JMLR: Workshop and Conference Proceedings vol 49:1–26, 2016

An Improved Gap-Dependency Analysis of the Noisy Power Method

Maria-Florina Balcan NINAMF@CS.CMU.EDU

Simon S. Du SSDU@CS.CMU.EDU

Yining Wang YININGWA@CS.CMU.EDU

Adams Wei Yu WEIYU@CS.CMU.EDU

Machine Learning Department, School of Computer Science, Carnegie Mellon University

Abstract

We consider the *noisy power method* algorithm, which has wide applications in machine learning and statistics, especially those related to principal component analysis (PCA) under resource (communication, memory or privacy) constraints. Existing analysis of the noisy power method (Hardt and Price, 2014; Li et al., 2016) shows an unsatisfactory dependency over the "consecutive" spectral gap $(\sigma_k - \sigma_{k+1})$ of an input data matrix, which could be very small and hence limits the algorithm's applicability. In this paper, we present a new analysis of the noisy power method that achieves improved gap dependency for both sample complexity and noise tolerance bounds. More specifically, we improve the dependency over $(\sigma_k - \sigma_{k+1})$ to dependency over $(\sigma_k - \sigma_{q+1})$, where q is an intermediate algorithm parameter and could be much larger than the target rank k. Our proofs are built upon a novel characterization of proximity between two subspaces that differ from canonical angle characterizations analyzed in previous works (Hardt and Price, 2014; Li et al., 2016). Finally, we apply our improved bounds to distributed private PCA and memory-efficient streaming PCA and obtain bounds that are superior to existing results in the literature.

Keywords: principal component analysis, noisy power method, spectral gap.

1. Introduction

Principal Component Analysis (PCA) is a fundamental problem in statistics and machine learning. The objective of PCA is to find a small number of orthogonal directions in the d-dimensional Euclidean space $\mathbb{R}^d$ that have the highest variance of a given sample set. Mathematically speaking, given a $d \times d$ positive semi-definite matrix $\mathbf{A}$ of interest ($\mathbf{A}$ is usually the sample covariance matrix $\mathbf{A} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i \boldsymbol{z}_i^\top$ for n data points $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n$), one wishes to find the top-k eigen-space of $\mathbf{A}$, where k is the number of principal directions of interest and is typically much smaller than the ambient dimension d. A popular algorithm for computing PCA is the *matrix power method*, which starts with a random $d \times p$ matrix ($p \geq k$) $\mathbf{X}_0$ with orthonormal columns and iteratively performs the following computation for $\ell = 1, \cdots, L$:

1. **Subspace iteration**: $\mathbf{Y}_\ell = \mathbf{A}\mathbf{X}_{\ell-1}$.
2. **QR factorization**: $\mathbf{Y}_\ell = \mathbf{X}_\ell \mathbf{R}_\ell$, where $\mathbf{X}_\ell \in \mathbb{R}^{d\times p}$ has orthonormal columns and $\mathbf{R}_\ell \in \mathbb{R}^{p\times p}$ is an upper-triangular matrix.

It is well-known that when the number of iterations L is sufficiently large, the span of the output $\mathbf{X}_L$ can be arbitrarily close to $\mathbf{U}_k$, the top-k eigen-space of $\mathbf{A}$; that is, $\|(\mathbf{I} - \mathbf{X}_L\mathbf{X}_L^\top)\mathbf{U}_k\|_2 \leq \epsilon$ for

arbitrarily small $\epsilon > 0$. One particular drawback of power method is that the rate of convergence depends on the *consecutive* eigengap $(\sigma_k - \sigma_{k+1})$ when $p = k$ (i.e., $\mathbf{X}_\ell$ has exactly the same number of columns as the target rank k). The consecutive eigengap could be very small for practical large-scale matrices. As a remedy, practitioners generally set p to be slightly larger than k for faster convergence and numerical stability (Musco and Musco, 2015). Gu (2015) formally justifies this process by proving that under mild conditions, the dependency on $(\sigma_k - \sigma_{k+1})$ could be improved to the "larger" spectral gap $(\sigma_k - \sigma_{q+1})$, for some $k \leq q \leq p$, which may be significantly larger than the consecutive gap even if q is at the same order of k. [1] Despite the wide applicability and extensive analysis of the (exact) matrix power method, in practice it is sometimes desired to analyze a *noisy* version of power method, where each subspace iteration computation is corrupted with noise. Such noise could come from resource constraints such as inherent machine precision or memory storage, or artificially imposed constraints for additional objectives such as data privacy preservation. In both cases, the noise model can be expressed as $\mathbf{Y}_\ell = \mathbf{A}\mathbf{X}_{\ell-1} + \mathbf{G}_\ell$, where $\mathbf{G}_\ell$ is a $d \times p$ noise matrix for iteration ℓ that can be either stochastic or deterministic (adversarial). Note that $\mathbf{G}_\ell$ could differ from iteration to iteration but the QR factorization step $\mathbf{Y}_\ell = \mathbf{X}_\ell \mathbf{R}_\ell$ is still assumed to be exact. The noisy power method has attracted increasing interest from both machine learning and theoretical computer science societies due to its simplicity and broad applicability (Hardt and Price, 2014; Li et al., 2016; Musco and Musco, 2015; Mitliagkas et al., 2013). In particular, (Hardt and Price, 2014) establishes both convergence guarantees and error tolerance (i.e., the largest magnitude of the noise matrix $\mathbf{G}_\ell$ the algorithm allows to produce consistent estimates of $\mathbf{U}_k$) of the noisy power method. (Hardt and Price, 2014) also applied their results to PCA with resource (privacy, memory) constraints and obtained improved bounds over existing results.

1.1. Our contributions

Improved gap dependency analysis of the noisy power method Our main contribution is a new analysis of the noisy power method with improved gap dependency. More specifically, we improve the prior gap dependency $(\sigma_k - \sigma_{k+1})$ to $(\sigma_k - \sigma_{q+1})$, where q is certain integer between the target rank k and the number of columns used in subspace iteration p. Our results partially solve a open question in (Hardt and Price, 2014), which conjectured that such improvement over gap dependency should be possible if p is larger than k. To our knowledge, our bounds are the first to remove dependency over the consecutive spectral gap $(\sigma_k - \sigma_{k+1})$ for the noisy power method.

Gap-independent bounds As a by-product of our improved gap dependency analysis, we apply techniques in a recent paper (Musco and Musco, 2015) to obtain *gap-independent* bounds for the approximation error $\|\mathbf{A} - \mathbf{X}_L \mathbf{X}_L^\top \mathbf{A}\|_2$. This partially addresses another conjecture in (Hardt and Price, 2014) regarding gap-independent approximation error bounds with slightly worse bounds on magnitude of error matrices $\mathbf{G}_\ell$.

Applications The PCA problem has been previously considered under various resource constraints. Two particularly important directions are private PCA (Hardt and Roth, 2013; Dwork et al., 2014; Chaudhuri et al., 2012; Hardt and Price, 2014), where privacy of the data matrix being analyzed is formally preserved, and distributed PCA (Balcan et al., 2014; Boutsidis et al., 2015) where data matrices are stored separately on several machines and communications among machines are

1. Sec. 2 provides such an example matrix with power-law decaying spectrum.

constrained. In this paper we propose a *distributed private PCA* problem that unifies these two settings. Our problem includes the entrywise private PCA setting in (Hardt and Roth, 2013; Hardt and Price, 2014) and distributed PCA setting in (Balcan et al., 2014) as special cases and we demonstrate improved bounds over existing results for both problems.

We also apply our results to the memory-efficient streaming PCA problem considered in (Hardt and Price, 2014; Li et al., 2016; Mitliagkas et al., 2013), where data points arrive in streams and the algorithm is only allowed to use memory proportional to the size of the final output. Built upon our new analysis of the noisy power method we improve state-of-the-art sample complexity bounds obtained in (Hardt and Price, 2014).

Proof techniques The noisy power method poses unique challenges for an improved gap dependency analysis. In the analysis of (Hardt and Price, 2014) the largest principal angle between $\mathbf{X}_\ell$ and $\mathbf{U}_k$ is considered for every iteration ℓ. However, such analysis cannot possibly remove the dependency over $(\sigma_k - \sigma_{k-1})$, as we discuss in Sec. 2.1. To overcome such difficulties, we propose in Eq. (3) a novel characterization between a rank-p subspace $\mathbf{X}_\ell$ and the rank-k target space $\mathbf{U}_k$ through an intermediate subspace $\mathbf{U}_q$, which we name as *rank-k perturbation on $\mathbf{U}_q$ by $\mathbf{X}_\ell$*. This quantity does not correspond to any principal angle between linear subspaces when $p > k$. Built upon the shrinkage behavior of the proposed quantity across iterations, we are able to obtain improved gap dependency for the noisy power method. We hope our proof could shed light to the analysis of an even broader family of numerical linear algebra algorithms that involve noisy power iterations.

1.2. Setup

For a $d \times d$ positive semi-definite matrix $\mathbf{A}$, we denote $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$ as its eigen-decomposition, where $\mathbf{U}$ is an orthogonal $d \times d$ matrix and $\mathbf{\Sigma} = \mathbf{diag}(\sigma_1, \cdots, \sigma_d)$ is a $d \times d$ diagonal matrix consisting eigenvalues of $\mathbf{A}$, sorted in descending order: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d \geq 0$. The spectral norm $\|\mathbf{A}\|_2$ and Frobenious norm $\|\mathbf{A}\|_F$ can then be expressed as $\|\mathbf{A}\|_2 = \sigma_1$ and $\|\mathbf{A}\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_d^2}$. For an integer $k \in [d]$, we define $\mathbf{U}_k$ as a $d \times k$ matrix with orthonormal columns, whose column space corresponds to the top-k eigen-space of $\mathbf{A}$. Similarly, $\mathbf{\Sigma}_k = \mathbf{diag}(\sigma_1, \cdots, \sigma_k)$ corresponds to the top-k eigenvalues of $\mathbf{A}$. Let $\mathbf{A}_k \in \operatorname{argmin}_{\mathbf{B}:\mathrm{rank}(\mathbf{B})\leq k} \|\mathbf{A} - \mathbf{B}\|_\xi$ be the optimal rank-k approximation of $\mathbf{A}$. It is well-known that $\mathbf{A}_k = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{U}_k^\top$ is the optimal approximation for both spectral norm ($\xi = 2$) and Frobenious norm ($\xi = F$) (Eckart and Young, 1936).

QR Factorization is a process to obtain an orthonormal column basis of a matrix. For a $d \times p$ matrix $\mathbf{Y}$, QR factorization gives us $\mathbf{Y} = \mathbf{X}\mathbf{R}$ where $\mathbf{X} \in \mathbb{R}^{d\times p}$ is orthonormal and $\mathbf{R} \in \mathbb{R}^{p\times p}$ is an upper triangular matrix (Trefethen and Bau III, 1997).

2. An improved analysis of the noisy power method

The noisy power method is described in Algorithm 1. (Hardt and Price, 2014) provides the first general-purpose analysis of the convergence rate and noise tolerance of Algorithm 1. We cite their main theoretical result below:

Theorem 2.1 (Hardt and Price (2014)) *Fix $\epsilon \in (0, 1/2)$ and let $k \leq p$. Let $\mathbf{U}_k \in \mathbb{R}^{d\times k}$ be the top-k eigenvectors of a positive semi-definite matrix $\mathbf{A}$ and let $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$ denote its*

Algorithm 1: The noisy matrix power method
Data: positive semi-definite data matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, target rank k, iteration rank $p \geq k$, number of iterations L.
Result: approximated eigen-space $\mathbf{X}_L \in \mathbb{R}^{d\times p}$, with orthonormal columns.
Initialization: orthonormal $\mathbf{X}_0 \in \mathbb{R}^{d\times p}$ by QR decomposition on a random Gaussian matrix $\mathbf{G}_0$;
for $\ell = 1$ *to* L **do**
  Observe $\mathbf{Y}_\ell = \mathbf{A}\mathbf{X}_{\ell-1} + \mathbf{G}_\ell$ for some noise matrix $\mathbf{G}_\ell$;
  QR factorization: $\mathbf{Y}_\ell = \mathbf{X}_\ell \mathbf{R}_\ell$, where $\mathbf{X}_\ell$ consists of orthonormal columns;
end

eigenvalues. Suppose at every iteration of the noisy power method the noise matrix $\mathbf{G}_\ell$ *satisfies*

$$5\|\mathbf{G}_\ell\|_2 \leq \epsilon(\sigma_k - \sigma_{k+1}) \quad \textit{and} \quad 5\|\mathbf{U}_k^\top \mathbf{G}_\ell\|_2 \leq (\sigma_k - \sigma_{k+1})\frac{\sqrt{p} - \sqrt{k-1}}{\tau\sqrt{d}}$$

for some fixed constant τ*. Assume in addition that the number of iterations* L *is lower bounded as*

$$L = \Omega\left(\frac{\sigma_k}{\sigma_k - \sigma_{k+1}} \log\left(\frac{d\tau}{\epsilon}\right)\right).$$

Then with probability at least $1 - \tau^{-\Omega(p+1-k)} - e^{-\Omega(d)}$ *we have* $\|(\mathbf{I} - \mathbf{X}_L\mathbf{X}_L^\top)\mathbf{U}_k\|_2 \leq \epsilon$.

Theorem 2.1 has one major drawback: both bounds for noise tolerance and convergence rate depend crucially on the "small" singular value gap $(\sigma_k - \sigma_{k+1})$. This gap could be extremely small for most data matrices in practice since it concerns the difference between two *consecutive* singular values. We show in later paragraphs an example where such gap-dependency could lead to significant deterioration in terms of both error tolerance and computing. A perhaps even more disappointing fact is that the dependency over $(\sigma_k - \sigma_{k+1})$ cannot be improved under the existing analytical framework by increasing p, the number of components maintained by $\mathbf{X}_\ell$ at each iteration. On the other hand, one expects the noisy power method to be more robust to per-iteration noise when p is much larger than k. This intuition has been formally established in (Gu, 2015) under the noiseless setting and was also articulated as a conjecture in (Hardt and Price, 2014):

Conjecture 2.1 (Hardt and Price (2014)) *The noise tolerance terms in Theorem 2.1 can be improved to*

$$5\|\mathbf{G}_\ell\|_2 \leq \epsilon(\sigma_k - \sigma_{p+1}) \quad \textit{and} \quad 5\|\mathbf{U}_k^\top \mathbf{G}_\ell\|_2 \leq \frac{\sqrt{p} - \sqrt{k-1}}{\tau\sqrt{d}}. \tag{1}$$

In this section, we provide a more refined theoretical analysis of the noisy matrix power method presented in Algorithm 1. Our analysis significantly improves the gap dependency over existing results in Theorem 2.1 and partially solves Conjecture 2.1 up to additional constant-level dependencies:

Theorem 2.2 (Improved gap-dependent bounds for noisy power method) *Let* $k \leq q \leq p$*. Let* $\mathbf{U}_q \in \mathbb{R}^{d\times q}$ *be the top-*q *eigenvectors of a positive semi-definite matrix* $\mathbf{A}$ *and let* $\sigma_1 \geq \cdots \geq \sigma_d \geq 0$ *denote its eigenvalues and fix any* ϵ *between* 0 *and* $O\left(\frac{\sigma_q}{\sigma_k} \cdot \min\left\{\frac{1}{\log\left(\frac{\sigma_k}{\sigma_q}\right)}, \frac{1}{\log(\tau d)}\right\}\right)$*. Suppose at*

every iteration of the noisy power method the noise matrix $\mathbf{G}_\ell$ *satisfies*

$$\|\mathbf{G}_\ell\|_2 = O\left(\epsilon(\sigma_k - \sigma_{q+1})\right) \quad \textit{and} \quad \|\mathbf{U}_q^\top \mathbf{G}_\ell\|_2 = O\left(\epsilon\left(\sigma_k - \sigma_{q+1}\right)\frac{\sqrt{p}-\sqrt{q-1}}{\tau\sqrt{d}}\right)$$

for some constant $\tau > 0$*. Then after*

$$L = \Theta\left(\frac{\sigma_k}{\sigma_k - \sigma_{q+1}}\log\left(\frac{\tau d}{\epsilon}\right)\right).$$

iterations, with probability at least $1 - \tau^{-\Omega(p+1-q)} - e^{-\Omega(d)}$*, we have*

$$\|(\mathbf{I} - \mathbf{X}_L\mathbf{X}_L^\top)\mathbf{U}_k\|_2 \leq \epsilon.$$

Furthermore,

$$\left\|\mathbf{A} - \mathbf{X}_L\mathbf{X}_L^\top\mathbf{A}\right\|_2^2 \leq \sigma_{k+1}^2 + \epsilon^2\sigma_k^2$$

$$\left\|\mathbf{A} - \mathbf{X}_L\mathbf{X}_L^\top\mathbf{A}\right\|_F^2 \leq \sum_{i=k+1}^{d}\sigma_i^2 + k\epsilon^2\sigma_k^2$$

Discussion Compared to existing bounds in Theorem 2.1, the noise tolerance as well as convergence rate of noisy power method is significantly improved in Theorem 2.2, where the main gap-dependent term $(\sigma_k - \sigma_{k+1})$ is improved to $(\sigma_k - \sigma_{q+1})$ for some intermediate singular value σ_q with $k \leq q \leq p$. Since the singular values are non-increasing, setting a large value of q in Theorem 2.2 would improve the bounds. However, q cannot be too close to p due to the presence of a $(\sqrt{p} - \sqrt{q-1})$ term. In addition, the convergence rate (i.e., bound on L) specified in Theorem 2.2 reproduces recent results in (Gu, 2015) for noisy power method under noiseless settings ($\mathbf{G}_\ell = \mathbf{0}$). There are three main differences between our theorems and the conjecture raised by (Hardt and Price, 2014). First, the strength of projected noise $\mathbf{U}_q^\top\mathbf{G}$ also depends on ϵ. However, in many applications, this assumption is implied by the $\|\mathbf{G}_\ell\|_2 = O\left(\epsilon(\sigma_k - \sigma_{q+1})\right)$ assumption. Second, we have $\left(\sqrt{p} - \sqrt{q-1}\right)$ instead of $\left(\sqrt{p} - \sqrt{k-1}\right)$ dependence. When $q = \Theta(k)$ and $p \geq 2q$, then this term is the at the same order as in the conjecture. Lastly, we notice that the second term of (1) is totally independent of σ_k, σ_{p+1} and their gap, which seems to be either a typo or unattainable result. Nonetheless, Theorem 2.2 has shown significant improvement on Theorem 2.1.

To further shed light on the nature of our obtained results, we consider the following example to get a more interpretable comparison between Theorem 2.2 and 2.1:

Example: power-law decaying spectrum We consider the example where the spectrum of the input data matrix $\mathbf{A}$ has *power-law* decay; that is, $\sigma_k \asymp k^{-\alpha}$ for some parameter $\alpha > 1$. Many data matrices that arise in practical data applications have such spectral decay property (Liu et al., 2015). The small eigengap $(\sigma_k - \sigma_{k+1})$ is on the order of $k^{-\alpha-1}$. As a result, the number of iterations L should be at least $\Omega(k\log(d/\epsilon))$, which implies a total running time of $O(dk^3\log(d/\epsilon))$. On the other hand, by setting $q = ck$ for some constant $c > 1$ the "large" spectral gap $(\sigma_k - \sigma_{q+1})$ is on the order of $k^{-\alpha}$. Consequently, the number of iterations L under the new theoretical analysis only needs to scale as $\Omega(\log(d/\epsilon))$ and the total number of flops is $O(dk^2\log(d/\epsilon))$. This is an $O(k)$ improvement over existing bounds for noisy power method.

Apart from convergence rates, our new analysis also improves the noise tolerance (i.e., bounds on $\|\mathbf{G}_\ell\|_2$) in an explicit way when the data matrix $\mathbf{A}$ is assumed to have power-law spectral decay. More specifically, old results in (Hardt and Price, 2014) requires the magnitude of the noise matrix $\|\mathbf{G}_\ell\|_2$ to be upper bounded by $O(\epsilon k^{-\alpha-1})$, while under the new analysis (Theorem 2.2) a bound of the form $\|\mathbf{G}_\ell\|_2 = O(\epsilon k^{-\alpha})$ suffices, provided that $q = ck$ for some constant $c > 1$ and ϵ is small. This is another $O(k)$ improvement in terms of bounds on the maximum tolerable amount of per-iteration noise.

2.1. Proof of Theorem 2.2

Before presenting our proof of the main theorem (Theorem 2.2), we first review the arguments in (Hardt and Price, 2014) and explain why straightforward adaptations of their analysis cannot lead to improved gap dependency. (Hardt and Price, 2014) considered the tangent of the kth principle angle between $\mathbf{U}_k$ and $\mathbf{X}_\ell$:

$$\tan\theta_k(\mathbf{U}_k, \mathbf{X}_\ell) = \left\|(\mathbf{U}_{d-k}^\top \mathbf{X}_\ell)(\mathbf{U}_k^\top \mathbf{X}_\ell)^\dagger\right\|_2, \tag{2}$$

where $\mathbf{U}_{d-k} \in \mathbb{R}^{d\times(d-k)}$ is the orthogonal complement of the top-k eigen-space $\mathbf{U}_k \in \mathbb{R}^{d\times k}$ of $\mathbf{A}$. It can then be shown that when both $\|\mathbf{G}_\ell\|_2$ and $\|\mathbf{U}_k^\top \mathbf{G}_\ell\|_2$ are properly bounded, the angle geometrically shrinks after each power iteration; that is, $\tan\theta_k(\mathbf{U}_k, \mathbf{X}_{\ell+1}) \le \rho \tan\theta_k(\mathbf{U}_k, \mathbf{X}_\ell)$ for some fixed $\rho \in (0,1)$. However, as pointed outed by (Hardt and Price, 2014), this geometric shrinkage might not hold with larger level of noise.

To overcome such difficulties, in our analysis we consider a different characterization between $\mathbf{U}_k$ (or $\mathbf{U}_q$) and $\mathbf{X}_\ell$ at each iteration. Let $\mathbf{U}_k \in \mathbb{R}^{d\times k}$, $\mathbf{U}_q \in \mathbb{R}^{d\times q}$ be the top k and top q eigenvectors of $\mathbf{X}$ and let $\mathbf{U}_{d-q} \in \mathbb{R}^{d\times(d-p)}$ be the remaining eigenvectors. For an orthonormal matrix $\mathbf{X}_\ell \in \mathbb{R}^{d\times p}$, define the *rank-$k$ perturbation on $\mathbf{U}_q$ by $\mathbf{X}_\ell$* as

$$h_\ell := \left\|(\mathbf{U}_{d-q}^\top \mathbf{X}_\ell)(\mathbf{U}_q^\top \mathbf{X}_\ell)^\dagger \begin{pmatrix} \mathbf{I}_{k\times k} \\ \mathbf{0} \end{pmatrix}\right\|_2. \tag{3}$$

Our definition of h_ℓ is motivated by Ming Gu's recent analysis on improved gap dependency of exact (noiseless) power method (Gu, 2015), which considered $\widetilde{\mathbf{H}}_\ell = \mathbf{\Sigma}_{d-q}^\ell \left(\mathbf{U}_{d-q}^\top \mathbf{X}_0\right)\left(\mathbf{U}_q^\top \mathbf{X}_0\right)^\dagger \begin{pmatrix} \mathbf{\Sigma}_k^{-\ell} \\ \mathbf{0} \end{pmatrix}$ as the reference matrix for their analysis. Here $\mathbf{X}_0$ is the initial matrix and ℓ is the number of iterations. Compared to the classical quantity $(\mathbf{U}_{d-k}^\top \mathbf{X}_0)(\mathbf{U}_k^\top \mathbf{X}_0)^\dagger$ in Eq. (3), $\widetilde{\mathbf{H}}_\ell$ consists of the enlarged top-q eigenspace and have the singular value matrices $\mathbf{\Sigma}_{d-q}^\ell$ and $\mathbf{\Sigma}_k^{-\ell}$ multiplied on both sides of the quantity. By analyzing properties of $\widetilde{\mathbf{H}}_\ell$, Gu (2015) demonstrated enlarged spectral gap $(\sigma_k - \sigma_{q+1})$ in power iteration convergence. However, $\widetilde{\mathbf{H}}_\ell$ is defined over the intial test matrix $\mathbf{X}_0$ and thus cannot handle a large amount of noise across power iterations. To adapt the analysis in Gu (2015) to *noisy* power method, we consider a variant of $\widetilde{\mathbf{H}}_\ell$: $\mathbf{H}_\ell = \left(\mathbf{U}_{d-q}^\top \mathbf{X}_\ell\right)\left(\mathbf{U}_q^\top \mathbf{X}_\ell\right)^\dagger \begin{pmatrix} \mathbf{I}_{k\times k} \\ \mathbf{0} \end{pmatrix}$, where the $\mathbf{\Sigma}_{d-q}$ and $\mathbf{\Sigma}_k$ terms are removed and $\mathbf{X}_0$ is replaced with $\mathbf{X}_\ell$. Because $\mathbf{H}_\ell$ is based on the possibly noisy test matrix $\mathbf{X}_\ell$ after ℓ iterations, it automatically adjusts itself towards the presence of noise across power iterations and thus leads to relaxed spectral gap bound for noisy power method. Note also that $\mathbf{H}_\ell$ reduces to $\widetilde{\mathbf{H}}_\ell$ when exact (noiseless) power iterations $\mathbf{X}_\ell = \mathbf{A}^\ell \mathbf{X}_0$ is carried out.

We can then show the following shrinkage results for h_ℓ across iterations:

Lemma 2.1 *If the noise matrix at each iteration satisfies*

$$\|\mathbf{G}_\ell\|_2 \leq c\epsilon\left(\sigma_k - \sigma_{q+1}\right), \quad \left\|\mathbf{U}_q^\top \mathbf{G}_\ell\right\|_2 \leq c \cdot \min\{\epsilon\left(\sigma_k - \sigma_{q+1}\right)\cos\theta_q(\mathbf{U}_q, \mathbf{X}_\ell), \sigma_q \cos\theta_q(\mathbf{U}_q, \mathbf{X}_\ell)\},$$

for some sufficiently small absolute constant $0 < c < 1$*, define*

$$\rho := \frac{\sigma_{q+1} + C\epsilon\left(\sigma_k - \sigma_{q+1}\right)}{\sigma_k}.$$

we then have

$$h_{\ell+1} - \frac{C\epsilon\left(\sigma_k - \sigma_{q+1}\right)}{\left(1-\rho\right)\sigma_k} \leq \rho\left(h_\ell - \frac{C\epsilon\left(\sigma_k - \sigma_{q+1}\right)}{\left(1-\rho\right)\sigma_k}\right),$$

for some sufficiently small global constant $0 < C < 1$.

The following lemma bounds the rank-k perturbation on $\mathbf{U}_q$ by $\mathbf{X}_0$ when it is initialized via QR decomposition on a random Gaussian matrix $\mathbf{G}_0$, as described in Algorithm 1.

Lemma 2.2 *With all but* $\tau^{-\Omega(p+1-q)} + e^{-\Omega(d)}$ *probability, we have that*

$$h_0 \leq \tan\theta_q(\mathbf{U}_q, \mathbf{X}_0) \leq \frac{\tau\sqrt{d}}{\sqrt{p} - \sqrt{q-1}}.$$

Finally, Lemma 2.3 shows that small h_L values imply small angles between $\mathbf{X}_L$ and $\mathbf{U}_k$.

Lemma 2.3 *For any* $\epsilon \in (0,1)$*, if* $h_L \leq \epsilon/4$ *then* $\tan\theta_k(\mathbf{U}_k, \mathbf{X}_L) \leq \epsilon$.

The proofs of Lemma 2.1, 2.2 and 2.3 involve some fairly technical matrix computations and is thus deferred to Appendix A. We are now ready to prove Theorem 2.2:

Proof [**Theorem** 2.2] First, the chosen ϵ ensures Corollary A.1 in Appendix A holds, therefore, the noise conditions in Theorem 2.2 imply those noise conditions in Lemma 2.1 with high probability. As a result, the following holds for all $\ell \in [L]$:

$$h_{\ell+1} - \frac{C\epsilon\left(\sigma_k - \sigma_{q+1}\right)}{\left(1-\rho\right)\sigma_k} \leq \rho\left(h_\ell - \frac{C\epsilon\left(\sigma_k - \sigma_{q+1}\right)}{\left(1-\rho\right)\sigma_k}\right), \tag{4}$$

where $\rho = \frac{\sigma_{q+1} + C\epsilon(\sigma_k - \sigma_{q+1})}{\sigma_k}$ and C is an absolute constant. Define $g_\ell := h_\ell - \frac{C\epsilon(\sigma_k - \sigma_{q+1})}{(1-\rho)\sigma_k}$. Eq. (4) is then equivalent to $g_{\ell+1} \leq \rho g_\ell$. In addition, Lemma 2.2 yields

$$g_0 \leq h_0 \leq \frac{\tau\sqrt{d}}{\sqrt{p} - \sqrt{q-1}}$$

with high probability. Consequently, with $L = O(\log(g_0/\epsilon)/\log(1/\rho))$ iterations we have $g_L \leq \epsilon/2$. h_L can then be bounded by

$$h_L = g_L + \frac{C\epsilon(\sigma_k - \sigma_{q+1})}{(1-\rho)\sigma_k} = \frac{\epsilon}{2} + \frac{C\epsilon\left(\sigma_k - \sigma_{q+1}\right)}{\sigma_k} \cdot \frac{\sigma_k}{\sigma_k - \sigma_{q+1} - C\epsilon\left(\sigma_k - \sigma_{q+1}\right)} \leq \epsilon.$$

Subsequently, invoking Lemma 2.3 we get $\|(\mathbf{I} - \mathbf{X}_L\mathbf{X}_L^\top)\mathbf{U}_k\|_2 = \sin\theta_k(\mathbf{U}_k, \mathbf{X}_\ell) \leq \tan\theta_k(\mathbf{U}_k, \mathbf{X}_\ell) \leq 8\epsilon = O(\epsilon)$, where we adopt the definition of $\sin\theta_k(\mathbf{U}_k, \mathbf{X}_\ell)$ from (Hardt and Price, 2014). By

Lemma A.5 and A.6 we also obtain the reconstruction error bounds. The constant in $O(\epsilon)$ can be absorbed into the bounds of $\mathbf{G}_\ell$ and L.

We next simplify the bound $L = O(\log(g_0/\epsilon)/\log(1/\rho))$. We first upper bound the shrinkage parameter ρ as follows:

$$\begin{aligned}\rho =& \frac{\sigma_{q+1}}{\sigma_k} + \frac{C\left(\sigma_k - \sigma_{q+1}\right)\epsilon}{\sigma_k} \leq \frac{\sigma_{q+1} + \left(\sigma_k - \sigma_{q+1}\right)\epsilon/4}{\sigma_{q+1} + \left(\sigma_k - \sigma_{q+1}\right)/2} \\ =& \frac{\sigma_{q+1} + \left(\sigma_k - \sigma_{q+1}\right)/4}{\sigma_{q+1} + \left(\sigma_k - \sigma_{q+1}\right)/2} \cdot \frac{\sigma_{q+1}}{\sigma_{q+1} + \left(\sigma_k - \sigma_{q+1}\right)/4} + \frac{\left(\sigma_k - \sigma_{q+1}\right)/4}{\sigma_{q+1} + \left(\sigma_k - \sigma_{q+1}\right)/2} \cdot \epsilon \\ \leq& \max\left(\frac{\sigma_{q+1}}{\sigma_{q+1} + \left(\sigma_k - \sigma_{q+1}\right)/4}, \epsilon\right),\end{aligned}$$

where the last inequality is due to that weighted mean is no larger than the maximum of two terms. Then we further have

$$\begin{aligned}\log(1/\rho) \geq& \log\left[\min\left(\frac{\sigma_{q+1} + (\sigma_k - \sigma_{q+1})/4}{\sigma_{q+1}}, \frac{1}{\epsilon}\right)\right] \geq \min\left(\log\frac{\sigma_k + 3\sigma_{q+1}}{4\sigma_{q+1}}, 1\right) \\ \geq& \min\left(1 - \frac{4\sigma_{q+1}}{\sigma_k + 3\sigma_{q+1}}, 1\right) = 1 - \frac{4\sigma_{q+1}}{\sigma_k + 3\sigma_{q+1}} = \frac{\sigma_k - \sigma_{q+1}}{\sigma_k + 3\sigma_{q+1}}\end{aligned}$$

where the last inequality results from $\log\frac{\sigma_k+3\sigma_{q+1}}{4\sigma_{q+1}} \geq 1 - \frac{4\sigma_{q+1}}{\sigma_k+3\sigma_{q+1}}$. Subsequently, $\log(g_0/\epsilon)/\log(1/\rho)$ can be upper bounded as

$$\frac{\log\left(g_0/\epsilon\right)}{\log\left(1/\rho\right)} = O\left(\frac{\log\left(\tan\theta_q\left(\mathbf{U}_q, \mathbf{X}_0\right)/\epsilon\right)}{(\sigma_k - \sigma_{q+1})/(\sigma_k + 3\sigma_{q+1})}\right) = O\left(\frac{\sigma_k}{\sigma_k - \sigma_{q+1}}\log\left(\frac{\tau d}{\epsilon}\right)\right),$$

where we use the fact that $g_0 \leq h_0 \leq \tan\theta_q\left(\mathbf{U}_q, \mathbf{X}_0\right)$ and the term $3\sigma_{q+1}$ is absorbed to σ_k. ■

2.2. Gap-independent bounds

We lead a slight astray here to consider *gap-independent* bounds for the noisy power method, which is a straightforward application of our derived gap-dependent bounds in Theorem 2.2. It is clear that the angle $\sin\theta_k(\mathbf{U}_k, \mathbf{X}_L) = \|(\mathbf{I} - \mathbf{X}_L\mathbf{X}_L^\top)\mathbf{U}_k\|_2$ cannot be gap-free, because the top-k eigen-space $\mathbf{U}_k$ is ill-defined when the spectral gap $(\sigma_k - \sigma_{k+1})$ or $(\sigma_k - \sigma_{q+1})$ is small. On the other hand, it is possible to derive gap-independent bounds for the approximation error $\|\mathbf{A} - \mathbf{X}_L\mathbf{X}_L^\top\mathbf{A}\|_2$ because $\mathbf{X}_L$ does not need to be close to $\mathbf{U}_k$ to achieve good approximation of the original data matrix $\mathbf{A}$. This motivates Hardt and Price to present the following conjecture on gap-independent bounds of noisy power method:

Conjecture 2.2 (Hardt and Price (2014)) [2] *Fix $\epsilon \in (0, 1)$, $p \geq 2k$ and suppose $\mathbf{G}_\ell$ satisfies*

$$\|\mathbf{G}_\ell\|_2 = O(\epsilon\sigma_{k+1}), \quad \|\mathbf{U}_k^\top\mathbf{G}_\ell\|_2 = O\left(\epsilon\sigma_{k+1}\sqrt{k/d}\right) \tag{5}$$

for all iterations $\ell = 1, \cdots, L$. Then with high probability, after $L = O(\frac{\log d}{\epsilon})$ iterations we have

$$\|\mathbf{A} - \mathbf{X}_L\mathbf{X}_L^\top\mathbf{A}\|_2 \leq (1 + O(\epsilon))\|\mathbf{A} - \mathbf{A}_k\|_2 = (1 + O(\epsilon))\sigma_{k+1}.$$

2. We rephrase the original conjecture to make ϵ not scale with singular values.

Built upon the gap-dependent bound we derived in the previous section and a recent technique introduced in (Musco and Musco, 2015) for the analysis of block Lanczos methods, we are able to prove the following theorem that partially solves Conjecture 2.2.

Theorem 2.3 *Fix* $0 < \epsilon < 1$ *and suppose the noise matrix satisfies*

$$\|\mathbf{G}_\ell\|_2 = O\left(\epsilon^2 \sigma_{k+1}\right) \qquad \text{and} \qquad \left\|\mathbf{U}_k^\top \mathbf{G}_\ell\right\|_2 = O\left(\frac{\epsilon^2\left(\sqrt{p}-\sqrt{k-1}\right)\sigma_{k+1}}{\tau\sqrt{d}}\right)$$

for some constant $\tau > 0$*. Then after*

$$L = \Theta\left(\frac{1}{\epsilon}\log\left(\frac{\tau d}{\epsilon}\right)\right)$$

iterations, with probability at least $1 - \tau^{-\Omega(p+1-q)} - e^{-\Omega(d)}$*, we have*

$$\left\|\mathbf{A} - \mathbf{X}_L\mathbf{X}_L^\top\mathbf{A}\right\|_2 \leq (1+\epsilon)\left\|\mathbf{A}-\mathbf{A}_k\right\|_2 = (1+\epsilon)\sigma_{k+1}.$$

The major difference between Theorem 2.3 and its targeted Conjecture 2.2 is an extra $O(\epsilon)$ term in the noise bound of both $\|\mathbf{G}_\ell\|_2$ and $\|\mathbf{U}_k^\top \mathbf{G}_\ell\|_2$. Whether such a gap can be closed remains an important open question. The main idea of the proof is to find $m = \max_{0\leq i\leq k}\{\sigma_i - \sigma_{k+1} \geq \epsilon\sigma_{k+1}\}$ and apply Theorem 2.2 with m as the new targeted rank and k as the intermediate rank q. A complete proof is deferred to Appendix B. We notice that there are recent works on eigengap independent bound for other numerical methods, such as stochastic gradient decent, which may achieve even better result on specific problem such as low rank least-square problem (Sa et al., 2015) and PCA (Shamir, 2015). However, those analysis could not be applied to the noisy power method framework and thus we deem such studies orthogonal to ours.

3. Application to distributed private PCA

Our main result can readily lead to improvement of several downstream applications, which will be highlighted in the this section and next. Specifically, we will discuss the benefit brought to distributed private PCA setting in this section, and memory-efficient streaming PCA in the next.

3.1. The model

In our distributed private PCA model there are $s \geq 1$ computing nodes, each storing a positive semi-definite $d \times d$ matrix $\mathbf{A}^{(i)}$. $\mathbf{A}^{(i)}$ can be viewed as the sample covariance matrix of data points stored on node i. There is also a central computing node, with no data stored. The objective is to approximately compute the top-k eigen-space $\mathbf{U}_k$ of the aggregated data matrix $\mathbf{A} = \sum_{i=1}^s \mathbf{A}^{(i)}$ without leaking information of each data matrix $\mathbf{A}^{(1)}, \cdots, \mathbf{A}^{(s)}$. Each of the s computing nodes can and only can communicate with the central node via a *public* channel, where all bits communicated are public to the other nodes as well as any malicious party. We are interested in algorithms that meet the following formal guarantees:

Privacy guarantee We adopt the concept of (ε, δ)-differential privacy proposed in (Dwork et al., 2006). Fix privacy parameters $\varepsilon, \delta \in (0, 1)$. Let D be all bits communicated via the public channels between the s computing nodes and the central node. For every $i \in \{1, \cdots, s\}$ and all $\mathbf{A}^{(i)'}$ that differs from $\mathbf{A}^{(i)}$ in at most one entry with absolute difference at most 1, the following holds

$$\Pr\left[D \in \mathcal{D} | \mathbf{A}^{(i)}, \mathbf{A}^{(-i)}\right] \leq e^{\varepsilon} \Pr\left[D \in \mathcal{D} | \mathbf{A}^{(i)'}, \mathbf{A}^{(-i)}\right] + \delta, \tag{6}$$

where $\mathbf{A}^{(-i)} = (\mathbf{A}^{(1)}, \cdots, \mathbf{A}^{(i-1)}, \mathbf{A}^{(i+1)}, \cdots, \mathbf{A}^{(s)})$ and $\mathcal{D}$ is any measurable set of D bits communicated.

Utility guarantee Suppose $\mathbf{X}_L$ is the $d \times p$ dimensional output matrix. It is required that

$$\sin\theta_k(\mathbf{U}_k, \mathbf{X}_L) = \|(\mathbf{I} - \mathbf{X}_L \mathbf{X}_L^\top)\mathbf{U}_k\|_2 \leq \epsilon$$

with probability at least 0.9, where ϵ characterizes the error level and $\mathbf{U}_k$ is the top-k eigen-space of the aggregated data matrix $\mathbf{A} = \mathbf{A}^{(1)} + \cdots + \mathbf{A}^{(s)}$.

Communication guarantee The total amount of bits communicated between the s computing nodes and the central node is constrained. More specifically, we assume only M real numbers can be communicated via the public channels.

The model we considered is very general and reduces to several existing models of private or communication constrained PCA as special cases. Below we give two such examples that were analyzed in prior literature.

Remark 3.1 (Reduction from private PCA) *Setting $s = 1$ in our distributed private PCA model we obtain the private PCA model previously considered in (Hardt and Price, 2014; Hardt and Roth, 2013),* [3] *where neighboring data matrices differ by one* entry *with bounded absolute difference.*

Remark 3.2 (Reduction from distributed PCA) *Setting $\varepsilon \to \infty$ and $\delta = 0$ we obtain the distributed PCA model previously considered in (Balcan et al., 2014), where columns (data points) are split and stored separately on different computing nodes.*

3.2. Algorithm and analysis

We say an algorithm solves the $(\varepsilon, \delta, \epsilon, M)$-distributed private PCA problem if it satisfies all three guarantees mentioned in Sec. 3.1 with corresponding parameters. Algorithm 2 describes the idea of executing the noisy power method with Gaussian noise in a distributed manner.

The following theorem shows that Algorithm 2 solves the $(\varepsilon, \delta, \epsilon, M)$-distributed private PCA problem with detailed characterization of the utility parameter ϵ and communication complexity M. Its proof is deferred to Appendix C.

Theorem 3.1 (Distributed private PCA) *Let s be the number of nodes and $\mathbf{A}^{(1)}, \cdots, \mathbf{A}^{(s)} \in \mathbb{R}^{d \times d}$ be data matrices stored separately on the s nodes. Fix target rank k, intermediate rank $q \geq k$ and iteration rank p with $2q \leq p \leq d$. Suppose the number of iterations L is set as*

3. The $s = 1$ case is actually harder than models considered in (Hardt and Price, 2014; Hardt and Roth, 2013) in that intermediate steps of noisy power method are released to the public as well. However this does not invalidate the analysis of noisy power method based private PCA algorithms because of the privacy composition rule.

Algorithm 2: Distributed private PCA via distributed noisy power method

Data: distributedly stored data matrices $\mathbf{A}^{(1)}, \cdots, \mathbf{A}^{(s)} \in \mathbb{R}^{d\times d}$, number of iterations L, target rank k, iteration rank $p \geq k$, private parameters ε, δ.

Result: approximated eigen-space $\mathbf{X}_L \in \mathbb{R}^{d\times p}$, with orthonormal columns.

Initialization: orthonormal $\mathbf{X}_0 \in \mathbb{R}^{d\times p}$ by QR decomposition on a random Gaussian matrix $\mathbf{G}_0$; noise variance parameter $\nu = 4\varepsilon^{-1}\sqrt{pL\log(1/\delta)}$;

for $\ell = 1$ *to* L **do**

1. The central node broadcasts $\mathbf{X}_{\ell-1}$ to all s computing nodes;
2. Computing node i computes $\mathbf{Y}_\ell^{(i)} = \mathbf{A}^{(i)}\mathbf{X}_{\ell-1} + \mathbf{G}_\ell^{(i)}$ with $\mathbf{G}_\ell^{(i)} \sim \mathcal{N}(0, \|\mathbf{X}_{\ell-1}\|_\infty^2\nu^2)^{d\times p}$ and sends $\mathbf{Y}_\ell^{(i)}$ back to the central node;
3. The central node computes $\mathbf{Y}_\ell = \sum_{i=1}^s \mathbf{Y}_\ell^{(i)}$ and QR factorization $\mathbf{Y}_\ell = \mathbf{X}_\ell\mathbf{R}_\ell$.

end

$L = \Theta(\frac{\sigma_k}{\sigma_k-\sigma_{q+1}}\log(d))$. *Let* $\varepsilon, \delta \in (0,1)$ *be privacy parameters. Then Algorithm 2 solves the* $(\varepsilon, \delta, \epsilon, M)$*-distributed PCA problem with*

$$\epsilon = O\left(\frac{\nu\sqrt{\mu(\mathbf{A})s\log d\log L}}{\sigma_k - \sigma_{q+1}}\right) \quad \text{and} \quad M = O(spdL) = O\left(\frac{\sigma_k}{\sigma_k-\sigma_{q+1}}spd\log d\right).$$

Here assuming conditions in Theorem 2.2 are satisfied, $\nu = \varepsilon^{-1}\sqrt{4pL\log(1/\delta)}$ *and* $\mu(\mathbf{A})$ *is the* incoherence *(Hardt and Roth, 2013) of the aggregate data matrix* $\mathbf{A} = \sum_{i=1}^s \mathbf{A}^{(i)}$*; more specifically,* $\mu(\mathbf{A}) = d\|\mathbf{U}\|_\infty$ *where* $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ *is the eigen-decomposition of* $\mathbf{A}$.

It is somewhat difficult to evaluate the results obtained in Theorem 3.1 because our work, to our knowledge, is the first to consider distributed private PCA with the public channel communication model. Nevertheless, on the two special cases of private PCA in Remark 3.1 and distributed PCA in Remark 3.2, our result does significantly improve existing analysis. More specifically, we have the following two corollaries based on Theorem 3.1 and Theorem 2.2.

Corollary 3.1 (Improved private PCA) *For the case of* $s = 1$ *and* $2p \leq q \leq d$*, Algorithm 2 is* (ε, δ)*-differentially private and* $\mathbf{X}_L$ *satisfies*

$$\|(\mathbf{I} - \mathbf{X}_L\mathbf{X}_L^\top)\mathbf{U}_k\|_2 \leq \epsilon = O\left(\frac{\nu\sqrt{\mu(\mathbf{A})\log d\log L}}{\sigma_k - \sigma_{q+1}}\right)$$

with probability at least 0.9. Here $\mathbf{U}_k$ *is the top-*k *eigen-space of input data matrix* $\mathbf{A} \in \mathbb{R}^{d\times d}$.

Corollary 3.2 (Improved distributed PCA) *Fix error tolerance parameter* $\epsilon \in (0,1)$ *and set* $\nu = 0$*,* $L = \Theta(\frac{\sigma_k}{\sigma_k-\sigma_{q+1}}\log(d/\epsilon))$ *in Algorithm 2. We then have with high probability,*

$$\|(\mathbf{I} - \mathbf{X}_L\mathbf{X}_L^\top)\mathbf{U}_k\|_2 \leq \epsilon.$$

Here $\mathbf{U}_k$ *is the top-*k *eigen-space of the aggregated matrix* $\mathbf{A} = \sum_{i=1}^s \mathbf{A}^{(i)}$.

The proofs of Corollary 3.1 and 3.2 are simple and deferred to Appendix C. We now compare them with existing results in the literature. For private PCA, our bound has better spectral-gap dependency compared to the $O(\frac{\nu\sqrt{\mu(\mathbf{A})\log d\log L}}{\sigma_k-\sigma_{k-1}})$ bound obtained in (Hardt and Price, 2014). For distributed PCA, our bound achieves an *exponential* improvement over the $O(spd/\epsilon)$ communication complexity bound obtained in (Balcan et al., 2014). [4]

4. Application to memory-efficient streaming PCA

In the streaming PCA setting a computing machine receives a stream of samples $\boldsymbol{z}_1, \cdots \boldsymbol{z}_n \in \mathbb{R}^d$ drawn i.i.d from an unknown underlying distribution $\mathcal{D}$. The objective is to compute the leading k eigenvectors of the population covariance matrix $\mathbf{E}_{\boldsymbol{z}\sim\mathcal{D}}[\boldsymbol{z}\boldsymbol{z}^\top]$ with memory space constrained to the output size $O(kd)$. (Mitliagkas et al., 2013) gave an algorithm for this problem based on the noisy power method. Algorithm 3 gives the details.

Algorithm 3: Memory-efficient Streaming PCA (Mitliagkas et al., 2013)

Data: data stream $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n \overset{i.i.d.}{\sim} \mathcal{D}$, target rank k, iteration rank $p \geq k$, number of iterations L.
Result: approximated eigen-space $\mathbf{X}_L \in \mathbb{R}^{d\times p}$, with orthonormal columns.
Initialization: uniformly sampled orthonormal matrix $\mathbf{X}_0 \in \mathbb{R}^{d\times p}$; $T = \lfloor n/L \rfloor$;
for $\ell = 1$ *to* L **do**
  Power update: $\mathbf{Y}_\ell = \mathbf{A}_\ell \mathbf{X}_{\ell-1}$, where $\mathbf{A}_\ell = \sum_{i=(\ell-1)T+1}^{\ell T} \boldsymbol{z}_i \boldsymbol{z}_i^\top$;
  QR factorization: $\mathbf{Y}_\ell = \mathbf{X}_\ell \mathbf{R}_\ell$, where $\mathbf{X}_\ell$ consists of orthonormal columns.
end

(Hardt and Price, 2014) are among the first ones that analyze Algorithm 3 for a broad class of distributions $\mathcal{D}$ based on their analysis of the noisy power method. More specifically, (Hardt and Price, 2014) analyzed a family of distributions that have fast tail decay and proved gap-dependent sample complexity bounds for the memory-efficient streaming PCA algorithm.

Definition 4.1 ((B,p)-round distributions, (Hardt and Price, 2014)) *A distribution $\mathcal{D}$ over $\mathbb{R}^d$ is $(B,p)-$round if for every p-dimension projection $\mathbf{\Pi}$ and all $t \geq 1$, we have that*

$$\max\left\{\Pr_{\boldsymbol{z}\sim\mathcal{D}}\left[\|\boldsymbol{z}\|_2 \geq t\right], \Pr_{\boldsymbol{z}\sim\mathcal{D}}\left[\|\mathbf{\Pi}\boldsymbol{z}\|_2 \geq t\sqrt{Bp/d}\right]\right\} \leq \exp(-t).$$

Theorem 4.1 ((Hardt and Price, 2014)) *Suppose $\mathcal{D}$ is a (B,p)-round distribution over $\mathbb{R}^d$. Let $\sigma_1 \geq \cdots \geq \sigma_d \geq 0$ be the singular values of the population covariance matrix $\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}}[\boldsymbol{z}\boldsymbol{z}^\top]$. If Algorithm 3 is run with $L = \Theta(\frac{\sigma_k}{\sigma_k-\sigma_{k+1}}\log(d/\epsilon))$ and n satisfies* [5]

$$n = \widetilde{\Omega}\left(\frac{\sigma_k B^2 p \log^2 d}{(\sigma_k-\sigma_{k+1})^3 d\epsilon^2}\right),$$

then with probability at least 0.9 we have that $\|(\mathbf{I}-\mathbf{X}_L\mathbf{X}_L^\top)\mathbf{U}_k\|_2 \leq \epsilon$, where $\mathbf{U}_k$ is the top-k eigen-space of $\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}}[\boldsymbol{z}\boldsymbol{z}^\top]$.

4. Lemma 8 of Balcan et al. (2014) gives a communication upper bound that depends on all singular values bigger than k. It is not obvious which bound is better, but in the worst case, their bound is still linear in $\frac{1}{\epsilon}$.
5. In the $\widetilde{\Omega}(\cdot)$ notation we omit poly-logarithmic terms.

Recently, (Li et al., 2016) proposed a modified power method that achieves a logarithmic sample complexity improvement with respect to $1/\epsilon$. Nevertheless, both bounds in (Hardt and Price, 2014) and (Li et al., 2016) depend on the consecutive spectral gap $(\sigma_k - \sigma_{k+1})$, which could be very small for real-world data distributions. We are also aware of some recent results directly on incremental (Balsubramani et al., 2013) or streaming PCA (Jain et al., 2016), whose analysis, however, seem not easy to be extended to the noise power method setting. Built upon our analysis for the noisy power method, we obtain the following result for streaming PCA with improved gap dependencies:

Theorem 4.2 *Fix $k \leq q \leq p \leq d$. Suppose $\mathcal{D}$ is a (B, p)-round distribution over $\mathbb{R}^d$. Let $\sigma_1 \geq \cdots \geq \sigma_d \geq 0$ be the singular values of the population covariance matrix $\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}}[\boldsymbol{z}\boldsymbol{z}^\top]$. If Algorithm 3 is run with $L = \Theta(\frac{\sigma_k}{\sigma_k - \sigma_{q+1}} \log(d/\epsilon))$ and n satisfies*

$$n = \widetilde{\Omega}\left(\frac{\sigma_k B^2 p \log^2 d}{(\sigma_k - \sigma_{q+1})^3 d\epsilon^2}\right),$$

then with probability at least 0.9 we have that $\|(\mathbf{I} - \mathbf{X}_L\mathbf{X}_L^\top)\mathbf{U}_k\|_2 \leq \epsilon$.

Proof Note that Algorithm 3 is a direct application of noisy power method with $\mathbf{G}_\ell = (\mathbf{A} - \mathbf{A}_\ell)\,\mathbf{X}_{\ell-1}$, where $\mathbf{A} = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}}[\boldsymbol{z}\boldsymbol{z}^\top]$ is the covariance matrix of the population distribution of interest. By Lemma 3.5 of (Hardt and Price, 2014), we have that

$$T = \widetilde{\Omega}\left(\frac{B^2 p \log(d)}{\epsilon^2 (\sigma_k - \sigma_{q+1})^2}\right),$$

is sufficient to guarantee that $\mathbf{G}_\ell$ satisfy the conditions in Theorem 2.2 with high probability. Therefore, in total we need $n = LT = \widetilde{\Omega}(\frac{\sigma_k B^2 p \log^2 d}{(\sigma_k - \sigma_{q+1})^3 d\epsilon^2})$ data points. ■

5. Conclusions and Future Work

In this paper we give a novel analysis of spectral gap dependency for noisy power method, which partially solves a conjecture raised in (Hardt and Price, 2014) with additional mild conditions. As a by product, we derive a spectral gap independent bound which partially solved another conjecture in (Hardt and Price, 2014). Furthermore, our analysis directly leads to improved utility guarantees and sample complexity for downstream applications such as distributed PCA, private PCA and streaming PCA problems.

To completely solve the two conjectures in (Hardt and Price, 2014), we need a finer robustness analysis of $\mathbf{U}_{p-k}$ space. (Wang et al., 2015) gave a related analysis, but only for the noiseless case. Potentially, we may define a new function (like Eq. (3) in our case) to characterize the convergence behavior, and show it shrinks multiplicatively at each iteration.

In parallel to power method based algorithms, Krylov iteration is another method shown to converge faster in the noiseless case (Musco and Musco, 2015). It is also interesting to give a noise tolerance analysis for Krylov iteration and apply it to downstream applications.

Acknowledgments

We thank Cameron Musco for pointing out an error in the original proof. We also thank an anonymous reviewer for providing an alternative and elegant proof for Lemma 2.3.

References

Maria-Florina Balcan, Vandana Kanchanapally, Yingyu Liang, and David Woodruff. Improved distributed principal component analysis. In *NIPS*, 2014.

Akshay Balsubramani, Sanjoy Dasgupta, and Yoav Freund. The fast convergence of incremental PCA. In *NIPS*, pages 3174–3182, 2013.

Christos Boutsidis, David Woodruff, and Peilin Zhong. Optimal principal component analysis in distributed and streaming models. *arXiv: 1504.06729*, 2015.

Kamalika Chaudhuri, Anand Sarwate, and Kaushik Sinha. Near-optimal algorithms for differentially private principal components. In *NIPS*, 2012.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, 2006.

Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *STOC*, 2014.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Ming Gu. Subspace iteration randomization and singular value problems. *SIAM Journal on Scientific Computing*, 37(3):A1139–A1173, 2015.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53 (2):217–288, 2011.

Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. In *NIPS*, 2014.

Moritz Hardt and Aaron Roth. Beyond worst-case analysis in private singular vector computation. In *STOC*, 2013.

Prateek Jain, Chi Jin, Sham M. Kakade, Praneeth Netrapalli, and Aaron Sidford. Matching matrix bernstein with little memory: Near-optimal finite sample guarantees for oja's algorithm. *arxiv:1602.06929*, 2016.

Chun-Liang Li, Hsuan-Tien Lin, and Chi-Jen Lu. Rivalry of two families of algorithms for memory-restricted streaming pca. In *AISTATS*, 2016.

Ziqi Liu, Yu-Xiang Wang, and Alex Smola. Fast differentially private matrix factorization. In *RecSys*, 2015.

Ioannis Mitliagkas, Constantine Caramanis, and Prateek Jain. Memory limited, streaming pca. In *Advances in Neural Information Processing Systems*, pages 2886–2894, 2013.

Cameron Musco and Christopher Musco. Stronger approximate singular value decomposition via the block lanczos and power methods. In *NIPS*, 2015.

Christopher De Sa, Christopher Re, and Kunle Olukotun. Global convergence of stochastic gradient descent for some non-convex matrix problems. In *ICML*, pages 2332–2341, 2015.

Ohad Shamir. Convergence of stochastic gradient descent for PCA. *arxiv:1509.09002*, 2015.

Lloyd N Trefethen and David Bau III. *Numerical linear algebra*, volume 50. Siam, 1997.

Shusen Wang, Zhihua Zhang, and Tong Zhang. Improved Analyses of the Randomized Power Method and Block Lanczos Method. *ArXiv e-prints: 1508.06429*, August 2015.

Appendix A. Proofs of technical lemmas in Sec. 2.1

Lemma A.1 (Lemma 2.1) *If the noise matrix at each iteration satisfies*

$$\|\mathbf{G}_\ell\|_2 \leq c\epsilon\left(\sigma_k - \sigma_{q+1}\right), \quad \left\|\mathbf{U}_q^\top \mathbf{G}_\ell\right\|_2 \leq c \cdot \min\{\epsilon\left(\sigma_k - \sigma_{q+1}\right)\cos\theta_q(\mathbf{U}_q, \mathbf{X}_\ell), \sigma_q \cos\theta_q(\mathbf{U}_q, \mathbf{X}_\ell)\},$$

for some sufficiently small absolute constant $0 < c < 1$*, define*

$$\rho := \frac{\sigma_{q+1} + C\epsilon\left(\sigma_k - \sigma_{q+1}\right)}{\sigma_k}.$$

we then have

$$h_{\ell+1} - \frac{C\epsilon\left(\sigma_k - \sigma_{q+1}\right)}{\left(1-\rho\right)\sigma_k} \leq \rho\left(h_\ell - \frac{C\epsilon\left(\sigma_k - \sigma_{q+1}\right)}{\left(1-\rho\right)\sigma_k}\right),$$

for some sufficiently small global constant $0 < C < 1$.

Proof First notice that

$$\mathbf{U}_q^\top\left(\mathbf{A}\mathbf{X}_\ell + \mathbf{G}_\ell\right)\mathbf{R}_{\ell+1}^{-1}\left(\mathbf{R}_{\ell+1}\left(\mathbf{U}_q^\top\left(\mathbf{A}\mathbf{X}_\ell + \mathbf{G}_\ell\right)\right)^\dagger\right) = \mathbf{I}_{q\times q}.$$

Therefore, the pseudo-inverse of $\mathbf{U}_q^\top\left(\mathbf{A}\mathbf{X}_\ell + \mathbf{G}_\ell\right)\mathbf{R}_{\ell+1}^{-1}$ is $\mathbf{R}_{\ell+1}\left(\mathbf{U}_q^\top\left(\mathbf{A}\mathbf{X}_\ell + \mathbf{G}_\ell\right)\right)^\dagger$. We can then write out $h_{\ell+1}$ explicitly:

$$\begin{aligned}
h_{\ell+1} &= \left\|\mathbf{U}_{d-q}^\top \mathbf{X}_{\ell+1}\left(\mathbf{U}_q^\top \mathbf{X}_{\ell+1}\right)^\dagger\begin{pmatrix}\mathbf{I}_{k\times k}\\ \mathbf{0}\end{pmatrix}\right\|_2 \\
&= \left\|\mathbf{U}_{d-q}^\top\left(\mathbf{A}\mathbf{X}_\ell + \mathbf{G}_\ell\right)\mathbf{R}_{\ell+1}^{-1}\left(\mathbf{U}_q^\top\left(\mathbf{A}\mathbf{X}_\ell + \mathbf{G}_\ell\right)\mathbf{R}_{\ell+1}^{-1}\right)^\dagger\begin{pmatrix}\mathbf{I}_{k\times k}\\ \mathbf{0}\end{pmatrix}\right\|_2 \\
&= \left\|\mathbf{U}_{d-q}^\top\left(\mathbf{A}\mathbf{X}_\ell + \mathbf{G}_\ell\right)\left(\mathbf{U}_q^\top\left(\mathbf{A}\mathbf{X}_\ell + \mathbf{G}_\ell\right)\right)^\dagger\begin{pmatrix}\mathbf{I}_{k\times k}\\ \mathbf{0}\end{pmatrix}\right\|_2
\end{aligned}$$

$$= \left\| \left(\mathbf{\Sigma}_{d-q} \mathbf{U}_{d-q}^\top \mathbf{X}_\ell + \mathbf{U}_{d-q}^\top \mathbf{G}_\ell \right) \left(\mathbf{\Sigma}_p \mathbf{U}_q^\top \mathbf{X}_\ell + \mathbf{U}_q \mathbf{G}_\ell \right)^\dagger \begin{pmatrix} \mathbf{I}_{k\times k} \\ \mathbf{0} \end{pmatrix} \right\|_2$$
$$= \left\| \left(\mathbf{\Sigma}_{d-q} \mathbf{U}_{d-q}^\top \mathbf{X}_\ell + \mathbf{U}_{d-q}^\top \mathbf{G}_\ell \right) \left(\mathbf{U}_q^\top \mathbf{X}_\ell + \mathbf{\Sigma}_q^{-1} \mathbf{U}_q \mathbf{G}_\ell \right)^\dagger \begin{pmatrix} \mathbf{\Sigma}_k^{-1} \\ \mathbf{0} \end{pmatrix} \right\|_2.$$

Now we focus on the pseudo-inverse in the above expression. Our analysis relies on the following singular value decomposition (SVD) of $\mathbf{U}_q^\top \mathbf{X}_\ell$:

$$\mathbf{U}_q^\top \mathbf{X}_\ell = \widetilde{\mathbf{U}} \widetilde{\mathbf{\Sigma}} \widetilde{\mathbf{V}}^\top \in \mathbf{R}^{p\times q}.$$

For simplicity, define

$$\widetilde{\mathbf{P}} = \widetilde{\mathbf{U}} \widetilde{\mathbf{\Sigma}}.$$

Subsequently, we have that

$$\mathbf{U}_q^\top \mathbf{X}_\ell = \widetilde{\mathbf{P}} \widetilde{\mathbf{V}}^\top \qquad \text{and} \qquad \mathbf{X}_\ell^\top \mathbf{U}_q \widetilde{\mathbf{P}}^{-\top} = \widetilde{\mathbf{V}}.$$

By definition of pseudo-inverse, we have

$$\left(\mathbf{U}_q^\top \mathbf{X}_\ell + \mathbf{\Sigma}_q^{-1} \mathbf{U}_q^\top \mathbf{G}_\ell \right)^\dagger$$
$$= \left(\mathbf{U}_q^\top \mathbf{X}_\ell + \mathbf{\Sigma}_q^{-1} \mathbf{U}_q^\top \mathbf{G}_\ell \right)^\top \left[\left(\mathbf{U}_q^\top \mathbf{X}_\ell + \mathbf{\Sigma}_q^{-1} \mathbf{U}_q^\top \mathbf{G}_\ell \right) \left(\mathbf{U}_q^\top \mathbf{X}_\ell + \mathbf{\Sigma}_q^{-1} \mathbf{U}_q^\top \mathbf{G}_\ell \right)^\top \right]^{-1}.$$

The inversion in the above expression can be related to our assumptions of noise:

$$\left[\left(\mathbf{U}_q^\top \mathbf{X}_\ell + \mathbf{\Sigma}_q^{-1} \mathbf{U}_q^\top \mathbf{G}_\ell \right) \left(\mathbf{U}_q^\top \mathbf{X}_\ell + \mathbf{\Sigma}_q^{-1} \mathbf{U}_q^\top \mathbf{G}_\ell \right)^\top \right]^{-1}$$
$$= \left[\left(\widetilde{\mathbf{P}} \widetilde{\mathbf{V}}^\top + \mathbf{\Sigma}_q \mathbf{U}_q^\top \mathbf{G}_\ell \right) \left(\widetilde{\mathbf{V}} \widetilde{\mathbf{P}}^\top + \mathbf{G}_\ell^\top \mathbf{U}_q \mathbf{\Sigma}_q^{-1} \right) \right]^{-1}$$
$$= \widetilde{\mathbf{P}}^{-\top} \left[\left(\widetilde{\mathbf{V}}^\top + \widetilde{\mathbf{P}}^{-1} \mathbf{\Sigma}_q^{-1} \mathbf{U}_q^\top \mathbf{G}_\ell \right) \left(\widetilde{\mathbf{V}} + \mathbf{G}_\ell^\top \mathbf{U}_q \mathbf{\Sigma}_q^{-1} \widetilde{\mathbf{P}}^{-\top} \right) \right]^{-1} \widetilde{\mathbf{P}}^{-1}$$
$$= \widetilde{\mathbf{P}}^{-\top} \left[\mathbf{I} + \widetilde{\mathbf{V}}^\top \mathbf{G}_\ell^\top \mathbf{U}_q \mathbf{\Sigma}_q^{-1} \widetilde{\mathbf{P}}^{-\top} + \widetilde{\mathbf{P}}^{-1} \mathbf{\Sigma}_q^{-1} \mathbf{U}_q^\top \mathbf{G}_\ell \widetilde{\mathbf{V}} + \widetilde{\mathbf{P}}^{-1} \mathbf{\Sigma}_q^{-1} \mathbf{U}_q^\top \mathbf{G}_\ell \mathbf{G}_\ell^\top \mathbf{U}_q \mathbf{\Sigma}_q^{-1} \widetilde{\mathbf{P}}^{-\top} \right]^{-1} \widetilde{\mathbf{P}}^{-1}$$
$$= \widetilde{\mathbf{P}}^{-\top} \left(\mathbf{I} - (\mathbf{I} + \mathbf{Y})^{-1} \mathbf{Y} \right) \widetilde{\mathbf{P}}^{-1},$$

where $\mathbf{Y} = \widetilde{\mathbf{V}}^\top \mathbf{G}_\ell^\top \mathbf{U}_q \mathbf{\Sigma}_q^{-1} \widetilde{\mathbf{P}}^{-\top} + \widetilde{\mathbf{P}}^{-1} \mathbf{\Sigma}_q^{-1} \mathbf{U}_q^\top \mathbf{G}_\ell \widetilde{\mathbf{V}} + \widetilde{\mathbf{P}}^{-1} \mathbf{\Sigma}_q^{-1} \mathbf{U}_q^\top \mathbf{G}_\ell \mathbf{G}_\ell^\top \mathbf{U}_q \mathbf{\Sigma}_q^{-1} \widetilde{\mathbf{P}}^{-\top}$ and the last equation is by Woodbury's identity. Based on our noise assumptions, we can bound $\mathbf{Y}$ as

$$\|\mathbf{Y}\|_2 \le 2 \frac{\left\| \mathbf{U}_q^\top \mathbf{G}_\ell \right\|_2}{\sigma_q \sigma_{\min} \left(\mathbf{U}_q^\top \mathbf{X}_\ell \right)} + \frac{\left\| \mathbf{U}_q^\top \mathbf{G}_\ell \right\|_2^2}{\sigma_q^2 \sigma_{\min}^2 \left(\mathbf{U}_q^\top \mathbf{X}_\ell \right)} \le c_1 \min \left\{ \frac{\epsilon \left(\sigma_k - \sigma_{q+1} \right)}{\sigma_q}, 1 \right\}, \tag{7}$$

for some constant $0 < c_1 < 1$. Subsequently, we have that

$$\left\| (\mathbf{I} + \mathbf{Y})^{-1} \mathbf{Y} \right\|_2 \le \frac{\|\mathbf{Y}\|_2}{1 - \|\mathbf{Y}\|_2} \le c_2 \frac{\epsilon \left(\sigma_k - \sigma_{q+1} \right)}{\sigma_q}, \tag{8}$$

for some constant $0 < c_2 < 1$. Applying triangle inequality we obtain upper bounds on $h_{\ell+1}$:

$$h_{\ell+1} \leq \left\| \mathbf{\Sigma}_{d-q} \mathbf{U}_{d-q}^\top \mathbf{X}_\ell \left(\mathbf{U}_q^\top \mathbf{X}_\ell + \mathbf{\Sigma}_q^{-1} \mathbf{U}_q^\top \mathbf{G}_\ell \right)^\dagger \begin{pmatrix} \mathbf{\Sigma}_k^{-1} \\ \mathbf{0} \end{pmatrix} \right\|_2 + \left\| \mathbf{U}_{d-q} \mathbf{G}_\ell \left(\mathbf{U}_q^\top \mathbf{X}_\ell + \mathbf{\Sigma}_q^{-1} \mathbf{U}_q^\top \mathbf{G}_\ell \right)^\dagger \begin{pmatrix} \mathbf{\Sigma}_k^{-1} \\ \mathbf{0} \end{pmatrix} \right\|_2.$$

We next bound the two terms in the right-hand side of the above inequality separately. For the first term, we have that

$$\begin{aligned}
&\left\| \mathbf{\Sigma}_{d-q} \mathbf{U}_{d-q}^\top \mathbf{X}_\ell \left(\mathbf{U}_q^\top \mathbf{X}_\ell + \mathbf{\Sigma}_q^{-1} \mathbf{U}_q^\top \mathbf{G}_\ell \right)^\dagger \begin{pmatrix} \mathbf{\Sigma}_k^{-1} \\ \mathbf{0} \end{pmatrix} \right\|_2 \\
=&\left\| \mathbf{\Sigma}_{d-q} \mathbf{U}_{d-q}^\top \mathbf{X}_\ell \left[\left(\mathbf{U}_q^\top \mathbf{X}_\ell \right)^\dagger + \mathbf{G}_\ell^\top \mathbf{U}_q \mathbf{\Sigma}_q^{-1} \widetilde{\mathbf{P}}^{-\top} \widetilde{\mathbf{P}}^{-1} \right.\right. \\
&\left.\left. + \left(\mathbf{U}_q^\top \mathbf{X}_\ell \right)^\top \widetilde{\mathbf{P}}^{-\top} \left(\mathbf{I} + \mathbf{Y} \right)^{-1} \mathbf{Y} \widetilde{\mathbf{P}}^{-1} + \mathbf{G}_\ell^\top \mathbf{U}_q \mathbf{\Sigma}_q \widetilde{\mathbf{P}}^{-\top} \left(\mathbf{I} + \mathbf{Y} \right)^{-1} \mathbf{Y} \widetilde{\mathbf{P}}^{-1} \begin{pmatrix} \mathbf{\Sigma}_k^{-1} \\ \mathbf{0} \end{pmatrix} \right] \right\|_2 \\
\leq&\frac{1}{\sigma_k} \left(\sigma_{q+1} h_\ell + \frac{c_1 \sigma_{q+1} \epsilon \left(\sigma_k - \sigma_{q+1} \right)}{\sigma_q} \left(1 + h_\ell \right) + \frac{c_2 \sigma_{q+1} \epsilon \left(\sigma_k - \sigma_{q+1} \right)}{\sigma_q} \left(1 + h_\ell \right) \right. \\
&\left. + \frac{c_1 \sigma_{q+1} \epsilon \left(\sigma_k - \sigma_{q+1} \right)}{\sigma_q} \frac{c_2 \epsilon \left(\sigma_k - \sigma_{q+1} \right)}{\sigma_q} \left(1 + h_\ell \right) \right) \\
\leq&\frac{\sigma_{q+1} + c_4 \epsilon \left(\sigma_k - \sigma_{q+1} \right)}{\sigma_k} h_\ell + \frac{c_4 \epsilon \left(\sigma_k - \sigma_{q+1} \right)}{\sigma_k},
\end{aligned}$$

for some constant $0 < c_4 < 1$. Here the second inequality is due to Eq. (7,8) and Lemma A.2, Similarly, for the second term related to $\mathbf{U}_{d-q} \mathbf{G}_\ell$ we have that

$$\left\| \mathbf{U}_{d-q} \mathbf{G}_\ell \left(\mathbf{U}_q^\top \mathbf{X}_\ell + \mathbf{\Sigma}_q^{-1} \mathbf{U}_q^\top \mathbf{G}_\ell \right)^\dagger \begin{pmatrix} \mathbf{I}_{k \times k} \\ \mathbf{0} \end{pmatrix} \right\|_2 \leq \frac{c_5 \epsilon \left(\sigma_k - \sigma_{q+1} \right)}{\sigma_k} h_\ell + \frac{c_5 \epsilon \left(\sigma_k - \sigma_{q+1} \right)}{\sigma_k},$$

for some constant $0 < c_5 < 1$. Merging these two bounds we arrive at our desired result. ■

Lemma A.2

$$\left\| \widetilde{\mathbf{P}}^{-1} \begin{pmatrix} \mathbf{I}_{k \times k} \\ \mathbf{0} \end{pmatrix} \right\|_2 \leq 1 + h_\ell.$$

Proof

$$\begin{aligned}
\left\| \widetilde{\mathbf{P}}^{-1} \begin{pmatrix} \mathbf{I}_{k \times k} \\ \mathbf{0} \end{pmatrix} \right\|_2 &= \left\| \left(\widetilde{\mathbf{U}} \widetilde{\mathbf{\Sigma}} \right)^{-1} \begin{pmatrix} \mathbf{I}_{k \times k} \\ \mathbf{0} \end{pmatrix} \right\|_2 = \left\| \widetilde{\mathbf{\Sigma}}^{-1} \widetilde{\mathbf{U}}^\top \begin{pmatrix} \mathbf{I}_{k \times k} \\ \mathbf{0} \end{pmatrix} \right\|_2 \\
&= \left\| \widetilde{\mathbf{V}}^\top \widetilde{\mathbf{V}} \widetilde{\mathbf{\Sigma}}^{-1} \widetilde{\mathbf{U}} \begin{pmatrix} \mathbf{I}_{k \times k} \\ \mathbf{0} \end{pmatrix} \right\|_2 \leq \left\| \widetilde{\mathbf{V}} \widetilde{\mathbf{\Sigma}}^{-1} \widetilde{\mathbf{U}} \begin{pmatrix} \mathbf{I}_{k \times k} \\ \mathbf{0} \end{pmatrix} \right\|_2 = \left\| \left(\mathbf{U}_q^\top \mathbf{X}_\ell^\top \right)^\dagger \begin{pmatrix} \mathbf{I}_{k \times k} \\ \mathbf{0} \end{pmatrix} \right\|_2 \\
&= \left\| \mathbf{X}_\ell^\top \mathbf{X}_\ell \left(\mathbf{U}_1^\top \mathbf{X}_l \right)^\dagger \begin{pmatrix} \mathbf{I}_{k \times k} \\ \mathbf{0} \end{pmatrix} \right\|_2 \leq \left\| \mathbf{X}_\ell \left(\mathbf{U}_1^\top \mathbf{X}_l \right)^\dagger \begin{pmatrix} \mathbf{I}_{k \times k} \\ \mathbf{0} \end{pmatrix} \right\|_2
\end{aligned}$$

$$= \left\| \left(\mathbf{U}_q \mathbf{U}_q^\top + \mathbf{U}_{d-q} \mathbf{U}_{d-q}^\top \right) \mathbf{X}_\ell \left(\mathbf{U}_1^\top \mathbf{X}_l \right)^\dagger \begin{pmatrix} \mathbf{I}_{k\times k} \\ \mathbf{0} \end{pmatrix} \right\|_2$$
$$\leq 1 + \left\| \mathbf{U}_2^\top \mathbf{X}_\ell \left(\mathbf{U}_1^\top \mathbf{X}_\ell \right)^{-1} \begin{pmatrix} \mathbf{I}_{k\times k} \\ \mathbf{0} \end{pmatrix} \right\|_2 = 1 + h_\ell.$$

■

Lemma A.3 (Lemma 2.2) *With all but $\tau^{-\Omega(p+1-q)} + e^{-\Omega(d)}$ probability, we have thta*

$$h_0 \leq \tan\theta_q(\mathbf{U}_q, \mathbf{X}_0) \leq \frac{\tau\sqrt{d}}{\sqrt{p} - \sqrt{q-1}}.$$

Proof Notice that $\mathbf{U}_{d-q}^\top \mathbf{X}_0 \left(\mathbf{U}_q^\top \mathbf{X}_0 \right)^\dagger \begin{pmatrix} \mathbf{I}_{k\times k} \\ \mathbf{0} \end{pmatrix}$ is a sub-matrix of $\mathbf{U}_{d-q}^\top \mathbf{X}_0 \left(\mathbf{U}_q^\top \mathbf{X}_0 \right)^\dagger$. Therefore,

$$h_0 = \left\| \mathbf{U}_{d-q}^\top \mathbf{X}_0 \left(\mathbf{U}_q^\top \mathbf{X}_0 \right)^\dagger \begin{pmatrix} \mathbf{I}_{k\times k} \\ \mathbf{0} \end{pmatrix} \right\|_2 \leq \left\| \mathbf{U}_{d-q}^\top \mathbf{X}_0 \left(\mathbf{U}_q^\top \mathbf{X}_0 \right)^\dagger \right\|_2 = \tan\theta_q \left(\mathbf{U}_q, \mathbf{X}_0 \right).$$

By $\mathbf{X}_0$ is the column space of a $d \times p$ random Gaussian matrix, Lemma 2.5 in (Hardt and Price, 2014) yields

$$\tan\theta_q \left(\mathbf{U}_q, \mathbf{X}_0 \right) \leq \frac{\tau\sqrt{d}}{\sqrt{p} - \sqrt{q-1}}$$

with all but $\tau^{-\Omega(p+1-q)} + e^{-\Omega(d)}$ probability. ■

Lemma A.4 (Lemma 2.3) *If $h_L \leq \epsilon/4$ then $\tan\theta_k(\mathbf{U}_k, \mathbf{X}_L) \leq \epsilon$.*

Proof First, we write $\mathbf{X}_L$ as

$$\mathbf{X}_L = \mathbf{U}\mathbf{U}^\top \mathbf{X}_L = \mathbf{U} \begin{pmatrix} \mathbf{U}_q^\top \mathbf{X}_L \\ \mathbf{U}_{d-q}^\top \mathbf{X}_L \end{pmatrix},$$

where $\mathbf{U}$ is the orthogonal space of $\mathbf{A}$. Next, consider a $p \times q$ matrix $\widehat{\mathbf{X}}$ that is orthogonal to $\left(\mathbf{U}_q^\top \mathbf{X}_L \right)$; that is, $\left(\mathbf{U}_q^\top \mathbf{X}_L \right) \widehat{\mathbf{X}} = \mathbf{0}$. Following the techniques introduced in (Gu, 2015; Halko et al., 2011), we consider the following matrix:

$$\mathbf{X} = \begin{pmatrix} \left(\mathbf{U}_q^\top \mathbf{X}_L \right)^\dagger & \widehat{\mathbf{X}}. \end{pmatrix}$$

By definition, we then have that

$$\mathbf{X}_L \mathbf{X} = \mathbf{U} \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{H}_1 & \mathbf{H}_2 & \mathbf{H}_3 \end{pmatrix},$$

where

$$\mathbf{H}_1 = \left(\mathbf{U}_{d-q}^\top \mathbf{X}_L\right)\left(\mathbf{U}_q^\top \mathbf{X}_L\right)^\dagger \begin{pmatrix}\mathbf{I}_{k\times k}\\ \mathbf{0}\end{pmatrix},$$
$$\mathbf{H}_2 = \left(\mathbf{U}_{d-q}^\top \mathbf{X}_L\right)\left(\mathbf{U}_q^\top \mathbf{X}_L\right)^\dagger \begin{pmatrix}\mathbf{0}\\ \mathbf{I}_{(q-k)\times(q-k)}\end{pmatrix},$$
$$\mathbf{H}_3 = \left(\mathbf{U}_{d-q}^\top \mathbf{X}_L\right)\widehat{\mathbf{X}}.$$

Note that $\|\mathbf{H}_1\|_2 = h_L$ by definition. Under the condition of the lemma $h_L \le \epsilon/4$, we have that $\|\mathbf{H}_1\|_2 \le \epsilon/4$. We next consider an alternative QR decomposition of $\mathbf{X}_L\mathbf{X}$:

$$\mathbf{X}_L\mathbf{X} = \widehat{\mathbf{Q}}\widehat{\mathbf{R}} = \begin{pmatrix}\widehat{\mathbf{Q}}_1 & \widehat{\mathbf{Q}}_2 & \widehat{\mathbf{Q}}_3\end{pmatrix}\begin{pmatrix}\widehat{\mathbf{R}}_{11} & \widehat{\mathbf{R}}_{12} & \widehat{\mathbf{R}}_{13}\\ & \widehat{\mathbf{R}}_{22} & \widehat{\mathbf{R}}_{23}\\ & & \widehat{\mathbf{R}}_{33}\end{pmatrix}.$$

Because the projection matrix $\widehat{\mathbf{Q}}$ is unique, we have $\widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^\top = \mathbf{X}_L\mathbf{X}_L^\top$. Also note that the above QR decomposition embeds another smaller one:

$$\mathbf{U}\begin{pmatrix}\mathbf{I}\\ \mathbf{0}\\ \mathbf{H}_1\end{pmatrix} = \widehat{\mathbf{Q}}_1\widehat{\mathbf{R}}_{11}.$$

The projection operator orthogonal to $\widehat{\mathbf{Q}}_1$ can be expressed as

$$\begin{aligned}\mathbf{I} - \widehat{\mathbf{Q}}_1\widehat{\mathbf{Q}}_1^\top &= \mathbf{U}\mathbf{U}^\top - \widehat{\mathbf{Q}}_1\widehat{\mathbf{Q}}_1^\top\\ &= \mathbf{U}\begin{pmatrix}\mathbf{I}\\ \mathbf{0}\\ \mathbf{H}_1\end{pmatrix}\widehat{\mathbf{R}}_{11}^{-1}\widehat{\mathbf{R}}_{11}^{-\top}\begin{pmatrix}\mathbf{I} & \mathbf{0} & \mathbf{H}_1^\top\end{pmatrix}\mathbf{U}^\top\\ &= \mathbf{U}\begin{pmatrix}\mathbf{I} - \left(\mathbf{I}+\mathbf{H}_1^\top\mathbf{H}_1\right)^{-1} & \mathbf{0} & -\left(\mathbf{I}+\mathbf{H}_1^\top\mathbf{H}_1\right)\mathbf{H}_1^\top\\ \mathbf{0} & \mathbf{I} & \mathbf{0}\\ -\mathbf{H}_1\left(\mathbf{I}+\mathbf{H}_1^\top\mathbf{H}_1\right)^{-1} & \mathbf{0} & \mathbf{I} - \mathbf{H}_1\left(\mathbf{I}+\mathbf{H}_1^\top\mathbf{H}_1\right)^{-1}\mathbf{H}_1^\top\end{pmatrix}\mathbf{U}^\top,\end{aligned}$$

where in the last equation we use the fact that $\widehat{\mathbf{R}}_{11}\widehat{\mathbf{R}}_{11} = \left(\mathbf{I}+\mathbf{H}_1^\top\mathbf{H}_1\right)^{-1}$. The principal angle $\theta_k(\mathbf{U}_k, \mathbf{X}_L)$ can then be bounded as

$$\begin{aligned}\sin\theta_k\left(\mathbf{U}_k, \mathbf{X}_L\right) &= \left\|\left(\mathbf{I} - \mathbf{X}_L\mathbf{X}_L^\top\right)\mathbf{U}_k\right\|_2\\ &= \left\|\left(\mathbf{I} - \widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^\top\right)\mathbf{U}_k\right\|_2\\ &\le \left\|\left(\mathbf{I} - \widehat{\mathbf{Q}}_1\widehat{\mathbf{Q}}_1^\top\right)\mathbf{U}_k\right\|_2\\ &= \left\|\mathbf{U}\begin{pmatrix}\mathbf{I} - \left(\mathbf{I}+\mathbf{H}_1^\top\mathbf{H}_1\right)^{-1} & \mathbf{0} & -\left(\mathbf{I}+\mathbf{H}_1^\top\mathbf{H}_1\right)\mathbf{H}_1^\top\\ \mathbf{0} & \mathbf{I} & \mathbf{0}\\ -\mathbf{H}_1\left(\mathbf{I}+\mathbf{H}_1^\top\mathbf{H}_1\right)^{-1} & \mathbf{0} & \mathbf{I} - \mathbf{H}_1\left(\mathbf{I}+\mathbf{H}_1^\top\mathbf{H}_1\right)^{-1}\mathbf{H}_1^\top\end{pmatrix}\mathbf{U}^\top\mathbf{U}_k\right\|_2\end{aligned}$$

$$
\begin{aligned}
&= \left\| \mathbf{U} \begin{pmatrix} \mathbf{I} - \left(\mathbf{I} + \mathbf{H}_1^\top \mathbf{H}_1\right)^{-1} & \mathbf{0} & -\left(\mathbf{I} + \mathbf{H}_1^\top \mathbf{H}_1\right) \mathbf{H}_1^\top \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ -\mathbf{H}_1 \left(\mathbf{I} + \mathbf{H}_1^\top \mathbf{H}_1\right)^{-1} & \mathbf{0} & \mathbf{I} - \mathbf{H}_1 \left(\mathbf{I} + \mathbf{H}_1^\top \mathbf{H}_1\right)^{-1} \mathbf{H}_1^\top \end{pmatrix} \begin{pmatrix} \mathbf{I}_{k\times k} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} \right\|_2 \\
&\leq \left\| \mathbf{I} - \left(\mathbf{I} + \mathbf{H}_1^\top \mathbf{H}_1\right)^{-1} \right\|_2 + \left\| \mathbf{H}_1 \left(\mathbf{I} + \mathbf{H}_1^\top \mathbf{H}_1\right)^{-1} \right\|_2,
\end{aligned}
$$

where the first inequality is due to the space projected by $\widehat{\mathbf{Q}}_1 \widehat{\mathbf{Q}}_1^\top$ is a subspace of that by $\widehat{\mathbf{Q}} \widehat{\mathbf{Q}}^\top$. By Woodbury's identity, we have that

$$
\left\| \mathbf{I} - \left(\mathbf{I} + \mathbf{H}_1^\top \mathbf{H}_1\right)^{-1} \right\|_2 = \left\| \mathbf{H}_1^\top \left(\mathbf{I} + \mathbf{H}_1 \mathbf{H}_1^\top\right) \mathbf{H}_1 \right\|_2 \leq \frac{(\epsilon/4)^2}{1 - (\epsilon/4)^2} \leq \epsilon/2.
$$

For the other term, we have

$$
\left\| \mathbf{H}_1 \left(\mathbf{I} + \mathbf{H}_1^\top \mathbf{H}_1\right)^{-1} \right\|_2 \leq \frac{\epsilon/4}{1 - (\epsilon/4)^2} \leq \epsilon/2.
$$

Combing these two inequalities, we get

$$
\sin \theta_k \left(\mathbf{U}_k, \mathbf{X}_L\right) \leq \epsilon.
$$

The proof is then completed by noting that $\sin \theta_k \left(\mathbf{U}_k, \mathbf{X}_L\right) \leq \epsilon/2$ yields

$$
\tan \theta_k \left(\mathbf{U}_k, \mathbf{X}_L\right) = \frac{\sin \theta_k \left(\mathbf{U}_k, \mathbf{X}_L\right)}{\sqrt{1 - \sin^2 \left(\mathbf{U}_k, \mathbf{X}_L\right)}} \leq \epsilon.
$$

■

An anonymous reviewer provides an much cleaner proof for Lemma 2.3.

Proof Since for any vector $w \in R^d$, we have $\left\| w - \mathbf{X}\mathbf{X}^\top w \right\|_2 \leq \left\| w - z \right\|_2$, for any vector $z \in \mathbf{Span}(\mathbf{X})$ if $\mathbf{X}$ is an orthonormal column matrix. Thus,

$$
\begin{aligned}
\left\| (\mathbf{I} - \mathbf{X}_\ell \mathbf{X}_\ell^\top) \mathbf{U}_k \right\|_2 &\leq \left\| \mathbf{U}_k - \mathbf{X}_\ell (\mathbf{U}_q^\top \mathbf{X}_\ell)^\dagger \begin{pmatrix} \mathbf{I}_{k\times k} \\ \mathbf{0} \end{pmatrix} \right\|_2 \\
&= \left\| \mathbf{U}_q^\top \left(\mathbf{U}_k - \mathbf{X}_\ell (\mathbf{U}_q^\top \mathbf{X}_\ell)^\dagger \begin{pmatrix} \mathbf{I}_{k\times k} \\ \mathbf{0} \end{pmatrix} \right) \right\|_2 + \left\| \mathbf{U}_{d-q}^\top \left(\mathbf{U}_k - \mathbf{X}_\ell (\mathbf{U}_q^\top \mathbf{X}_\ell)^\dagger \begin{pmatrix} \mathbf{I}_{k\times k} \\ \mathbf{0} \end{pmatrix} \right) \right\|_2 \\
&= \left\| \mathbf{U}_{d-q}^\top \mathbf{X}_\ell (\mathbf{U}_q^\top \mathbf{X}_\ell)^\dagger \begin{pmatrix} \mathbf{I}_{k\times k} \\ \mathbf{0} \end{pmatrix} \right\|_2 \\
&\leq \epsilon/2.
\end{aligned}
$$

Therefore, again

$$
\tan \theta_k \left(\mathbf{U}_k, \mathbf{X}_L\right) = \frac{\sin \theta_k \left(\mathbf{U}_k, \mathbf{X}_L\right)}{\sqrt{1 - \sin^2 \left(\mathbf{U}_k, \mathbf{X}_L\right)}} \leq \epsilon.
$$

■

Lemma A.5 *Under the same assumption as Theorem 2.1, if $h_\ell \le \epsilon$, then*

$$\left\|\mathbf{A} - \mathbf{X}_{\ell+1}\mathbf{X}_{\ell+1}^\top \mathbf{A}\right\|_2^2 \le \sigma_{k+1}^2 + \epsilon^2 \sigma_k^2$$

Proof We consider $(\ell+1)$-th iteration:

$$\begin{aligned}\mathbf{Y}_\ell &= \mathbf{A}\mathbf{X}_\ell + \mathbf{G}_\ell \\ &= \mathbf{U}\begin{pmatrix}\boldsymbol{\Sigma}_q\mathbf{U}_q^\top\mathbf{X}_\ell + \mathbf{U}_q^\top\mathbf{G}_\ell \\ \boldsymbol{\Sigma}_{d-q}\mathbf{U}_{d-q}^\top\mathbf{X}_\ell + \mathbf{U}_{d-q}^\top\mathbf{G}_\ell.\end{pmatrix}\end{aligned}$$

We use perturbation theory for analysis. Consider the following matrix

$$\mathbf{X} = \begin{pmatrix}\left(\boldsymbol{\Sigma}_q\mathbf{U}_q^\top\mathbf{X}_\ell + \mathbf{U}_q^\top\mathbf{G}_\ell\right)^\dagger & \widehat{\mathbf{X}}\end{pmatrix},$$

where $\left(\boldsymbol{\Sigma}_q\mathbf{U}_q^\top\mathbf{X}_\ell + \mathbf{U}_q^\top\mathbf{G}_\ell\right)\widehat{\mathbf{X}} = \mathbf{0}$. We then have

$$\mathbf{Y}_\ell\mathbf{X} = \mathbf{U}\begin{pmatrix}\mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{H}_1 & \mathbf{H}_2 & \mathbf{H}_3\end{pmatrix},$$

where

$$\begin{aligned}\mathbf{H}_1 &= \left(\boldsymbol{\Sigma}_{d-q}\mathbf{U}_q^\top\mathbf{X}_\ell + \mathbf{U}_{d-q}^\top\mathbf{G}_\ell\right)\left(\boldsymbol{\Sigma}_q\mathbf{U}_q^\top\mathbf{X}_\ell + \mathbf{U}_q^\top\mathbf{G}_\ell\right)^\dagger\begin{pmatrix}\mathbf{I}_{k\times k} \\ \mathbf{0}\end{pmatrix} \\ \mathbf{H}_2 &= \left(\boldsymbol{\Sigma}_{d-q}\mathbf{U}_q^\top\mathbf{X}_\ell + \mathbf{U}_{d-q}^\top\mathbf{G}_\ell\right)\left(\boldsymbol{\Sigma}_q\mathbf{U}_q^\top\mathbf{X}_\ell + \mathbf{U}_q^\top\mathbf{G}_\ell\right)^\dagger\begin{pmatrix}\mathbf{0} \\ \mathbf{I}_{(q-k)\times(q-k)}\end{pmatrix} \\ \mathbf{H}_3 &= \left(\boldsymbol{\Sigma}_{d-q}\mathbf{U}_q^\top\mathbf{X}_\ell + \mathbf{U}_{d-q}^\top\mathbf{G}_\ell\right)\widehat{\mathbf{X}}.\end{aligned}$$

Similar to the proof of Lemma 2.3, the reconstruction error can be bounded in terms of $\mathbf{H}_q$:

$$\begin{aligned}\left\|\left(\mathbf{I} - \mathbf{X}_{\ell+1}\mathbf{X}_{\ell+1}^\top\right)\mathbf{A}\right\|_2^2 &= \left\|\mathbf{A}\left(\mathbf{I} - \mathbf{X}_{\ell+1}\mathbf{X}_{\ell+1}^\top\right)\mathbf{A}\right\|_2 \\ &\le \left\|\boldsymbol{\Sigma}\begin{pmatrix}\mathbf{I} - \left(\mathbf{I} + \mathbf{H}_1^\top\mathbf{H}_1\right)^{-1} & \mathbf{0} & -\left(\mathbf{I} + \mathbf{H}_1^\top\mathbf{H}_1\right)\mathbf{H}_1^\top \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ -\mathbf{H}_1\left(\mathbf{I} + \mathbf{H}_1^\top\mathbf{H}_1\right)^{-1} & \mathbf{0} & \mathbf{I} - \mathbf{H}_1\left(\mathbf{I} + \mathbf{H}_1^\top\mathbf{H}_1\right)^{-1}\mathbf{H}_1^\top\end{pmatrix}\boldsymbol{\Sigma}\right\|_2.\end{aligned}$$

By Proposition 8.2 of (Halko et al., 2011), we have

$$\begin{aligned}&\begin{pmatrix}\mathbf{I} - \left(\mathbf{I} + \mathbf{H}_1^\top\mathbf{H}_1\right)^{-1} & \mathbf{0} & -\left(\mathbf{I} + \mathbf{H}_1^\top\mathbf{H}_1\right)\mathbf{H}_1^\top \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ -\mathbf{H}_1\left(\mathbf{I} + \mathbf{H}_1^\top\mathbf{H}_1\right)^{-1} & \mathbf{0} & \mathbf{I} - \mathbf{H}_1\left(\mathbf{I} + \mathbf{H}_1^\top\mathbf{H}_1\right)^{-1}\mathbf{H}_1^\top\end{pmatrix} \\ &\preccurlyeq \begin{pmatrix}\mathbf{H}_1^\top\mathbf{H}_1 & \mathbf{0} & -\left(\mathbf{I} + \mathbf{H}_1^\top\mathbf{H}_1\right)\mathbf{H}_1^\top \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ -\mathbf{H}_1\left(\mathbf{I} + \mathbf{H}_1^\top\mathbf{H}_1\right)^{-1} & \mathbf{0} & \mathbf{I}\end{pmatrix}.\end{aligned}$$

Thus by Proposition 8.3 of (Halko et al., 2011),

$$\begin{aligned}&\left\|\left(\mathbf{I}-\mathbf{X}_{\ell+1}\mathbf{X}_{\ell+1}^{\top}\right)\mathbf{A}\right\|_2^2\\ \leq&\left\|\mathbf{\Sigma}_{d-k}\right\|_2^2+\left\|\mathbf{\Sigma}_k\mathbf{H}_1^{\top}\mathbf{H}_1\mathbf{\Sigma}_k\right\|_2\\ =&\left\|\mathbf{\Sigma}_{d-k}\right\|_2^2+\left\|\mathbf{H}_1\mathbf{\Sigma}_k\right\|_2^2.\end{aligned}$$

Thus we only need to bound $\left\|\mathbf{H}_1\mathbf{\Sigma}_k\right\|_2$. By definition of $\mathbf{H}_1$, we have

$$\begin{aligned}\left\|\mathbf{H}_1\mathbf{\Sigma}_k\right\|_2&=\left\|\left(\mathbf{\Sigma}_{d-q}\mathbf{U}_q^{\top}\mathbf{X}_\ell+\mathbf{U}_{d-q}^{\top}\mathbf{G}_\ell\right)\left(\mathbf{\Sigma}_q\mathbf{U}_q^{\top}\mathbf{X}_\ell+\mathbf{U}_q^{\top}\mathbf{G}_\ell\right)^{\dagger}\begin{pmatrix}\mathbf{\Sigma}_k\\ \mathbf{0}\end{pmatrix}\right\|_2\\ &=\left\|\left(\mathbf{\Sigma}_{d-q}\mathbf{U}_q^{\top}\mathbf{X}_\ell+\mathbf{U}_{d-q}^{\top}\mathbf{G}_\ell\right)\left(\mathbf{U}_q^{\top}\mathbf{X}_\ell+\mathbf{\Sigma}_q^{-1}\mathbf{U}_q^{\top}\mathbf{G}_\ell\right)^{\dagger}\begin{pmatrix}\mathbf{I}_{k\times k}\\ \mathbf{0}\end{pmatrix}\right\|_2 \qquad (9)\end{aligned}$$

Notice the similar forms of Eqn. (9) and $h_{\ell+1}$ in the proof of Lemma 2.1. $\left\|\mathbf{H}_1\mathbf{\Sigma}_k\right\|_2$ can be bounded using the exactly same argument, so based on assumption on the noise and h_ℓ, we have:

$$\left\|\mathbf{H}_1\mathbf{\Sigma}_k\right\|_2=O\left(\epsilon\left(\sigma_k-\sigma_{q+1}\right)\right)+O\left(\epsilon\sigma_{q+1}\right)=O\left(\epsilon\sigma_k\right).$$

■

Lemma A.6 *Under the same assumption as Theorem 2.1, if $h_\ell \leq \epsilon$, then*

$$\left\|\mathbf{A}-\mathbf{X}_{\ell+1}\mathbf{X}_{\ell+1}^{\top}\mathbf{A}\right\|_F^2\leq\sum_{i=k+1}^{d}\sigma_i^2+\epsilon^2k\sigma_k^2$$

Proof *By the proof of Theorem 4.4 of (Gu, 2015), we have*

$$\left\|\left(\mathbf{I}-\mathbf{X}_{\ell+1}\mathbf{X}_{\ell+1}^{\top}\right)\mathbf{A}\right\|_F^2\leq\left\|\mathbf{\Sigma}_{d-k}\right\|_F^2+k\left\|\mathbf{H}_1\mathbf{\Sigma}_k\right\|_2^2,$$

where $\mathbf{H}_1$ is defined similarly as in the proof of Lemma A.5. ■

Lemma A.7 *Fix $0<\gamma<1$. If at each iteration ℓ the noise matrix $\mathbf{G}_\ell$ satisfies*

$$\left\|\mathbf{G}_\ell\right\|_2=O\left(\gamma\sigma_q\right)\quad\text{and}\quad\left\|\mathbf{U}_q^{\top}\mathbf{G}_\ell\right\|_2=O\left(\frac{\sqrt{p}-\sqrt{q-1}}{\tau\sqrt{d}}\cdot\gamma\sigma_q\right),$$

then for all $\ell=O\left(1/\gamma\right)$, the following holds with probability all but $\tau^{-\Omega(p+1-q)}+e^{-\Omega(d)}$ probability:

$$\tan\theta_q\left(\mathbf{U}_q,\mathbf{X}_\ell\right)=O\left(\frac{\tau\sqrt{d}}{\sqrt{p}-\sqrt{q-1}}\right),\qquad\cos\theta_q\left(\mathbf{U}_q,\mathbf{X}_\ell\right)=\Omega\left(\frac{\sqrt{p}-\sqrt{q-1}}{\tau\sqrt{d}}\right).$$

Proof By Lemma 2.2, the tangent of the qth principal angle between $\mathbf{U}_q$ and $\mathbf{X}_0$ can be bounded as

$$\tan\theta_q(\mathbf{U}_q,\mathbf{X}_0)\leq\frac{\tau\sqrt{d}}{\sqrt{p}-\sqrt{q-1}} \tag{10}$$

with high probability. We also consider the following inequality that upper bounds $\tan\theta_q(\mathbf{U}_q,\mathbf{X}_\ell)$ in terms of $\tan\theta_q(\mathbf{U}_q,\mathbf{X}_0)$:

$$\tan\theta_q\left(\mathbf{U}_q,\mathbf{X}_\ell\right)+\frac{c_1}{c_1+c_3}\leq\left(\frac{1+c_1\gamma}{1-c_3\gamma}\right)^\ell\left(\tan\theta_q\left(\mathbf{U}_q,\mathbf{X}_0\right)+\frac{c_1}{c_1+c_3}\right). \tag{11}$$

Here $c_1,c_2,c_3>0$ are universal constants. Eq. (10) and Eq. (11) imply $\tan\theta_q(\mathbf{U}_q,\mathbf{X}_\ell)=O(\frac{\tau\sqrt{d}}{\sqrt{p}-\sqrt{q-1}})$ for all $\ell=O(1/\gamma)$ because

$$\left(\frac{1+c_1\gamma}{1-c_3\gamma}\right)^\ell=\left(1+\frac{(c_1+c_3)\,\gamma}{1-c_3\gamma}\right)^{\frac{(c_1+c_3)\gamma}{1-c_3\gamma}\cdot\left(\frac{1-c_3\gamma}{(c_1+c_3)\gamma}\right)\cdot\ell}\leq\exp\left(\frac{1-c_3\gamma}{(c_1+c_3)\,\gamma}\cdot\ell\right)=O(1),$$

if $\ell=O(1/\gamma)$. $\cos\theta_q(\mathbf{U}_q,\mathbf{X}_\ell)$ can subsequently be lower bounded as

$$\cos\left(\mathbf{U}_q,\mathbf{X}_\ell\right)\geq\frac{1}{1+\tan\left(\mathbf{U}_q,\mathbf{X}_\ell\right)}=\Omega\left(\frac{\sqrt{p}-\sqrt{q-1}}{\tau\sqrt{d}}\right).$$

The rest of the proof is dedicated to prove Eq. (11) via mathematical induction. When $\ell=0$, the statement is trivially true. Suppose for Eq. (11) is true for all $\ell=1,\cdots,s$. We want to prove that Eq. (11) is also true for $\ell=s+1$. By definition,

$$\tan\theta_q\left(\mathbf{U}_q,\mathbf{X}_\ell\right)=\min_{\mathbf{\Pi}\in\mathcal{P}_p}\max_{\|\mathbf{w}\|=1,\mathbf{\Pi}\mathbf{w}=\mathbf{w}}\frac{\left\|\mathbf{U}_{d-q}^\top\mathbf{X}_\ell\mathbf{w}\right\|}{\left\|\mathbf{U}_q^\top\mathbf{X}_\ell\mathbf{w}\right\|}=\max_{\|\mathbf{w}\|=1,\mathbf{\Pi}^\star\mathbf{w}=\mathbf{w}}\frac{\left\|\mathbf{U}_{d-q}^\top\mathbf{X}_\ell\mathbf{w}\right\|}{\left\|\mathbf{U}_q^\top\mathbf{X}_\ell\mathbf{w}\right\|}.$$

Here $\mathcal{P}_p$ denotes the set of all projection matrices on $\mathbb{R}^p$ and $\mathbf{\Pi}^*$ is the projection matrix that achieves the minimum value in the second term. We then have

$$\begin{aligned}
\tan\theta_q\left(\mathbf{U}_q,\mathbf{X}_{\ell+1}\right)&=\tan\theta_q\left(\mathbf{U}_q,\mathbf{A}\mathbf{X}_\ell+\mathbf{G}_\ell\right)\\
&=\min_{\mathbf{\Pi}\in\mathcal{P}_p}\max_{\|\mathbf{w}\|_2=1,\mathbf{\Pi}\mathbf{w}=\mathbf{w}}\frac{\left\|\mathbf{U}_{d-q}^\top\left(\mathbf{A}\mathbf{X}_\ell+\mathbf{G}_\ell\right)\mathbf{w}\right\|}{\left\|\mathbf{U}_q^\top\left(\mathbf{A}\mathbf{X}_\ell+\mathbf{G}_\ell\right)\mathbf{w}\right\|}\\
&\leq\max_{\|\mathbf{w}\|_2=1,\mathbf{\Pi}^\star\mathbf{w}=\mathbf{w}}\frac{\left\|\mathbf{\Sigma}_{d-q}\mathbf{U}_{d-q}^\top\mathbf{X}_\ell\mathbf{w}\right\|_2+\left\|\mathbf{U}_{d-q}\mathbf{G}_\ell\mathbf{w}\right\|_2}{\left\|\mathbf{\Sigma}_q\mathbf{U}_q^\top\mathbf{X}_\ell\mathbf{w}\right\|_2-\left\|\mathbf{U}_q^\top\mathbf{G}_\ell\mathbf{w}\right\|_2}\\
&\leq\max_{\|\mathbf{w}\|_2=1,\mathbf{\Pi}^\star\mathbf{w}=\mathbf{w}}\frac{\sigma_{q+1}\left\|\mathbf{U}_{d-q}^\top\mathbf{X}_\ell\mathbf{w}\right\|_2/\left\|\mathbf{U}_q^\top\mathbf{X}_\ell w\right\|_2+\|\mathbf{G}_\ell\|_2/\left\|\mathbf{U}_q^\top\mathbf{X}_\ell\mathbf{w}\right\|_2}{\sigma_q-\left\|\mathbf{U}_q^\top\mathbf{G}_\ell\mathbf{w}\right\|_2/\left\|\mathbf{U}_q^\top\mathbf{X}_\ell\mathbf{w}\right\|_2}
\end{aligned} \tag{12}$$

By definition of the principal angles, we have

$$\max_{\|\mathbf{w}\|_2=1,\mathbf{\Pi}^\star\mathbf{w}=\mathbf{w}}\left\|\mathbf{U}_{d-q}^\top\mathbf{X}_\ell\mathbf{w}\right\|_2/\left\|\mathbf{U}_q^\top\mathbf{X}_\ell w\right\|_2=\tan\left(\mathbf{U}_q,\mathbf{X}_\ell\right),$$

$$\max_{\|\mathbf{w}\|_2=1,\mathbf{\Pi}^\star\mathbf{w}=\mathbf{w}} \frac{1}{\left\|\mathbf{U}_q^\top \mathbf{X}_\ell \mathbf{w}\right\|_2} = \frac{1}{\cos\left(\mathbf{U}_q, \mathbf{X}_\ell\right)} \leq 1 + \tan\left(\mathbf{U}_q, \mathbf{X}_\ell\right).$$

Also, conditions on the noise matrices $\mathbf{G}_\ell$ read

$$\|\mathbf{G}_\ell\|_2 =\leq c_1\gamma\sigma_q, \qquad \left\|\mathbf{U}_q^\top \mathbf{G}_\ell\right\|_2 \leq c_3\gamma\sigma_q \cos\left(\mathbf{U}_q, \mathbf{X}_\ell\right).$$

Plugging these inequalities into Eq. (12), we obtain

$$\begin{aligned}
\tan\left(\mathbf{U}_q, \mathbf{X}_{\ell+1}\right) &\leq \frac{\sigma_{q+1}\tan\left(\mathbf{U}_q, \mathbf{X}_\ell\right) + c_1\gamma\left(1+\tan\left(\mathbf{U}_q, \mathbf{X}_\ell\right)\right)}{\sigma_q - c_3\gamma\sigma_q} \\
&\leq \left(\frac{1+c_1\gamma}{1-c_3\gamma}\right)\tan\left(\mathbf{U}_q, \mathbf{X}_\ell\right) + \frac{c_1\gamma}{1-c_3\gamma} \\
&\leq \left(\frac{1+c_1\gamma}{1-c_3\gamma}\right)^\ell \left(\tan\theta_q\left(\mathbf{U}_q, \mathbf{X}_0\right) + \frac{c_1}{c_1+c_3}\right),
\end{aligned}$$

where the last inequality is due to induction hypothesis placed on Eq. (11). ■

Corollary A.1 *Fix* $\epsilon = O\left(\frac{\sigma_q}{\sigma_k}\cdot\min\left\{\frac{1}{\log\left(\frac{\sigma_k}{\sigma_q}\right)}, \frac{1}{\log(\tau d)}\right\}\right)$. *Suppose at each iteration the noise matrix* $\mathbf{G}_\ell$ *satisfies*

$$\|\mathbf{G}_\ell\|_2 = O\left(\epsilon\left(\sigma_k - \sigma_{q+1}\right)\right) \quad \textit{and} \quad \left\|\mathbf{U}_q^\top \mathbf{G}_\ell\right\| = O\left(\frac{\sqrt{p}-\sqrt{q-1}}{\tau\sqrt{d}}\cdot\min\{\epsilon\left(\sigma_k-\sigma_{q+1}\right), \sigma_q\}\right),$$

then for all $\ell = O\left(\frac{\sigma_k}{\sigma_k-\sigma_{q+1}}\log\left(\frac{\tau d}{\epsilon}\right)\right)$ *the following holds with all but* $\tau^{-\Omega(p+1-q)} + e^{-\Omega(d)}$ *probability:*

$$\tan\theta_q\left(\mathbf{U}_q, \mathbf{X}_\ell\right) = O\left(\frac{\tau\sqrt{d}}{\sqrt{p}-\sqrt{q-1}}\right), \qquad \cos\theta_q\left(\mathbf{U}_q, \mathbf{X}_\ell\right) = \Omega\left(\frac{\sqrt{p}-\sqrt{q-1}}{\tau\sqrt{d}}\right).$$

Proof Apply Lemma A.7 with $\gamma = \min\{\frac{\epsilon(\sigma_k-\sigma_{q+1})}{\sigma_q}, 1\}$. ■

Appendix B. Proof of Theorem 2.3

Proof Define $m = \text{argmax}_i\{\sigma_i - \sigma_{k+1} \geq \epsilon\sigma_{k+1}\}$. If $m = 0$, then we are done since $\left\|\mathbf{A} - \mathbf{X}_L\mathbf{X}_L^\top\mathbf{A}\right\|_2 \leq \|\mathbf{A}\|_2 \leq \sigma_1 \leq (1+\epsilon)\sigma_{k+1} = (1+\epsilon)\|\mathbf{A}-\mathbf{A}_k\|_2$. Otherwise, consider the case that our target rank is m, and the leading rank-k subspace. By our definition on m and noise conditions, we have

$$\begin{aligned}
\|\mathbf{G}\|_2 &= O\left(\epsilon^2\sigma_{k+1}\right) = O\left(\epsilon\left(\sigma_m - \sigma_{k+1}\right)\right); \\
\left\|\mathbf{U}_k^\top\mathbf{G}\right\|_2 &= O\left(\frac{\epsilon^2\left(\sqrt{p}-\sqrt{k-1}\right)\sigma_{k+1}}{\tau\sqrt{d}}\right) = O\left(\frac{\epsilon\left(\sqrt{p}-\sqrt{k-1}\right)\left(\sigma_m-\sigma_{k+1}\right)}{\tau\sqrt{d}}\right).
\end{aligned}$$

Next, by Lemma B.1, for all $\ell = O\left(\frac{1}{\epsilon^2}\right)$ the cosine principal angle $\cos\theta_q(\mathbf{U}_q, \mathbf{X}_\ell)$ can be lower bounded as

$$\cos\left(\mathbf{U}_k, \mathbf{X}_\ell\right) = \Omega\left(\frac{\sqrt{p}-\sqrt{k-1}}{\tau\sqrt{d}}\right).$$

Note also that $\frac{\sigma_m}{\sigma_m-\sigma_{k+1}}\log\left(\frac{\tau d}{\epsilon}\right) \lesssim \frac{1}{\epsilon}\log\left(\frac{\tau d}{\epsilon}\right) \lesssim L$.

Using the same argument as in the proof of Lemma A.5, we have

$$\left\|\mathbf{A} - \mathbf{X}_{L+1}\mathbf{X}_{L+1}^\top\mathbf{A}\right\|_2^2 \le \|\mathbf{\Sigma}_{d-m}\|_2^2 + \|\mathbf{H}_1\mathbf{\Sigma}_m\|_2^2,$$

where

$$\mathbf{H}_1 = \left(\mathbf{\Sigma}_{d-k}\mathbf{U}_k^\top\mathbf{X}_L + \mathbf{U}_{d-m}^\top\mathbf{G}_L\right)\left(\mathbf{\Sigma}_k\mathbf{U}_k^\top\mathbf{X}_L + \mathbf{U}_k^\top\mathbf{G}_L\right)^\dagger\begin{pmatrix}\mathbf{I}_{m\times m}\\ \mathbf{0}\end{pmatrix}.$$

Again by the same argument in the proof of 2.1, $\|\mathbf{H}_1\mathbf{\Sigma}_m\|_2 \le \epsilon\sigma_{k+1}$. Lastly, by the definition of m we obtain the desired result. ■

Lemma B.1 *Fix $\epsilon = O\left(1/\log\left(\tau d\right)\right)$. If at each iteration the noise matrix $\mathbf{G}_\ell$ satisfies*

$$\|\mathbf{G}_\ell\|_2 = O\left(\epsilon^2\sigma_k\right) \quad \textit{and} \quad \left\|\mathbf{U}_q^\top\mathbf{G}_\ell\right\| = O\left(\frac{\sqrt{p}-\sqrt{q-1}}{\tau\sqrt{d}}\cdot\epsilon^2\sigma_k\right),$$

then for all $\ell = O\left(1/\epsilon^2\right)$ the following holds with all but $\tau^{-\Omega(p+1-q)} + e^{-\Omega(d)}$ probability:

$$\tan\theta_q\left(\mathbf{U}_q, \mathbf{X}_\ell\right) = O\left(\frac{\tau\sqrt{d}}{\sqrt{p}-\sqrt{q-1}}\right), \qquad \cos\theta_q\left(\mathbf{U}_q, \mathbf{X}_\ell\right) = \Omega\left(\frac{\sqrt{p}-\sqrt{q-1}}{\tau\sqrt{d}}\right).$$

Proof Apply Lemma A.7 with $p = k$ and $\gamma = \epsilon^2$. ■

Appendix C. Proof of results for distributed private PCA

Theorem C.1 (Distributed private PCA, Theorem 3.1) *Let s be the number of computing nodes and $\mathbf{A}^{(1)}, \cdots, \mathbf{A}^{(s)} \in \mathbb{R}^{d\times d}$ be data matrices stored separately on the s nodes. Fix target rank k, intermediate rank $q \ge k$ and iteration rank p with $2q \le p \le d$. Suppose the number of iterations L is set as $L = \Theta(\frac{\sigma_k}{\sigma_k-\sigma_{q+1}}\log(d))$. Let $\varepsilon, \delta \in (0,1)$ be privacy parameters. Then Algorithm 2 solves the $(\varepsilon, \delta, \epsilon, M)$-distributed PCA problem with*

$$\epsilon = O\left(\frac{\nu\sqrt{\mu(\mathbf{A})s\log d\log L}}{\sigma_k-\sigma_{q+1}}\right) \quad \textit{and} \quad M = O(spdL) = O\left(\frac{\sigma_k}{\sigma_k-\sigma_{q+1}}spd\log d\right).$$

Here assuming conditions in Theorem 2.2 are satisfied, $\nu = \varepsilon^{-1}\sqrt{4pL\log(1/\delta)}$ and $\mu(\mathbf{A})$ is the incoherence *(Hardt and Roth, 2013) of the aggregate data matrix $\mathbf{A} = \sum_{i=1}^s \mathbf{A}^{(i)}$; more specifically, $\mu(\mathbf{A}) = d\|\mathbf{U}\|_\infty$ where $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ is the eigen-decomposition of $\mathbf{A}$.*

Proof We prove privacy, utility and communication guarantees of Algorithm 2 separately.

Privacy guarantee By Claim 4.2 in (Hardt and Price, 2014), Algorithm 2 satisfies (ε, δ)-differential privacy with respect to data matrix $\mathbf{A}^{(i)}$ on each computing node i. Because information of each data matrix $\mathbf{A}^{(i)}$ is only released by the corresponding computing node i via the public communication channel, we immediately have that Algorithm 2 is (ε, δ)-differentially private in terms of the definition in Eq. (6).

Utility guarantee Let $\mathbf{G}_\ell = \mathbf{G}_\ell^{(1)} + \cdots + \mathbf{G}_\ell^{(s)}$. Because $\mathbf{G}_\ell^{(1)}, \cdots, \mathbf{G}_\ell^{(s)} \overset{i.i.d.}{\sim} \mathcal{N}(0, \|\mathbf{X}_{\ell-1}\|_\infty^2 \nu^2)^{d\times p}$, we have that $\mathbf{G}_\ell \sim \mathcal{N}(0, \|\mathbf{X}_{\ell-1}\|_\infty^2 \tilde{\nu}^2)^{d\times p}$ for $\tilde{\nu} = \nu\sqrt{s}$. Properties of Gaussian matrices (e.g., Lemma A.2 in (Hardt and Price, 2014)) show that with high probability $\mathbf{G}_\ell$ satisfies the noise conditions in Theorem 2.2 with $\epsilon = \frac{\nu \max_\ell \|\mathbf{X}_\ell\|_\infty \sqrt{ds \log L}}{\sigma_k - \sigma_{q+1}}$. In addition, Theorem 4.9 in (Hardt and Price, 2014) shows that $\max_\ell \|\mathbf{X}_\ell\|_\infty^2 = O(\mu(\mathbf{A}) \log d/d)$ with high probability. The utility guarantee then holds by applying Theorem 2.2 with bounds on ϵ and $\max_\ell \|\mathbf{X}_\ell\|_\infty^2$.

Communication guarantee For each iteration ℓ, the central node broadcasts $\mathbf{X}_{\ell-1}$ to each computing node and receives $\mathbf{A}_\ell^{(i)}\mathbf{X}_{\ell-1} + \mathbf{G}_\ell^{(i)}$ from computing node i, for each $i = 1, \cdots, s$. Both matrices communicated on the public channel between the central node and each computing node is $d \times p$, which yields a per-iteration communication complexity of $O(spd)$. As a result, the total amount of communication is $O(spdL)$, where L is the number of iterations carried out in Algorithm 2. Because L is set as $L = \Theta(\frac{\sigma_k}{\sigma_k - \sigma_{q+1}} \log d)$, we have that $M = O(spdL) = O\left(\frac{\sigma_k}{\sigma_k - \sigma_{q+1}} spd \log d\right)$. ■

Corollary C.1 (Corollary 3.1) *For the case of $s = 1$ and $2p \le q \le d$, Algorithm 2 is (ε, δ)-differentially private and $\mathbf{X}_L$ satisfies*

$$\|(\mathbf{I} - \mathbf{X}_L \mathbf{X}_L^\top)\mathbf{U}_k\|_2 \le \epsilon = O\left(\frac{\nu\sqrt{\mu(\mathbf{A}) \log d \log L}}{\sigma_k - \sigma_{q+1}}\right)$$

with probability at least 0.9. Here $\mathbf{U}_k$ is the top-k eigen-space of input data matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$.

Proof Setting $s = 1$ in Theorem 3.1 we immediately get this corollary. ■

Corollary C.2 (Corollary 3.2) *Fix error tolerance parameter $\epsilon \in (0, 1)$ and set $\nu = 0$, $L = \Theta(\frac{\sigma_k}{\sigma_k - \sigma_{q+1}} \log(d/\epsilon))$ in Algorithm 2. We then have that with probability 1*

$$\|(\mathbf{I} - \mathbf{X}_L \mathbf{X}_L^\top)\mathbf{U}_k\|_2 \le \epsilon.$$

Here $\mathbf{U}_k$ is the top-k eigen-space of the aggregated matrix $\mathbf{A} = \sum_{i=1}^s \mathbf{A}^{(i)}$.

Proof Because $\nu = 0$, we are not adding any amount of noise in Algorithm 2; that is, $\mathbf{G}_\ell = \mathbf{0}$. Applying Theorem 2.2 with $\mathbf{G}_\ell = \mathbf{0}$ and $L = \Theta(\frac{\sigma_k}{\sigma_k - \sigma_{q+1}} \log(d/\epsilon))$ we have $\|(\mathbf{I} - \mathbf{X}_L \mathbf{X}_L^\top)\mathbf{U}_k\|_2 \le \epsilon$ with high probability. ■