The Noisy Power Method:
A Meta Algorithm with Applications

Moritz Hardt[*] Eric Price[†]

July 8, 2014

Abstract

We provide a new robust convergence analysis of the well-known power method for computing the dominant singular vectors of a matrix that we call the *noisy power method*. Our result characterizes the convergence behavior of the algorithm when a significant amount noise is introduced after each matrix-vector multiplication. The noisy power method can be seen as a meta-algorithm that has recently found a number of important applications in a broad range of machine learning problems including alternating minimization for matrix completion, streaming principal component analysis (PCA), and privacy-preserving spectral analysis. Our general analysis subsumes several existing ad-hoc convergence bounds and resolves a number of open problems in multiple applications:

Streaming PCA. A recent work of Mitliagkas et al. (NIPS 2013) gives a space-efficient algorithm for PCA in a streaming model where samples are drawn from a gaussian spiked covariance model. We give a simpler and more general analysis that applies to arbitrary distributions confirming experimental evidence of Mitliagkas et al. Moreover, even in the spiked covariance model our result gives quantitative improvements in a natural parameter regime. It is also notably simpler and follows easily from our general convergence analysis of the noisy power method together with a matrix Chernoff bound.

Private PCA. We provide the first nearly-linear time algorithm for the problem of differentially private principal component analysis that achieves nearly tight worst-case error bounds. Complementing our worst-case bounds, we show that the error dependence of our algorithm on the matrix dimension can be replaced by an essentially tight dependence on the *coherence* of the matrix. This result resolves the main problem left open by Hardt and Roth (STOC 2013). The coherence is always bounded by the matrix dimension but often substantially smaller thus leading to strong average-case improvements over the optimal worst-case bound.

1 Introduction

Computing the dominant singular vectors of a matrix is one of the most important algorithmic tasks underlying many applications including low-rank approximation, PCA, spectral clustering, dimensionality reduction, matrix completion and topic modeling. The classical problem is well-understood, but many recent applications in machine learning face the fundamental problem of approximately finding singular vectors in the presence of noise. Noise can enter the computation

[*] IBM Research Almaden. Email: mhardt@us.ibm.com

[†] IBM Research Almaden. Email: ecprice@mit.edu

through a variety of sources including sampling error, missing entries, adversarial corruptions and privacy constraints. It is desirable to have one robust method for handling a variety of cases without the need for ad-hoc analyses. In this paper we consider the *noisy power method*, a fast general purpose method for computing the dominant singular vectors of a matrix when the target matrix can only be accessed through inaccurate matrix-vector products.

Figure 1 describes the method when the target matrix A is a symmetric $d \times d$ matrix—a generalization to asymmetric matrices is straightforward. The algorithm starts from an initial matrix $X_0 \in \mathbb{R}^{d\times p}$ and iteratively attempts to perform the update rule $X_\ell \to AX_\ell$. However, each such matrix product is followed by a possibly adversarially and adaptively chosen perturbation G_ℓ leading to the update rule $X_\ell \to AX_\ell + G_\ell$. It will be convenient though not necessary to maintain that X_ℓ has orthonormal columns which can be achieved through a QR-factorization after each update.

```
Input: Symmetric matrix $A \in \mathbb{R}^{d\times d}$, number of iterations $L$, dimension $p$
  1. Choose $X_0 \in \mathbb{R}^{d\times p}$.
  2. For $\ell = 1$ to $L$:
       (a) $Y_\ell \leftarrow AX_{\ell-1} + G_\ell$ where $G_\ell \in \mathbb{R}^{d\times p}$ is some perturbation
       (b) Let $Y_\ell = X_\ell R_\ell$ be a QR-factorization of $Y_\ell$
Output: Matrix $X_L$
```

Figure 1: Noisy Power Method (NPM)

The noisy power method is a meta algorithm that when instantiated with different settings of G_ℓ and X_0 adapts to a variety of applications. In fact, there have been a number of recent surprising applications of the noisy power method:

1. Jain et al. [**?**] observe that the update rule of the well-known alternating least squares heuristic for matrix completion can be considered as an instance of NPM. This lead to the first provable convergence bounds for this important heuristic.

2. Mitgliakas et al. [**?**] observe that NPM applies to a streaming model of principal component analysis (PCA) where it leads to a space-efficient and practical algorithm for PCA in settings where the covariance matrix is too large to process directly.

3. Hardt and Roth [**?**] consider the power method in the context of privacy-preserving PCA where noise is added to achieve differential privacy.

In each setting there has so far only been an ad-hoc analysis of the noisy power method. In the first setting, only local convergence is argued, that is, X_0 has to be cleverly chosen. In the second setting, the analysis only holds for the spiked covariance model of PCA. In the third application, only the case $p = 1$ was considered.

In this work we give a completely general analysis of the noisy power method that overcomes limitations of previous analyses. Our result characterizes the global convergence properties of the algorithm in terms of the noise G_ℓ and the initial subspace X_0. We then consider the important case where X_0 is a randomly chosen orthonormal basis. This case is rather delicate since the initial correlation between a random matrix X_0 and the target subspace is vanishing in the dimension d for small p. Another important feature of the analysis is that it shows how X_ℓ converges towards the first $k \leqslant p$ singular vectors. Choosing p to be larger than the target dimension leads to a quantitatively stronger result. Theorem 2.4 formally states our convergence bound. Here we highlight one useful corollary to illustrate our more general result.

Corollary 1.1. *Let $k \leqslant p$. Let $U \in \mathbb{R}^{d\times k}$ represent the top k singular vectors of A and let $\sigma_1 \geqslant \cdots \geqslant \sigma_n \geqslant 0$ denote its singular values. Suppose X_0 is an orthonormal basis of a random p-dimensional subspace. Further suppose that at every step of* NPM *we have*

$$5\|G_\ell\| \leqslant \varepsilon(\sigma_k - \sigma_{k+1}) \quad \textit{and} \quad 5\|U^\top G_\ell\| \leqslant (\sigma_k - \sigma_{k+1})\frac{\sqrt{p}-\sqrt{k-1}}{\tau\sqrt{d}}$$

for some fixed parameter τ and $\varepsilon < 1/2$. Then with all but $\tau^{-\Omega(p+1-k)} + e^{-\Omega(d)}$ probability, there exists an $L = O(\frac{\sigma_k}{\sigma_k - \sigma_{k+1}}\log(d\tau/\varepsilon))$ so that after L steps we have that $\left\|(I - X_L X_L^\top)U\right\| \leqslant \varepsilon$.

The corollary shows that the algorithm converges in the strong sense that the entire spectral norm of U up to an ε error is contained in the space spanned by X_L. To achieve this the result places two assumptions on the magnitude of the noise. The total spectral norm of G_ℓ must be bounded by ε times the separation between σ_k and σ_{k+1}. This dependence on the singular value separation arises even in the classical perturbation theory of Davis-Kahan [**?**]. The second condition is specific to the power method and requires that the noise term is proportionally smaller when projected onto the space spanned by the top k singular vectors. This condition ensures that the correlation between X_ℓ and U that is initially very small is not destroyed by the noise addition step. If the noise term has some spherical properties (e.g. a Gaussian matrix), we expect the projection onto U to be smaller by a factor of $\sqrt{k/d}$, since the space U is k-dimensional. In the case where $p = k + \Omega(k)$ this is precisely what the condition requires. When $p = k$ the requirement is stronger by a factor of k. This phenomenon stems from the fact that the smallest singular value of a random $p \times k$ gaussian matrix behaves differently in the square and the rectangular case.

We demonstrate the usefulness of our convergence bound with several novel results in some of the aforementioned applications.

1.1 Application to memory-efficient streaming PCA

In the streaming PCA setting we receive a stream of samples $z_1, z_2, \ldots z_n \in \mathbb{R}^d$ drawn i.i.d. from an unknown distribution $\mathcal{D}$ over $\mathbb{R}^d$. Our goal is to compute the dominant k eigenvectors of the covariance matrix $A = \mathbb{E}_{z\sim\mathcal{D}} zz^\top$. The challenge is to do this in space linear in the output size, namely $O(kd)$. Recently, Mitgliakas et al. [**?**] gave an algorithm for this problem based on the noisy power method. We analyze the same algorithm, which we restate here and call SPM:

```
Input: Stream of samples $z_1, z_2, \ldots, z_n \in \mathbb{R}^d$, iterations $L$, dimension $p$
  1. Let $X_0 \in \mathbb{R}^{d\times p}$ be a random orthonormal basis. Let $T = \lfloor m/L \rfloor$
  2. For $\ell = 1$ to $L$:
      (a) Compute $Y_\ell = A_\ell X_{\ell-1}$ where $A_\ell = \sum_{i=(\ell-1)T+1}^{\ell T} z_i z_i^\top$
      (b) Let $Y_\ell = X_\ell R_\ell$ be a QR-factorization of $Y_\ell$
Output: Matrix $X_L$
```

Figure 2: Streaming Power Method (SPM)

The algorithm can be executed in space $O(pd)$ since the update step can compute the $d \times p$ matrix $A_\ell X_{\ell-1}$ incrementally without explicitly computing A_ℓ. The algorithm maps to our setting by defining $G_\ell = (A_\ell - A)X_{\ell-1}$. With this notation $Y_\ell = AX_{\ell-1} + G_\ell$. We can apply Corollary 1.1 directly once we have suitable bounds on $\|G_\ell\|$ and $\|U^\top G_\ell\|$.

The result of [?] is specific to the spiked covariance model. The spiked covariance model is defined by an orthonormal basis $U \in \mathbb{R}^{d\times k}$ and a diagonal matrix $\Lambda \in \mathbb{R}^{k\times k}$ with diagonal entries $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_k > 0$. The distribution $\mathcal{D}(U,\Lambda)$ is defined as the normal distribution $\mathrm{N}(0,(U\Lambda^2U^\top + \sigma^2\mathrm{Id}_{d\times d}))$. Without loss of generality we can scale the examples such that $\lambda_1 = 1$. One corollary of our result shows that the algorithm outputs X_L such that $\left\|(I - X_LX_L^\top)U\right\| \leqslant \varepsilon$ with probability 9/10 provided $p = k + \Omega(k)$ and the number of samples satisfies

$$n = \Theta\left(\frac{\sigma^6+1}{\varepsilon^2\lambda_k^6}kd\right).$$

Previously, the same bound[1] was known with a quadratic dependence on k in the case where $p = k$. Here we can strengthen the bound by increasing p slightly.

While we can get some improvements even in the spiked covariance model, our result is substantially more general and applies to any distribution. The sample complexity bound we get varies according to a technical parameter of the distribution. Roughly speaking, we get a near linear sample complexity if the distribution is either "round" (as in the spiked covariance setting) or is very well approximated by a k dimensional subspace. To illustrate the latter condition, we have the following result without making any assumptions other than scaling the distribution:

Corollary 1.2. *Let $\mathcal{D}$ be any distribution scaled so that $\mathbb{P}\mathrm{r}\{\|z\| > t\} \leqslant \exp(-t)$ for every $t \geqslant 1$. Let U represent the top k eigenvectors of the covariance matrix $\mathbb{E}zz^\top$ and $\sigma_1 \geqslant \cdots \geqslant \sigma_d \geqslant 0$ its eigenvalues. Then,* SPM *invoked with $p = k + \Omega(k)$ outputs a matrix X_L such with probability* 9/10 *we have $\left\|(I - X_LX_L^\top)U\right\| \leqslant \varepsilon$ provided* SPM *receives n samples where n satisfies $n = \tilde{O}\left(\frac{\sigma_k}{\varepsilon^2k(\sigma_k-\sigma_{k+1})^3}\cdot d\right)$.*

The corollary establishes a sample complexity that's linear in d provided that the spectrum decays quickly, as is common in applications. For example, if the spectrum follows a power law so that $\sigma_j \approx j^{-c}$ for a constant $c > 1/2$, the bound becomes $n = \tilde{O}(k^{2c+2}d/\varepsilon^2)$.

1.2 Application to privacy-preserving spectral analysis

Many applications of singular vector computation are plagued by the fact that the underlying matrix contains sensitive information about individuals. A successful paradigm in privacy-preserving data analysis rests on the notion of *differential privacy* which requires all access to the data set to be randomized in such a way that the presence or absence of a single data item is hidden. The notion of data item varies and could either refer to a single entry, a single row, or a rank-1 matrix of bounded norm. More formally, Differential Privacy requires that the output distribution of the algorithm changes only slightly with the addition or deletion of a single data item. This requirement often necessitates the introduction of significant levels of noise that make the computation of various objectives challenging. Differentially private singular vector computation has been studied actively since the work of Blum et al. [?]. There are two main objectives. The first is computational efficiency. The second objective is to minimize the amount of error that the algorithm introduces.

In this work, we give a fast algorithm for differentially private singular vector computation based on the noisy power method that leads to nearly optimal bounds in a number of settings

[1] That the bound stated in [?] has a σ^6 dependence is not completely obvious. There is a $O(\sigma^4)$ in the numerator and $\log((\sigma^2 + 0.75\lambda_k^2)/(\sigma^2 + 0.5\lambda_k^2))$ in the denominator which simplifies to $O(1/\sigma^2)$ for constant λ_k and $\sigma^2 \geqslant 1$.

that were considered in previous work. The algorithm is described in Figure 3. It's a simple instance of NPM in which each noise matrix G_ℓ is a gaussian random matrix scaled so that the algorithm achieves (ε,δ)-differential privacy (as formally defined in Definition 4.1). It is easy to see that the algorithm can be implemented in time nearly linear in the number of nonzero entries of the input matrix (input sparsity). This will later lead to strong improvements in running time compared with several previous works.

```
Input: Symmetric $A \in \mathbb{R}^{d\times d}$, $L$, $p$, privacy parameters $\varepsilon,\delta > 0$
  1. Let $X_0$ be a random orthonormal basis and put $\sigma = \varepsilon^{-1}\sqrt{4pL\log(1/\delta)}$
  2. For $\ell = 1$ to $L$:
      (a) $Y_\ell \leftarrow AX_{\ell-1} + G_\ell$ where $G_\ell \sim \mathrm{N}(0, \|X_{\ell-1}\|_\infty^2\sigma^2)^{d\times p}$.
      (b) Compute the QR-factorization $Y_\ell = X_\ell R_\ell$
Output: Matrix $X_L$
```

Figure 3: Private Power Method (PPM). Here $\|X\|_\infty = \max_{ij}|X_{ij}|$.

We first state a general purpose analysis of PPM that follows from Corollary 1.1.

Theorem 1.3. *Let $k \leqslant p$. Let $U \in \mathbb{R}^{d\times k}$ represent the top k singular vectors of A and let $\sigma_1 \geqslant \cdots \geqslant \sigma_d \geqslant 0$ denote its singular values. Then,* PPM *satisfies (ε,δ)-differential privacy and after $L = O(\frac{\sigma_k}{\sigma_k-\sigma_{k+1}}\log(d))$ iterations we have with probability 9/10 that*

$$\left\|(I - X_L X_L^\top)U\right\| \leqslant O\left(\frac{\sigma \max\|X_\ell\|_\infty\sqrt{d\log L}}{\sigma_k - \sigma_{k+1}} \cdot \frac{\sqrt{p}}{\sqrt{p}-\sqrt{k-1}}\right).$$

When $p = k + \Omega(k)$ the trailing factor becomes a constant. If $p = k$ it creates a factor k overhead. In the worst-case we can always bound $\|X_\ell\|_\infty$ by 1 since X_ℓ is an orthonormal basis. However, in principle we could hope that a much better bound holds provided that the target subspace U has small coordinates. Hardt and Roth [?, ?] suggested a way to accomplish a stronger bound by considering a notion of *coherence* of A, denoted as $\mu(A)$. Informally, the coherence is a well-studied parameter that varies between 1 and n, but is often observed to be small. Intuitively, the coherence measures the correlation between the singular vectors of the matrix with the standard basis. Low coherence means that the singular vectors have small coordinates in the standard basis. Many results on matrix completion and robust PCA crucially rely on the assumption that the underlying matrix has low coherence [?, ?, ?] (though the notion of coherence here will be somewhat different).

Theorem 1.4. *Under the assumptions of Theorem 1.3, we have the conclusion*

$$\left\|(I - X_L X_L^\top)U\right\| \leqslant O\left(\frac{\sigma\sqrt{\mu(A)\log d\log L}}{\sigma_k - \sigma_{k+1}} \cdot \frac{\sqrt{p}}{\sqrt{p}-\sqrt{k-1}}\right).$$

Hardt and Roth proved this result for the case where $p = 1$. The extension to $p > 1$ lost a factor of $\sqrt{d}$ in general and therefore gave no improvement over Theorem 1.3. Our result resolves the main problem left open in their work. The strength of Theorem 1.4 is that the bound is essentially dimension-free under a natural assumption on the matrix and never worse than our worst-case result. It is also known that in general the dependence on d achieved in Theorem 1.3 is best possible in the worst case (see discussion in [?]) so that further progress requires making stronger assumptions. Coherence is a natural such assumption. The proof of Theorem 1.4 proceeds by showing that each iterate X_ℓ satisfies $\|X_\ell\|_\infty \leqslant O(\sqrt{\mu(A)\log(d)/d})$ and applying Theorem 1.3. To do this we exploit a non-trivial symmetry of the algorithm that we discuss in Section 4.3.

Other variants of differential privacy. Our discussion above applied to (ε,δ)-differential privacy under changing a single entry of the matrix. Several works consider other variants of differential privacy. It is generally easy to adapt the power method to these settings by changing the noise distribution or its scaling. To illustrate this aspect, we consider the problem of privacy-preserving principal component analysis as recently studied by [**?**, **?**]. Both works consider an algorithm called *exponential mechanism*. The first work gives a heuristic implementation that may not converge, while the second work gives a provably polynomial time algorithm though the running time is more than cubic. Our algorithm gives strong improvements in running time while giving nearly optimal accuracy guarantees as it matches a lower bound of [**?**] up to a $\tilde{O}(\sqrt{k})$ factor. We also improve the error dependence on k by polynomial factors compared to previous work. Moreover, we get an accuracy improvement of $O(\sqrt{d})$ for the case of (ε,δ)-differential privacy, while these previous works only apply to $(\varepsilon,0)$-differential privacy. Section 4.2 provides formal statements.

1.3 Related Work

Numerical Analysis. One might expect that a suitable analysis of the noisy power method would have appeared in the numerical analysis literature. However, we are not aware of a reference and there are a number of points to consider. First, our noise model is adaptive thus setting it apart from the classical perturbation theory of the singular vector decomposition [**?**]. Second, we think of the perturbation at each step as large making it conceptually different from floating point errors. Third, research in numerical analysis over the past decades has largely focused on faster Krylov subspace methods. There is some theory of *inexact Krylov methods*, e.g., [**?**] that captures the effect of noisy matrix-vector products in this context. Related to our work are also results on the perturbation stability of the QR-factorization since those could be used to obtain convergence bounds for subspace iteration. Such bounds, however, must depend on the condition number of the matrix that the QR-factorization is applied to. See Chapter 19.9 in [**?**] and the references therein for background. Our proof strategy avoids this particular dependence on the condition number.

Streaming PCA. PCA in the streaming model is related to a host of well-studied problems that we cannot survey completely here. We refer to [**?**, **?**] for a thorough discussion of prior work. Not mentioned therein is a recent work on incremental PCA [**?**] that leads to space efficient algorithms computing the top singular vector; however, it's not clear how to extend their results to computing multiple singular vectors.

Privacy. There has been much work on differentially private spectral analysis starting with Blum et al. [**?**] who used an algorithm known as *Randomized Response* which adds a single noise matrix N either to the input matrix A or the covariance matrix $AA^\top$. This approach appears in a number of papers, e.g. [**?**]. While often easy to analyze it has the disadvantage that it converts sparse matrices to dense matrices and is often impractical on large data sets. Chaudhuri et al. [?] and Kapralov-Talwar [?] use the so-called *exponential mechanism* to sample approximate eigenvectors of the matrix. The sampling is done using a heuristic approach without convergence polynomial time convergence guarantees in the first case and using a polynomial time algorithm in the second. Both papers achieve a tight dependence on the matrix dimension d (though the dependence on k is suboptimal in general). Most closely related to our work are the results of Hardt and Roth [?, ?] that introduced matrix coherence as a way to circumvent existing worst-case lower bounds on the error. They also analyzed a natural noisy variant of power iteration for the case of computing the dominant eigenvector of A. When multiple eigenvectors are needed, their algorithm uses the well-known deflation technique. However, this step loses control of the coherence of the original matrix and hence results in suboptimal bounds. In fact, a $\sqrt{\mathrm{rank}(A)}$ factor is lost.

1.4 Organization

All proofs can be found in the supplementary material. In the remaining space, we limit ourselves to a more detailed discussion of our convergence analysis and the application to streaming PCA. The entire section on privacy is in the supplementary materials in Section 4.

2 Convergence of the noisy power method

Figure 1 presents our basic algorithm that we analyze in this section. An important tool in our analysis are principle angles, which are useful in analyzing the convergence behavior of numerical eigenvalue methods. Roughly speaking, we will show that the tangent of the k-th principal angle between X and the top k eigenvectors of A decreases as σ_{k+1}/σ_k in each iteration of the noisy power method.

Definition 2.1 (Principal angles)**.** Let $\mathcal{X}$ and $\mathcal{Y}$ be subspaces of $\mathbb{R}^d$ of dimension at least k. The *principal angles* $0 \leqslant \theta_1 \leqslant \cdots \leqslant \theta_k$ between $\mathcal{X}$ and $\mathcal{Y}$ and associated *principal vectors* $x_1, \ldots, x_k$ and $y_1, \ldots, y_k$ are defined recursively via

$$\theta_i(\mathcal{X}, \mathcal{Y}) = \min \left\{ \arccos\left(\frac{\langle x, y \rangle}{\|x\|_2 \|y\|_2} \right) : x \in \mathcal{X}, y \in \mathcal{Y}, x \perp x_j, y \perp y_j \text{ for all } j < i \right\}$$

and x_i, y_i are the x and y that give this value. For matrices X and Y, we use $\theta_k(X, Y)$ to denote the kth principal angle between their ranges.

2.1 Convergence argument

Fix parameters $1 \leqslant k \leqslant p \leqslant d$. In this section we consider a symmetric $d \times d$ matrix A with singular values $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_d$. We let $U \in \mathbb{R}^{d \times k}$ contain the first k eigenvectors of A. Our main lemma shows that $\tan \theta_k(U, X)$ decreases multiplicatively in each step.

Lemma 2.2. *Let U contain the largest k eigenvectors of a symmetric matrix $A \in \mathbb{R}^{d \times d}$, and let $X \in \mathbb{R}^{d \times p}$ for $p \geqslant k$. Let $G \in \mathbb{R}^{d \times p}$ satisfy*

$$\begin{aligned} 4\|U^\top G\| &\leqslant (\sigma_k - \sigma_{k+1}) \cos \theta_k(U, X) \\ 4\|G\| &\leqslant (\sigma_k - \sigma_{k+1}) \varepsilon. \end{aligned}$$

for some $\varepsilon < 1$. Then

$$\tan\theta_k(U, AX+G) \leqslant \max\left(\varepsilon, \max\left(\varepsilon, \left(\frac{\sigma_{k+1}}{\sigma_k}\right)^{1/4}\right)\tan\theta_k(U,X)\right).$$

We can inductively apply the previous lemma to get the following general convergence result.

Theorem 2.3. *Let U represent the top k eigenvectors of the matrix A and $\gamma = 1 - \sigma_{k+1}/\sigma_k$. Suppose that the initial subspace X_0 and noise G_ℓ is such that*

$$\begin{aligned} 5\|U^\top G_\ell\| &\leqslant (\sigma_k - \sigma_{k+1})\cos\theta_k(U, X_0) \\ 5\|G_\ell\| &\leqslant \varepsilon(\sigma_k - \sigma_{k+1}) \end{aligned}$$

at every stage ℓ, for some $\varepsilon < 1/2$. Then there exists an $L \lesssim \frac{1}{\gamma}\log\left(\frac{\tan\theta_k(U,X_0)}{\varepsilon}\right)$ such that for all $\ell \geqslant L$ we have $\tan\theta(U, X_L) \leqslant \varepsilon$.

2.2 Random initialization

The next lemma essentially follows from bounds on the smallest singular value of gaussian random matrices [**?**].

Claim 2.4. *For an arbitrary orthonormal U and random subspace X, we have*

$$\tan\theta_k(U,X) \leqslant \tau\frac{\sqrt{d}}{\sqrt{p} - \sqrt{k-1}}$$

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

Proof of Corollary 1.1. By Claim 2.5, with the desired probability we have $\tan\theta_k(U, X_0) \leqslant \frac{\tau\sqrt{d}}{\sqrt{p}-\sqrt{k-1}}$. Hence $\cos\theta_k(U, X_0) \geqslant 1/(1 + \tan\theta_k(U, X_0)) \geqslant \frac{\sqrt{p}-\sqrt{k-1}}{2\cdot\tau\sqrt{d}}$. Rescale τ and apply Theorem 2.4 to get that $\tan\theta_k(U, X_L) \leqslant \varepsilon$. Then $\|(I - X_L X_L^\top)U\| = \sin\theta_k(U, X_L) \leqslant \tan\theta_k(U, X_L) \leqslant \varepsilon$. ■

3 Memory efficient streaming PCA

In the streaming PCA setting we receive a stream of samples $z_1, z_2, \cdots \in \mathbb{R}^d$. Each sample is drawn i.i.d. from an unknown distribution $\mathcal{D}$ over $\mathbb{R}^d$. Our goal is to compute the dominant k eigenvectors of the covariance matrix $A = \mathbb{E}_{z\sim\mathcal{D}} zz^\top$. The challenge is to do this with small space, so we cannot store the d^2 entries of the sample covariance matrix. We would like to use $O(dk)$ space, which is necessary even to store the output.

The streaming power method (Figure 2, introduced by [**?**]) is a natural algorithm that performs streaming PCA with $O(dk)$ space. The question that arises is how many samples it requires to achieve a given level of accuracy, for various distributions $\mathcal{D}$. Using our general analysis of the noisy power method, we show that the streaming power method requires fewer samples and applies to more distributions than was previously known.

We analyze a broad class of distributions:

Definition 3.1. A distribution $\mathcal{D}$ over $\mathbb{R}^d$ is (B,p)*-round* if for every p-dimensional projection P and all $t \geqslant 1$ we have $\mathbb{P}\mathrm{r}_{z\sim\mathcal{D}}\{\|z\| > t\} \leqslant \exp(-t)$ and $\mathbb{P}\mathrm{r}_{z\sim\mathcal{D}}\left\{\|Pz\| > t \cdot \sqrt{Bp/d}\right\} \leqslant \exp(-t)$.

The first condition just corresponds to a normalization of the samples drawn from $\mathcal{D}$. Assuming the first condition holds, the second condition always holds with $B = d/p$. For this reason our analysis in principle applies to any distribution, but the sample complexity will depend quadratically on B.

Let us illustrate this definition through the example of the spiked covariance model studied by [**?**]. The spiked covariance model is defined by an orthonormal basis $U \in \mathbb{R}^{d\times k}$ and a diagonal matrix $\Lambda \in \mathbb{R}^{k\times k}$ with diagonal entries $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_k > 0$. The distribution $\mathcal{D}(U,\Lambda)$ is defined as the normal distribution $\mathrm{N}(0,(U\Lambda^2U^\top + \sigma^2\mathrm{Id}_{d\times d})/D)$ where $D = \Theta(d\sigma^2 + \sum_i \lambda_i^2)$ is a normalization factor chosen so that the distribution satisfies the norm bound. Note that the the i-th eigenvalue of the covariance matrix is $\sigma_i = (\lambda_i^2 + \sigma^2)/D$ for $1 \leqslant i \leqslant k$ and $\sigma_i = \sigma^2/D$ for $i > k$. We show in Lemma 3.6 that the spiked covariance model $\mathcal{D}(U,\Lambda)$ is indeed (B,p)-round for $B = O(\frac{\lambda_1^2+\sigma^2}{\mathrm{tr}(\Lambda)/d+\sigma^2})$, which is constant for $\sigma \gtrsim \lambda_1$.

We have the following main theorem.

Theorem 3.2. *Let $\mathcal{D}$ be a (B,p)-round distribution over $\mathbb{R}^d$ with covariance matrix A whose eigenvalues are $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_d \geqslant 0$. Let $U \in \mathbb{R}^{d\times k}$ be an orthonormal basis for the eigenvectors corresponding to the first k eigenvalues of A. Then, the streaming power method* SPM *returns an orthonormal basis $X \in \mathbb{R}^{d\times p}$ such that $\tan\theta(U,X) \leqslant \varepsilon$ with probability $9/10$ provided that* SPM *receives n samples from $\mathcal{D}$ for some n satisfying*

$$n \leqslant \tilde{O}\left(\frac{B^2\sigma_k k\log^2 d}{\varepsilon^2(\sigma_k - \sigma_{k+1})^3 d}\right)$$

if $p = k + \Theta(k)$. More generally, for all $p \geqslant k$ one can get the slightly stronger result

$$n \leqslant \tilde{O}\left(\frac{Bp\sigma_k \max\{1/\varepsilon^2, Bp/(\sqrt{p} - \sqrt{k-1})^2\}\log^2 d}{(\sigma_k - \sigma_{k+1})^3 d}\right).$$

Instantiating with the spiked covariance model gives the following:

Corollary 3.3. *In the spiked covariance model $\mathcal{D}(U,\Lambda)$ the conclusion of Theorem 3.2 holds for $p = 2k$ with*

$$n = \tilde{O}\left(\frac{(\lambda_1^2 + \sigma^2)^2(\lambda_k^2 + \sigma^2)}{\varepsilon^2\lambda_k^6} dk\right).$$

When $\lambda_1 = O(1)$ and $\lambda_k = \Omega(1)$ this becomes $n = \tilde{O}\left(\frac{\sigma^6+1}{\varepsilon^2} \cdot dk\right)$.

We can apply Theorem 3.2 to all distributions that have exponentially concentrated norm by setting $B = d/p$. This gives

Corollary 3.4. *Let $\mathcal{D}$ be any distribution scaled such that $\mathrm{Pr}_{z\sim\mathcal{D}}[\|z\| > t] \leqslant \exp(-t)$ for all $t \geqslant 1$. Then the conclusion of Theorem 3.2 holds for $p = 2k$ with*

$$n = \tilde{O}\left(\frac{\sigma_k}{\varepsilon^2 k(\sigma_k - \sigma_{k+1})^3} \cdot d\right).$$

If the eigenvalues follow a power law, $\sigma_j \approx j^{-c}$ for a constant $c > 1/2$, this gives an $n = \tilde{O}(k^{2c+2}d/\varepsilon^2)$ bound on the sample complexity.

A Proofs from the convergence analyis

We will make use of a non-recursive expression for the principal angles, defined in terms of the set $\mathcal{P}_k$ of $p \times p$ projection matrices Π from p dimensions to k dimensional subspaces:

Claim A.1. *Let $U \in \mathbb{R}^{d\times k}$ have orthonormal columns and $X \in \mathbb{R}^{d\times p}$ have independent columns, for $p \geqslant k$. Then*

$$\cos\theta_k(U,X) = \min_{\Pi\in\mathcal{P}_k} \min_{\substack{x\in\mathrm{range}(X\Pi)\\ \|x\|_2=1}} \|U^\top x\| = \min_{\Pi\in\mathcal{P}_k} \min_{\substack{\|w\|_2=1\\ \Pi w=w}} \frac{\|U^\top Xw\|}{\|Xw\|}.$$

For $V = U^\perp$, we have

$$\tan\theta_k(U,X) = \min_{\Pi\in\mathcal{P}_k} \max_{x\in\mathrm{range}(X\Pi)} \frac{\|V^\top x\|}{\|U^\top x\|} = \min_{\Pi\in\mathcal{P}_k} \max_{\substack{\|w\|_2=1\\ \Pi w=w}} \frac{\|V^\top Xw\|}{\|U^\top Xw\|}.$$

*Proof of Lemma **??**.* Let Π^* be the matrix projecting onto the smallest k principal angles of X, so that

$$\tan\theta_k(U,X) = \max_{\substack{\|w\|_2=1\\ \Pi^* w=w}} \frac{\|V^\top Xw\|}{\|U^\top Xw\|}.$$

We have that

$$\begin{aligned}
\tan\theta_k(U,AX+G) &= \min_{\Pi\in\mathcal{P}_k} \max_{\substack{\|w\|_2=1\\ \Pi w=w}} \frac{\|V^\top(AX+G)w\|}{\|U^\top(AX+G)w\|}\\
&\leqslant \max_{\substack{\|w\|_2=1\\ \Pi^* w=w}} \frac{\|V^\top AXw\| + \|V^\top Gw\|}{\|U^\top AXw\| - \|U^\top Gw\|}\\
&\leqslant \max_{\substack{\|w\|_2=1\\ \Pi^* w=w}} \frac{1}{\|U^\top Xw\|} \cdot \frac{\sigma_{k+1}\|V^\top Xw\| + \|V^\top Gw\|}{\sigma_k - \|U^\top Gw\|/\|U^\top Xw\|} \qquad (1)
\end{aligned}$$

Define $\Delta = (\sigma_k - \sigma_{k+1})/4$. By the assumption on G,

$$\max_{\substack{\|w\|_2=1\\ \Pi^* w=w}} \frac{\|U^\top Gw\|}{\|U^\top Xw\|} \leqslant \|U^\top G\|/\cos\theta_k(U,X) \leqslant (\sigma_k - \sigma_{k+1})/4 = \Delta.$$

Similarly, and using that $1/\cos\theta \leqslant 1 + \tan\theta$ for any angle θ,

$$\max_{\substack{\|w\|_2=1\\ \Pi^* w=w}} \frac{\|V^\top Gw\|}{\|U^\top Xw\|} \leqslant \|G\|/\cos\theta_k(U,X) \leqslant \varepsilon\Delta(1+\tan\theta_k(U,X)).$$

Plugging back into (1) and using $\sigma_k = \sigma_{k+1} + 4\Delta$,

$$\begin{aligned}
\tan\theta_k(U,AX+G) &\leqslant \max_{\substack{\|w\|_2=1\\ \Pi^* w=w}} \frac{\|V^\top Xw\|}{\|U^\top Xw\|} \cdot \frac{\sigma_{k+1}}{\sigma_{k+1}+3\Delta} + \frac{\varepsilon\Delta(1+\tan\theta_k(U,X))}{\sigma_{k+1}+3\Delta}.\\
&= \frac{\sigma_{k+1}+\varepsilon\Delta}{\sigma_{k+1}+3\Delta}\tan\theta_k(U,X) + \frac{\varepsilon\Delta}{\sigma_{k+1}+3\Delta}\\
&= (1 - \frac{\Delta}{\sigma_{k+1}+3\Delta})\frac{\sigma_{k+1}+\varepsilon\Delta}{\sigma_{k+1}+2\Delta}\tan\theta_k(U,X) + \frac{\Delta}{\sigma_{k+1}+3\Delta}\varepsilon\\
&\leqslant \max(\varepsilon, \frac{\sigma_{k+1}+\varepsilon\Delta}{\sigma_{k+1}+2\Delta}\tan\theta_k(U,X))
\end{aligned}$$

where the last inequality uses that the weighted mean of two terms is less than their maximum. Finally, we have that

$$\frac{\sigma_{k+1} + \varepsilon\Delta}{\sigma_{k+1} + 2\Delta} \leqslant \max(\frac{\sigma_{k+1}}{\sigma_{k+1} + \Delta}, \varepsilon)$$

because the left hand side is a weighted mean of the components on the right. Since $\frac{\sigma_{k+1}}{\sigma_{k+1}+\Delta} \leqslant (\frac{\sigma_{k+1}}{\sigma_{k+1}+4\Delta})^{1/4} = (\sigma_{k+1}/\sigma_k)^{1/4}$, this gives the result. ■

Proof of Theorem 2.4. We will see that at every stage ℓ of the algorithm,

$$\tan\theta_k(U, X_\ell) \leqslant \max(\varepsilon, \tan\theta_k(U, X_0))$$

which implies for $\varepsilon \leqslant 1/2$ that

$$\cos\theta_k(U, X_\ell) \geqslant \min(1 - \varepsilon^2/2, \cos\theta_k(U, X_0)) \geqslant \frac{7}{8}\cos\theta_k(U, X_0)$$

so Lemma 2.3 applies at every stage. This means that

$$\tan\theta_k(U, X_{\ell+1}) = \tan\theta_k(U, AX_\ell + G) \leqslant \max(\varepsilon, \delta\tan\theta_k(U, X_\ell))$$

for $\delta = \max(\varepsilon, (\sigma_{k+1}/\sigma_k)^{1/4})$. After

$$L = \log_{1/\delta}\frac{\tan\theta_k(U, X_0)}{\varepsilon}$$

iterations the tangent will reach ε and remain there. Observing that

$$\log(1/\delta) \gtrsim \min(\log(1/\varepsilon), \log(\sigma_k/\sigma_{k+1})) \geqslant \min(1, \log\frac{1}{1-\gamma}) \geqslant \min(1, \gamma) = \gamma$$

gives the result. ■

A.1 Random initialization

Proof of Lemma **??**. Consider the singular value decomposition $U^\top X = A\Sigma B^\top$ of $U^\top X$. Setting Π to be matrix projecting onto the first k columns of B, we have that

$$\tan\theta_k(U, X) \leqslant \max_{\substack{\|w\|_2=1 \\ \Pi w = w}} \frac{\|V^\top X w\|}{\|U^\top X w\|} \leqslant \|V^\top X\| \max_{\substack{\|w\|_2=1 \\ \Pi w = w}} \frac{1}{\|\Sigma B^\top w\|} = \|V^\top X\| \max_{\substack{\|w\|_2=1 \\ \mathrm{supp}(w)\in[k]}} \frac{1}{\|\Sigma w\|} = \frac{\|V^\top X\|}{\sigma_k(U^\top X)}.$$

Let $X \sim N(0, I_{d\times p})$ represent the random subspace. Then $Y := U^\top X \sim N(0, I_{k\times p})$. By [?], for any ε, the smallest singular value of Y is at least $(\sqrt{p} - \sqrt{k-1})/\tau$ with all but $\tau^{-\Omega(p+1-k)} + e^{-\Omega(p)}$ probability. On the other hand, $\|X\| \lesssim \sqrt{d}$ with all but $e^{-\Omega(d)}$ probability. Hence

$$\tan\theta_k(U, X) \lesssim \tau\frac{\sqrt{d}}{\sqrt{p} - \sqrt{k-1}}$$

with the desired probability. Rescaling τ gets the result. ■

B Proofs for streaming PCA

B.1 Error term analysis

Fix an orthonormal basis $X \in \mathbb{R}^{d\times k}$. Let $z_1,\dots,z_n \sim \mathcal{D}$ be samples from a distribution $\mathcal{D}$ with covariance matrix A and consider the matrix

$$G = \left(A - \widehat{A}\right)X\,,$$

where $\widehat{A} = \frac{1}{n}\sum_{i=1}^n z_i z_i^\top$ is the empirical covariance matrix on n samples. Then, we have that $\widehat{A}X = AX + G$. In other words, one update step of the power method executed on $\widehat{A}$ can be expressed as an update step on A with noise matrix G. This simple observation allows us to apply our analysis of the noisy power method to this setting after obtaining suitable bounds on $\|G\|$ and $\|U^\top G\|$.

Lemma B.1. *Let $\mathcal{D}$ be a (B,p)-round distribution with covariance matrix M. Then with all but $O(1/n^2)$ probability,*

$$\|G\| \lesssim \sqrt{\frac{Bp\log^4 n\log d}{dn}} + \frac{1}{n^2} \quad \textit{and} \quad \|U^\top G\| \lesssim \sqrt{\frac{B^2p^2\log^4 n\log d}{d^2 n}} + \frac{1}{n^2}$$

Proof. We will use a matrix Chernoff bound to show that

1. $\Pr\left\{\|G\| > Ct\log(n)^2\sqrt{Bp/d} + O(1/n^2)\right\} \leqslant d\exp(-t^2 n) + 1/n^2$
2. $\Pr\left\{\|U^\top G\| > Ct\log(n)^2 Bp/d + O(1/n^2)\right\} \leqslant d\exp(-t^2 n) + 1/n^2$

setting $t = \sqrt{\frac{2}{n}\log d}$ gives the result. However, matrix Chernoff inequality requires the distribution to satisfy a norm bound with probability 1. We will therefore create a closely related distribution $\tilde{\mathcal{D}}$ that satisfies such a norm constraint and is statistically indistinguishable up to small error on n samples. We can then work with $\tilde{\mathcal{D}}$ instead of $\mathcal{D}$. This truncation step is standard and works because of the concentration properties of $\mathcal{D}$.

Indeed, let $\tilde{\mathcal{D}}$ be the distribution obtained from $\mathcal{D}$ be replacing a sample z with 0 if

$$\|z\| > C\log(n) \quad \text{or} \quad \|U^\top z\| \geqslant C\log(n)\sqrt{Bp/d} \quad \text{or} \quad \|z^\top X\| > C\log(n)\sqrt{Bp/d}\,.$$

For sufficiently large constant C, it follows from the definition of (B,p)-round that the probability that one or more of n samples from $\mathcal{D}$ get zeroed out is at most $1/n^2$. In particular, the two product distributions $\mathcal{D}^{(n)}$ and $\tilde{\mathcal{D}}^{(n)}$ have total variation distance at most $1/n^2$. Furthermore, we claim that the covariance matrices of the two distributions are at most $O(1/n^2)$ apart in spectral norm. Formally,

$$\left\|\mathop{\mathbb{E}}_{z\sim\mathcal{D}} zz^\top - \mathop{\mathbb{E}}_{\tilde{z}\sim\tilde{\mathcal{D}}} \tilde{z}\tilde{z}^\top\right\| \leqslant \frac{1}{n^2}\cdot O\left(\int_{t\geqslant 1} C^2t^2\log^2(n)\exp(-t)\mathrm{d}t\right) \leqslant O(1/n^2)\,.$$

In the first inequality we use the fact that z only gets zeroed out with probability $1/n^2$. Conditional on this event, the norm of z is larger than $tC\log(n)$ with probability at most $n^2\exp(-\frac{1}{2}tC\log n) \leqslant \exp(-t)$. Assuming the norm is at most $tC\log(n)$ we have $\|zz^\top\| \leqslant t^2C^2\log^2(n)$ and this bounds the contribution to the spectral norm of the difference.

Now let $\tilde{G}$ be the error matrix defined as G except that we replace the samples $z_1,\dots,z_n$ by n samples $\tilde{z}_1,\dots,\tilde{z}_n$ from the truncated distribution $\tilde{\mathcal{D}}$. By our preceding discussion, it now suffices to show that

1. $\mathbb{P}\mathrm{r}\left\{\|\tilde{G}\| > Ct\log^2(n)\sqrt{Bp/d}\right\} \leqslant d\exp(-t^2 n)$
2. $\mathbb{P}\mathrm{r}\left\{\|U^\top \tilde{G}\| > Ct\log^2(n)Bp/d\right\} \leqslant d\exp(-t^2 n)$

To see this, let $S_i = \tilde{z}_i\tilde{z}_i^\top X$. We have

$$\|S_i\| \leqslant \|\tilde{z}_i\| \cdot \left\|\tilde{z}_i^\top X\right\| \leqslant C^2\log^2(n)\cdot\sqrt{Bp/d}$$

Similarly,

$$\left\|U^\top S_i\right\| \leqslant \|U^\top \tilde{z}_i\| \cdot \left\|\tilde{z}_i^\top X\right\| \leqslant C^2\log^2(n)\cdot\frac{Bp}{d}.$$

The claims now follow directly from the matrix Chernoff bound stated in Lemma A.4. ■

B.2 Proof of Theorem 3.2

Given Lemma 3.5 we will choose n such that the error term in each iteration satisfies the assumptions of Theorem 2.4. Let G_ℓ denote the instance of the error term G arising in the ℓ-th iteration of the algorithm. We can find an n satisfying

$$\frac{n}{\log(n)^4} = O\left(\frac{Bp\max\left\{1/\varepsilon^2, Bp/(\sqrt{p}-\sqrt{k-1})^2\right\}\log d}{(\sigma_k-\sigma_{k+1})^2 d}\right)$$

such that by Lemma 3.5 we have that with probability $1-O(1/n^2)$,

$$\|G_\ell\| \leqslant \frac{\varepsilon(\sigma_k-\sigma_{k+1})}{5} \quad\text{and}\quad \|U^\top G_\ell\| \leqslant \frac{\sigma_k-\sigma_{k+1}}{5}\frac{\sqrt{p}-\sqrt{k-1}}{\sqrt{d}}.$$

Here we used that by definition $1/n \ll \varepsilon$ and $1/n \ll \sigma_k - \sigma_{k+1}$ and so the $1/n^2$ term in Lemma 3.5 is of lower order.

With this bound, it follows from Theorem 2.4 that after $L = O(\log(d/\varepsilon)/(1-\sigma_{k+1}/\sigma_k))$ iterations we have with probability $1-\max\{1, L/n^2\}$ that $\tan\theta(U, X_L) \leqslant \varepsilon$. The over all sample complexity is therefore

$$Ln = \tilde{O}\left(\frac{Bp\sigma_k\max\left\{1/\varepsilon^2, Bp/(\sqrt{p}-\sqrt{k-1})^2\right\}\log^2 d}{(\sigma_k-\sigma_{k+1})^3 d}\right).$$

Here we used that $1-\sigma_{k+1}/\sigma_k = (\sigma_k-\sigma_{k+1})/\sigma_k$. This concludes the proof of Theorem 3.2.

B.3 Proof of Lemma 3.6 and Corollary 3.4

Lemma B.2. *The spiked covariance model $\mathcal{D}(U,\Lambda)$ is (B,k)-round for $B = O(\frac{\lambda_1^2+\sigma^2}{\mathrm{tr}(\Lambda)/d+\sigma^2})$.*

Proof. Note that an example $z \sim \mathcal{D}(U,\Lambda)$ is distributed as $U\Lambda g + g'$ where $g \sim \mathrm{N}(0,1/D)^k$ is a standard gaussian and $g' \sim \mathrm{N}(0,\sigma^2/D)^d$. is a noise term. Recall, that D is the normalization term. Let P be any projection operator onto a k-dimensional space. Then,

$$\|Pz\| = \|PU\Lambda g + Pg'\| \leqslant \|PU\Lambda g\| + \|Pg'\| \leqslant \|\Lambda g\| + \|Pg'\| \leqslant \lambda_1\|g\| + \|Pg'\|.$$

By rotational invariance of g', we may assume that P is the projection onto the first k coordinates. Hence, $\|Pg'\|$ is distributed like the norm of $\mathrm{N}(0,\sigma^2/D)^k$. Using standard tail bounds for the norm of a gaussian random variables, we can see that $\|Pz\|^2 = O(t(k\lambda_1^2 + k\sigma^2)/D)$ with probability $1-\exp(-t)$. On the other hand, $D = \Theta(\sum_{i=1}^k \lambda_i^2 + d\sigma^2)$. We can now solve for B by setting

$$\Theta(\frac{k\lambda_1^2 + k\sigma^2}{\sum_{i=1}^k \lambda_i^2 + d\sigma^2}) = \frac{Bk}{d} \quad \Leftrightarrow \quad B = \Theta(\frac{\lambda_1^2 + \sigma^2}{\frac{1}{d}\sum_{i=1}^k \lambda_i^2 + \sigma^2}).$$

■

Corollary 3.4 follows by plugging in the bound on B and the eigenvalues of the covariance matrix into our main theorem.

Proof of Corollary 3.4. In the spiked covariance model $\mathcal{D}(U,\Lambda)$ we have

$$B = \frac{\lambda_1^2 + \sigma^2}{D}, \quad \sigma_k = \frac{\lambda_k^2 + \sigma^2}{D}, \quad \sigma_{k+1} = \frac{\sigma^2}{D}, \quad D = O(\mathrm{tr}(\Lambda^2) + d\sigma^2).$$

Hence,

$$\frac{B^2\sigma_k}{(\sigma_k - \sigma_{k+1})^3 d} = \frac{(\lambda_1^2+\sigma^2)^2(\lambda_k^2+\sigma^2)}{\lambda_k^6 d} \leqslant \frac{(\lambda_1^2+\sigma^2)^3}{\lambda_k^6 d}$$

Plugging this bound into Theorem 3.2 gives Corollary 3.4. ■

C Privacy-preserving singular vector computation

In this section we prove our results about privacy-preserving singular vector computation. We begin with a standard definition of differential privacy, sometimes referred to as *entry-level differential privacy*, as it hides the presence or absence of a single entry.

Definition C.1 (Differential Privacy)**.** A randomized algorithm $M\colon \mathbb{R}^{d\times d'} \to R$ (where R is some arbitrary abstract range) is *(ε,δ)-differentially private* if for all pairs of matrices $A, A' \in \mathbb{R}^{d\times d'}$ differing in only one entry by at most 1 in absolute value, we have that for all subsets of the range $S \subseteq R$, the algorithm satisfies: $\mathbb{P}\mathrm{r}\{M(A)\in S\} \leqslant \exp(\varepsilon)\mathbb{P}\mathrm{r}\{M(A')\in S\} + \delta$.

The definition is most meaningful when A has entries in $[0,1]$ so that the above definition allows for a single entry to change arbitrarily within this range. However, this is not a requirement for us. The privacy guarantee can be strengthened by decreasing $\varepsilon > 0$.

For our choice of σ in Figure 3 the algorithm satisfies (ε,δ)-differential privacy as follows easily from properties of the Gaussian distribution:

Claim C.2. PPM *satisfies (ε,δ)-differential privacy.*

See, for example, [**?**] for a proof.

It is straightforward to prove Theorem 1.3 by invoking our convergence analysis of the noisy power method together with suitable error bounds. The error bounds are readily available as the noise term is just gaussian.

Proof of Theorem 1.3. Let $m = \max\|X_\ell\|_\infty$. By Lemma A.2 the following bounds hold with probability 99/100:

1. $\max_{\ell=1}^{L} \|G_\ell\| \lesssim \sigma m \sqrt{d \log L}$
2. $\max_{\ell=1}^{L} \|U^\top G_\ell\| \lesssim \sigma m \sqrt{k \log L}$

Let

$$\varepsilon' = \frac{\sigma m \sqrt{d \log L}}{\sigma_k - \sigma_{k+1}} \gtrsim \frac{5 \max_{\ell=1}^{L} \|G_\ell\|}{\sigma_k - \sigma_{k+1}}.$$

By Corollary 1.1, if we also have that $\max_{\ell=1}^{L} \|U^\top G_\ell\| \leqslant (\sigma_k - \sigma_{k+1}) \frac{\sqrt{p} - \sqrt{k-1}}{\tau \sqrt{d}}$ for a sufficiently large constant τ, then we will have that

$$\|(I - X_L X_L^\top) U\| \leqslant \varepsilon' \leqslant \frac{\sigma m \sqrt{d \log L}}{\sigma_k - \sigma_{k+1}}$$

after the desired number of iterations, giving the theorem. Otherwise,

$$(\sigma_k - \sigma_{k+1}) \frac{\sqrt{p} - \sqrt{k-1}}{\tau \sqrt{d}} \leqslant \max_{\ell=1}^{L} \|U^\top G_\ell\| \lesssim \varepsilon' (\sigma_k - \sigma_{k+1}) \sqrt{k/d},$$

so it is trivially true that

$$\frac{\sigma m \sqrt{d \log L}}{\sigma_k - \sigma_{k+1}} \frac{\sqrt{p}}{\sqrt{p} - \sqrt{k-1}} \geqslant \varepsilon' \frac{\sqrt{k}}{\sqrt{p} - \sqrt{k-1}} \gtrsim 1 \geqslant \|(I - X_L X_L^\top) U\|.$$

■

C.1 Low-rank approximation

Our results readily imply that we can compute accurate differentially private low-rank approximations. The main observation is that, assuming X_L and U have the same dimension, $\tan \theta(U, X_L) \leqslant \alpha$ implies that the matrix X_L also leads to a good low-rank approximation for A in the spectral norm. In particular

$$\|(I - X_L X_L^\top) A\| \leqslant \sigma_{k+1} + \alpha \sigma_1 . \tag{2}$$

Moreover the projection step of computing $X_L X_L^\top A$ can be carried out easily in a privacy-preserving manner. It is again the ℓ_∞-norm of the columns of X_L that determine the magnitude of noise that is needed. Since A is symmetric, we have $X^\top A = (AX)^\top$. Hence, to obtain a good low-rank approximation it suffices to compute the product AX_L privately as $AX_L + G_L$. This leads to the following corollary.

Corollary C.3. *Let $A \in \mathbb{R}^{d \times d}$ be a symmetric matrix with singular values $\sigma_1 \geqslant \ldots \geqslant \sigma_d$ and let $\gamma = 1 - \sigma_{k+1}/\sigma_k$. There is an (ε, δ)-differentially private algorithm that given A and k, outputs a rank 2k matrix B such that with probability* 9/10,

$$\|A - B\| \leqslant \sigma_{k+1} + \tilde{O}\left(\frac{\sigma_1 \sqrt{(k/\gamma) d \log d \log(1/\delta)}}{\varepsilon (\sigma_k - \sigma_{k+1})} \right).$$

The $\tilde{O}$-notation hides the factor $O\left(\sqrt{\log(\log(d)/\gamma)}\right)$.

Proof. Apply Theorem 1.3 with $p = 2k$ and run the algorithm for $L+1$ steps with $L = O(\gamma^{-1}\log d)$. This gives the bound

$$\alpha = \|(I - X_L X_L^\top)A\| \leqslant O\left(\frac{\sqrt{(k/\gamma)d\log d\log(\log(d)/\gamma)\log(1/\delta)}}{\varepsilon(\sigma_k - \sigma_{k+1})}\right).$$

Moreover, the algorithm has computed $Y_{L+1} = AX_L + G_L$ and we have $B = X_L Y_{L+1}^\top = X_L X_L^\top A + X_L G_L^\top$. Therefore

$$\|A - B\| \leqslant \sigma_{k+1} + \alpha\sigma_1 + \left\|X_L G_L^\top\right\|$$

where $\left\|X_L G_L^\top\right\| \leqslant \|G_L\|$. By definition of the algorithm and Lemma A.2, we have

$$\|G_L\| \leqslant O\left(\sqrt{\sigma^2 d}\right) = O\left(\frac{1}{\varepsilon}\sqrt{(k/\gamma)d\log(d)\log(1/\delta)}\right).$$

Given that the α-term gets multiplied by σ_1, this bound on $\|G_L\|$ is of lower order and the corollary follows. ■

C.2 Principal Component Analysis

Here we illustrate that our bounds directly imply results for the privacy notion studied by Kapralov and Talwar [**?**]. The notion is particularly relevant in a setting where we think of A as a sum of rank 1 matrices each of bounded spectral norm.

Definition C.4. A randomized algorithm $M\colon \mathbb{R}^{d\times d'} \to R$ (where R is some arbitrary abstract range) is (ε,δ)-*differentially private under unit spectral norm changes* if for all pairs of matrices $A, A' \in \mathbb{R}^{d\times d'}$ satisfying $\|A - A'\|_2 \leqslant 1$, we have that for all subsets of the range $S \subseteq R$, the algorithm satisfies: $\mathbb{P}\mathrm{r}\{M(A) \in S\} \leqslant \exp(\varepsilon)\mathbb{P}\mathrm{r}\{M(A') \in S\} + \delta$.

Lemma C.5. *If* PPM *is executed with each G_ℓ sampled independently as $G_\ell \sim N(0,\sigma^2)^{d\times p}$ with $\sigma = \varepsilon^{-1}\sqrt{4pL\log(1/\delta)}$, then* PPM *satisfies (ε,δ)-differential privacy under unit spectral norm changes.*

If G_ℓ is sampled with i.i.d. Laplacian entries $G_\ell \sim \mathrm{Lap}(0,\lambda)^{n\times k}$ where $\lambda = 10\varepsilon^{-1}pL\sqrt{d}$, then PPM *satisfies $(\varepsilon, 0)$-differential privacy under unit spectral norm changes.*

Proof. The first claim follows from the privacy proof in [**?**]. We sketch the argument here for completeness. Let D be any matrix with $\|D\|_2 \leqslant 1$ (thought of as $A - A'$ in Definition 4.4) and let $\|x\| = 1$ be any unit vector which we think of as one of the columns of $X = X_{\ell-1}$. Then, we have $\|Dx\| \leqslant \|D\|\cdot\|x\| \leqslant 1$, by definition of the spectral norm. This shows that the "ℓ_2-sensitivity" of one matrix-vector multiplication in our algorithm is bounded by 1. It is well-known that it suffices to add Gaussian noise scaled to the ℓ_2-sensitivity of the matrix-vector product in order to achieve differential privacy. Since there are kL matrix-vector multiplications in total we need to scale the noise by a factor of $\sqrt{kL}$.

The second claim follows analogously. Here however we need to scale the noise magnitude to the "ℓ_1-sensitivity" of the matrix-vector product which be bound by $\sqrt{n}$ using Cauchy-Schwarz. The claim then follows using standard properties of the Laplacian mechanism. ■

Given the previous lemma it is straightforward to derive the following corollaries.

Corollary C.6. *Let $A \in \mathbb{R}^{d\times d}$ be a symmetric matrix with singular values $\sigma_1 \geqslant \ldots \geqslant \sigma_d$ and let $\gamma = 1 - \sigma_{k+1}/\sigma_k$. There is an algorithm that given a A and parameter k, preserves (ε,δ)-differentially privacy under unit spectral norm changes and outputs a rank $2k$ matrix B such that with probability* 9/10,

$$\|A - B\| \leqslant \sigma_{k+1} + \tilde{O}\left(\frac{\sigma_1\sqrt{(k/\gamma)d\log d\log(1/\delta)}}{\varepsilon(\sigma_k - \sigma_{k+1})}\right).$$

The $\tilde{O}$-notation hides the factor $O\left(\sqrt{\log(\log(d)/\gamma)}\right)$.

Proof. The proof is analogous to the proof of Corollary 4.3. ■

A similar corollary applies to $(\varepsilon,0)$-differential privacy.

Corollary C.7. *Let $A \in \mathbb{R}^{d\times d}$ be a symmetric matrix with singular values $\sigma_1 \geqslant \ldots \geqslant \sigma_d$ and let $\gamma = 1 - \sigma_{k+1}/\sigma_k$. There is an algorithm that given a A and parameter k, preserves (ε,δ)-differentially privacy under unit spectral norm changes and outputs a rank $2k$ matrix B such that with probability* 9/10,

$$\|A - B\| \leqslant \sigma_{k+1} + \tilde{O}\left(\frac{\sigma_1 k^{1.5} d\log(d)\log(d/\gamma)}{\varepsilon\gamma(\sigma_k - \sigma_{k+1})}\right).$$

Proof. We invoke PPM with $p = 2k$ and Laplacian noise with the scaling given by Lemma 4.5 so that the algorithm satisfies $(\varepsilon,0)$-differential privacy. Specifically, $G_\ell \sim \mathrm{Lap}(0,\lambda)^{d\times p}$ where $\lambda = 10\varepsilon^{-1}pL\sqrt{d}$. Lemma A.3. Indeed, with probability 99/100, we have

1. $\max_{\ell=1}^{L}\|G_\ell\| \leqslant O\left(\lambda\sqrt{kd}\log(kdL)\right) = O\left((1/\varepsilon\gamma)k^{1.5}d\log(d)\log(kdL)\right)$
2. $\max_{\ell=1}^{L}\|U^\top G_\ell\| \leqslant O(\lambda k\log(kL)) = O\left((1/\varepsilon\gamma)k^2\sqrt{d}\log(d)\log(kL)\right)$

We can now plug these error bounds into Corollary 1.1 to obtain the bound

$$\left\|(I - X_L X_L^\top)U\right\| \leqslant O\left(\frac{k^{1.5}d\log(d)\log(d/\gamma)}{\varepsilon\gamma(\sigma_k - \sigma_{k+1})}\right)$$

Repeating the argument from the proof of Corollary 4.3 gives the stated guarantee for low-rank approximation. ■

The bound above matches a lower bound shown by Kapralov and Talwar [**?**] up to a factor of $\tilde{O}(\sqrt{k})$. We believe that this factor can be eliminated from our bounds by using a quantitatively stronger version of Lemma A.3. Compared to the upper bound of [**?**] our algorithm is faster by a more than a quadratic factor in d. Moreover, previously only bounds for $(\varepsilon,0)$-differential privacy were known for the spectral norm privacy notion, whereas our bounds strongly improve when going to (ε,δ)-differential privacy.

C.3 Dimension-free bounds for incoherent matrices

The guarantee in Theorem 1.3 depends on the quantity $\|X_\ell\|_\infty$ which could in principle be as small as $\sqrt{1/d}$. Yet, in the above theorems, we use the trivial upper bound 1. This in turn resulted in a dependence on the dimensions of A in our theorems. Here, we show that the dependence on the dimension can be replaced by an essentially tight dependence on the *coherence* of the input matrix. In doing so, we resolve the main open problem left open by Hardt and Roth [**?**]. The definition of coherence that we will use is formally defined as follows.

Definition C.8 (Matrix Coherence)**.** We say that a matrix $A \in \mathbb{R}^{d \times d'}$ with singular value decomposition $A = U\Sigma V^\top$ has *coherence*

$$\mu(A) \overset{\text{def}}{=} \left\{ d\|U\|_\infty^2, d'\|V\|_\infty^2 \right\}.$$

Here $\|U\|_\infty = \max_{ij} |U_{ij}|$ denotes the largest entry of U in absolute value.

Our goal is to show that the ℓ_∞-norm of the vectors arising in PPM is closely related to the coherence of the input matrix. We obtain a nearly tight connection between the coherence of the matrix and the ℓ_∞-norm of the vectors that PPM computes.

Theorem C.9. *Let $A \in \mathbb{R}^{d \times d}$ be symmetric. Suppose* NPM *is invoked on A, and $L \leqslant n$, with each G_ℓ sampled from $N(0, \sigma_\ell^2)^{d \times p}$ for some $\sigma_\ell > 0$. Then, with probability* $1 - 1/n$*,*

$$\max_{\ell=1}^{L} \|X_\ell\|_\infty^2 \leqslant O\left(\frac{\mu(A) \log(d)}{d} \right).$$

Proof. Fix $\ell \in [L]$. Let $A = \sum_{i=1}^n \sigma_i u_i u_i^\top$ be given in its eigendecomposition. Note that

$$B = \max_{i=1}^{d} \|u_i\|_\infty \leqslant \sqrt{\frac{\mu(A)}{d}}.$$

We may write any column x of X_ℓ as $x = \sum_{i=1}^d s_i \alpha_i u_i$ where α_i are non-negative scalars such that $\sum_{i=1}^d \alpha_i^2 = 1$, and $s_i \in \{-1, 1\}$ where $s_i = sign(\langle x, u_i \rangle)$. Hence, by Lemma 4.13 (shown below), the signs $(s_1, \ldots, s_d)$ are distributed uniformly at random in $\{-1, 1\}^d$. Hence, by Lemma 4.14 (shown below), it follows that $\Pr\left\{ \|x\|_\infty > 4B\sqrt{\log d} \right\} \leqslant 1/n^3$. By a union bound over all $p \leqslant d$ columns it follows that $\Pr\left\{ \|X_\ell\|_\infty > 4B\sqrt{\log d} \right\} \leqslant 1/d^2$. Another union bound over all $L \leqslant d$ steps completes the proof. ■

The previous theorem states that no matter what the scaling of the Gaussian noise is in each step of the algorithm, so long as it is Gaussian the algorithm will maintain that X_ℓ has small coordinates. We cannot hope to have coordinates smaller than $\sqrt{\mu(A)/d}$, since eventually the algorithm will ideally converge to U. This result directly implies the theorem we stated in the introduction.

Proof of Theorem 1.4. The claim follows directly from Theorem 1.3 after applying Theorem 4.9 which shows that with probability $1 - 1/n$,

$$\max_{\ell=1}^{L} \|X_\ell\|_\infty^2 \leqslant O\left(\frac{\mu(A) \log(d)}{d} \right). \qquad ■$$

C.4 Proofs of supporting lemmas

We will now establish Lemma 4.13 and Lemma 4.14 that were needed in the proof of the previous theorem. For that purpose we need some basic symmetry properties of the QR-factorization. To establish these properties we recall the Gram-Schmidt algorithm for computing the QR-factorization.

Definition C.10 (Gram-Schmidt)**.** The *Gram-Schmidt orthonormalization* algorithm, denoted GS, is given an input matrix $V \in \mathbb{R}^{d\times p}$ with columns $v_1,\dots,v_p$ and outputs an orthonormal matrix $Q \in \mathbb{R}^{d\times p}$ with the same range as V. The columns $q_1,\dots,q_p$ of Q are computed as follows:

```
For $i = 1$ to $p$ do:
  – $r_{ii} \leftarrow \|v_i\|$
  – $q_i \leftarrow v_i/r_{ii}$
  – For $j = i+1$ to $p$ do:
      – $r_{ij} \leftarrow \langle q_i, v_j\rangle$
      – $v_j \leftarrow v_j - r_{ij}q_i$
```

The first states that the Gram-Schmidt operation commutes with an orthonormal transformation of the input.

Lemma C.11. *Let $V \in \mathbb{R}^{d\times p}$ and let $O \in \mathbb{R}^{d\times d}$ be an orthonormal matrix. Then,* $\mathrm{GS}(OV) = O\times\mathrm{GS}(V)$*.*

Proof. Let $\{r_{ij}\}_{ij\in[p]}$ denote the scalars computed by the Gram-Schmidt algorithm as specified in Definition 4.10. Notice that each of the numbers $\{r_{ij}\}_{ij\in[p]}$ is invariant under an orthonormal transformation of the vectors $v_1,\dots,v_p$. This is because $\|Ov_i\| = \|v_i\|$ and $\langle Ov_i, Ov_j\rangle = \langle v_i, v_j\rangle$. Moreover, The output Q of Gram-Schmidt on input of V satisfies $Q = VR$, where R is an upper right triangular matrix which only depends on the numbers $\{r_{ij}\}_{i,j\in[p]}$. Hence, the matrix R is identical when the input is OV. Thus, $\mathrm{GS}(OV) = OVR = O\times\mathrm{GS}(V)$. ■

Given i.i.d. Gaussian matrices $G_0, G_1, \dots, G_L \sim N(0,1)^{d\times p}$, we can describe the behavior of our algorithm by a deterministic function $f(G_0, G_1, \dots, G_L)$ which executes subspace iteration starting with G_0 and then suitably scales G_ℓ in each step. The next lemma shows that this function is distributive with respect to orthonormal transformations.

Lemma C.12. *Let $f\colon (\mathbb{R}^{d\times p})^L \to \mathbb{R}^{n\times p}$ denote the output of* PPM *on input of a matrix $A \in \mathbb{R}^{n\times n}$ as a function of the noise matrices used by the algorithm as described above. Let O be an orthonormal matrix with the same eigenbasis as A. Then,*

$$f(OG_0, OG_1, \dots, OG_L) = O\times f(G_0,\dots,G_L). \tag{3}$$

Proof. For ease of notation we will denote by $X_0,\dots,X_L$ the iterates of the algorithm when the noise matrices are $G_0,\dots,G_L$, and we denote by $Y_0,\dots,Y_L$ the iterates of the algorithm when the noise matrices are $OG_0,\dots,OG_L$. In this notation, our goal is to show that $Y_L = OX_L$.

We will prove the claim by induction on L. For $L = 0$, the base case follows from Lemma 4.11. Indeed,

$$Y_0 = \mathrm{GS}(OG_0) = O\times\mathrm{GS}(G_0) = OX_0\,.$$

Let $\ell \geqslant 1$. We assume the claim holds for $\ell - 1$ and show that it holds for ℓ. We have,

$$\begin{aligned}
Y_\ell &= \mathrm{GS}(AY_{\ell-1} + OG_\ell) \\
&= \mathrm{GS}(AOX_{\ell-1} + OG_\ell) && \text{(by induction hypothesis)} \\
&= \mathrm{GS}(O(AX_{\ell-1} + G_\ell)) && \text{(A and O commute)} \\
&= O\times\mathrm{GS}(AX_{\ell-1} + G_\ell) && \text{(Lemma 4.11)} \\
&= OX_\ell\,.
\end{aligned}$$

Note that A and O commute, since they share the same eigenbasis by the assumption of the lemma. This is what we needed to prove. ■

The previous lemmas lead to the following result characterizing the distribution of signs of inner products between the columns of X_ℓ and the eigenvectors of A.

Lemma C.13 (Sign Symmetry). *Let A be a symmetric matrix given in its eigendecomposition as $A = \sum_{i=1}^d \lambda_i u_i u_i^\top$. Let $\ell \geqslant 0$ and let x be any column of X_ℓ, where X_ℓ is the iterate of* PPM *on input of A. Put $S_i = sign(\langle u_i, x\rangle)$ for $i \in [d]$. Then $(S_1, \dots, S_d)$ is uniformly distributed in $\{-1,1\}^d$.*

Proof. Let $(z_1, \dots, z_d) \in \{-1,1\}^d$ be a uniformly random sign vector. Let $O = \sum_{i=1}^d z_i u_i u_i^\top$. Note that O is an orthonormal transformation. Clearly, any column Ox of OX_ℓ satisfies the conclusion of the lemma, since $\langle u_i, Ox\rangle = z_i\langle u_i, x\rangle$. Since the Gaussian distribution is rotationally invariant, we have that OG_ℓ and G_ℓ follow the same distribution. In particular, denoting by Y_ℓ the matrix computed by the algorithm if $OG_0, \dots, OG_\ell$ were chosen, we have that Y_ℓ and X_ℓ are identically distributed. Finally, by Lemma 4.12, we have that $Y_\ell = OX_\ell$. By our previous observation this means that Y_ℓ satisfies the conclusion of the lemma. As Y_ℓ and X_ℓ are identically distributed, the claim also holds for X_ℓ. ■

We will use the previous lemma to bound the ℓ_∞-norm of the intermediate matrices X_ℓ arising in power iteration in terms of the coherence of the input matrix. We need the following large deviation bound.

Lemma C.14. *Let $\alpha_1, \dots, \alpha_d$ be scalars such that $\sum_{i=1}^d \alpha_i^2 = 1$ and $u_1, \dots, u_d$ are unit vectors in $\mathbb{R}^n$. Put $B = \max_{i=1}^d \|u_i\|_\infty$. Further let $(s_1, \dots, s_d)$ be chosen uniformly at random in $\{-1,1\}^d$. Then,*

$$\mathbb{P}\mathrm{r}\left\{\left\|\sum_{i=1}^d s_i\alpha_i u_i\right\|_\infty > 4B\sqrt{\log d}\right\} \leqslant 1/d^3 .$$

Proof. Let $X = \sum_{i=1}^d X_i$ where $X_i = s_i\alpha_i u_i$. We will bound the deviation of X in each entry and then take a union bound over all entries. Consider $Z = \sum_{i=1}^d Z_i$ where Z_i is the first entry of X_i. The argument is identical for all other entries of X. We have $\mathbb{E}Z = 0$ and $\mathbb{E}Z^2 = \sum_{i=1}^d \mathbb{E}Z_i^2 \leqslant B^2\sum_{i=1}^d \alpha_i^2 = B^2$. Hence, by Theorem A.1 (Chernoff bound),

$$\mathbb{P}\mathrm{r}\left\{|Z| > 4B\sqrt{\log(d)}\right\} \leqslant \exp\left(-\tfrac{16B^2\log(d)}{4B^2}\right) \leqslant \exp(-4\log(d)) = \tfrac{1}{d^4} .$$

The claim follows by taking a union bound over all d entries of X. ■

D Deferred Concentration Inequalities

Theorem D.1 (Chernoff bound). *Let the random variables $X_1, \dots, X_m$ be independent random variables such that for every i, $X_i \in [-1,1]$ almost surely. Let $X = \sum_{i=1}^m X_i$ and let $\sigma^2 = \mathbb{V}X$. Then, for any $t > 0$, $\mathbb{P}\mathrm{r}\{|X - \mathbb{E}X| > t\} \leqslant \exp\left(-\frac{t^2}{4\sigma^2}\right)$.*

The next lemma follows from standard concentration properties of the Gaussian distribution.

Lemma D.2. *Let $U \in \mathbb{R}^{d\times k}$ be a matrix with orthonormal columns. Let $G_1, \dots, G_L \sim N(0,\sigma^2)^{d\times p}$ with $k \leqslant p \leqslant d$ and assume that $L \leqslant d$. Then, with probability $1 - 10^{-4}$,*

$$\max_{\ell\in[L]} \|U^\top G_\ell\| \leqslant O\left(\sigma\sqrt{p + \log L}\right) .$$

Proof. By rotational invariance of G_ℓ the spectral norm $\|U^\top G_\ell\|$ is distributed like largest singular value of a random draw from $k\times p$ gaussian matrix $\mathrm{N}(0,\sigma^2)^{k\times p}$. Since $p\geqslant k$, the largest singular value strongly concentrates around $O(\sigma\sqrt{p})$ with a gaussian tail. By the gaussian concentration of Lipschitz functions of gaussians, taking the maximum over L gaussian matrices introduces an additive $O(\sigma\sqrt{\log L})$ term. ■

We also have an analogue of the previous lemma for the Laplacian distribution.

Lemma D.3. *Let $U\in\mathbb{R}^{n\times k}$ be a matrix with orthonormal columns. Let $G_1,\dots,G_L\sim\mathrm{Lap}(0,\lambda)^{d\times p}$ with $k\leqslant p\leqslant d$ and assume that $L\leqslant d$. Then, with probability $1-10^{-4}$,*

$$\max_{\ell\in[L]}\|U^\top G_\ell\|\leqslant O\left(\lambda\sqrt{pk}\log(Lpk)\right).$$

Proof. We claim that with probability $1-10^{-4}$ for every $\ell\in[L]$, every entry of $U^\top G_\ell$ is bounded by $O(\lambda\log(Lpk))$ in absolute value. This follows because each entry has variance λ^2 and is a weighted sum of n independent Laplacian random variables $\mathrm{Lap}(0,\lambda)$. Assuming this event occurs, we have

$$\max_{\ell\in[L]}\|U^\top G_\ell\|\leqslant\max_{\ell\in[L]}\|U^\top G_\ell\|_F\leqslant O\left(\lambda\sqrt{pk}\log(Lpk)\right).$$ ■

Lemma D.4 (Matrix Chernoff)**.** *Let $X_1,\dots,X_n\sim\mathcal{X}$ be i.i.d. random matrices of maximum dimension d and mean μ, uniformly bounded by $\|X\|\leqslant R$. Then for all $t\leqslant 1$,*

$$\mathbb{P}\mathrm{r}\left\{\left\|\tfrac{1}{n}\textstyle\sum_i X_i-\mathbb{E}X_1\right\|\geqslant tR\right\}\leqslant de^{-\Omega(mt^2)}$$

E Reduction to symmetric matrices

For all our purposes it suffices to consider symmetric $n\times n$ matrices. Given a non-symmetric $m\times n$ matrix B we may always consider the $(m+n)\times(m+n)$ matrix $A=[\,0\,B\,|\,B^\top\,0\,]$. This transformation preserves all the parameters that we are interested in as was argued in [**?**] more formally. This allows us to discuss symmetric eigendecompositions rather than singular vector decompositions and therefore simplify our presentation below.