsymbol{\delta}_k^\top\right)}_{\boldsymbol{\mathcal{E}}_{1,k}} \\
&\quad - \underbrace{\left(2\lambda_k \mathbf{u}_k^\top \boldsymbol{\delta}_k + \boldsymbol{\delta}_k^\top \mathbf{\Sigma}_k \boldsymbol{\delta}_k\right)(\mathbf{u}_k + \boldsymbol{\delta}_k)(\mathbf{u}_k + \boldsymbol{\delta}_k)^\top}_{\boldsymbol{\mathcal{E}}_{2,k}}
\end{aligned}
$$

For the convenience of the analysis, we let $\Delta_k = \|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\|_F$. In this way, we can write

$$
\Delta_{k+1} = \left\|\tilde{\mathbf{\Sigma}}_{k+1} - \mathbf{\Sigma}_{k+1}^* - \boldsymbol{\mathcal{E}}_{1,k} - \boldsymbol{\mathcal{E}}_{2,k}\right\|_F \leq \left\|\tilde{\mathbf{\Sigma}}_{k+1} - \mathbf{\Sigma}_{k+1}^*\right\|_F + \|\boldsymbol{\mathcal{E}}_{1,k}\|_F + \|\boldsymbol{\mathcal{E}}_{2,k}\|_F \tag{26}
$$

Recall the definition of $\mathbf{\Sigma}_{k+1}^*$ from (4). By Lemma C.1, we have $\mathbf{\Sigma}_{k+1}^* = \mathbf{\Sigma}_k^* - \lambda_k^* \mathbf{u}_k^* \mathbf{u}_k^{*\top}$. Thus, the first term in (26) can be bounded as

$$
\begin{aligned}
\left\|\tilde{\mathbf{\Sigma}}_{k+1} - \mathbf{\Sigma}_{k+1}^*\right\|_F &= \left\|\mathbf{\Sigma}_k - \lambda_k \mathbf{u}_k \mathbf{u}_k^\top - \mathbf{\Sigma}_k^* + \lambda_k^* \mathbf{u}_k^* \mathbf{u}_k^{*\top}\right\|_F \\
&\leq \|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\|_F + \left\|\lambda_k \mathbf{u}_k \mathbf{u}_k^\top - \lambda_k^* \mathbf{u}_k^* \mathbf{u}_k^{*\top}\right\|_F \\
&\leq \Delta_k + \left\|(\lambda_k - \lambda_k^*)\mathbf{u}_k \mathbf{u}_k^\top - \lambda_k^* \left(\mathbf{u}_k^* \mathbf{u}_k^{*\top} - \mathbf{u}_k \mathbf{u}_k^\top\right)\right\|_F \\
&\leq 2\Delta_k + \lambda_k^* \left\|\mathbf{u}_k \mathbf{u}_k^\top - \mathbf{u}_k^* \mathbf{u}_k^{*\top}\right\|_F
\end{aligned}
$$

We use Weyl's inequality in the last inequality to obtain that $|\lambda_k - \lambda_k^*| \leq \Delta_k$. The last term above can be further bounded as

$$
\left\|\mathbf{u}_k \mathbf{u}_k^\top - \mathbf{u}_k^* \mathbf{u}_k^{*\top}\right\|_F = \left\|\mathbf{u}_k (\mathbf{u}_k - \mathbf{u}_k^*)^\top + (\mathbf{u}_k - \mathbf{u}_k^*)\mathbf{u}_k^{*\top}\right\|_F \leq 2\|\mathbf{u}_k - \mathbf{u}_k^*\|_2 = 2\theta_k
$$

Therefore, we have

$$
\left\|\tilde{\mathbf{\Sigma}}_{k+1} - \mathbf{\Sigma}_{k+1}^*\right\|_F \leq 2\Delta_k + 2\lambda_k^* \theta_k \tag{27}
$$

Now, we focus on $\boldsymbol{\mathcal{E}}_{1,k}$. It can be bounded as

$$
\|\boldsymbol{\mathcal{E}}_{1,k}\|_F = \lambda_k \left\|\boldsymbol{\delta}_k \mathbf{u}_k^\top + \mathbf{u}_k \boldsymbol{\delta}_k^\top + \boldsymbol{\delta}_k \boldsymbol{\delta}_k^\top\right\|_F \leq \lambda_k \left(2\|\boldsymbol{\delta}_k\|_2 + \|\boldsymbol{\delta}_k\|_2^2\right)
$$

Given that $\lambda_k \leq \lambda_k^* + |\lambda_k - \lambda_k| \leq \lambda_k + \Delta_k$, we have

$$
\|\boldsymbol{\mathcal{E}}_{1,k}\|_F \leq (\lambda_k^* + \Delta_k)\left(2\|\boldsymbol{\delta}_k\|_2 + \|\boldsymbol{\delta}_k\|_2^2\right) \tag{28}
$$

Similarly, for $\boldsymbol{\mathcal{E}}_{2,k}$ we have

$$
\begin{aligned}
\|\boldsymbol{\mathcal{E}}_{2,k}\|_F &\leq \left|2\lambda_k \mathbf{u}_k^\top \boldsymbol{\delta}_k + \boldsymbol{\delta}_k^\top \mathbf{\Sigma}_k \boldsymbol{\delta}_k\right| \|\mathbf{u}_k + \boldsymbol{\delta}_k\|_2^2 \\
&\leq \left(2(\lambda_k^* + \Delta_k)\|\boldsymbol{\delta}_k\|_2 + (\lambda_k^* + \Delta_k)\|\boldsymbol{\delta}_k\|_2^2\right)(1 + \|\boldsymbol{\delta}_k\|_2)^2 \\
&= (\lambda_k^* + \Delta_k)\left(2\|\boldsymbol{\delta}_k\|_2 + \|\boldsymbol{\delta}_k\|_2^2\right)(1 + \|\boldsymbol{\delta}_k\|_2)^2
\end{aligned}
\tag{29}
$$

where the first inequality follows from $\|\Sigma_k\|_2 \le \lambda_k$ since, by definition, λ_k is the largest eigenvalue of Σ_k. Plugging (27), (28), and (29) into (26), we have

$$\Delta_{k+1} \le 2\Delta_k + 2\lambda_k^* \theta_k + (\lambda_k^* + \Delta_k) \left(2\|\boldsymbol{\delta}_k\|_2 + \|\boldsymbol{\delta}_k\|_2^2\right)$$

$$+ (\lambda_k^* + \Delta_k) \left(2\|\boldsymbol{\delta}_k\|_2 + \|\boldsymbol{\delta}_k\|_2^2\right) (1 + \|\boldsymbol{\delta}_k\|_2)^2$$

$$\le 2\Delta_k + 2\lambda_k^* \theta_k + (\lambda_k^* + \Delta_k) \left(2\|\boldsymbol{\delta}_k\|_2 + \|\boldsymbol{\delta}_k\|_2^2\right) \left(1 + (1 + \|\boldsymbol{\delta}_k\|_2)^2\right)$$

$$= 2\Delta_k + 2\lambda_k^* \theta_k + (\lambda_k^* + \Delta_k) \|\boldsymbol{\delta}_k\|_2 \left(2 + \|\boldsymbol{\delta}_k\|_2\right) \left(1 + (1 + \|\boldsymbol{\delta}_k\|_2)^2\right)$$

When $\|\boldsymbol{\delta}_k\|_2 \le \frac{1}{6}$, we will have that

$$(2 + \|\boldsymbol{\delta}_k\|_2) \left(1 + (1 + \|\boldsymbol{\delta}_k\|_2)^2\right) = \frac{13}{6} \cdot \left(1 + \left(\frac{7}{6}\right)^2\right) = \frac{1014}{216} \le 5$$

Thus, we have

$$\Delta_{k+1} \le 2\Delta_k + 2\lambda_k^* \theta_k + 5(\lambda_k^* + \Delta_k) \|\boldsymbol{\delta}_k\|_2 \le 3\Delta_k + 2\lambda_k^* \theta_k + 5\lambda_k^* \|\boldsymbol{\delta}_k\|_2$$

A.2. Proof of Lemma 3.5

To prove Lemma 3.5, we first recall the Davis-Kahan $\sin\Theta$ theorem (Davis & Kahan, 1970; Eldridge et al., 2017).

Theorem A.1 (Davis-Kahan $\sin\Theta$ theorem). *Let $\mathbf{M}^* \in \mathbb{R}^{d \times d}$ and let $\mathbf{M} = \mathbf{M}^* + \mathbf{H}$. Let $\mathbf{a}_1^*$ be the top eigenvector of $\mathbf{M}^*$ and $\mathbf{a}_1$ be the top eigenvector of $\mathbf{M}$. Then we have*

$$\sin\angle\{\mathbf{a}_1^*, \mathbf{a}_1\} \le \frac{\|\mathbf{H}\|_2}{\min_{j \ne k} |\sigma_k^* - \sigma_j|}.$$

To start, we notice that, under the assumption that $\|\Sigma_k - \Sigma_k^*\|_F \le \frac{1}{4}\mathcal{T}_k$, by the Weyl's inequality, we have

$$\min_{j \ne k} |\sigma_k^* - \sigma_j| \ge \min_{j \ne k} |\sigma_k^* - \sigma_j^*| - |\sigma_j - \sigma_j^*| \ge \mathcal{T}_k - \|\Sigma_k - \Sigma_k^*\|_F \ge \frac{3}{4}\mathcal{T}_k$$

By Lemma C.1, we have that $\mathbf{u}_k^*$ is the top eigenvector of Σ_k^* with corresponding eigenvector λ_k^*. Moreover, as defined in (10), $\mathbf{u}_k$ is the top eigenvector of Σ_k. Therefore, a direct application of Theorem A.1 gives

$$\sin\angle\{\mathbf{u}_k, \mathbf{u}_k^*\} \le \frac{4\|\Sigma_k - \Sigma_k^*\|_F}{3\mathcal{T}_k} \tag{30}$$

Moreover, notice that

$$\sin\angle\{\mathbf{u}_k, \mathbf{u}_k^*\}^2 = 1 - \cos\angle\{\mathbf{u}_k, \mathbf{u}_k^*\}^2 = 1 - \left(\mathbf{u}_k^\top \mathbf{u}_k^*\right)^2$$

which gives that $\left(\mathbf{u}_k^\top \mathbf{u}_k^*\right)^2 = 1 - \sin\angle\{\mathbf{u}_k, \mathbf{u}_k^*\}^2$. Therefore

$$\|\mathbf{u}_k - \mathbf{u}_k^*\|_2^2 = 2 - 2\mathbf{u}_k^\top \mathbf{u}_k^* \le 2 - 2\left(\mathbf{u}_k^\top \mathbf{u}_k^*\right)^2 = 2\sin\angle\{\mathbf{u}_k, \mathbf{u}_k^*\}^2$$

Here the first inequality follows from the fact that $\left|\mathbf{u}_k^\top \mathbf{u}_k^*\right| \le 1$. Applying (30) gives

$$\|\mathbf{u}_k - \mathbf{u}_k^*\|_2 \le \sqrt{2}\sin\angle\{\mathbf{u}_k, \mathbf{u}_k^*\} \le \frac{4\sqrt{2}}{3\mathcal{T}_k} \|\Sigma_k - \Sigma_k^*\|_F \le \frac{2}{\mathcal{T}_k} \|\Sigma_k - \Sigma_k^*\|_F$$

A.3. Proof of Theorem 3.1

Let $\hat{K}$ be the smallest integer such that $\left\|\Sigma_{\hat{K}} - \Sigma_{\hat{K}}^*\right\|_2 > \frac{1}{4}\mathcal{T}_{K,\min}$, and let $K' = \min\left\{\hat{K}, K+1\right\}$. By definition, $\mathcal{T}_{K,\min} = \min_{k \in [K]} \mathcal{T}_k$. Therefore, for all $k < K'$, we shall have that $\|\Sigma_k - \Sigma_k^*\|_2 \le \frac{1}{4}\mathcal{T}_k$. Therefore, for all $k \in [K']$, we can invoke Lemma 3.5 to get that

$$\|\mathbf{u}_k - \mathbf{u}_k^*\|_2 \le \frac{2}{\mathcal{T}_k} \|\Sigma_k - \Sigma_k^*\|_F \tag{31}$$

By the assumption of Theorem 3.1, we have

$$\|\boldsymbol{\delta}_k\|_2 \leq \frac{\mathcal{T}_{K,\min}}{20K} \prod_{j=k+1}^{K-1} \left(3 + \frac{2\lambda_j^*}{\mathcal{T}_j}\right)^{-1} \leq \frac{1}{6}$$

since $\mathcal{T}_{K,\min} \leq \lambda_1^* = 1$ and $K \geq 1$. Therefore, we also have that $\|\boldsymbol{\delta}_k\|_2 \leq \frac{1}{6}$ for all $k \in [K'-1]$. This allows us to invoke Lemma 3.4 to get that

$$\left\|\boldsymbol{\Sigma}_{k+1} - \boldsymbol{\Sigma}_{k+1}^*\right\|_F \leq 3\left\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\right\|_F + 2\lambda_k^* \|\mathbf{u}_k - \mathbf{u}_k^*\|_2 + 5\lambda_k^* \|\boldsymbol{\delta}_k\|_2 \tag{32}$$

Plugging (31) into (32) gives

$$\left\|\boldsymbol{\Sigma}_{k+1} - \boldsymbol{\Sigma}_{k+1}^*\right\|_F \leq \left(3 + \frac{2\lambda_k^*}{\mathcal{T}_k}\right)\left\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\right\|_F + 5\lambda_k^* \|\boldsymbol{\delta}_k\|_2$$

Let the sequence $\{Q_k\}_{k=1}^{K'}$ be defined as

$$Q_{k+1} = a_k Q_k + b_k; \quad Q_0 = 0; \quad a_k = 3 + \frac{2\lambda_k^*}{\mathcal{T}_k}; \quad b_k = 5\lambda_k^* \|\boldsymbol{\delta}_k\|_2$$

Then we must have that $\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_F \leq Q_k$ for all $k \in [K']$. Invoking Lemma C.4 thus gives

$$\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_F \leq 5 \sum_{k'=1}^{k-1} \lambda_{k'}^* \|\boldsymbol{\delta}_{k'}\|_2 \prod_{j=k'+1}^{k-1} \left(3 + \frac{2\lambda_j^*}{\mathcal{T}_j}\right) \tag{33}$$

Combining with (31) and the fact that $\|\mathbf{v}_k - \mathbf{u}_k^*\|_2 \leq \|\boldsymbol{\delta}_k\|_2 + \|\mathbf{u}_k - \mathbf{u}_k^*\|_2$ gives

$$\|\mathbf{v}_k - \mathbf{u}_k^*\|_2 \leq \|\boldsymbol{\delta}_k\|_2 + \frac{2}{\mathcal{T}_k} \|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_F \leq \|\boldsymbol{\delta}_k\|_2 + \frac{10}{\mathcal{T}_k} \sum_{k'=1}^{k-1} \lambda_{k'}^* \|\boldsymbol{\delta}_{k'}\|_2 \prod_{j=k'+1}^{k-1} \left(3 + \frac{2\lambda_j^*}{\mathcal{T}_j}\right) \tag{34}$$

Notice that

$$\frac{10}{\mathcal{T}_k} \leq \frac{5}{\lambda_k^*} \cdot \left(3 + \frac{2\lambda_k^*}{\mathcal{T}_k}\right)$$

Therefore, (34) becomes

$$\|\mathbf{v}_k - \mathbf{u}_k^*\|_2 \leq \|\boldsymbol{\delta}_k\|_2 + 5\sum_{k'=1}^{k-1} \frac{\lambda_{k'}^*}{\lambda_k^*} \|\boldsymbol{\delta}_{k'}\|_2 \prod_{j=k'+1}^{k}\left(3 + \frac{2\lambda_j^*}{\mathcal{T}_j}\right) \leq 5\sum_{k'=1}^{k} \frac{\lambda_{k'}^*}{\lambda_k^*} \|\boldsymbol{\delta}_{k'}\|_2 \prod_{j=k'+1}^{k}\left(3 + \frac{2\lambda_j^*}{\mathcal{T}_j}\right)$$

for all $k \in [K']$. It thus remains to show that $\hat{K} > K$. In this way, we will have $K' = K$, and the theorem is then proved. For the sake of contradiction, that $\hat{K} \leq K$. Then we shall have that there exists $k \in [K]$ such that $\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_2 > \frac{1}{4}\mathcal{T}_{K,\min}$. To reach a contradiction, it thus remains to show that for all $k \in [K]$

$$\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_2 \leq \frac{1}{8}\mathcal{T}_{K,\min}$$

By (33), it suffices to guarantee that

$$5 \sum_{k'=1}^{k-1} \lambda_{k'}^* \|\boldsymbol{\delta}_{k'}\|_2 \prod_{j=k'+1}^{k-1}\left(3 + \frac{2\lambda_j^*}{\mathcal{T}_j}\right) \leq \frac{1}{4}\mathcal{T}_{K,\min} \tag{35}$$

Indeed, by the assumption of Theorem 3.1, we require

$$\|\boldsymbol{\delta}_k\|_2 \leq \frac{\mathcal{T}_{K,\min}}{20K} \prod_{j=k+1}^{K-1}\left(3 + \frac{2\lambda_j^*}{\mathcal{T}_j}\right)^{-1}$$

Plugging this requirement into (33) would guarantee that $\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_2 \leq \frac{1}{4}\mathcal{T}_{K,\min}$ and thus achieves the contradiction. This shows that $\hat{K} > K$ and finishes the proof.

B. Missing Proofs from Section 4

B.1. Proof of Lemma 4.3

Since each $\mathbf{x}_{0,k}$ is generated independently, it suffices to show that for a general $\mathbf{x}_0$ generated for unit sphere and some unit vector $\mathbf{u}$, we have $\left|\mathbf{x}_0^\top \mathbf{u}\right| \geq$ with high probability, and then apply a union bound for all $k \in [K]$. To start, for a fixed $\mathbf{u}$, we let $\mathbf{U}$ be an orthogonal matrix with $\mathbf{u}$ being its first row. Since $\mathbf{x}_0$ is generated uniformly at random from a unit sphere, we must have that $\mathbf{b} = \mathbf{U}\mathbf{x}$ is also generated uniformly at random from a unit sphere. Let b_j denote the jth entry of $\mathbf{b}$, then $b_1 = \mathbf{x}_0^\top \mathbf{u}$. We first aim at showing

$$\mathbb{P}\left(\sum_{j=1}^d \left(\frac{b_j}{b_1}\right) \leq 2d^3\right) \geq 1 - \frac{2}{3d}$$

Let $\mathcal{B}_d(1)$ denote the unit ball in $\mathbb{R}^d$. Define the indicator function $\mathbb{I}(\mathbf{b})$ as

$$\mathbb{I}(\mathbf{b}) = \begin{cases} 1 & \text{if } \sum_{j=1}^d \left(\frac{b_j}{b_1}\right) \leq 2d^3 \\ 0 & \text{otherwise} \end{cases}$$

Then we have

$$\mathbb{P}\left(\sum_{j=1}^d \left(\frac{b_j}{b_1}\right) \leq 2d^3\right) = \int_{\|\mathbf{x}\|_2=1} \mathbb{I}(\mathbf{b}) d\mu(\mathbf{b})$$

where μ is the probability measure over the unit sphere. Since $\mathbb{I}(\mathbf{b})$ satisfies $\mathbb{I}(\alpha\mathbf{b}) = \mathbb{I}(\mathbf{b})$ for any $\alpha \neq 0$ and $\mathbb{I}(\mathbf{b})$ does not depend on the signs of of b_j's, then we can use Remark 7.2 from (Kuczynski & Woźniakowski, 1992) to get that

$$\int_{\|\mathbf{x}\|_2=1} \mathbb{I}(\mathbf{b}) d\mu(\mathbf{b}) = \frac{1}{V_d} \int_{\mathcal{B}_d(1)} \mathbb{I}(\mathbf{b}) d\mathbf{b}$$

$$= 2\int_0^1 \left(\int_{\sum_{j=2}^d b_j^2 \leq \min\{1-b_1^2, 2b_1 d^3\}} db_2 \ldots db_d\right) db_1$$

$$= \frac{2V_{d-1}}{V_d} \int_0^1 \min\{1-b_1^2, 2b_1 d^3\}^{\frac{d-1}{2}} db_1$$

where the last equality follows from the observation that the inner integral is exactly the volume of the ball with radius $\min\{1-b_1^2, 2b_1 d^3\}$. We can lower-bound the last integral as

$$\int_0^1 \min\{1-b_1^2, 2b_1 d^3\}^{\frac{d-1}{2}} db_1 \geq \int_{\sqrt{\frac{1}{1+2d^3}}}^1 \left(1-b_1^2\right)^{\frac{d-1}{2}} db_1$$

$$= \int_0^1 \left(1-b_1^2\right)^{\frac{d-1}{2}} db_1 - \int_0^{\sqrt{\frac{1}{1+2d^3}}} \left(1-b_1^2\right)^{\frac{d-1}{2}} db_1$$

$$\geq \frac{V_d}{2V_{d-1}} - \sqrt{\frac{1}{1+2d^3}}$$

This gives

$$\mathbb{P}\left(\sum_{j=1}^d \left(\frac{b_j}{b_1}\right) \leq 2d^3\right) \geq 1 - \frac{2V_{d-1}}{V_d}\sqrt{\frac{1}{1+2d^3}}$$

By (Kuczynski & Woźniakowski, 1992), Eq (13), we have $\frac{V_{d-1}}{V_d} \leq 0.412\sqrt{d}$ when $d \geq 8$. Thus, we have

$$\mathbb{P}\left(\sum_{j=1}^d \left(\frac{b_j}{b_1}\right) \leq 2d^3\right) \geq 1 - 0.824\sqrt{\frac{d}{1+2d^3}} \geq 1 - \frac{2}{3d}$$

Notice that $\sum_{j=1}^{d} b_j^2 = 1$. Then we have that

$$\sum_{j=2}^{d} \left(\frac{b_j}{b_1}\right)^2 = \frac{1}{b_1^2} - 1 \Rightarrow |b_1| = \left(\sum_{j=2}^{d} \left(\frac{b_j}{b_1}\right)^2 + 1\right)^{-\frac{1}{2}}$$

Thus, we have

$$\mathbb{P}\left(|b_1| \geq \frac{1}{\sqrt{1 + 2d^3}}\right) \geq 1 - \frac{2}{3d}$$

Apply a union bound over all k gives that with probability at least $1 - \frac{2K}{3d}$, we have

$$\left|\mathbf{x}_{0,k}^{\top} \mathbf{u}_k\right| \geq \frac{1}{\sqrt{1 + 2d^3}}$$

B.2. Proof of Lemma 4.4

For convenience, we define the sequence $\{\hat{\mathbf{x}}_k\}_{k=0}^{\infty}$ as

$$\hat{\mathbf{x}}_t = \mathbf{M}\hat{\mathbf{x}}_{t-1}; \quad \hat{\mathbf{x}}_0 = \mathbf{x}_0$$

Then $\mathbf{x}_t$ from (14) can be written as

$$\mathbf{x}_t = \frac{\hat{\mathbf{x}}_t}{\|\hat{\mathbf{x}}_t\|_2} = \frac{\hat{\mathbf{x}}_t}{\|\mathbf{M}^t \mathbf{x}_0\|_2}$$

Focusing on $\hat{\mathbf{x}}_t$, since $\mathbf{M} = \mathbf{M}^* + \mathbf{H}$, we have

$$\hat{\mathbf{x}}_t = \mathbf{M}\hat{\mathbf{x}}_{t-1} = \mathbf{M}^*\hat{\mathbf{x}}_{t-1} + \mathbf{H}\hat{\mathbf{x}}_{t-1}$$

Since $\sigma_j^*, \mathbf{a}_j^*$ are the jth eigenvalue and eigenvector of $\mathbf{M}^*$, we have that

$$\hat{\mathbf{x}}_t^{\top} \mathbf{a}_j^* = \hat{\mathbf{x}}_{t-1}^{\top} \mathbf{M}^* \mathbf{a}_j^* + \hat{\mathbf{x}}_{t-1}^{\top} \mathbf{H} \mathbf{a}_j^* = \sigma_j^* \hat{\mathbf{x}}_{t-1}^{\top} \mathbf{a}_j^* + \hat{\mathbf{x}}_{t-1}^{\top} \mathbf{H} \mathbf{a}_j^*$$

Unrolling the iterates gives

$$\hat{\mathbf{x}}_t^{\top} \mathbf{a}_j^* = \sigma_i^{*t} \mathbf{x}_0^{\top} \mathbf{a}_j^* + \sum_{t'=0}^{t-1} \sigma_j^{*t-t'} \hat{\mathbf{x}}_{t'}^{\top} \mathbf{H} \mathbf{a}_j^* = \sigma_j^{*t}\left(\mathbf{x}_0^{\top} \mathbf{a}_j^* + \sum_{t'=0}^{t-1} \frac{1}{\sigma_j^{*t'}} \hat{\mathbf{x}}_{t'}^{\top} \mathbf{H} \mathbf{a}_j^*\right)$$

Recalling that $\hat{\mathbf{x}}_t = \mathbf{M}^t \mathbf{x}_0$, we can see that

$$\hat{\mathbf{x}}_t^{\top} \mathbf{a}_j^* = \sigma_j^{*t}\left(\mathbf{x}_0^{\top} \mathbf{a}_j^* + \sum_{t'=0}^{t-1} \left(\frac{\mathbf{M}^{t'}}{\sigma_j^{*t'}} \hat{\mathbf{x}}_0\right)^{\top} \mathbf{H} \mathbf{a}_j^*\right) = \sigma_j^{*t} \mathbf{x}_0^{\top}\left(\mathbf{I} + \sum_{t'=0}^{t-1} \frac{\mathbf{M}^{t'}}{\sigma_j^{*t'}} \cdot \mathbf{H}\right) \mathbf{a}_j^*$$

Plugging in $\mathbf{x}_t = \frac{\hat{\mathbf{x}}_t}{\|\mathbf{M}^t \mathbf{x}_0\|_2}$ gives the first conclusion. Next, without loss of generality, we assume that $\|\mathbf{x}_0\|_2 = 1$, since $\left|\mathbf{x}_t^{\top} \mathbf{a}_j^*\right|$ does not depend on $\|\mathbf{x}_0\|_2$. Since $\|\mathbf{x}_0\|_2 = \|\mathbf{a}_j^*\|_2 = 1$, it holds that

$$\left|\hat{\mathbf{x}}_t^{\top} \mathbf{a}_j^*\right| \leq \sigma_j^{*t}\left(1 + \left\|\sum_{t'=0}^{t-1} \frac{\mathbf{M}^{t'}}{\sigma_j^{*t'}} \cdot \mathbf{H} \mathbf{a}_j^*\right\|_2\right)$$

$$\leq \sigma_j^{*t}\left(1 + \sum_{t'=0}^{t-1} \frac{\sigma_1^{t'}}{\sigma_j^{*t'}} \left\|\mathbf{H} \mathbf{a}_j^*\right\|_2\right)$$

$$\leq \sigma_j^{*t}\left(1 + \frac{\left(\frac{\sigma_1}{\sigma_j^*}\right)^t - 1}{\frac{\sigma_1}{\sigma_j^*} - 1} \left\|\mathbf{H} \mathbf{a}_j^*\right\|_2\right)$$

$$\leq \sigma_j^{*k} + \frac{\sigma_1^t - \sigma_i^{*t}}{\frac{\sigma_1}{\sigma_j^*} - 1} \left\|\mathbf{H} \mathbf{a}_j^*\right\|_2$$

$$\leq \sigma_j^{*t} + \frac{\sigma_j^* \sigma_1^t}{\sigma_1 - \sigma_j^*} \left\|\mathbf{H} \mathbf{a}_j^*\right\|_2$$

Since there exists some $\left| \mathbf{x}_0^\top \mathbf{a}_1 \right| \geq c_0^{-1} > 0$, we can bound $\left\| \mathbf{M}^t \mathbf{x}_0 \right\|_2$ as

$$\left\| \mathbf{M}^t \mathbf{x}_0 \right\|_2 = \left(\sum_{j=1}^r \sigma_j^{2t} \left(\mathbf{x}_0^\top \mathbf{a}_j \right)^2 \right)^{\frac{1}{2}} \geq \sigma_1^t \left| \mathbf{x}_0^\top \mathbf{a}_j \right| \geq \sigma_1^t c_0^{-1}$$

Therefore, we have

$$\left| \mathbf{x}_t^\top \mathbf{a}_j^* \right| \leq \left\| \mathbf{M}^t \mathbf{x}_0 \right\|_2^{-1} \left| \hat{\mathbf{x}}_t^\top \mathbf{a}_j^* \right| \leq c_0 \left(\left(\frac{\sigma_j^*}{\sigma_1} \right)^t + \frac{\sigma_j^*}{\sigma_1 - \sigma_j^*} \left\| \mathbf{H} \mathbf{a}_j^* \right\|_2 \right)$$

This shows the second conclusion.

B.3. Proof of Lemma 4.5

Recall the definition of the set of empirical deflated matrices

$$\mathbf{\Sigma}_{k+1} = \mathbf{\Sigma}_k - \mathbf{v}_k \mathbf{v}_k^\top \mathbf{\Sigma}_k \mathbf{v}_k \mathbf{v}_k^\top$$

and the set of ground-truth deflated matrices

$$\mathbf{\Sigma}_{k+1}^* = \mathbf{\Sigma}_k^* - \mathbf{u}_k^* \mathbf{u}_k^{*\top} \mathbf{\Sigma}_k^* \mathbf{u}_k^* \mathbf{u}_k^{*\top}$$

By Lemma C.1, we have

$$\mathbf{\Sigma}_k^* = \sum_{j=k}^n \lambda_j^* \mathbf{u}_j^* \mathbf{u}_j^{*\top}$$

Therefore, by the orthogonality between $\mathbf{u}_i^*$ and $\mathbf{u}_j^*$ when $i \neq j$, we have

$$\mathbf{\Sigma}_k^* \mathbf{u}_j^* = \sum_{i=k}^n \lambda_i^* \mathbf{u}_i^* \left(\mathbf{u}_i^{*\top} \mathbf{u}_j^* \right) = \begin{cases} \lambda_j^* \mathbf{u}_j^* & \text{if } j \geq k \\ \mathbf{0} & \text{if } j < k \end{cases} \tag{36}$$

Next, we will focus on $\mathbf{\Sigma}_k \mathbf{u}_j^*$. Notice that

$$\mathbf{\Sigma}_{k+1} \mathbf{u}_j^* = \mathbf{\Sigma}_k \mathbf{u}_j^* - \mathbf{v}_k \mathbf{v}_k^\top \mathbf{\Sigma}_k \mathbf{v}_k \mathbf{v}_k^\top \mathbf{u}_j^*$$

Therefore

$$\mathbf{\Sigma}_k \mathbf{u}_j^* = \mathbf{\Sigma} \mathbf{u}_j^* - \sum_{k'=1}^{k-1} \mathbf{v}_{k'}^\top \mathbf{u}_j^* \cdot \mathbf{v}_{k'} \mathbf{v}_{k'}^\top \mathbf{\Sigma}_{k'} \mathbf{v}_{k'} = \lambda_j^* \mathbf{u}_j^* - \sum_{k'=1}^{k-1} \mathbf{v}_{k'}^\top \mathbf{u}_j^* \cdot \mathbf{v}_{k'}^\top \mathbf{\Sigma}_{k'} \mathbf{v}_{k'} \cdot \mathbf{v}_{k'} \tag{37}$$

Therefore, combining (36) and (37) gives

$$\left(\mathbf{\Sigma}_k^* - \mathbf{\Sigma}_k \right) \mathbf{u}_j^* = \begin{cases} \sum_{k'=1}^{k-1} \mathbf{v}_{k'}^\top \mathbf{u}_j^* \cdot \mathbf{v}_{k'}^\top \mathbf{\Sigma}_{k'} \mathbf{v}_{k'} \cdot \mathbf{v}_{k'} & \text{if } j \geq k \\ \sum_{k'=1}^{k-1} \mathbf{v}_{k'}^\top \mathbf{u}_j^* \cdot \mathbf{v}_{k'}^\top \mathbf{\Sigma}_{k'} \mathbf{v}_{k'} \cdot \mathbf{v}_{k'} - \lambda_j^* \mathbf{u}_j & \text{if } j < k \end{cases}$$

This shows the first conclusion. Next, we will put more emphasis on the case where $j \geq k$. Under this scenario, we can first show that

$$\mathbf{v}_{k'}^\top \mathbf{\Sigma}_{k'} \mathbf{v}_{k'} \leq \lambda_{k'} \left\| \mathbf{v}_{k'} \right\|_2^2 = \lambda_{k'}$$

Therefore,

$$\left\| \left(\mathbf{\Sigma}_k^* - \mathbf{\Sigma}_k \right) \mathbf{u}_j^* \right\|_2 \leq \sum_{k'=1}^{k-1} \lambda_{k'} \left| \mathbf{v}_{k'}^\top \mathbf{u}_j^* \right| \cdot \left\| \mathbf{v}_k' \right\|_2 = \sum_{k'=1}^{k-1} \lambda_{k'} \left| \mathbf{v}_{k'}^\top \mathbf{u}_j^* \right|$$

17

B.4. Proof of Lemma 4.6

Let us consider $\mathbf{a}_i^\top (\mathbf{M}^* + \mathbf{H}) \mathbf{a}_j^*$. On one hand, since $\mathbf{a}_i$ is an eigenvector of $\mathbf{M}^* + \mathbf{H}$, we have

$$\mathbf{a}_i^\top (\mathbf{M}^* + \mathbf{H}) \mathbf{a}_j^* = \sigma_i \mathbf{a}_i^\top \mathbf{a}_j^*$$

On the other hand, since $\mathbf{a}_i^*$ is the ith eigenvector of $\mathbf{M}^*$, we have

$$\mathbf{a}_i^\top (\mathbf{M}^* + \mathbf{H}) \mathbf{a}_j^* = \sigma_j^* \mathbf{a}_i^\top \mathbf{a}_j^* + \mathbf{a}_i^\top \mathbf{H} \mathbf{a}_j^*$$

Therefore,

$$\sigma_i \mathbf{a}_i^\top \mathbf{a}_j^* = \sigma_j^* \mathbf{a}_i^\top \mathbf{a}_j^* + \mathbf{a}_i^\top \mathbf{H} \mathbf{a}_j^*$$

which implies that

$$\left(\sigma_i - \sigma_j^*\right) \mathbf{a}_i^\top \mathbf{a}_j^* = \mathbf{a}_i^\top \mathbf{H} \mathbf{a}_j^*$$

Thus, when $\sigma_i \neq \sigma_j^*$, we have

$$\mathbf{a}_i^\top \mathbf{a}_j^* = \frac{\mathbf{a}_i^\top \mathbf{H} \mathbf{a}_j^*}{\sigma_i - \sigma_j^*}$$

B.5. Proof of Lemma 4.7

The proof of Lemma 4.7 is based on the Neumann expansion (Eldridge et al., 2017; Chen et al., 2020), which we state below

Lemma B.1 (Theorem 7 of (Eldridge et al., 2017)). *Let $\mathbf{M}^* \in \mathbb{R}^{d \times d}$ be a rank-r matrix, and let $\mathbf{M} = \mathbf{M}^* + \mathbf{H}$ for some $\mathbf{H} \in \mathbb{R}^{d \times d}$. Let $\sigma_i^*, \mathbf{a}_i^*$ be the i-th eigenvalue and eigenvector of $\mathbf{M}^*$, and $\sigma_i, \mathbf{a}_i$ be the i-th eigenvalue and eigenvector of $\mathbf{M}$. Suppose that $\|\mathbf{H}\|_2 \leq |\sigma_i|$. Then,*

$$\mathbf{a}_i = \sum_{j=1}^r \frac{\sigma_j^*}{\sigma_i} \mathbf{a}_i^\top \mathbf{a}_j^* \left(\sum_{s=0}^\infty \frac{1}{\sigma_i^s} \mathbf{H}^s \mathbf{a}_j^*\right).$$

To start, using the Neumann expansion (Lemma B.1), we prove an auxiliary lemma, as stated below.

Lemma B.2. *Under the same setting of Lemma B.1, we further assume that $\sigma_i^* \geq 0$ for all $i \in [r]$, and that $\|\mathbf{H}\|_2 \leq \frac{1}{8}\sigma_1^*$. Then we have*

$$\left\|\mathbf{a}_1 - \sum_{j=1}^r \frac{\sigma_j^*}{\sigma_1} \mathbf{a}_1^\top \mathbf{a}_j^* \cdot \mathbf{a}_j^*\right\|_2 \leq \frac{4}{3} \cdot \frac{\|\mathbf{H}\|_2}{\sigma_1}$$

Proof. Applying Lemma B.1 with $i = 1$, we have

$$\mathbf{a}_1 = \sum_{j=1}^r \frac{\sigma_j^*}{\sigma_1} \mathbf{a}_1^\top \mathbf{a}_j^* \left(\mathbf{a}_j^* + \sum_{s=1}^\infty \frac{1}{\sigma_1^s} \mathbf{H}^s \mathbf{a}_j^*\right) = \sum_{j=1}^r \frac{\sigma_j^*}{\sigma_1} \mathbf{a}_1^\top \mathbf{a}_j^* \cdot \mathbf{a}_j^* + \sum_{j=1}^r \frac{\sigma_j^*}{\sigma_1} \mathbf{a}_1^\top \mathbf{a}_j^* \left(\sum_{s=1}^\infty \frac{1}{\sigma_i^s} \mathbf{H}^s \mathbf{a}_j^*\right)$$

Therefore, our quantity of interest can be written as

$$
\left\| \mathbf{a}_1 - \sum_{j=1}^{r} \frac{\sigma_j^*}{\sigma_1} \mathbf{a}_1^\top \mathbf{a}_j^* \cdot \mathbf{a}_j^* \right\|_2 = \left\| \sum_{j=1}^{r} \frac{\sigma_j^*}{\sigma_1} \mathbf{a}_1^\top \mathbf{a}_j^* \left(\sum_{s=1}^{\infty} \frac{1}{\sigma_1^s} \mathbf{H}^s \mathbf{a}_j^* \right) \right\|_2
$$

$$
= \left\| \sum_{s=1}^{\infty} \frac{\mathbf{H}^s}{\sigma_1^{s+1}} \sum_{j=1}^{r} \sigma_j^* \mathbf{a}_1^\top \mathbf{a}_j^* \cdot \mathbf{a}_j^* \right\|_2
$$

$$
= \left\| \sum_{s=1}^{\infty} \frac{\mathbf{H}^s}{\sigma_1^{s+1}} \begin{bmatrix} \mathbf{a}_1^* & \cdots & \mathbf{a}_r^* \end{bmatrix} \begin{bmatrix} \sigma_1^* \mathbf{a}_1^\top \mathbf{a}_1^* \\ \vdots \\ \sigma_r^* \mathbf{a}_1^\top \mathbf{a}_r^* \end{bmatrix} \right\|_2 \tag{38}
$$

$$
\leq \left\| \sum_{s=1}^{\infty} \frac{\mathbf{H}^s}{\sigma_1^{s+1}} \right\|_2 \cdot \left\| \begin{bmatrix} \mathbf{a}_1^* & \cdots & \mathbf{a}_r^* \end{bmatrix} \right\|_2 \cdot \left\| \begin{bmatrix} \sigma_1^* \mathbf{a}_1^\top \mathbf{a}_1^* \\ \vdots \\ \sigma_r^* \mathbf{a}_1^\top \mathbf{a}_r^* \end{bmatrix} \right\|_2
$$

Now, we have a product of three terms to analyze. Recall that $\sigma_1^* > 0$. To start, by Weyl's inequality, we have $\sigma_1 \geq \sigma_1^* - \|\mathbf{H}\|_2$. Since $\|\mathbf{H}\|_2 \leq \frac{1}{8}\sigma_1^*$, we must have that $\|\mathbf{H}\|_2 \leq \sigma_1$. Applying the triangle inequality and the standard result of the geometric series gives

$$
\left\| \sum_{s=1}^{\infty} \frac{\mathbf{H}^s}{\sigma_1^{s+1}} \right\|_2 \leq \sum_{s=1}^{\infty} \frac{\|\mathbf{H}\|_2^s}{\sigma_1^{s+1}} = \frac{\|\mathbf{H}\|_2}{\sigma_1 (\sigma_1 - \|\mathbf{H}\|_2)} \tag{39}
$$

Next, since $\mathbf{a}_1^*, \ldots, \mathbf{a}_r^*$ are orthonormal vectors, we must have that

$$
\left\| \begin{bmatrix} \mathbf{a}_1^* & \cdots & \mathbf{a}_r^* \end{bmatrix} \right\|_2 \leq 1 \tag{40}
$$

For the same reason, and since $\sigma_i^* \geq 0$ for all $i \in [r]$, we also have that

$$
\left\| \begin{bmatrix} \sigma_1^* \mathbf{a}_1^\top \mathbf{a}_1^* \\ \vdots \\ \sigma_r^* \mathbf{a}_1^\top \mathbf{a}_r^* \end{bmatrix} \right\|_2 \leq \sigma_1^* \left\| \begin{bmatrix} \mathbf{a}_1^\top \mathbf{a}_1^* \\ \vdots \\ \mathbf{a}_1^\top \mathbf{a}_r^* \end{bmatrix} \right\|_2 = \sigma_1^* \left\| \mathbf{a}_1^\top \begin{bmatrix} \mathbf{a}_1^*, \ldots, \mathbf{a}_r^* \end{bmatrix} \right\|_2 \leq \sigma_1^* \tag{41}
$$

To put things together, we plug (39), (40), and (41) into (38) to get that

$$
\left\| \mathbf{a}_1 - \sum_{j=1}^{r} \frac{\sigma_j^*}{\sigma_1} \mathbf{a}_1^\top \mathbf{a}_j^* \cdot \mathbf{a}_j^* \right\|_2 \leq \frac{\|\mathbf{H}\|_2}{\sigma_1 (\sigma_1 - \|\mathbf{H}\|_2)} \cdot 1 \cdot \sigma_1^* = \frac{\sigma_1^* \|\mathbf{H}\|_2}{\sigma_1 (\sigma_1 - \|\mathbf{H}\|_2)} \tag{42}
$$

Apply again $\|\mathbf{H}\|_2 \leq \frac{1}{8}\sigma_1^*$, we have

$$
\sigma_1 - \|\mathbf{H}\|_2 \geq \sigma_1^* - 2\|\mathbf{H}\|_2 \geq \frac{3}{4}\sigma_1^*
$$

Therefore, (42) becomes

$$
\left\| \mathbf{a}_1 - \sum_{j=1}^{r} \frac{\sigma_j^*}{\sigma_1} \mathbf{a}_1^\top \mathbf{a}_j^* \cdot \mathbf{a}_j^* \right\|_2 \leq \frac{4}{3} \cdot \frac{\|\mathbf{H}\|_2}{\sigma_1} \tag{43}
$$

$\square$

Now, we will use Lemma B.2 to prove the desired statements. Notice that the Euclidean projection of $\mathbf{a}_1$ onto the subspace spanned by $\{\mathbf{a}_j^*\}_{j=1}^r$ (namely $\mathcal{S}^*$) is given by $\sum_{j=1}^r \mathbf{a}_1^\top \mathbf{a}_j^* \cdot \mathbf{a}_j^*$. Therefore, we have

$$
\begin{aligned}
\left(1 - \sum_{j=1}^r \left(\mathbf{a}_1^\top \mathbf{a}_j^*\right)^2\right)^{\frac{1}{2}} &= \left\|\mathbf{a}_1 - \sum_{j=1}^r \mathbf{a}_1^\top \mathbf{a}_j^* \cdot \mathbf{a}_j^*\right\|_2 \\
&= \min_{\mathbf{z} \in \mathcal{S}^*} \|\mathbf{a}_1 - \mathbf{z}\|_2 \\
&\leq \left\|\mathbf{a}_1 - \sum_{j=1}^r \frac{\sigma_j^*}{\sigma_1} \mathbf{a}_1^\top \mathbf{a}_j^* \cdot \mathbf{a}_j^*\right\|_2
\end{aligned}
\tag{44}
$$

where the inequality follows from the fact that $\sum_{j=1}^r \frac{\sigma_j^*}{\sigma_1} \mathbf{a}_1^\top \mathbf{a}_j^* \cdot \mathbf{a}_j^* \in \mathcal{S}^*$. With the assumption of the lemma, we can invoke Lemma B.2 to get that

$$
\left\|\mathbf{a}_1 - \sum_{j=1}^r \frac{\sigma_j^*}{\sigma_1} \mathbf{a}_1^\top \mathbf{a}_j^* \cdot \mathbf{a}_j^*\right\|_2 \leq \frac{4}{3} \cdot \frac{\|\mathbf{H}\|_2}{\sigma_1} \leq \frac{32}{21} \cdot \frac{\|\mathbf{H}\|_2}{\sigma_1^*}
$$

Combining with (44) gives

$$
\sum_{j=1}^r \left(\mathbf{a}_1^\top \mathbf{a}_j^*\right)^2 \geq 1 - \left(\frac{32}{21} \cdot \frac{\|\mathbf{H}\|_2}{\sigma_1^*}\right)^2 \geq 1 - 2.4 \cdot \frac{\|\mathbf{H}\|_2^2}{\sigma_1^{*2}}
$$

Moving $\sum_{j=2}^r \left(\mathbf{a}_1^\top \mathbf{a}_j^*\right)^2$ to the right-hand side of the inequality gives

$$
\left(\mathbf{a}_1^\top \mathbf{a}_1^*\right)^2 \geq 1 - 2.4 \cdot \frac{\|\mathbf{H}\|_2^2}{\sigma_1^{*2}} - \sum_{j=2}^r \left(\mathbf{a}_1^\top \mathbf{a}_j^*\right)^2
$$

The proof then conclude by letting $\mathbf{a}_1 = \mathbf{u}_k, \sigma_1^* = \lambda_k^*, \mathbf{a}_j^* = \mathbf{u}_{k+j-1}^*$ for all $j \in [d - k + 1]$, and $\mathbf{H} = \mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*$.

B.6. Proof of Lemma 4.8

To start, by the definition of c_0, we invoke Lemma 4.4 with $\mathbf{M} = \mathbf{\Sigma}_k, \mathbf{M}^* = \mathbf{\Sigma}_k^*$ to obtain that

$$
\left|\mathbf{v}_k^\top \mathbf{u}_j^*\right| \leq c_0 \left(\left(\frac{\lambda_j^*}{\lambda_k}\right)^t + \frac{\lambda_j^*}{\lambda_k - \lambda_i^*} \left\|\left(\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\right)\mathbf{u}_j^*\right\|_2\right)
\tag{45}
$$

Also, invoking Lemma 4.5, we have that when $j \geq k$,

$$
\left\|\left(\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\right)\mathbf{u}_j^*\right\|_2 \leq \sum_{k'=1}^{k-1} \lambda_{k'} \left|\mathbf{v}_{k'}^\top \mathbf{u}_j^*\right|
\tag{46}
$$

Plugging the upper bound of $\left\|\left(\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\right)\mathbf{u}_j^*\right\|_2$ from (46) into (45), we have that when $j \geq k$

$$
\left|\mathbf{v}_k^\top \mathbf{u}_j^*\right| \leq c_0 \left(\left(\frac{\lambda_j^*}{\lambda_k}\right)^t + \frac{\lambda_j^*}{\lambda_k - \lambda_j^*} \sum_{k'=1}^{k-1} \lambda_{k'} \left|\mathbf{v}_{k'}^\top \mathbf{u}_j^*\right|\right)
$$

Equivalently, we can multiply both sides of the inequality with λ_k to obtain that

$$
\lambda_k \left|\mathbf{v}_k^\top \mathbf{u}_j^*\right| \leq \lambda_k c_0 \left(\left(\frac{\lambda_j^*}{\lambda_k}\right)^t + \frac{\lambda_j^*}{\lambda_k - \lambda_j^*} \sum_{k'=1}^{k-1} \lambda_{k'} \left|\mathbf{v}_{k'}^\top \mathbf{u}_j^*\right|\right)
\tag{47}
$$

Notice that, the form of the inequality in (47) allows us to invoke Lemma C.3 to analyze the growth of $\lambda_k \left|\mathbf{v}_k^\top \mathbf{u}_j^*\right|$. In particular, setting $a_k = c_0 \lambda_k \left(\frac{\lambda_j^*}{\lambda_k}\right)^t$ for $k > 1$ and $a_1 = \lambda_1 \left|\mathbf{v}_1^\top \mathbf{u}_j^*\right|$, and $b_k = \frac{c_0 \lambda_k \lambda_j^*}{\lambda_k - \lambda_j^*}$, we can have that $\lambda_k \left|\mathbf{v}_k^\top \mathbf{u}_j^*\right| \leq Q_k$

for all k for Q_k defined in Lemma C.3. Therefore, Lemma C.3 implies that

$$\sum_{k'=1}^{k-1} \lambda_{k'} \left| \mathbf{v}_{k'}^\top \mathbf{u}_j^* \right| \le c_0 \sum_{k'=1}^{k-1} \lambda_{k'} \left(\frac{\lambda_j^*}{\lambda_k} \right)^t \prod_{\ell=k'+1}^{k-1} \left(1 + \frac{c_0 \lambda_\ell \lambda_j^*}{\lambda_\ell - \lambda_j^*} \right) \tag{48}$$

Now, we focus on the term $\left(1 + \frac{c_0 \lambda_\ell \lambda_j^*}{\lambda_\ell - \lambda_j^*} \right)$ in (48). In particular, we should notice that $\left(1 + \frac{c_0 \lambda_\ell \lambda_j^*}{\lambda_\ell - \lambda_j^*} \right)$ is only used in (48) for $\ell < k$. We can observe that

$$\frac{c_0 \lambda_\ell \lambda_j^*}{\lambda_\ell - \lambda_j^*} = c_0 \left(\frac{1}{\lambda_j^*} - \frac{1}{\lambda_\ell} \right)^{-1}$$

This implies that $\left(1 + \frac{c_0 \lambda_\ell \lambda_j^*}{\lambda_\ell - \lambda_j^*} \right)$ increases monotonically as λ_ℓ decrease. Recall that $|\lambda_\ell^* - \lambda_\ell| \le \|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\|_F \le \frac{1}{8} \mathcal{T}_{k,\min}$. This implies that for all $\ell < k$, we must have that $\lambda_\ell \ge \lambda_{\ell+1}^* \ge \lambda_{k+1}^* \ge \lambda_k^*$. Thus, we must have that

$$1 + \frac{c_0 \lambda_\ell \lambda_j^*}{\lambda_\ell - \lambda_j^*} \le 1 + \frac{c_0 \lambda_k^* \lambda_j^*}{\lambda_k^* - \lambda_j^*} \le 1 + \frac{c_0 \lambda_k^* \lambda_{k+1}^*}{\lambda_k^* - \lambda_{k+1}^*} \tag{49}$$

Recall that in the statement of Lemma 4.8 we defined $\mathcal{G}_k = 1 + \frac{c_0 \lambda_k^* \lambda_{k+1}^*}{\lambda_k^* - \lambda_{k+1}^*}$. Then (49) implies that (48) can be simplified Suppose that $\lambda_k \ge \lambda_j^*$, then we have

$$\sum_{k'=1}^{k-1} \lambda_{k'} \left| \mathbf{v}_{k'}^\top \mathbf{u}_j^* \right| \le c_0 \sum_{k'=1}^{k-1} \lambda_{k'} \mathcal{G}_k^{k-k'-1} \left(\frac{\lambda_j^*}{\lambda_k} \right)^t \le c_0 \sum_{k'=1}^{k-1} \mathcal{G}_k^{k-k'-1} \left(\frac{\lambda_j^*}{\lambda_k} \right)^t \tag{50}$$

where the last inequality follows from the fact that

$$\lambda_{k'} \le \lambda_{k'}^* + \|\mathbf{\Sigma}_{k'} - \mathbf{\Sigma}_{k'}^*\|_2 \le \lambda_2^* + \frac{1}{8} \mathcal{T}_{k',\min} \le \lambda_1^* = 1$$

for all $k \ge 2$, and $\lambda_1 = \lambda_1^* = 1$ by definition. Plugging (50) into the bound on $\left\| (\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*) \mathbf{u}_j^* \right\|_2$ in (46) gives

$$\left\| (\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*) \mathbf{u}_j^* \right\|_2 \le c_0 \sum_{k'=1}^{k-1} \mathcal{G}_k^{k-k'-1} \left(\frac{\lambda_j^*}{\lambda_k} \right)^t \tag{51}$$

Now, we can invoke Lemma 4.6 with $\mathbf{M} = \mathbf{\Sigma}_k, \mathbf{M}^* = \mathbf{\Sigma}_k^*$, and correspondingly $\mathbf{a}_1 = \mathbf{u}_k, \mathbf{a}_j = \mathbf{u}_j^*$ to get that

$$\left| \mathbf{u}_k^\top \mathbf{u}_j^* \right| = \left| \frac{\mathbf{u}_k^\top (\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*) \mathbf{u}_j^*}{\lambda_k - \lambda_j^*} \right| \le \frac{\left\| (\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*) \mathbf{u}_j^* \right\|_2}{|\lambda_k - \lambda_j^*|} \tag{52}$$

This allows us to plug (51) into (52) to obtain that

$$\left| \mathbf{u}_k^\top \mathbf{u}_j^* \right| \le \frac{c_0}{|\lambda_k - \lambda_j^*|} \sum_{k'=1}^{k-1} \mathcal{G}_k^{k-k'-1} \left(\frac{\lambda_j^*}{\lambda_k} \right)^t$$

Since $\|\mathbf{\Sigma}_{k'} - \mathbf{\Sigma}_{k'}^*\|_F \le \frac{1}{8} \mathcal{T}_{k,\min}$ for all $k' \in [k-1]$, then we have that $\lambda_{k-1} \ge \lambda_{k-1}^* - \frac{1}{8} \mathcal{T}_{k-1} \ge \lambda_k^* + \frac{1}{8} \mathcal{T}_{k-1} \ge \lambda_k$. Apply Lemma C.5, we have that as long as $t \ge \frac{\log \mathcal{G}_k}{\log \lambda_{k'} - \lambda_{k'+1}}$ for all $k' \in [k-1]$,

$$\begin{aligned} \left| \mathbf{u}_k^\top \mathbf{u}_j^* \right| &\le \frac{c_0}{|\lambda_k - \lambda_j^*|} \sum_{k'=1}^{k-1} \mathcal{G}_k^{k-k'-1} \left(\frac{\lambda_j^*}{\lambda_k} \right)^t \\ &= \frac{c_0 \lambda_j^{*t} \mathcal{G}_k^{k-1}}{|\lambda_k - \lambda_j^*|} \sum_{k'=1}^{k-1} \frac{\left(\lambda_{k'}^{-1} \right)^t}{\mathcal{G}_k^{k'}} \\ &\le \frac{c_0 \lambda_j^{*t} \mathcal{G}^{k-1}}{|\lambda_k - \lambda_j^*|} \cdot \left(1 - \mathcal{G}_k \cdot \max_{k' \in [k-1]} \left(\frac{\lambda_{k'+1}}{\lambda_{k'}} \right)^t \right)^{-1} \frac{\lambda_{k-1}^{-t}}{\mathcal{G}_k^{k-1}} \\ &= \frac{c_0}{|\lambda_k - \lambda_j^*|} \left(1 - \mathcal{G}_k \cdot \max_{k' \in [k-1]} \left(\frac{\lambda_{k'+1}}{\lambda_{k'}} \right)^t \right)^{-1} \left(\frac{\lambda_j^*}{\lambda_{k-1}} \right)^t \end{aligned}$$

In fact, as long as $t \geq \frac{\log 2\mathcal{G}_k}{\log \lambda_{k'} - \lambda_{k'+1}}$, we must have that for all $k' \in [k-1]$,

$$\mathcal{G}_k \cdot \max_{k' \in [k-1]} \left(\frac{\lambda_{k'+1}}{\lambda_{k'}} \right)^t \leq \frac{1}{2}$$

which implies that for all $j > k$

$$\left| \mathbf{u}_k^\top \mathbf{u}_j^* \right| \leq \frac{2c_0}{\left| \lambda_k - \lambda_j^* \right|} \left(\frac{\lambda_j^*}{\lambda_{k-1}} \right)^t \leq \frac{16c_0}{7\mathcal{T}_k} \left(\frac{\lambda_j^*}{\lambda_k^*} \right)^t \tag{53}$$

where the last inequality follows from

$$\left| \lambda_k - \lambda_j^* \right| \geq \left| \lambda_k^* - \frac{1}{8}\mathcal{T}_k - \lambda_j^* \right| \geq \frac{7}{8}\mathcal{T}_k$$

since $\lambda_k \geq \lambda_k^* - \|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_2 \geq \lambda_k^* - \frac{1}{8}\mathcal{T}_k$, and $\lambda_k^* \geq \lambda_j^* + \mathcal{T}_k$ for $j > k$. Applying Lemma 4.7, we have that

$$\left(\mathbf{u}_k^\top \mathbf{u}_k^* \right)^2 \geq 1 - \frac{2.4}{\lambda_k^{*2}} \|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_2^2 - \sum_{j=k+1}^{d} \left(\mathbf{u}_k^\top \mathbf{u}_j^* \right)^2$$

Therefore, $\|\mathbf{u}_k - \mathbf{u}_k^*\|_2^2$ can be bounded as

$$\|\mathbf{u}_k - \mathbf{u}_k^*\|_2^2 \leq 2 \left(1 - \left(\mathbf{u}_k^\top \mathbf{u}_k^* \right)^2 \right) \leq \frac{4.8}{\lambda_k^*} \|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_2^2 + \sum_{j=k+1}^{d} \left(\mathbf{u}_k^\top \mathbf{u}_j^* \right)^2 \tag{54}$$

We finally plug (53) into (54) to get

$$\|\mathbf{u}_k - \mathbf{u}_k^*\|_2^2 \leq \frac{4.8}{\lambda_k^*} \|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_2^2 + \frac{11c_0^2}{2\mathcal{T}_k^2} \sum_{j=k+1}^{d} \left(\frac{\lambda_j^*}{\lambda_k^*} \right)^{2t}$$

as desired. Lastly, we should notice that since $\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_F \leq \frac{1}{8}\mathcal{T}_{k,\min} \leq \lambda_{k'}^* - \lambda_{k'+1}^*$ for all $k' \in [k]$, it must hold that $\lambda_{k'} \geq \lambda_{k'}^* - \|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_F \geq \frac{1}{8}\lambda_{k'+1}^* + \frac{7}{8}\lambda_{k'}^*$ and $\lambda_{k'+1} \leq \lambda_{k'+1}^* + \|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_F \leq \frac{1}{8}\lambda_{k'}^* + \frac{7}{8}\lambda_{k'+1}^*$ for all $k' \in [k-1]$. Thus, the condition that $t \geq \frac{\log 2\mathcal{G}_k}{\log \lambda_{k'} - \lambda_{k'+1}}$ can be satisfied as long as

$$t \geq \log 2\mathcal{G}_k \left(\log \frac{\frac{1}{8}\lambda_{k'+1}^* + \frac{7}{8}\lambda_{k'}^*}{\frac{1}{8}\lambda_{k'}^* + \frac{7}{8}\lambda_{k'+1}^*} \right)^{-1} = \log 2\mathcal{G}_k \left(\log \frac{\lambda_{k'+1}^* + 7\lambda_{k'}^*}{\lambda_{k'}^* + 7\lambda_{k'+1}^*} \right)^{-1}$$

which is provided as an assumption in the lemma. This then finishes the proof.

B.7. Proof of Theorem 4.1

Let $\hat{K}$ be the smallest integer such that $\left\| \boldsymbol{\Sigma}_{\hat{K}} - \boldsymbol{\Sigma}_{\hat{K}}^* \right\|_2 > \frac{1}{8}\mathcal{T}_{K,\min}$, and let $K' = \min\left\{ \hat{K}, K+1 \right\}$. Therefore, for all $k < K'$, we shall have that $\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_2 \leq \frac{1}{8}\mathcal{T}_{K,\min}$. Moreover, since $t \geq \log 2\mathcal{G}_k \left(\log \frac{\lambda_{k'+1}^* + 7\lambda_{k'}^*}{\lambda_{k'}^* + 7\lambda_{k'+1}^*} \right)^{-1}$, we can invoke Lemma 4.8 to have that

$$\|\mathbf{u}_k - \mathbf{u}_k^*\|_2^2 \leq \frac{4.8}{\lambda_k^*} \|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_2^2 + \frac{11c_0^2}{2\mathcal{T}_k^2} \sum_{j=k+1}^{d} \left(\frac{\lambda_j^*}{\lambda_k^*} \right)^{2t} \tag{55}$$

Applying Lemma C.5 again with $g = 1$ and $p_{k'} = \lambda_{d-k'+1}^*$ yields that, as long as $t \geq \frac{1}{\log \lambda_j^* - \lambda_{j+1}^*}$ for all $j > k$, it holds that

$$\sum_{j=k+1}^{d} \lambda_j^{*2t} \leq \left(1 - \max_{j>k} \left(\frac{\lambda_j^*}{\lambda_{j+1}^*} \right)^{2t} \right) \lambda_{k+1}^{*2t} \leq 2\lambda_{k+1}^{*2t}$$

This transforms (55) into

$$\|\mathbf{u}_k - \mathbf{u}_k^*\|_2^2 \leq \frac{4.8}{\lambda_k^*} \|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_2^2 + \frac{11c_0^2}{\mathcal{T}_k^2} \left(\frac{\lambda_{k+1}^*}{\lambda_k^*}\right)^{2t} \tag{56}$$

Now, we invoke Lemma 3.4 to obtain that, as long as $\|\boldsymbol{\delta}_k\| \leq \frac{1}{6}$, we shall have that

$$\left\|\boldsymbol{\Sigma}_{k+1} - \boldsymbol{\Sigma}_{k+1}^*\right\|_F \leq 3 \|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_F + 2\lambda_k^* \|\mathbf{u}_k - \mathbf{u}_k^*\|_2 + 5\lambda_k^* \|\boldsymbol{\delta}_k\|_2 \tag{57}$$

We first plug (56) into (57) by applying $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ to have that

$$\left\|\boldsymbol{\Sigma}_{k+1} - \boldsymbol{\Sigma}_{k+1}^*\right\|_F \leq 8 \|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_F + \lambda_k^* \left(5 \|\boldsymbol{\delta}_k\|_2 + \frac{7c_0}{\mathcal{T}_k} \left(\frac{\lambda_{k+1}^*}{\lambda_k^*}\right)^t\right) \tag{58}$$

Now, we can apply Lemma C.4 to unroll (58) to get that

$$\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_F \leq \sum_{k'=1}^{k-1} 8^{k-k'} \lambda_{k'}^* \left(5 \|\boldsymbol{\delta}_{k'}\|_2 + \frac{7c_0}{\mathcal{T}_{k'}} \left(\frac{\lambda_{k'+1}^*}{\lambda_{k'}^*}\right)^t\right) \tag{59}$$

Next, we can plug (59) into (56) to get that

$$\|\mathbf{u}_k - \mathbf{u}_k^*\|_2 \leq 3 \sum_{k'=1}^{k-1} 8^{k-k'} \frac{\lambda_{k'}^*}{\lambda_k^*} \left(5 \|\boldsymbol{\delta}_{k'}\|_2 + \frac{7c_0}{\mathcal{T}_{k'}} \left(\frac{\lambda_{k'+1}^*}{\lambda_{k'}^*}\right)^t\right) + \frac{4c_0}{\mathcal{T}_k} \left(\frac{\lambda_{k+1}^*}{\lambda_k^*}\right)^t \tag{60}$$

Finally, recall the definition of $\mathbf{v}_k = \mathbf{u}_k + \boldsymbol{\delta}_k$. Therefore, we have

$$\|\mathbf{v}_k - \mathbf{u}_k^*\|_2 \leq 3 \sum_{k'=1}^{k-1} 8^{k-k'} \frac{\lambda_{k'}^*}{\lambda_k^*} \left(5 \|\boldsymbol{\delta}_{k'}\|_2 + \frac{7c_0}{\mathcal{T}_{k'}} \left(\frac{\lambda_{k'+1}^*}{\lambda_{k'}^*}\right)^t\right) + \frac{4c_0}{\mathcal{T}_k} \left(\frac{\lambda_{k+1}^*}{\lambda_k^*}\right)^t + \|\boldsymbol{\delta}_k\|_2$$

$$\leq 3 \sum_{k'=1}^{k} 8^{k-k'} \frac{\lambda_{k'}^*}{\lambda_k^*} \left(5 \|\boldsymbol{\delta}_{k'}\|_2 + \frac{7c_0}{\mathcal{T}_{k'}} \left(\frac{\lambda_{k'+1}^*}{\lambda_{k'}^*}\right)^t\right)$$

for all $k \in [K']$. It thus remains to show that $\hat{K} > K$. In this way, we will have $K' = K$ and the theorem is then proved. Assume, for the sake of contradiction, that $\hat{K} \leq K$. Then we shall have that there exists $k \in [K]$ such that $\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_2 > \frac{1}{8}\mathcal{T}_{K,\min}$. To reach a contradiction, it thus remains to show that for all $k \in [K]$

$$\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|_2 \leq \frac{1}{8}\mathcal{T}_{K,\min}$$

By (59), it suffices to guarantee that

$$\sum_{k'=1}^{K-1} 8^{K-k'} \lambda_{k'}^* \left(5 \|\boldsymbol{\delta}_{k'}\|_2 + \frac{7c_0}{\mathcal{T}_{k'}} \left(\frac{\lambda_{k'+1}^*}{\lambda_{k'}^*}\right)^t\right) \leq \frac{1}{8}\mathcal{T}_{K,\min} \tag{61}$$

Moreover, recall that $\mathbf{v}_k$ is the output of the kth power iteration procedure. Therefore, we can invoke Lemma C.6 to have that

$$\|\boldsymbol{\delta}_k\|_2 \leq 2 \left(\frac{\lambda_2\left(\boldsymbol{\Sigma}_k\right)}{\lambda_k}\right)^t$$

Let $r_k = \max\left\{\frac{\lambda_2(\boldsymbol{\Sigma}_k)}{\lambda_k}, \frac{\lambda_{k+1}^*}{\lambda_k^*}\right\}$. Therefore,

$$5 \|\boldsymbol{\delta}_{k'}\|_2 + \frac{7c_0}{\mathcal{T}_{k'}} \left(\frac{\lambda_{k'+1}^*}{\lambda_{k'}^*}\right)^t \leq \frac{17c_0}{\mathcal{T}_{k'}} r_{k'}^t$$

Thus, to guarantee (61), it suffices to guarantee that

$$\sum_{k'=1}^{K-1} 8^{K-k'} \frac{\lambda_{k'}^*}{\mathcal{T}_{k'}} r_{k'}^t \leq \frac{\mathcal{T}_{K,\min}}{140c_0}$$

which, since $\lambda_k \leq \lambda_1 = \lambda_1^* = 1$, is satisfied as long as

$$t \geq \frac{3(K-k) + \log \frac{140c_0 K}{\mathcal{T}_{K,\min}^2}}{\log r_k^{-1}}$$

Since $\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_k^*\|_F \leq \frac{1}{8}\mathcal{T}_{k,\min}$, we have that $r_k \leq \frac{7\lambda_{k+1}^* + \lambda_k}{7\lambda_k^* + \lambda_{k+1}^*}$. Therefore, the requirement on t becomes

$$t \geq \left(\log \frac{7\lambda_k^* + \lambda_{k+1}^*}{7\lambda_{k+1}^* + \lambda_k^*} \right)^{-1} \left(3(K-k) + \log \frac{140c_0 K}{\mathcal{T}_{K,\min}^2} \right)$$

for all $k \in [K]$, which is provided as an assumption of the Theorem. This finishes the proof.

C. Auxiliary Lemmas

Lemma C.1. *Let $\{\mathbf{\Sigma}_k^*\}_{k=1}^d$ be defined in (4). For all $k \in [d]$, we have*

$$\mathbf{\Sigma}_k^* = \sum_{j=k}^d \lambda_j^* \mathbf{u}_j^* \mathbf{u}_j^{*\top}; \quad \mathbf{\Sigma}_k^* \mathbf{u}_\ell^* = \begin{cases} \lambda_\ell^* \mathbf{u}_\ell^* & \text{if } \ell \geq k \\ \mathbf{0} & \text{if } \ell < k \end{cases}$$

Proof. We shall prove the statement by induction. For the base case, when $k = 1$, we have $\mathbf{\Sigma}_k^* = \mathbf{\Sigma}$, and, by the form of the eigen-decomposition, we have

$$\mathbf{\Sigma}_k^* = \sum_{j=1}^d \lambda_j^* \mathbf{u}_j^* \mathbf{u}_j^{*\top}; \quad \mathbf{u}_k^* \mathbf{u}_k^{*\top} \mathbf{\Sigma}_k^* \mathbf{u}_k^* \mathbf{u}_k^{*\top} = \lambda_k^* \mathbf{u}_k^* \mathbf{u}_k^{*\top}$$

Now, suppose that we have

$$\mathbf{\Sigma}_k^* = \sum_{j=k}^d \lambda_j^* \mathbf{u}_j^* \mathbf{u}_j^{*\top}$$

We can see that

$$\mathbf{u}_k^* \mathbf{u}_k^{*\top} \mathbf{\Sigma}_k^* \mathbf{u}_k^* \mathbf{u}_k^{*\top} = \sum_{j=k}^d \lambda_j^* \left(\mathbf{u}_k^{*\top} \mathbf{u}_j^* \right)^2 \mathbf{u}_k^* \mathbf{u}_k^{*\top} = \lambda_k^* \mathbf{u}_k^* \mathbf{u}_k^{*\top}$$

where the last equality follows from the orthogonality between $\mathbf{u}_k^*$ and $\mathbf{u}_j^*$ when $j \neq k$. Therefore

$$\mathbf{\Sigma}_{k+1}^* = \sum_{j=k}^d \lambda_j^* \mathbf{u}_j^* \mathbf{u}_j^{*\top} - \lambda_k^* \mathbf{u}_k^* \mathbf{u}_k^{*\top} = \sum_{j=k+1}^d \lambda_j^* \mathbf{u}_j^* \mathbf{u}_j^{*\top}$$

which shows the inductive step and thus proves the first statement. To see the second statement, we notice that $\mathbf{u}_j^{*\top} \mathbf{u}_\ell^* = 0$ for all $j \neq \ell$ and $\mathbf{u}_j^{*\top} \mathbf{u}_\ell^* = 0$ if $j = \ell$. Therefore

$$\mathbf{\Sigma}_k^* \mathbf{u}_\ell^* = \sum_{j=k}^d \mathbf{u}_j^{*\top} \mathbf{u}_\ell^* \cdot \lambda_j^* \mathbf{u}_j^* = \begin{cases} \lambda_\ell^* \mathbf{u}_\ell^* & \text{if } \ell \geq k \\ \mathbf{0} & \text{if } \ell < k \end{cases}$$

$\square$

Lemma C.2. *Consider the power iteration procedure in (14). Let $\{\sigma_j\}_{j=1}^r$ be the eigenvalues of $\mathbf{M}$ in (14). Then we have*

$$\min_{s \in \{\pm 1\}} \|s \cdot \mathbf{x}_t - \mathbf{a}_1\|_2^2 \leq 2 \left(\frac{\sigma_2}{\sigma_1} \right)^t$$

Proof. For convenience, we define the sequence $\{\hat{\mathbf{x}}_k\}_{k=0}^{\infty}$ as

$$\hat{\mathbf{x}}_t = \mathbf{M}\hat{\mathbf{x}}_{t-1}; \quad \hat{\mathbf{x}}_0 = \mathbf{x}_0$$

Then $\mathbf{x}_t$ from (14) can be written as

$$\mathbf{x}_t = \frac{\hat{\mathbf{x}}_t}{\|\hat{\mathbf{x}}_t\|_2} = \frac{\hat{\mathbf{x}}_t}{\|\mathbf{M}^t\mathbf{x}_0\|_2}$$

Let $\{\mathbf{a}_j\}_{j=1}^d$ be an extended orthogonal basis of $\{\mathbf{a}_j\}_{j=1}^r$. Then we can write

$$\mathbf{x}_0 = \sum_{j=1}^d \mathbf{x}_0^\top \mathbf{a}_j \cdot \mathbf{a}_j$$

Let $\sigma_j = 0$ for $j \geq r + 1$. Then we have that $\mathbf{M}\mathbf{a}_j = \sigma_j \mathbf{a}_j$. Therefore

$$\mathbf{M}^t\mathbf{x}_0 = \sum_{j=1}^d \mathbf{x}_0^\top \mathbf{a}_j \cdot \mathbf{M}^t \mathbf{a}_j = \sum_{j=1}^d \mathbf{x}_0^\top \mathbf{a}_j \cdot \sigma_j^t \mathbf{a}_j = \sum_{j=1}^r \mathbf{x}_0^\top \mathbf{a}_j \cdot \sigma_j^t \mathbf{a}_j$$

Since $\|\mathbf{M}^t\mathbf{x}_0\|_2 \leq \sigma_1^t$, we must have that

$$\left|\mathbf{x}_t^\top \mathbf{a}_j\right| = \frac{|\hat{\mathbf{x}}_t^\top \mathbf{a}_j|}{\|\mathbf{M}^t\mathbf{x}_0\|_2} \leq \left(\frac{\sigma_j}{\sigma_1}\right)^t \left|\mathbf{x}_0^\top \mathbf{a}_j\right|$$

Since $\{\mathbf{a}_j\}_{j=1}^d$ is an orthogonal basis of $\mathbb{R}^d$, we must have that $\|\mathbf{x}_t\|_2^2 = \sum_{j=1}^d \left(\mathbf{x}_t^\top \mathbf{a}_j\right)^2$. Thus

$$\left(\mathbf{x}_t^\top \mathbf{a}_1\right)^2 = 1 - \sum_{j=2}^d \left(\mathbf{x}_t^\top \mathbf{a}_j\right)^2 \geq 1 - \sum_{j=2}^d \left(\frac{\sigma_j}{\sigma_1}\right)^{2t} \left(\mathbf{x}_0^\top \mathbf{a}_j\right)^2 \geq 1 - \left(\frac{\sigma_2}{\sigma_1}\right)^{2t}$$

where the last inequality follows from $\sigma_j \leq \sigma_2$ for $j \geq 2$ and $\sum_{j=1}^d \left(\mathbf{x}_0^\top \mathbf{a}_j\right)^2 = \|\mathbf{x}_0\|_2^2 = 1$. Therefore, we can conclude that

$$\min_{s \in \{\pm 1\}} \|s \cdot \mathbf{x}_t - \mathbf{a}_1\|_2^2 \leq 2 \left(1 - \left(\mathbf{x}_t^\top \mathbf{a}_1\right)^2\right) \leq 2 \left(\frac{\sigma_2}{\sigma_1}\right)^{2t}$$

Taking a square root for both sides gives the desired result. $\qquad \square$

Lemma C.3. *Consider a sequence of quantities $\{Q_k\}_{k=1}^{\infty}$ satisfying*

$$Q_k = a_k + b_k \sum_{k'=1}^{k-1} Q_{k'}$$

with some $a_k, b_k \geq 0$ for all $k \in \mathbb{Z}_+$. Setting $Q_1 = a_1$, then we shall have that

$$Q_k = a_k + b_k \sum_{k'=1}^{k-1} a_{k'} \prod_{\ell=k'+1}^{k-1} (1 + b_\ell)$$

Proof. We shall prove by induction start from $Q_1 = a_1$. Suppose that $Q_1, \ldots Q_k$ satisfies the characterization. To start, we can see that

$$\sum_{k'=1}^{k-1} Q_{k'} = \frac{1}{b_k} (Q_k - a_k) = \sum_{k'=1}^{k-1} a_{k'} \prod_{\ell=k'+1}^{k-1} (1 + b_\ell)$$

25

Then, by the inductive hypothesis, we have

$$\sum_{k'=1}^{k} Q_{k'} = Q_k + \sum_{k'=1}^{k-1} Q_{k'}$$

$$= a_k + (b_k + 1) \sum_{k'=1}^{k-1} a_{k'} \prod_{\ell=k'+1}^{k-1} (1 + b_\ell)$$

$$= a_k + \sum_{k'=1}^{k-1} a_{k'} \prod_{\ell=k'+1}^{k} (1 + b_\ell)$$

$$= \sum_{k'=1}^{k} a_{k'} \prod_{\ell=k'+1}^{k} (1 + b_\ell)$$

This implies that

$$Q_{k+1} = a_{k+1} + b_{k+1} \sum_{k'=1}^{k} Q_{k'} = a_{k+1} + b_{k+1} \sum_{k'=1}^{k} a_{k'} \prod_{\ell=k'+1}^{k} (1 + b_\ell)$$

which finishes the proof. $\qquad\square$

Lemma C.4. *Consider a sequence of quantities* $\{Q_k\}_{k=1}^{\infty}$ *satisfying*

$$Q_{k+1} = a_k Q_k + b_k$$

with some $a_k, b_k \geq 0$ *for all* $k \in \mathbb{Z}_+$. *Set* $b_0 = Q_1$. *Then we have that*

$$Q_k = \sum_{k'=0}^{k-1} b_{k'} \prod_{j=k'+1}^{k-1} a_j$$

Proof. We shall prove by induction. For the base case, let $k = 1$. In this case, we have that

$$Q_1 = \sum_{k'=0}^{0} b_{k'} \prod_{j=k'+1}^{0} a_j = b_0 = Q_1$$

For the inductive case, assume that the property holds for k. Then we have that

$$Q_{k+1} = a_k Q_k + b_k = a_k \cdot \sum_{k'=0}^{k-1} b_{k'} \prod_{j=k'+1}^{k-1} a_j + b_k = \sum_{k'=0}^{k} b_{k'} \prod_{j=k'+1}^{k} a_j$$

This proves the inductive step and finishes the proof. $\qquad\square$

Lemma C.5. *Let* $\{p_k\}_{k=1}^{\infty}$ *be a positive increasing sequence, and* $g, t > 0$. *Let the series* S_k *be defined as*

$$S_k = \sum_{k'=1}^{k-1} \frac{p_{k'}^t}{g^{k'}}$$

As long as $t > \frac{\log g}{\log p_{k'} - \log p_{k'+1}}$ *for all* $k' \in [k-1]$, *we shall have that*

$$S_k \leq \left(1 - \max_{k' \in [k-1]} g \left(\frac{p_{k'}}{p_{k'+1}}\right)^t\right)^{-1} \frac{p_{k-1}^t}{g^{k-1}}$$

Proof. Set $\gamma = \max_{k' \in [k-1]} g\left(\frac{p_{k'}}{p_{k'+1}}\right)^t$. Since $t > \frac{\log g}{\log p_{k'+1} - \log p_{k'}}$ for all $k' \in [k-1]$, we must have that $\gamma < 1$. Moreover, by definition of γ, it must holds that for all $k' \in [k-1]$

$$\frac{p_{k'}^t}{g^{k'}} \leq \gamma \cdot \frac{p_{k'+1}^t}{g^{k'+1}}$$

Thus, we have that for all $k' \in [k-1]$

$$\frac{p_{k'}^t}{g^{k'}} \leq \gamma^{k-k'-1} \cdot \frac{p_{k-1}^t}{g^{k-1}}$$

Therefore S_k can be upper-bounded as

$$S_k \leq \sum_{k'=1}^{k-1} \gamma^{k-k'-1} \cdot \frac{p_{k-1}^t}{g^{k-1}} = \frac{p_{k-1}^t}{g^{k-1}} \sum_{k'=0}^{k-2} \gamma^{k'} \leq \frac{p_{k-1}^t}{g^{k-1}(1-\gamma)}$$

$\square$

Lemma C.6. *Let $\mathbf{x}_t$ be the result of running power iteration starting from $\mathbf{x}_0$ for t iterations, as defined in (14). Let $\sigma_j, \mathbf{a}_j$ be the j-th eigenvalue and eigenvector of $\mathbf{M}$ in (14). Assume that there exists $c_0 > 0$ such that $\left|\mathbf{x}_0^\top \mathbf{a}_1\right| \geq c_0^{-1}$. Then for all $j = 2, \ldots, r$, we have:*

$$\left|\mathbf{x}_t^\top \mathbf{a}_j\right| = \left|(\mathbf{x}_t - \mathbf{a}_1)^\top \mathbf{a}_j\right| \leq c_0 \left(\frac{\sigma_j}{\sigma_1}\right)^t.$$

Proof. We notice that for $j \geq 2$, it holds that $(\mathbf{x}_t - \mathbf{a}_1)^\top \mathbf{a}_j = \mathbf{x}_t^\top \mathbf{a}_j$. Since $\mathbf{M}$ is symmetric, we have that

$$\mathbf{x}_t^\top \mathbf{a}_j = \alpha \mathbf{x}_0^\top \mathbf{M}^t \mathbf{a}_j = \alpha \sigma_j^t \mathbf{x}_0^\top \mathbf{a}_j = \frac{\sigma_j^t}{\|\mathbf{M}^t \mathbf{x}_0\|_2} \mathbf{x}_0^\top \mathbf{a}_j$$

This brings the focus to $\|\mathbf{M}^t \mathbf{x}_0\|_2$. Let r be the rank of $\mathbf{M}$. Let $\{\mathbf{a}_j\}_{j=1}^d$ be an extended orthogonal basis of $\{\mathbf{a}_j\}_{j=1}^r$. Then we can write

$$\mathbf{x}_0 = \sum_{j=1}^d \mathbf{x}_0^\top \mathbf{a}_j \cdot \mathbf{a}_j$$

Let $\sigma_j = 0$ for $j \geq r+1$. Then we have that $\mathbf{M}\mathbf{a}_j = \sigma_j \mathbf{a}_j$. Therefore

$$\mathbf{M}^t \mathbf{x}_0 = \sum_{j=1}^d \mathbf{x}_0^\top \mathbf{a}_j \cdot \mathbf{M}^t \mathbf{a}_j = \sum_{j=1}^d \mathbf{x}_0^\top \mathbf{a}_j \cdot \sigma_j^t \mathbf{a}_j = \sum_{j=1}^r \mathbf{x}_0^\top \mathbf{a}_j \cdot \sigma_j^{*t} \mathbf{a}_j$$

where the last equality follows from $\sigma_j^t = 0$ when $t > 0$. Therefore, according to the Pythagorean Theorem,

$$\|\mathbf{M}^k \mathbf{x}_0\|_2 = \left(\sum_{j=1}^r \sigma_j^{2t} \left(\mathbf{x}_0^\top \mathbf{a}_j\right)^2\right)^{\frac{1}{2}} \geq \sigma_1^t \left|\mathbf{x}_0^\top \mathbf{a}_j\right|$$

Recall that $\left|\mathbf{x}_0^\top \mathbf{a}_j\right| \geq c_0^{-1} > 0$. Thus, we have

$$\left|(\mathbf{x}_t - \mathbf{a}_1)^\top \mathbf{a}_j\right| = \frac{\sigma_j^t}{\|\mathbf{M}^t \mathbf{x}_0\|_2} \left|\mathbf{x}_0^\top \mathbf{a}_j\right| \leq \left(\frac{\sigma_j}{\sigma_1}\right)^t \cdot \left|\frac{\mathbf{x}_0^\top \mathbf{a}_j}{\mathbf{x}_0^\top \mathbf{a}_1}\right| \leq \left|\mathbf{x}_0^\top \mathbf{a}_1\right|^{-1} \left(\frac{\sigma_j}{\sigma_1}\right)^t \leq c_0 \left(\frac{\sigma_j}{\sigma_1}\right)^t$$

$\square$

D. Experiment Details of Figure 1

We conduct spectral clustering (Belkin & Niyogi, 2003) on a subset of the MNIST dataset (Deng, 2012), namely $\mathcal{S} = \{\mathbf{x}_i\}_{i=1}^n$. In particular, our experiment starts with building a weighted graph $\mathcal{G} = (V, E)$ where each node in the graph $\mathcal{G}$ represents a sample image, and each edge weight is chosen according to

$$e_{ij} = \begin{cases} \exp\left(-\frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_j\|_2^2\right) & \text{if } \mathbf{x}_i \in \text{rNN}(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \text{rNN}(\mathbf{x}_i) \\ 0 & \text{otherwise} \end{cases}$$

Here $\text{rNN}(\mathbf{x}_i)$ represents the set of r Nearest Neighbors of $\mathbf{x}_i$. To be more specific, two nodes in the graph are considered adjacent if the two corresponding samples are top-r nearest neighbors, and the weight of the edge between the two nodes is computed using the RBF kernel with unit variance. Next, we compute the degree-normalized graph Laplacian as our similarity matrix

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$$

where $\mathbf{D} \in \mathbb{R}^{n \times n}$ is the diagonal degree matrix, and $\mathbf{A} \in \mathbb{R}^{n \times n}$ is the adjacency matrix. We extract the top-k eigenvectors of the similarity matrix $\mathbf{L}$ with the deflation algorithm in Algorithm 1 with t power iteration steps in each call to the PCA sub-routine, which results in the approximate eigenvectors $\mathbf{u}_1, \dots, \mathbf{u}_k$. We form a matrix $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_k] \in \mathbb{R}^{n \times k}$, and assign the ith sample with a new feature vector $\mathbf{v}_i$ as the ith row of $\mathbf{U}$. Lastly, we perform k'-means clustering with the newly extracted features $\{\mathbf{v}_i\}_{i=1}^n$, forming clusters $C = \{C_j\}_{j=1}^{k'}$, where $C_j \subseteq [n]$. The clustering result is evaluated using the mutual information metric (Hubert & Arabie, 1985).

$$\text{MI}(C, C^*) = \sum_{j,j'=1}^{k'} \mathbb{P}\left(C_j \cap C_{j'}^*\right) \log \frac{\mathbb{P}\left(C_j \cap C_{j'}^*\right)}{\mathbb{P}(C_j)\,\mathbb{P}\left(C_{j'}^*\right)} \tag{62}$$

where $\mathbb{P}(C_j) = \frac{|C_j|}{n}$ and ,similarly, $\mathbb{P}\left(C_j \cap C_{j'}^*\right) = \frac{|C_j \cap C_j^*|}{n}$. Here $C^* = \{C_j^*\}_{j=1}^{k'}$ is the ground-truth clustering with $C_j^* = \{i \in [n] : \text{the label of sample } i \text{ is } j\}$.

In our experiment, we use a subset of MNIST with $n = 1000$ randomly drawn samples. When constructing the graph, we use $r = 10$, that is, we connect each node to ten of its nearest neighbors. We cluster the samples into $k' = 10$ clusters. In Figure 1, the x-axis represents the number of power iteration steps t in each call to the sub-routine PCA, and the y-axis represents the mutual information score in (62). Each line in the figure represents different k, which is the number of eigenvectors extracted from $\mathbf{L}$.