
Adaptive Sampling for Low-Rank Matrix Approximation in the Matrix-Vector Product Model

Arvind Rathnashyam

Department of Mathematics
Rensselaer Polytechnic Institute
Troy, NY 12180, USA
rathna@rpi.edu

Nicolas Boullé

Department of Mathematics
and Theoretical Physics
University of Cambridge
Cambridge, CB3 0WA, UK
nb690@cam.ac.uk

Alex Townsend

Department of Mathematics
University of Cornell
Ithaca, NY 14853, USA
townsend@cornell.edu

Abstract

We consider the problem of low-rank matrix approximation when the matrix \mathbf{A} is accessible only via matrix-vector products. The Randomized Singular Value Decomposition (rSVD) is an effective algorithm for obtaining the low rank representation of a matrix developed by Halko, Martinsson, and Tropp [19]. Recently, Boullé and Townsend [6] generalized the rSVD to Hilbert-Schmidt Operators where functions are sampled from non-standard Covariance Matrices when there is already prior information on the right singular vectors within the column space of the target matrix, \mathbf{A} . We consider the open problem posed by Boullé and Townsend in [4] of adaptive covariance matrix selection. We study adaptive sampling in the matrix-vector product model and develop adaptive sampling algorithm which samples vectors in rounds. We develop an adaptive sampling algorithm that learns a low-rank approximation with lower Frobenius norm errors than the generalized randomized SVD without prior knowledge on the dominant right singular space of the target matrix.

1 Introduction

Obtaining the Low-Rank Matrix Approximation by sketching has been a problem of interest at the intersection of computational linear algebra and machine learning for the past two decades. In many real-world applications, it is often not possible to run experiments in parallel. Consider the following setting, there are a set of n inputs and m outputs, and there exists a PDE such it maps any set of inputs in $\mathbb{C}^m \rightarrow \mathbb{C}^n$. However, to run experiments, it takes hours for set up, execution, or it is expensive, e.g. aerodynamics [17], fluid dynamics [21]. Thus, after each experimental run, we want to sample a function such that in expectation, we will be exploring an area of the PDE which we have the least knowledge of. For Low-Rank Approximation the Randomized SVD, [19], has been theoretically analyzed and used in various applications. Even more recently, [3] discovered if we have prior information on the right singular vectors of \mathbf{A} , we can modify the Covariance Matrix such that the sampled vectors are within the column space of \mathbf{A} . They extended the theory for Randomized SVD where the covariance matrix is now a general PSD matrix. The basis of our analysis is the idea of sampling vectors in the Null-Space of the Low-Rank Approximation. This idea has been introduced recently in Machine Learning in [31] for training neural networks for sequential tasks. In a Bayesian sense, we want to maximize the expected information gain of the PDE in each iteration by sampling in the space where we have no information. This leads to the formulation of our iterative algorithm for sampling vectors for the Low-Rank Approximation. The current state of the art algorithms for low-rank matrix approximation in the matrix-vector product model used a fixed covariance matrix structure. In this paper, we consider the adaptive setting where the algorithm \mathcal{A} chooses a vector $\omega^{(k)}$ with access to the previous query vectors $\omega^{(1)}, \dots, \omega^{(k-1)}$, the matrix-vector

products $\mathbf{A}\omega^{(1)}, \dots, \mathbf{A}\omega^{(k-1)}$, and the intermediate low-rank matrix approximations, $\mathbf{Q}^{(k)}\mathbf{Q}^{(k)*}\mathbf{A}$, where $\mathbf{Q}^{(k)}\mathbf{R}^{(k)}$ is the economized QR decomposition of $\mathbf{A}\Omega^{(k)}$, where $\Omega^{(k)}$ is the concatenation of vectors $\omega^{(1)}, \dots, \omega^{(k)}$.

Adaptive Sampling techniques for Low-Rank Matrix Approximation first appeared in CUR Matrix Decomposition in [18]. Optimal column-sampling for the CUR Matrix Decomposition received much attention as can be seen in the works [20, 14, 15]. More recently, [25] gave an algorithm for sampling the rows for CUR-Matrix Factorization and proved it is possible to improve upon any relative-error Column Subset Selection Problem by adaptive sampling.

Adaptively sampling vectors for matrix problems has been studied in detail in [29]. The theoretical properties of adaptively sampled matrix vector queries for estimating the minimum eigenvalue of a Wishart matrix have been studied in [10]. Their bounds are used in [1] to develop adaptive bounds for their low-rank matrix approximation method using Krylov Subspaces. To our knowledge, we are the first paper to give an algorithm for low-rank approximation in the non-symmetric matrix low-rank approximation in the matrix-vector product model. Our algorithm utilizes the SVD computation of the low-rank approximation at each step to sample the next vector. Although there are runtime limitations, both in theory under certain conditions and many real-world matrices, our algorithm obtains a closer estimate to the optimal in the Frobenius Norm.

We will now clearly state our contributions.

Main Contributions.

1. We develop a novel adaptive sampling algorithm for Low-Rank Matrix Approximation problem in the matrix-vector product model which does not utilize prior information of \mathbf{A} .
2. We provide a novel theoretical analysis for adaptive sampling in the matrix-vector product query model.
3. We perform Numerical Experiments on real-world and synthetic matrices that confirm our theoretical claims.

2 Background and related works

The Randomized SVD is a method to find an orthonormal matrix that captures the range of the top left singular space of \mathbf{A} by multiplying the matrix \mathbf{A} with a Matrix with Standard Normal entries[22]. One first samples $k + p$ gaussian vectors $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for $i \in [k + p]$ where k is the target rank and p is the oversampling parameter. One then calculates an orthonormal basis for the range by the economized QR decomposition and obtains $\mathbf{QR} = \mathbf{AX}$. It is then proved by Halko et al. [19] that \mathbf{Q} is a good approximation of the range of the top right singular space of \mathbf{A} and thus the approximation $\mathbf{QQ}^*\mathbf{A}$ is close to \mathbf{A} in both the spectral and Frobenius Norms. The analysis has been extended to SRTT matrices by Boutsidis and Gittens [9]. More recently, Boullé and Townsend studied the Randomized SVD when the columns of the Gaussian Matrix are sampled from a Correlated Gaussian Matrix. Their results indicate that there exist Covariance Matrices that are able to obtain better approximation bounds than the standard rSVD.

3 Adaptive Sampling

The adaptive range finder algorithm was proposed in [19] as a method to guarantee a high accuracy guarantee for the low-rank approximation by sampling vectors one at a time. It does not modify the distribution of the sampling vectors to reduce sample complexity or error. To clarify the distinction in the meaning of the term *adaptive*, we give a formal definition of the problem.

Definition 1. *Given access to $r(k + p)$ right matrix-vector products. Sample ω_i for $i \in [r(k + p)]$ such that, after each matrix-vector query, one has access to the low-rank approximation $\mathbf{Q}_t\mathbf{Q}_t^*\mathbf{A}$ for $t \in [r]$. Obtain a matrix $\mathbf{Q} \in \mathbb{O}_{m, r(k+p)}$ using maximally $r(k + p)$ left and right matrix-vector product queries to \mathbf{A} to find \mathbf{Q} s.t.*

$$\|\mathbf{A} - \mathbf{QQ}^*\mathbf{A}\|_F \leq (1 + \epsilon) \min_{\mathbf{U}: \mathbf{U}^*\mathbf{U} = \mathbf{I}_{r\ell}} \|\mathbf{A} - \mathbf{UU}^*\mathbf{A}\|_F \quad (1)$$

Our definition of *adaptive* is standard in the Theoretical Computer Science literature (see e.g. [25, 15]), in that we are using our previous samples to inform our future choices.

3.1 Algorithm

The Pseudo Code for the optimal function sampling is given in Algorithm 1. Algorithm 1 is developed with the goal of minimizing the number of matrix-vector products necessary to obtain the same accuracy as the randomized SVD or generalized randomized SVD. We follow the convention in [25] to perform adaptive sampling in rounds.

: our input is a batch size = $k + p$

Algorithm 1 Adaptive sampling for low-rank matrix approximation

input: Target matrix $\mathbf{A} \in \mathbb{C}^{m \times n}$, target rank k , oversampling parameter p , number of rounds r

output: Low-rank approximation $\hat{\mathbf{A}}$ of \mathbf{A} to minimize $\|\hat{\mathbf{A}} - \mathbf{A}\|_F$

- 1: $\hat{\mathbf{A}}^{(0)} \leftarrow \mathbf{0}$
- 2: **for** $t \in [r]$ **do**
- 3: Form $\Omega^{(t)} \in \mathbb{C}^{n \times (k+p)}$ with columns sampled i.i.d from $\mathcal{N}(\mathbf{0}, \mathbf{C}^{(t)})$
- 4: $\mathbf{Y} \leftarrow [\mathbf{Y}, \mathbf{A}\Omega^{(t)}] \in \mathbb{C}^{m \times t(k+p)}$ and obtain the economized QR decomposition $\mathbf{Y} = \mathbf{Q}\mathbf{R}$
- 5: Update the low-rank approximation: $\hat{\mathbf{A}}^{(t)} = \hat{\mathbf{A}}^{(t-1)} + \mathbf{Q}_-^{t(k+p)} \mathbf{Q}_-^{t(k+p)*} \mathbf{A}$
- 6: Calculate the economized SVD: $\hat{\mathbf{U}} \hat{\Sigma} \hat{\mathbf{V}}^* = \hat{\mathbf{A}}^{(t)}$
- 7: Update the covariance matrix: $\mathbf{C}_t^{1/2} \leftarrow [\hat{\mathbf{V}}_{t(k+p)} \mathbf{0}]$

return: $\mathbf{Q}^{r(k+p)} \mathbf{Q}^{r(k+p)*} \mathbf{A}$

Algorithm 1 runs in multiple rounds. In each round, we use the information we have learned from the matrix to update the covariance matrix.

Covariance Update in Algorithm 1. In each round, we update our covariance matrix to be the projection matrix of the singular space of the low-rank approximation. It can be calculated either from an SVD calculation or from a pseudoinverse, as we have from an expansion of the SVD,

$$\hat{\mathbf{V}}_{t(k+p)} \hat{\mathbf{V}}_{t(k+p)}^* = \hat{\mathbf{V}}^{(t)} \hat{\Sigma}^{(t)} + \hat{\mathbf{U}}^{(t)*} \hat{\mathbf{U}}^{(t)} \hat{\Sigma}^{(t)} \hat{\mathbf{V}}^{(t)*} = \hat{\mathbf{A}}^{(t)} + \hat{\mathbf{A}}^{(t)} \quad (2)$$

Therefore, one can call a pseudoinverse procedure of SVD procedure to update the covariance matrix. **Algorithm 1 runtime analysis.** In total we sample $k + p$ Gaussian vectors from t different Gaussian Distributions and therefore the matrix-matrix product $\mathbf{A}\Omega$ scales in $O(tmn(k + p))$. We perform t QR factorizations which scale in $O(tmn(k + p))$. We then must perform the SVD decomposition on the low-rank approximation $\mathbf{Q}\mathbf{Q}^* \mathbf{A} \in \mathbb{C}^{m \times n}$ a total of t times which can be done on the order of $O(tmn^2)$ as well. Thus the total complexity can be observed as $O(tmn^2)$. The dominating runtime is the economized SVD calculation.

4 Theoretical Analysis

In this section we will first give the mathematical setup for the theoretical analysis. We first provide improvements to the Generalized Randomized SVD Approximation Bounds. We utilize our improved approximation bounds to derive approximation bounds for simple adaptive sampling. The proofs of all results presented in this section are deferred to Appendix A.

4.1 Notation

For any integer t , We define $[t]$ as the set of integers $\{1, \dots, t\}$. Let $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$ for any two real numbers a, b . We define $\mathbb{O}_{n,k}$ as the set of all $n \times k$ matrices with orthonormal columns, i.e. $\{\mathbf{V} : \mathbf{V}^* \mathbf{V} = \mathbf{I}_{k \times k}\}$. We define $\mathbb{S}^{d-1} = \{\mathbf{x} \in \mathbb{C}^d : \|\mathbf{x}\|_2 = 1\}$. We use Big-O notation, $y = O(x)$, to denote there exists a constant x_0 and a positive constant C such that $y \leq Cx$ for all $x \geq x_0$. We define $\mathbf{E}[X]$ as expectation of random variable X , $\Pr\{A\}$ as probability of event A occurring, and $\mathbf{Var}(X)$ as variance of a random variable X .

4.2 Linear Algebra

The pseudoinverse of a matrix \mathbf{X} is given by \mathbf{X}^+ .

Definition 2. The Moore-Penrose Inverse of \mathbf{X} given as \mathbf{X}^+ satisfies the following four conditions,

$$\mathbf{X}\mathbf{X}^+\mathbf{X} = \mathbf{X}, \quad \mathbf{X}^+\mathbf{X}\mathbf{X}^+ = \mathbf{X}^+, \quad (\mathbf{X}\mathbf{X}^+)^* = \mathbf{X}\mathbf{X}^+, \quad (\mathbf{X}^+\mathbf{X})^* = \mathbf{X}^+\mathbf{X}$$

If \mathbf{X} has full row-rank, then the pseudo-inverse can be given explicitly as $\mathbf{X}^+ = (\mathbf{X}^*\mathbf{X})^{-1}\mathbf{X}^*$. For any matrix $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^*$, the pseudoinverse is given as $\mathbf{X}^+ = \mathbf{U}\Sigma^+\mathbf{V}^*$. An orthogonal projector matrix is a Hermitian Matrix and satisfies $\mathbf{P}^2 = \mathbf{P}$. This property of the orthogonal projector matrix implies $\mathbf{0} \preceq \mathbf{P} \preceq \mathbf{I}$. Suppose \mathbf{Q} represents an orthonormal basis of the column space of a matrix \mathbf{Y} , then $\mathbf{P} = \mathbf{Q}\mathbf{Q}^*$ represents the unique projection matrix on to the range of \mathbf{Y} . We follow a similar setup as previous literature. We factorize \mathbf{A} as follows,

$$\mathbf{A} = \begin{bmatrix} \mathbf{U}_k & \mathbf{U}_{k,\perp} \end{bmatrix} \begin{bmatrix} \Sigma_k & \mathbf{0} \\ \mathbf{0} & \Sigma_{k,\perp} \end{bmatrix} \begin{bmatrix} \mathbf{V}_k^* \\ \mathbf{V}_{k,\perp}^* \end{bmatrix} = \sum_{i=1}^n \sigma_i \mathbf{u}_i \mathbf{v}_i^*$$

The trace of a matrix is given as $\text{Tr}(\mathbf{A}) = \sum_{i \in [n]} \sigma_i(\mathbf{A})$. We furthermore define for $\xi \in \{2, \text{F}\}$,

$$\mathbf{A}_k = \min_{\mathbf{B}: \text{rank}(\mathbf{B})=k} \|\mathbf{A} - \mathbf{B}\|_\xi$$

The Eckart-Young-Mirsky Theorem [16, ?] tells us for the spectral and Frobenius norms this matrix is equivalent to the k dominant singular matrix of \mathbf{A} . Throughout the paper, we utilize the following norms,

$$\text{Spectral Norm: } \|\mathbf{A}\|_2^2 = \|\mathbf{A}^* \mathbf{A}\|_2 = \sigma_1^2(\mathbf{A}) \quad (3)$$

$$\text{Frobenius Norm: } \|\mathbf{A}\|_{\text{F}}^2 = \text{Tr}(\mathbf{A}^* \mathbf{A}) = \sum_{i=1}^n \sigma_i^2(\mathbf{A}) \quad (4)$$

Our theoretical arguments rely on the following proposition given by Halko et al. [19].

Proposition 3 (Conjugation Rule). *Suppose that $\mathbf{M} \succeq \mathbf{0}$. Then for any conformal matrix \mathbf{A} ,*

$$\mathbf{M} \preceq \mathbf{N} \implies \mathbf{A}^* \mathbf{M} \mathbf{A} \preceq \mathbf{A}^* \mathbf{N} \mathbf{A}$$

We also utilize a corollary of Weyl's Inequality, a classical result in Perturbation Theory.

Lemma 4 (Weyl's Inequality). *Suppose \mathbf{A}, \mathbf{B} have real eigenvalues (e.g. they are Hermitian), then for any $i \in [n]$,*

$$\lambda_i(\mathbf{A} + \mathbf{B}) \leq \lambda_i(\mathbf{A}) + \lambda_1(\mathbf{B})$$

4.3 Generalized Randomized SVD

In this subsection we present our result for the Frobenius norm approximation bounds for the Generalized Randomized SVD presented by Boullé and Townsend [6].

Theorem 5. *Let $\mathbf{A} \in \mathbb{C}^{m \times n}$, set $k \geq 1$ an integer, an oversampling parameter $p \geq 4$. Let $\Omega \in \mathbb{R}^{n \times k+p}$ represent the test matrix with columns sampled from $\mathcal{N}(\mathbf{0}, \mathbf{C})$. Then, let $\mathbf{Q}\mathbf{R} = \mathbf{A}\Omega$ represent the economized QR decomposition of $\mathbf{A}\Omega$, then with probability at least $1 - \delta - t^{-p}$,*

$$\begin{aligned} \|\mathbf{A} - \mathbf{Q}\mathbf{Q}^* \mathbf{A}\|_{\text{F}} &\leq \|\Sigma_{k,\perp}\|_{\text{F}} \left(1 + \sqrt{6t \log(n(k+p)/\delta)} \|\mathbf{K}_{22}\|_2^{1/2} \sqrt{\text{Tr}(\mathbf{K}_{11}^{-1})} \sqrt{\frac{1}{p+1}} \right) \\ \|\mathbf{A} - \mathbf{Q}\mathbf{Q}^* \mathbf{A}\|_2 &\leq \|\Sigma_{k,\perp}\|_2 \left(1 + \sqrt{2t \log((n-k)(k+p)/\delta)} \|\mathbf{K}_{22}\|_2^{1/2} \|\mathbf{K}_{11}^{-1}\|_2^{1/2} \frac{e\sqrt{k+p}}{p+1} \right) \end{aligned}$$

where

$$\mathbf{K} = \mathbf{V}^* \mathbf{C} \mathbf{V} = \begin{bmatrix} \mathbf{V}_k^* \mathbf{C} \mathbf{V}_k & \mathbf{V}_k^* \mathbf{C} \mathbf{V}_{k,\perp} \\ \mathbf{V}_{k,\perp}^* \mathbf{C} \mathbf{V}_k & \mathbf{V}_{k,\perp}^* \mathbf{C} \mathbf{V}_{k,\perp} \end{bmatrix} = \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix}$$

The proof of Theorem 5 can be derived with the deterministic error bound of Theorem 9.1 in [19] and combining Lemma 11 with both relations of Proposition 19. From Theorem 5, we infer that we obtain a better theoretical bound than the randomized SVD when \mathbf{C} has better alignment with the top right singular space of \mathbf{A} than the identity. As example, choosing $\mathbf{C} = \mathbf{V}_k \mathbf{V}_k^*$, we obtain $\mathbf{K}_{22} = \mathbf{0}$ and we then obtain the optimal $\|\Sigma_{k,\perp}\|_{\text{F}}^2$ Frobenius norm approximation error. The weakness in the bound when compared to the probabilistic bound in [19] is the $O(\log(n(k+p)))$ term that is

a result of Proposition 19. Our bounds presented in Theorem 5 are the strongest relative spectral and Frobenius norm error approximations to the generalized Randomized SVD presented in the literature to our knowledge. The difference between our bounds and the bounds derived in Theorem 2.4 of [26] is that we do not have an additive error term, giving us the desired $(1 + \epsilon)$ approximation results for low-rank approximation.

Corollary 6. *Let $\mathbf{A} \in \mathbb{C}^{m \times n}$, set $k \geq 1$ an integer, an oversampling parameter $p \geq 4$. Let $\mathbf{\Omega} \in \mathbb{R}^{n \times (k+p)}$ represent the test matrix with columns sampled from $\mathcal{N}(\mathbf{0}, \mathbf{C})$. Then, let $\mathbf{QR} = \mathbf{A}\mathbf{\Omega}$ represent the economized QR decomposition of $\mathbf{A}\mathbf{\Omega}$. Then if*

$$p \geq (12/\epsilon) \log(n^2/\delta) \|\mathbf{K}_{22}\|_2 \text{Tr}(\mathbf{K}_{11}^{-1})$$

With probability exceeding $1 - \delta - 0.0625$,

$$\|\mathbf{A} - \mathbf{QQ}^*\mathbf{A}\|_F^2 \leq (1 + \epsilon) \|\mathbf{A} - \mathbf{A}_k\|_F^2$$

Corollary 6 develops a relative Frobenius norm error bound for the Generalized Randomized SVD and can be derived from elementary algebraic manipulations after an application of the first relation in Theorem 5.

4.4 Adaptive Sampling Theory

In this section, we will discuss the theoretical motivation for adaptive sampling. Adaptive sampling in works such as [15] and [25] for the column subselection problem (see e.g. [12] § 1 for a problem definition) sample columns from the residual matrix, $\mathbf{A} - \mathbf{Q}^{(t)}\mathbf{Q}^{(t)*}\mathbf{A}$, at each iteration t instead of directly from \mathbf{A} . Our first result is to formalize this statement and prove that indeed, adaptive low rank matrix approximation is at each iteration equivalent to low-rank approximation on the Residual Matrix. We require the following Lemma, which gives us theoretical insight on where we can derive advantages with adaptive sampling.

Lemma 7. *Let $\mathbf{\Omega}_+ = [\mathbf{\Omega}, \mathbf{\Omega}_-]$, $\mathbf{Y} = \mathbf{A}\mathbf{\Omega}$, $\mathbf{Q} = \text{orth}(\mathbf{Y})$, and $\mathbf{Q}_+ = \text{orth}([\mathbf{Y}, \mathbf{A}\mathbf{\Omega}_-])$. Finally, for shorthand, define $\mathbf{P}_\perp = \mathbf{I} - \mathbf{QQ}^*$, then for $\xi \in \{2, F\}$,*

$$\|\mathbf{A} - \mathbf{Q}_+\mathbf{Q}_+^*\mathbf{A}\|_\xi = \|\mathbf{P}_\perp\mathbf{A} - \mathbf{Q}_-\mathbf{Q}_-^*\mathbf{P}_\perp\mathbf{A}\|_\xi \quad (5)$$

From Lemma 7, to minimize the RHS of Equation (5) we observe that \mathbf{Q}_- is equal to the dominant left singular space of $\mathbf{P}_\perp\mathbf{A}$. From this result, we then see that at each iteration, the optimal vector query is the top right singular vector of $\mathbf{P}_\perp\mathbf{A}$. Moving forward, we then discuss how we can use the intermediate low rank approximations $\mathbf{Q}_t\mathbf{Q}_t^*\mathbf{A}$ for $t \in [k + p]$ to obtain sampling vectors closer to the top right singular vector of $\mathbf{P}_\perp\mathbf{A}$.

We will now show where we can have improvement over the randomized SVD with normal samples.

Theorem 8. *Let $\mathbf{A} \in \mathbb{C}^{m \times n}$ and let $\mathbf{C} \in \mathbb{C}^{n \times n}$ be a PSD covariance matrix. Let the sampling matrix be decomposed as $\mathbf{\Omega} = [\mathbf{\Omega}, \mathbf{\Omega}_-]$, the matrix-matrix products $\mathbf{Y} = \mathbf{A}\mathbf{\Omega}$, $\mathbf{Y}_- = \mathbf{A}\mathbf{\Omega}_-$, then let \mathbf{Q}_+ be the orthonormal matrix from an economized QR decomposition of $[\mathbf{A}\mathbf{\Omega}_-, \mathbf{Y}]$. Let $p \geq 4$ be the oversampling parameter, then with probability exceeding $1 - \delta - t^{-p}$,*

$$\|\mathbf{A} - \mathbf{Q}_+\mathbf{Q}_+^*\mathbf{A}\|_F^2 \leq \epsilon(1 + \epsilon) \|\mathbf{A} - \mathbf{A}_k\|_F^2 + (1 + \epsilon) \|\mathbf{A} - \mathbf{A}_{2k}\|_F^2$$

From Theorem 8, we observe that our choice of covariance matrix reduces the number of matrix-vector products required to obtain the same accuracy as the Randomized SVD or even generalized randomized SVD.

Comparison to Continued Sampling. Consider the case where $t = 2$. In the continued sampling model, we sample all the $2k$ vectors initially. Continued sampling incurs Frobenius Norm Error

$$\|\mathbf{A} - \mathbf{Q}_+\mathbf{Q}_+^*\mathbf{A}\|_F^2 \leq \left(1 + \frac{\epsilon}{2}\right) \|\mathbf{A} - \mathbf{A}_k\|_F^2$$

Next, we consider the bound given in Theorem 8. From some simple algebraic manipulations, we obtain when

$$\|\mathbf{A} - \mathbf{A}_{2k}\|_F^2 \leq \left(\frac{1 - \epsilon/2 - \epsilon^2}{1 + \epsilon}\right) \|\mathbf{A} - \mathbf{A}_k\|_F^2 \leq (1 - \epsilon) \|\mathbf{A} - \mathbf{A}_k\|_F^2$$

The adaptive bound is stronger than the classical bound given sufficient singular decay when the goal is a $(1 + \epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2$ approximation. Even stronger, is we can note there exists an ϵ for which the adaptive bound is always stronger than the classical bound,

$$\sum_{i=2k+1}^n \sigma_i^2 \leq (1 - \epsilon) \sum_{i=k+1}^n \sigma_i^2 \iff \epsilon \sum_{i=k+1}^n \sigma_i^2 \leq \sum_{i=k+1}^{2k} \sigma_i^2$$

Therefore, we see there exists a sufficiently small ϵ that gives a stronger $(1 + \epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F$ approximation. We now consider the analysis for T steps.

Theorem 9. *Let $\mathbf{A} \in \mathbb{C}^{m \times n}$ and let $\mathbf{C}_i \in \mathbb{C}^{n \times n}$ be PSD covariance matrices. The sampling matrix is given as $\mathbf{\Omega} = [\mathbf{C}_1 \mathbf{G}_1, \dots, \mathbf{C}_t \mathbf{G}_t]$. Let \mathbf{Q} be the orthonormal basis of $\mathbf{A}\mathbf{\Omega}$, then with probability exceeding $1 - T(\delta - t^{-p})$,*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^* \mathbf{A}\|_F^2 \leq \|\mathbf{A} - \mathbf{A}_{tk}\|_F^2 + \epsilon \sum_{i=1}^{T-1} (1 + \epsilon)^{T-i} \|\mathbf{A} - \mathbf{A}_{ik}\|_F^2$$

4.5 Analysis of Algorithm 1

Our covariance update in Algorithm 1 of Algorithm 1 does not align with the theory we have developed Section 4.4. Consider Lemma 7, it is optimal to sample in the dominant right singular subspace of $\mathbf{A} - \mathbf{P}_Y \mathbf{A}$. However, in the implicit matrix problem, we do not have access to $\mathbf{A} - \mathbf{P}_Y \mathbf{A}$. We first show for one round of our adaptive procedure in Algorithm 1, it is sufficient to sample in our estimation of the dominant right singular subspace of \mathbf{A} .

Theorem 10. *Consider one round of Algorithm 1 with a covariance matrix $\mathbf{C} \in \mathbb{C}^{n \times n}$. Then,*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^* \mathbf{A}\|_F^2 \lesssim \|\mathbf{A} - \mathbf{A}_{2k}\|_F^2 \left(1 + 2 \log((k+p)^2/\delta) \left\| \mathbf{V}_{k,\perp}^* \mathbf{C}^{1/2} \right\|_2^2 \left\| (\mathbf{V}_k^* \mathbf{C}^{1/2})^+ \right\|_2^2 \frac{e^2(k+p)}{(p+1)^2} \right)$$

We can now note that $(1 + \epsilon)^2 = 1 + O(\epsilon)$. Then, from ideas in subspace perturbation theory (see e.g. [24]), the sample complexity to obtain a $(1 + \epsilon)\|\mathbf{A} - \mathbf{A}_{2k}\|_F$ is reduced.

4.6 Randomized Nyström Approximation

When the matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric positive semi-definite (SPSD), the *randomized Nyström Approximation* is a stronger method for low-rank approximation when compared to the Randomized SVD [30]. The Nyström Approximation is given as,

$$\hat{\mathbf{A}} = \mathbf{A}\mathbf{\Omega}(\mathbf{\Omega}^* \mathbf{A} \mathbf{\Omega})^+ \mathbf{\Omega}^* \mathbf{A} \approx \mathbf{A}$$

Adaptive Sampling for the Nyström Approximation is known to be difficult to study theoretically (see discussion in e.g. [?]). Recently, Persson et al. studied the Nyström Approximation in the Matrix-Vector Product Query Model with the columns of $\mathbf{\Omega}$ sampled from a Gaussian matrix with correlated covariance matrix [26]. It is known for $\xi \in \{2, F\}$ that

$$\left\| \mathbf{A} - \hat{\mathbf{A}} \right\|_{\xi} \leq \left\| \mathbf{\Sigma}_{k,\perp} \right\|_{\xi} + \left\| (\mathbf{\Sigma}_2^{1/2} \mathbf{\Omega}_2 \mathbf{\Omega}_1^+)^* \mathbf{\Sigma}_{k,\perp}^{1/2} \mathbf{\Omega}_2 \mathbf{\Omega}_1^+ \right\|_{\xi}$$

From Equation (3), our relative spectral error bound in Theorem 5 still applies, however we would need a new Frobenius Relative Error Bound. Then we might hope to prove the same ideas as the previous subsection, in that we are able to sample from the range of our low-rank matrix approximation we can improve the results of the Nyström Approximation.

5 Numerical experiments

NB: One thing we might want is a comparison against a Power method, computational timings.

In this section, we will test various Synthetic Matrices and Differential Operators in real-world applications with our framework and compare against state-of-the-art non-adaptive approaches for

low-rank matrix approximation. All experiments are run in MATLAB on a 3.60 GHz processor and 16.0 GB RAM. All points in the plots for Fig. 1 are the average over 10 randomized runs. We define

$$\text{OPT} = \min_{\mathbf{Z} \in \mathbb{C}^{m \times t(k+p)}} \|\mathbf{A} - \mathbf{Z}\mathbf{Z}^* \mathbf{A}\|_{\text{F}} \quad (6)$$

From an application of the Eckart-Young-Mirsky Theorem for the Frobenius Norm [16, 23], we find that the minimizing $\mathbf{Z} = \mathbf{U}_{k+p}$ and therefore,

$$\text{OPT} = \sqrt{\sum_{i=t(k+p)+1}^n \sigma_i^2(\mathbf{A})} \quad (7)$$

Experiment 1: Learning an Inverse Differential Operator. In our first experiment we attempt to learn the discretized $10^3 \times 10^3$ matrix of the inverse of the following differential operator:

$$\mathcal{L}u = \frac{\partial^2 u}{\partial x^2} - 100 \sin(5\pi x) u, \quad x \in [0, 1] \quad (8)$$

Learning the inverse operator of a PDE is equivalent to learning the Green’s Function of a PDE. This has been theoretically proven for certain classes of PDEs (Linear Parabolic [5, 7]) as the inverse Differential operator is compact. Furthermore, there are multiple works suggesting the learning of inverse differential operators is provably efficient [3, 4].

Experiment 2: Matrix with Polynomial Singular Value Decay. We sample \mathbf{U} uniformly randomly from the Haar Measure over $\mathbb{O}_{m,m}$ and \mathbf{V} uniformly randomly from the Haar Measure over $\mathbb{O}_{n,n}$. We then form the matrix \mathbf{A} in the scheme described in Equation (9).

$$\mathbf{A} = \sum_{i \in [n]} i^{-p} \cdot \mathbf{U}_{(:,i)} \mathbf{V}_{(:,i)}^*, \quad \mathbf{U} \in \mathbb{O}_{m,m}, \mathbf{V} \in \mathbb{O}_{n,n} \quad (9)$$

Experiment 3: Learning an Inverse Differential Operator in the Real World. We attempt to learn the inverse of the discretized differential operator given in matrix `pde2961`, sourced from the TAMU Matrix Suite [13]. `pde2961` is the matrix associated with a Model Partial Differential Equations Problem.

Experiment 4: Learning a forward matrix in the Real World. We learn the a low rank matrix approximation of `fs-680-1`, also sourced from the TAMU Sparse Matrix Suite [13]. `fs-680-1` is the associated matrix for a chemical kinetics problem.

Observation 1: We observe in Fig. 1(a), even without prior knowledge of the dominant right singular space of \mathbf{A} , after approximately 150 matrix-vector products adaptive sampling learns a better low-rank approximation with respect to the Frobenius norm.

Observation 2: In Fig. 1(b), we find that Algorithm 1 obtains nearly-optimal Frobenius norm error from 175 to 650 matrix-vector product queries.

Observation 3: Algorithm 1 performs well in large matrices after considering the results in Fig. 1(c). The `pde2961` matrix is of size 2961×2961 and Algorithm 1 outperforms rSVD and the generalized rSVD.

Observation 4: Algorithm 1 works well in a wide variety of singular value spectra. This can be concluded from observing that Fig. 1 display a diverse range of singular value spectra and cover both real and synthetic scenarios.

Our experiments on real-world matrices are promising and indicate that our algorithm and implementation can be used in real-world applications of learning low-rank approximations of matrices that are only accessible via matrix-vector products.

6 Conclusions

We have theoretically and empirically analyzed a novel Covariance Update to iteratively construct the sampling matrix, $\mathbf{\Omega}$ in the Randomized SVD algorithm. We introduce a new adaptive sampling framework for low-rank matrix approximation when the matrix is only accessible by matrix-vector products by giving the algorithm access to intermediate low-rank matrix approximations. Our covariance update for generating sampling vectors and functions can find use various PDE learning

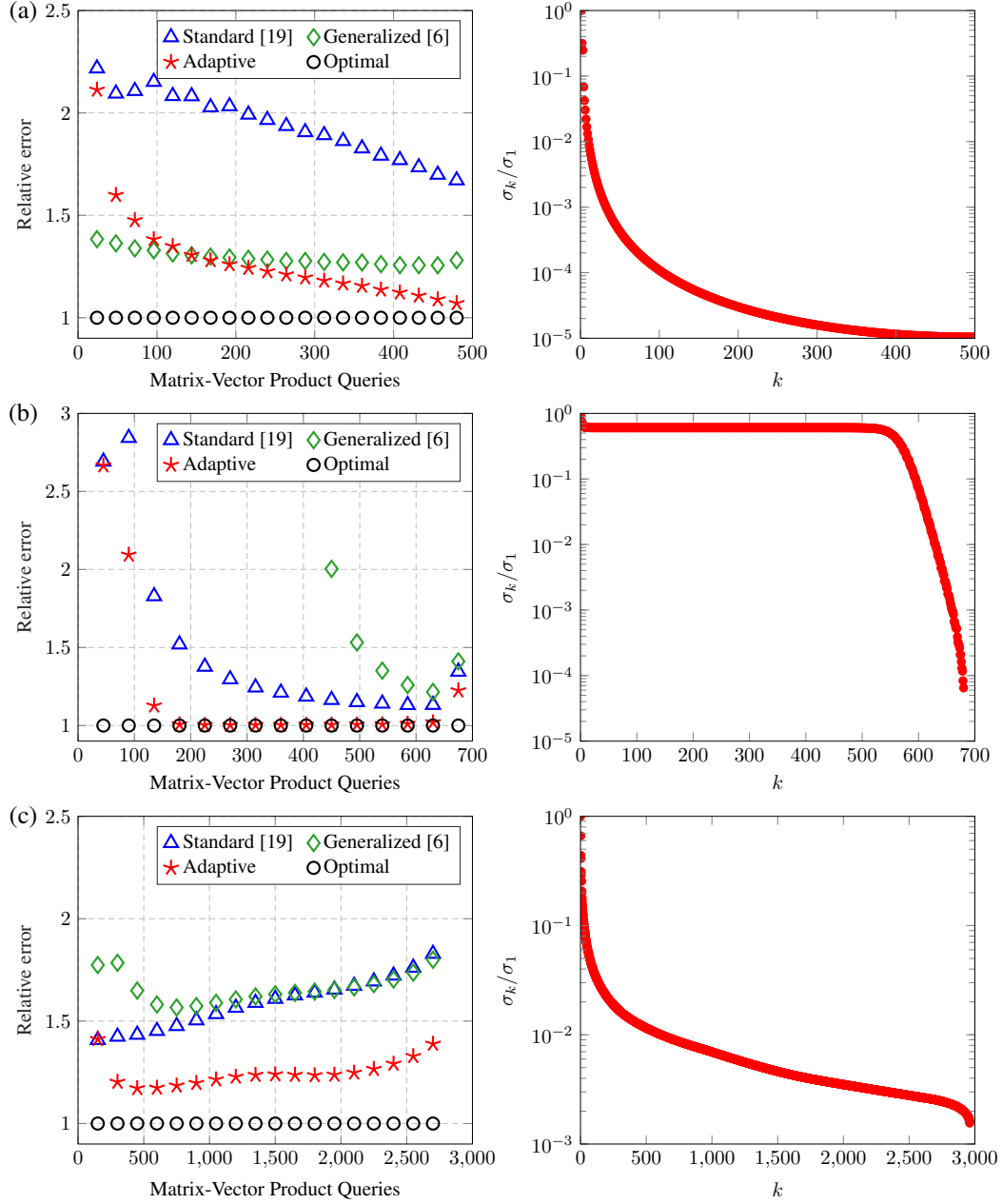


Figure 1: Left panels: Relative error (in the Frobenius norm) of the low-rank matrix approximation with respect to the number of matrix-vector products. Right panels: Singular value decay of the matrix. (a) Inverse of a discretized Laplacian differential operator (see Eq. (8) and [6, Fig. 2]), $k = 8, p = 16, t = 20$. (b) Matrix fs-680-1 from [13], $k = 8, p = 8, t = 20$. (c) Matrix pde2961 from [13], $k = 50, p = 100, t = 18$. **Need to compare against best $k + p$ -approximation, need to call batch size = $k + p$**

applications, [3, 11]. Numerical Experiments indicate without prior knowledge of the matrix, we are able to obtain superior performance to the Randomized SVD and generalized Randomized SVD with covariance matrix utilizing prior information of the PDE. Theoretically, we provide an analysis of our update extended to k -steps and show in expectation, under certain singular value decay conditions, we obtain better performance expectation.

Acknowledgments and Disclosure of Funding

The paper originated from the Cornell 2023 Math REU. A.R. and A.T. were supported by NSF RTG (DMS-1645643). N.B. and A.T. were supported by the Office of Naval Research (ONR), under grant N00014-23-1-2729. A.T. was partially supported by NSF CAREER (DMS-2045646). The authors thank Alex Gittens and Christopher Wang for the helpful discussions.

References

- [1] A. BAKSHI, K. L. CLARKSON, AND D. P. WOODRUFF, *Low-rank approximation with $1/\varepsilon^3$ matrix-vector products*, in 54th Annual ACM Symposium on Theory of Computing, 2022, pp. 1130–1143.
- [2] R. L. BERGER AND G. CASELLA, *Statistical inference*, Duxbury, 2001.
- [3] N. BOULLÉ, C. J. EARLS, AND A. TOWNSEND, *Data-driven discovery of Green’s functions with human-understandable deep learning*, Sci. Rep., 12 (2022), p. 4824.
- [4] N. BOULLÉ, D. HALIKIAS, AND A. TOWNSEND, *Elliptic PDE learning is provably data-efficient*, Proc. Natl. Acad. Sci. U.S.A., 120 (2023), p. e2303904120.
- [5] N. BOULLÉ, S. KIM, T. SHI, AND A. TOWNSEND, *Learning Green’s functions associated with time-dependent partial differential equations*, J. Mach. Learn. Res., 23 (2022), pp. 9797–9830.
- [6] N. BOULLÉ AND A. TOWNSEND, *A generalization of the randomized singular value decomposition*, in International Conference on Learning Representations, 2022.
- [7] N. BOULLÉ AND A. TOWNSEND, *Learning elliptic partial differential equations with randomized linear algebra*, Found. Comput. Math., 23 (2023), pp. 709–739.
- [8] C. BOUTSIDIS, P. DRINEAS, AND M. MAGDON-ISMAIL, *Near-optimal column-based matrix reconstruction*, SIAM J. Comput., 43 (2014), pp. 687–717.
- [9] C. BOUTSIDIS AND A. GITTENS, *Improved matrix algorithms via the subsampled randomized Hadamard transform*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 1301–1340.
- [10] M. BRAVERMAN, E. HAZAN, M. SIMCHOWITZ, AND B. WOODWORTH, *The gradient complexity of linear regression*, in Conference on Learning Theory, 2020, pp. 627–647.
- [11] S. L. BRUNTON, B. R. NOACK, AND P. KOUMOUTSAKOS, *Machine learning for fluid mechanics*, Annu. Rev. Fluid Mech., 52 (2020), pp. 477–508.
- [12] A. CIVRIL, *Column subset selection problem is ug-hard*, J. Comput. Syst. Sci., 80 (2014), pp. 849–859.
- [13] T. A. DAVIS AND Y. HU, *The University of Florida sparse matrix collection*, ACM Trans. Math. Softw., 38 (2011), pp. 1–25.
- [14] A. DESHPANDE, L. RADEMACHER, S. S. VEMPALA, AND G. WANG, *Matrix approximation and projective clustering via volume sampling*, Theory Comput., 2 (2006), pp. 225–247.
- [15] A. DESHPANDE AND S. VEMPALA, *Adaptive sampling and fast low-rank matrix approximation*, in International Workshop on Approximation Algorithms for Combinatorial Optimization, Springer, 2006, pp. 292–303.
- [16] C. ECKART AND G. YOUNG, *The approximation of one matrix by another of lower rank*, Psychometrika, 1 (1936), pp. 211–218.
- [17] H.-Y. FAN, G. S. DULIKRAVICH, AND Z.-X. HAN, *Aerodynamic data modeling using support vector machines*, Inverse Probl. Sci. En., 13 (2005), pp. 261–278.
- [18] A. FRIEZE, R. KANNAN, AND S. VEMPALA, *Fast monte-carlo algorithms for finding low-rank approximations*, J. ACM, 51 (2004), pp. 1025–1041.
- [19] N. HALKO, P. G. MARTINSSON, AND J. A. TROPP, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Rev., 53 (2011), pp. 217–288.
- [20] S. HAR-PELED, *Low rank matrix approximation in linear time*, arXiv preprint arXiv:1410.8802, (2014).
- [21] H. LOMAX, T. H. PULLIAM, D. W. ZINGG, T. H. PULLIAM, AND D. W. ZINGG, *Fundamentals of computational fluid dynamics*, Springer, 2001.

- [22] P.-G. MARTINSSON AND J. A. TROPP, *Randomized numerical linear algebra: Foundations and algorithms*, Acta Numerica, 29 (2020), pp. 403–572.
- [23] L. MIRSKY, *Symmetric gauge functions and unitarily invariant norms*, Q. J. Math., 11 (1960), pp. 50–59.
- [24] S. O’ROURKE, V. VU, AND K. WANG, *Random perturbation of low rank matrices: Improving classical bounds*, Linear Algebra and its Applications, 540 (2018), pp. 26–59.
- [25] S. PAUL, M. MAGDON-ISMAIL, AND P. DRINEAS, *Column selection via adaptive sampling*, in Advances in Neural Information Processing Systems, vol. 28, 2015.
- [26] D. PERSSON, N. BOULLÉ, AND D. KRESSNER, *Randomized Nyström approximation of non-negative self-adjoint operators*, arXiv preprint arXiv:2404.00960, (2024).
- [27] P. RIGOLLET AND J.-C. HÜTTER, *High-dimensional statistics*, arXiv preprint arXiv:2310.19244, (2023).
- [28] E. SCHMIDT, *Zur theorie der linearen und nichtlinearen integralgleichungen*, Math. Ann., 63 (1907), pp. 433–476.
- [29] X. SUN, D. P. WOODRUFF, G. YANG, AND J. ZHANG, *Querying a matrix through matrix-vector products*, ACM Transactions on Algorithms (TALG), 17 (2021), pp. 1–19.
- [30] J. A. TROPP AND R. J. WEBBER, *Randomized algorithms for low-rank matrix approximation: Design, analysis, and applications*, arXiv preprint arXiv:2306.12418, (2023).
- [31] S. WANG, X. LI, J. SUN, AND Z. XU, *Training networks in null space of feature covariance for continual learning*, in Conference on Computer Vision and Pattern Recognition, 2021, pp. 184–193.

A Proofs

In this section we give proofs for results we deferred from the main text.

A.1 Distribution of $\mathbf{V}^*\Omega$

We devote this section to the matrix $\mathbf{V}^*\Omega$. We will derive various concentration inequalities which allow us to give the main theorems.

Lemma 11. *Let $\Omega = [\omega_1, \dots, \omega_\ell] \in \mathbb{C}^{n \times \ell}$ such that $\omega_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ for all $i \in [\ell]$ and $\mathbf{C} \succeq 0$ is symmetric. Then, let $\mathbf{V} \in \mathbb{O}_{n \times k}$, it then follows the columns of $\mathbf{V}^*\Omega$ are sampled from a centered multivariate Gaussian Distribution with second-moment matrix $\mathbf{K} = \mathbf{V}^*\mathbf{C}\mathbf{V}$.*

Proof. The matrix $\mathbf{V}^*\Omega$ can be decomposed as follows,

$$\mathbf{V}^*\Omega = \begin{bmatrix} \mathbf{v}_1^*\omega_1 & \cdots & \mathbf{v}_1^*\omega_\ell \\ \vdots & \ddots & \vdots \\ \mathbf{v}_k^*\omega_1 & \cdots & \mathbf{v}_k^*\omega_\ell \end{bmatrix}$$

Let $\mathcal{D} = \mathcal{N}(\mathbf{0}, \mathbf{I})$. We will first show that the entries of each column of $\mathbf{V}^*\Omega$ are Gaussian. From the fact that \mathbf{C} is symmetric, we have that $\mathbf{C} = \mathbf{U}\Sigma\mathbf{U}$ for a unitary \mathbf{U} and diagonal $\Sigma \succeq 0$. Then for any $i \in [k]$ and $j \in [\ell]$, we have for $\mathbf{x} \sim \mathcal{D}$,

$$\mathbf{v}_i^*\omega_j = \mathbf{v}_i^*\mathbf{C}^{1/2}\mathbf{x} = \mathbf{v}_i^*\mathbf{U}\Sigma^{1/2}\mathbf{x} = \sum_{k \in [n]} \mathbf{v}_i^*\mathbf{u}_k \sqrt{\lambda_k(\mathbf{C})} x_k$$

In the above, we have that each $[\mathbf{V}^*\Omega]_{i,j}$ is Gaussian for $(i, j) \in [k] \times [\ell]$ as a linear combination of Gaussians is Gaussian. We will calculate the mean and covariance. We first calculate the mean of a column of $\mathbf{V}^*\Omega$. For any $(i, j) \in [k] \times [\ell]$,

$$\begin{aligned} \mathbb{E}_{\omega_j \sim \mathcal{N}(\mathbf{0}, \mathbf{C})} [\mathbf{v}_i^*\omega_j] &= \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} [\mathbf{v}_i^*\mathbf{C}^{1/2}\mathbf{x}] = \mathbf{v}_i^*\mathbf{C}^{1/2} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} [\mathbf{x}] \\ &= \sum_{p \in [n]} \mathbf{v}_i^*\mathbf{u}_p \sqrt{\lambda_p(\mathbf{C})} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} [x_p] = 0 \end{aligned}$$

Now we calculate the covariance matrix. Let $\mathbf{v} \in \mathbb{S}^{n-1}$, then for any $(i, i', j) \in [k] \times [k] \times [\ell]$, and $i \neq i'$, we have

$$\begin{aligned} \mathbb{E}_{\omega_j \sim \mathcal{N}(\mathbf{0}, \mathbf{C})} \left[(\mathbf{v}_i^*\omega_j - \mathbb{E}_{\omega_j \sim \mathcal{N}(\mathbf{0}, \mathbf{C})} [\mathbf{v}_i^*\omega_j]) (\mathbf{v}_{i'}^*\omega_j - \mathbb{E}_{\omega_j \sim \mathcal{N}(\mathbf{0}, \mathbf{C})} [\mathbf{v}_{i'}^*\omega_j]) \right] \\ = \mathbb{E}_{\omega_j \sim \mathcal{N}(\mathbf{0}, \mathbf{C})} [\mathbf{v}_i^*\omega_j \omega_j^* \mathbf{v}_{i'}] = \mathbf{v}_i^* \mathbf{C} \mathbf{v}_{i'} \end{aligned} \quad (10)$$

For the diagonal covariance elements, we have

$$\mathbf{E}_{\omega_j \sim \mathcal{N}(\mathbf{0}, \mathbf{C})}[(\mathbf{v}_i^* \omega_j - \mathbf{E}[\mathbf{v}_i^* \omega_j])^2] = \mathbf{E}_{\omega_k \sim \mathcal{N}(\mathbf{0}, \mathbf{C})}[\mathbf{v}_i^* \omega_j \omega_j^* \mathbf{v}_i] = \mathbf{v}_i^* \mathbf{C} \mathbf{v}_i \quad (11)$$

Then combining Equations (10) and (11), we have

$$\mathbf{K} = \mathbf{V}^* \mathbf{C} \mathbf{V}$$

Our proof is complete. ■

A.2 Proof of Theorem 5

We now give an improvement to the deterministic error bound given in Theorem 9.1 of [19] by Boutsidis et al. [8].

Lemma 12 (Lemma 3.2 in [8]). *Let $\mathbf{A} \in \mathbb{R}^{m \times n}$, and let \mathbf{Q} be an orthonormal basis of $\mathbf{A}\Omega$ for $\Omega \in \mathbb{R}^{n \times (k+p)}$, then*

$$\|\mathbf{A} - \mathbf{Q}(\mathbf{Q}\mathbf{A})_k\|_{\text{F}}^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_{\text{F}}^2 + \|\Sigma_{k,\perp} \mathbf{V}_{k,\perp} \Omega (\mathbf{V}_k \Omega)^+\|_{\text{F}}^2$$

Before we give our improvement to the Generalized Randomized SVD, we present the following necessary lemma on the concentration of $\|(\mathbf{V}^* \Omega)^+\|_{\text{F}}$ for Ω with columns sampled from $\mathcal{N}(\mathbf{0}, \mathbf{C})$.

Lemma 13 (Lemma 3 in [7]). *Suppose $\mathbf{V}_k \in \mathbb{O}^{n,k}$ and $\Omega \in \mathbb{C}^{n \times k+p}$ with columns sampled i.i.d from $\mathcal{N}(\mathbf{0}, \mathbf{K})$ and $p \geq 4$. Then with probability at least $1 - t^{-p}$,*

$$\|(\mathbf{V}_k^* \Omega)^+\|_{\text{F}} \leq \sqrt{\frac{3 \text{Tr}(\mathbf{K}^{-1})}{p+1}} \cdot t^2$$

For our improvement to the Generalized Randomized SVD for the spectral norm, we present the following necessary lemma.

Lemma 14 (Proposition 10.4 in [19]). *Let $\mathbf{G} \in \mathbb{R}^{k \times (k+p)}$ have elements sampled i.i.d from $\mathcal{N}(0, 1)$ for $p \geq 4$. Then for $t \geq 1$, with probability exceeding $1 - t^{-(p+1)}$,*

$$\|\mathbf{G}^+\|_2 \leq \frac{e\sqrt{k+p}}{p+1} \cdot t$$

We are now ready to prove our relative Spectral and Frobenius Error Norm Bounds for the Generalized Randomized SVD originally presented in [6].

Proof.[Proof of Theorem 5] Recall $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^*$, $\Omega_k = \mathbf{V}_k^* \Omega$ and $\Omega_{k,\perp} = \mathbf{V}_{k,\perp}^* \Omega$ where the columns of $\Omega \in \mathbb{R}^{n \times (k+p)}$ are sampled from $\mathcal{N}(\mathbf{0}, \mathbf{C})$. Let

$$\mathbf{K} = \begin{bmatrix} \mathbf{V}_k^* \mathbf{C} \mathbf{V}_k & \mathbf{V}_k^* \mathbf{C} \mathbf{V}_{k,\perp} \\ \mathbf{V}_{k,\perp}^* \mathbf{C} \mathbf{V}_k & \mathbf{V}_{k,\perp}^* \mathbf{C} \mathbf{V}_{k,\perp} \end{bmatrix} = \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix}$$

We then have the following manipulations,

$$\begin{aligned} \|\mathbf{A} - \mathbf{Q}(\mathbf{Q}^* \mathbf{A})_k\|_{\text{F}} &\leq \left(\|\Sigma_{k,\perp}\|_{\text{F}}^2 + \|\Sigma_{k,\perp} \Omega_{k,\perp} \Omega_k^+\|_{\text{F}}^2 \right)^{1/2} \\ &= \left(\|\Sigma_{k,\perp}\|_{\text{F}}^2 + \left\| \Sigma_{k,\perp} \mathbf{V}_{k,\perp}^* \mathbf{C}^{1/2} \mathbf{X} (\mathbf{V}_k^* \mathbf{C}^{1/2} \mathbf{X})^+ \right\|_{\text{F}}^2 \right)^{1/2} \\ &\leq \left(\|\Sigma_{k,\perp}\|_{\text{F}}^2 + 2 \log((n-k)(k+p)/\delta) \left\| \Sigma_{k,\perp} \mathbf{V}_{k,\perp}^* \mathbf{C}^{1/2} \right\|_{\text{F}}^2 \left\| (\mathbf{V}_k^* \mathbf{C}^{1/2} \mathbf{X})^+ \right\|_{\text{F}}^2 \right)^{1/2} \\ &\leq \|\Sigma_{k,\perp}\|_{\text{F}} \left(1 + 6t^2 \log(n(k+p)/\delta) \|\mathbf{K}_{22}\|_2 \frac{\text{Tr}(\mathbf{K}_{11}^{-1})}{p+1} \right)^{1/2} \end{aligned}$$

In the above, the first inequality follows from the deterministic error bound in Lemma 12, the equality follows from noting that $\Omega = \mathbf{C}^{1/2} \mathbf{X}$ where the entries of \mathbf{X} are sampled i.i.d from $\mathcal{N}(0, 1)$, the

second inequality follows from Lemma 13 and Proposition 19 for the Frobenius norm with probability exceeding $1 - \delta - t^{-p}$, and the final inequality follows from noting by the sub-multiplicativity of the Frobenius Norm that states for any conformal matrices \mathbf{X}, \mathbf{Y} that $\|\mathbf{XY}\|_F \leq \|\mathbf{X}\|_F \|\mathbf{Y}\|_2$. Our proof for the probabilistic bound for the Frobenius norm is complete. We now can prove the spectral norm bound,

$$\begin{aligned}
\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^*\mathbf{A}\|_2 &\leq \left(\|\Sigma_{k,\perp}\|_2^2 + \|\Sigma_{k,\perp}\mathbf{\Omega}_{k,\perp}\mathbf{\Omega}_k^+\|_2^2 \right)^{1/2} \\
&= \left(\|\Sigma_{k,\perp}\|_2^2 + \left\| \Sigma_{k,\perp} \mathbf{V}_{k,\perp}^* \mathbf{C}^{1/2} \mathbf{X} (\mathbf{V}_k^* \mathbf{C}^{1/2} \mathbf{X})^+ \right\|_2^2 \right)^{1/2} \\
&= \left(\|\Sigma_{k,\perp}\|_2^2 + \left\| \Sigma_{k,\perp} \mathbf{K}_{22}^{1/2} \tilde{\mathbf{X}} (\mathbf{V}_k^* \mathbf{C}^{1/2} \mathbf{X})^+ \right\|_2^2 \right)^{1/2} \\
&\leq \left(\|\Sigma_{k,\perp}\|_2^2 + 2 \log((n-k)(k+p)/\delta) \left\| \Sigma_{k,\perp} \mathbf{K}_{22}^{1/2} \right\|_2^2 \left\| (\mathbf{V}_k^* \mathbf{C}^{1/2} \mathbf{X})^+ \right\|_2^2 \right)^{1/2} \\
&\leq \|\Sigma_{k,\perp}\|_2 \left(1 + \sqrt{2t \log((n-k)(k+p)/\delta)} \|\mathbf{K}_{22}\|_2^{1/2} \|\mathbf{K}_{11}^{-1}\|_2^{1/2} \frac{e\sqrt{k+p}}{p+1} \right)
\end{aligned}$$

The first inequality follows from the deterministic error bound in Theorem 9.1 of [19]. The second relation follows from noting $\mathbf{\Omega} = \mathbf{C}^{1/2} \mathbf{X}$ where \mathbf{X} is a Gaussian matrix. The third relation follows from Proposition 19 for the spectral norm and noting $\mathbf{X} \in \mathbb{R}^{n \times (k+p)}$. The final relation follows from noting by Cauchy-Schwarz,

$$\left\| \Sigma_{k,\perp} \mathbf{V}_{k,\perp}^* \mathbf{C}^{1/2} \right\|_2^2 \leq \|\Sigma_{k,\perp}\|_2^2 \left\| \mathbf{V}_{k,\perp}^* \mathbf{C}^{1/2} \right\|_2^2 = \|\Sigma_{k,\perp}\|_2^2 \|\mathbf{K}_{22}\|_2$$

Furthermore, from noting that for any matrix $\|\mathbf{B}^+\|_2 = \sigma_k^{-1}(\mathbf{B})$, we have

$$\left\| (\mathbf{V}_k^* \mathbf{C}^{1/2} \mathbf{X})^+ \right\|_2^2 = \sigma_k^{-2} \left(\mathbf{V}_k^* \mathbf{C}^{1/2} \mathbf{X} \right) \leq \sigma_k^{-2} \left(\mathbf{V}_k^* \mathbf{C}^{1/2} \right) \sigma_k^{-2}(\mathbf{X}) = \|\mathbf{K}_{11}^{-1}\|_2 \|\mathbf{X}^+\|_2^2$$

Then, applying the bound in Lemma 14 with probability exceeding $1 - t^{-p+1}$ completes the proof. ■

A.3 Proof of Lemma 7

Proof. Recall that \mathbf{Q}_{12} is an orthonormal basis of $[\mathbf{A}\mathbf{\Omega}_1, \mathbf{A}\mathbf{\Omega}_2]$. Let \mathbf{Q}_1 be an orthonormal basis of $\mathbf{A}\mathbf{\Omega}_1$, it thus follows that \mathbf{Q}_{12} is also an orthonormal basis of $[\mathbf{A}\mathbf{\Omega}_1, (\mathbf{I} - \mathbf{Q}_1 \mathbf{Q}_1^*) \mathbf{A}\mathbf{\Omega}_2]$. This can be seen from the Classical Gram-Schmidt Procedure [28]. Thus, we have

$$\mathbf{Q}_{12} \mathbf{Q}_{12}^* = \mathbf{Q}_1 \mathbf{Q}_1^* + \mathbf{Q}_{2|1} \mathbf{Q}_{2|1}^*$$

Then, expanding out the residual matrix, we have

$$\mathbf{A} - \mathbf{Q}_{12} \mathbf{Q}_{12}^* \mathbf{A} = \mathbf{A} - \mathbf{Q}_1 \mathbf{Q}_1^* \mathbf{A} - \mathbf{Q}_{2|1} \mathbf{Q}_{2|1}^* \mathbf{A}$$

Then from noting, $\mathbf{Q}_1^* \mathbf{Q}_{2|1} = \mathbf{0}$, we have

$$\mathbf{A} - \mathbf{Q}_1 \mathbf{Q}_1^* \mathbf{A} - \mathbf{Q}_{2|1} \mathbf{Q}_{2|1}^* \mathbf{A} = (\mathbf{A} - \mathbf{Q}_1 \mathbf{Q}_1^* \mathbf{A}) - \mathbf{Q}_{2|1} \mathbf{Q}_{2|1}^* (\mathbf{A} - \mathbf{Q}_1 \mathbf{Q}_1^* \mathbf{A})$$

Our proof is complete. ■

We now require the following linear algebraic lemma.

Lemma 15. Let $\mathbf{A} \in \mathbb{C}^{m \times n}$ and $\mathbf{B} \in \mathbb{C}^{m \times n}$ and $\mathbf{Q} \in \mathbb{O}_{m,\ell}$ for $\ell > k$. Then for any $i \in [n]$,

$$\sigma_i^2(\mathbf{A} - \mathbf{Q}\mathbf{Q}^*\mathbf{A}) \leq \sigma_i^2(\mathbf{A} - \mathbf{Q}(\mathbf{Q}^*\mathbf{A})_k)$$

Proof. We can note that $\sigma_i(\mathbf{Q}\mathbf{Q}^*\mathbf{A}) \leq \sigma_i(\mathbf{A})$ for all $i \in [n]$ by noting $\mathbf{Q}\mathbf{Q}^* \preceq \mathbf{I}$ as $\mathbf{Q} \in \mathbb{O}_{m,k}$. We first note that for any matrix $\mathbf{B} \in \mathbb{C}^{m \times n}$, the singular values and eigenvalues are related by $\sigma_i^2(\mathbf{B}) = \lambda_i(\mathbf{B}^* \mathbf{B})$ for any $i \in [n]$. Then let us consider any $i \in [n]$.

$$\lambda_i((\mathbf{A} - \mathbf{Q}\mathbf{Q}^*\mathbf{A})^*(\mathbf{A} - \mathbf{Q}\mathbf{Q}^*\mathbf{A})) = \lambda_i(\mathbf{A}^* \mathbf{A} - \mathbf{A}^* \mathbf{Q}\mathbf{Q}^* \mathbf{A})$$

Let us now consider the matrix in the RHS of the Lemma statement,

$$\begin{aligned}
& \lambda_i((\mathbf{A} - \mathbf{Q}(\mathbf{Q}^* \mathbf{A})_k)^*(\mathbf{A} - \mathbf{Q}(\mathbf{Q}^* \mathbf{A})_k)) \\
&= \lambda_i(\mathbf{A}^* \mathbf{A} - \mathbf{A}^* \mathbf{Q}(\mathbf{Q}^* \mathbf{A})_k - (\mathbf{Q}^* \mathbf{A})_k^* \mathbf{Q}^* \mathbf{A} + (\mathbf{Q}^* \mathbf{A})_k^* \mathbf{Q}^* \mathbf{Q}(\mathbf{Q}^* \mathbf{A})_k) \\
&= \lambda_i(\mathbf{A}^* \mathbf{A} - (\mathbf{A}^* \mathbf{Q} \mathbf{Q}^* \mathbf{A})_k - (\mathbf{A}^* \mathbf{Q} \mathbf{Q}^* \mathbf{A})_k + (\mathbf{A}^* \mathbf{Q} \mathbf{Q}^* \mathbf{A})_k) \\
&= \lambda_i(\mathbf{A}^* \mathbf{A} - (\mathbf{A}^* \mathbf{Q} \mathbf{Q}^* \mathbf{A})_k)
\end{aligned} \tag{12}$$

In the above, the final relation follows noting $\mathbf{Q}^* \mathbf{Q} = \mathbf{I}$ because $\mathbf{Q} \in \mathbb{O}_{m,\ell}$, and for any matrix \mathbf{B} , we have that $\mathbf{B}_k^* \mathbf{B} = (\mathbf{B}^* \mathbf{B})_k$ by an expansion of the SVD. We can now conclude our Lemma,

$$\begin{aligned}
\sigma_i^2(\mathbf{A} - \mathbf{Q} \mathbf{Q}^* \mathbf{A}) &= \lambda_i(\mathbf{A}^* \mathbf{A} - \mathbf{A}^* \mathbf{Q} \mathbf{Q}^* \mathbf{A}) \\
&= \lambda_i(\mathbf{A}^* \mathbf{A} - (\mathbf{A}^* \mathbf{Q} \mathbf{Q}^* \mathbf{A})_k - (\mathbf{A}^* \mathbf{Q} \mathbf{Q}^* \mathbf{A})_{k,\perp}) \\
&\leq \lambda_i(\mathbf{A}^* \mathbf{A} - (\mathbf{A}^* \mathbf{Q} \mathbf{Q}^* \mathbf{A})_k) \quad \text{Lemma 4} \\
&= \sigma_i^2(\mathbf{A} - \mathbf{A}(\mathbf{Q}^* \mathbf{A})_k) \quad \text{Equation (12)}
\end{aligned}$$

In the above, the third relation follows from noting $(\mathbf{A}^* \mathbf{Q} \mathbf{Q}^* \mathbf{A})_{k,\perp} \succeq \mathbf{0}$. Our proof is complete. ■

A.4 Proof of Theorem 8

Proof. Consider the alternative update of the form $\mathbf{Y} = [\mathbf{Y}^{(t)}, \mathbf{Y}]$, then we have

$$\|\mathbf{A} - \mathbf{Q}_+ \mathbf{Q}_+^* \mathbf{A}\|_F = \|(\mathbf{I} - \mathbf{Q}_- \mathbf{Q}_-^*) \mathbf{A} - \mathbf{Q} \mathbf{Q}^* (\mathbf{I} - \mathbf{Q}_- \mathbf{Q}_-^*) \mathbf{A}\|_F$$

The above equation follows from the same argument as Lemma 7. Then from our relative-error accuracy bound in Corollary 6, we have

$$\|(\mathbf{I} - \mathbf{Q}_- \mathbf{Q}_-^*) \mathbf{A} - \mathbf{Q} \mathbf{Q}^* (\mathbf{I} - \mathbf{Q}_- \mathbf{Q}_-^*) \mathbf{A}\|_F^2 \leq (1 + \epsilon) \|(\mathbf{A} - \mathbf{Q}_- \mathbf{Q}_-^* \mathbf{A})_{k,\perp}\|_F^2$$

We then have from Lemma 15,

$$\begin{aligned}
\|(\mathbf{A} - \mathbf{Q}_- \mathbf{Q}_-^* \mathbf{A})_{k,\perp}\|_F^2 &= \sum_{i \in [n] \setminus [k]} \sigma_i^2(\mathbf{A} - \mathbf{Q}_- \mathbf{Q}_-^* \mathbf{A}) \quad \text{Equation (4)} \\
&\leq \sum_{i \in [n] \setminus [k]} \sigma_i^2(\mathbf{A} - \mathbf{Q}_- (\mathbf{Q}_-^* \mathbf{A})_k) \quad \text{Lemma 15} \\
&= \|(\mathbf{A} - \mathbf{Q}_- (\mathbf{Q}_-^* \mathbf{A})_k)_{k,\perp}\|_F^2 \quad \text{Equation (4)}
\end{aligned}$$

Then, from expanding out the Frobenius Norm, we have

$$\|(\mathbf{A} - \mathbf{Q}_- (\mathbf{Q}_-^* \mathbf{A})_k)_{k,\perp}\|_F^2 = \|\mathbf{A} - \mathbf{Q}_- (\mathbf{Q}_-^* \mathbf{A})_k\|_F^2 - \sum_{i \in [k]} \sigma_i^2(\mathbf{A} - \mathbf{Q}_- (\mathbf{Q}_-^* \mathbf{A})_k)$$

We now leverage the Eckart-Young-Mirsky Theorem [16, 23], and obtain

$$\sum_{i \in [k]} \sigma_i^2(\mathbf{A} - \mathbf{Q}_- (\mathbf{Q}_-^* \mathbf{A})_k) \geq \inf_{\substack{\mathbf{B} \in \mathbb{C}^{m \times n} \\ \text{rank}(\mathbf{B})=k}} \sum_{i \in [k]} \sigma_i^2(\mathbf{A} - \mathbf{B}) = \|\mathbf{A}_k - \mathbf{A}_{2k}\|_F^2 \tag{13}$$

Then, from our improved Generalized Randomized SVD error bound in Theorem 5, we have with probability at least $0.99 - t^{-p}$,

$$\|\mathbf{A} - \mathbf{Q}_- (\mathbf{Q}_-^* \mathbf{A})_k\|_F^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 \left(1 + 6t^2 \log(100n\ell) \left\| \mathbf{V}_{k,\perp}^* \mathbf{C}^{1/2} \right\|_2^2 \frac{\text{Tr}((\mathbf{V}_k^* \mathbf{C} \mathbf{V}_k)^{-1})}{p+1} \right) \tag{14}$$

Then, from choosing our oversampling parameter sufficiently large in accordance to Corollary 6,

$$p \geq (6t^2/\epsilon) \left\| \mathbf{V}_{k,\perp}^* \mathbf{C}^{1/2} \right\|_2^2 \text{Tr}((\mathbf{V}_k^* \mathbf{C} \mathbf{V}_k)^{-1})$$

We can then combine our results in Equations (13) and (14) to obtain

$$\begin{aligned}
\|\mathbf{A} - \mathbf{Q}_+ \mathbf{Q}_+^* \mathbf{A}\|_F^2 &\leq (1 + \epsilon)^2 \|\mathbf{A} - \mathbf{A}_k\|_F^2 - (1 + \epsilon) \|\mathbf{A}_k - \mathbf{A}_{2k}\|_F^2 \\
&= \epsilon(1 + \epsilon) \|\mathbf{A} - \mathbf{A}_k\|_F^2 + (1 + \epsilon) \|\mathbf{A} - \mathbf{A}_{2k}\|_F^2
\end{aligned}$$

Our proof is complete. ■

A.5 Proof of Theorem 10

We first give necessary Lemmas for our proof.

Lemma 16 (Lemma 5.3 in [9]). *Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{Y} \in \mathbb{R}^{m \times \ell}$, then for all $\mathbf{X} \in \mathbb{R}^{\ell \times n}$,*

$$\|\mathbf{A} - \mathbf{Y}\mathbf{Y}^+\mathbf{A}\|_{\text{F}} \leq \|\mathbf{A} - \mathbf{Y}\mathbf{X}\|_{\text{F}} \quad (15)$$

Lemma 17 (Matrix Pythagoras). *Let \mathbf{X}, \mathbf{Y} be conformal matrices, then if $\mathbf{X}^*\mathbf{Y} = \mathbf{0}$, then*

$$\|\mathbf{X} - \mathbf{Y}\|_{\text{F}}^2 = \|\mathbf{X}\|_{\text{F}}^2 + \|\mathbf{Y}\|_{\text{F}}^2$$

Proof. The proof follows directly from Equation (4).

$$\|\mathbf{X} - \mathbf{Y}\|_{\text{F}}^2 = \text{Tr}(\mathbf{X}^*\mathbf{X} - \mathbf{X}^*\mathbf{Y} - \mathbf{Y}^*\mathbf{X} + \mathbf{Y}^*\mathbf{Y}) = \text{Tr}(\mathbf{X}^*\mathbf{X}) + \text{Tr}(\mathbf{Y}^*\mathbf{Y}) = \|\mathbf{X}\|_{\text{F}}^2 + \|\mathbf{Y}\|_{\text{F}}^2$$

In the above, we use the fact that $\text{Tr}(\mathbf{A} + \mathbf{B}) = \text{Tr}(\mathbf{A}) + \text{Tr}(\mathbf{B})$ for any conformal matrices, \mathbf{A}, \mathbf{B} .

Proof.[Proof of Theorem 10] Our argument relies on the definition of \mathbf{Q} . Since we have that \mathbf{Q} is an orthonormal basis of $[\mathbf{Y}_1 \ \mathbf{Y}_2]$. Then we have that \mathbf{Q} is also the unique orthonormal basis of $[\mathbf{Y}_1\mathbf{Y}_1^+\mathbf{Y}_1 \ (\mathbf{I} - \mathbf{Y}_1\mathbf{Y}_1^+)\mathbf{Y}_2]$ by the definition of the Moore-Penrose Inverse. Let $\mathbf{X} \in \mathbb{R}^{2\ell \times n}$, then we have from Lemma 16,

$$\begin{aligned} \|\mathbf{A} - \mathbf{Q}\mathbf{Q}^*\mathbf{A}\|_{\text{F}}^2 &\leq \left\| \mathbf{A} - [\mathbf{P}_{\mathbf{Y}_1}\mathbf{A}\mathbf{G}_1 \ (\mathbf{I} - \mathbf{P}_{\mathbf{Y}_1})\mathbf{A}\mathbf{C}^{1/2}\mathbf{G}_2] \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \right\|_{\text{F}}^2 \\ &= \underbrace{\|\mathbf{P}_{\mathbf{Y}_1}\mathbf{A} - \mathbf{P}_{\mathbf{Y}_1}\mathbf{A}\mathbf{G}_1\mathbf{X}_1\|_{\text{F}}^2}_{T_1} + \underbrace{\|(\mathbf{A} - \mathbf{P}_{\mathbf{Y}_1}\mathbf{A}) - (\mathbf{A} - \mathbf{P}_{\mathbf{Y}_1}\mathbf{A})\mathbf{C}^{1/2}\mathbf{G}_2\mathbf{X}_2\|_{\text{F}}^2}_{T_2} \end{aligned}$$

In the above, the second relation follows from Matrix Pythagoras (see Lemma 17). We then choose $\mathbf{X}_1 = (\widehat{\mathbf{V}}_{\ell}^*\mathbf{G}_1)^+\widehat{\mathbf{V}}_{\ell}^*$. Then we can note that $\widehat{\mathbf{V}}_{\ell}^*\mathbf{G}_1$ is invertible almost surely with probability 1. We then obtain,

$$T_1 = \|\mathbf{P}_{\mathbf{Y}_1}\mathbf{A} - \mathbf{P}_{\mathbf{Y}_1}\mathbf{A}\mathbf{G}_1\mathbf{X}_1\|_{\text{F}}^2 = \|\mathbf{P}_{\mathbf{Y}_1}\mathbf{A} - \mathbf{P}_{\mathbf{Y}_1}\mathbf{A}\|_{\text{F}}^2 = 0$$

We then have for the second term,

$$\begin{aligned} T_2 &= \left\| (\mathbf{A} - \mathbf{P}_{\mathbf{Y}_1}\mathbf{A}) - (\mathbf{A} - \mathbf{P}_{\mathbf{Y}_1}\mathbf{A})\mathbf{C}^{1/2}\mathbf{G}_2\mathbf{X}_2 \right\|_{\text{F}}^2 \\ &= \left\| (\mathbf{A} - \mathbf{A}\mathbf{C}^{1/2}\mathbf{G}_2\mathbf{X}_2) - (\mathbf{P}_{\mathbf{Y}_1}\mathbf{A} - \mathbf{P}_{\mathbf{Y}_1}\mathbf{A}\mathbf{C}^{1/2}\mathbf{G}_2\mathbf{X}_2) \right\|_{\text{F}}^2 \end{aligned}$$

Then, setting $\mathbf{X}_2 = (\mathbf{V}_k^*\mathbf{C}^{1/2}\mathbf{G}_2)^+\mathbf{V}_k^*$, we obtain

$$\begin{aligned} T_2 &\leq \|\mathbf{A}_{k,\perp} - \mathbf{P}_{\mathbf{Y}_1}\mathbf{A}_{k,\perp}\|_{\text{F}}^2 + \left\| (\mathbf{A}_{k,\perp} - \mathbf{P}_{\mathbf{Y}_1}\mathbf{A}_{k,\perp})\mathbf{C}^{1/2}\mathbf{G}_2(\mathbf{V}_k^*\mathbf{C}^{1/2}\mathbf{G}_2)^+ \right\|_{\text{F}}^2 \\ &\leq \|\mathbf{A}_{k,\perp} - \mathbf{P}_{\mathbf{A}_{k,\perp}\mathbf{G}_1}\mathbf{A}_{k,\perp}\|_{\text{F}}^2 + \left\| (\mathbf{A}_{k,\perp} - \mathbf{P}_{\mathbf{A}_{k,\perp}\mathbf{G}_1}\mathbf{A}_{k,\perp})\mathbf{C}^{1/2}\mathbf{G}_2(\mathbf{V}_k^*\mathbf{C}^{1/2}\mathbf{G}_2)^+ \right\|_{\text{F}}^2 \\ &\leq \|\mathbf{A}_{k,\perp} - \mathbf{P}_{\mathbf{A}_{k,\perp}\mathbf{G}_1}\mathbf{A}_{k,\perp}\|_{\text{F}}^2 \left(1 + \left\| \mathbf{V}_{k,\perp}^*\mathbf{C}^{1/2}\mathbf{G}_2(\mathbf{V}_k^*\mathbf{C}^{1/2}\mathbf{G}_2)^+ \right\|_2^2 \right) \\ &\leq (1 + \epsilon) \|\mathbf{A} - \mathbf{A}_{2k}\|_{\text{F}}^2 \left(1 + 2\log((k+p)^2/\delta) \left\| \mathbf{V}_{k,\perp}^*\mathbf{C}^{1/2} \right\|_2^2 \left\| (\mathbf{V}_k^*\mathbf{C}^{1/2})^+ \right\|_2^2 \frac{e^2(k+p)}{(p+1)^2} \right) \end{aligned}$$

In the above, the second inequality follows from the Conjugation Rule (see Proposition 3) with $\mathbf{P}_{\mathbf{A}_{k,\perp}\mathbf{G}_1} \preceq \mathbf{P}_{\mathbf{A}\mathbf{G}_1}$, suppose $\text{range}(\mathbf{P}_1) \subset \text{range}(\mathbf{P}_2)$, then $\mathbf{P}_1 \preceq \mathbf{P}_2$, and we have

$$\begin{aligned} \|(\mathbf{I} - \mathbf{P}_2)\mathbf{A}\|_{\text{F}}^2 &= \sum_{i \in [n]} \|(\mathbf{I} - \mathbf{P}_2)\mathbf{A}\mathbf{e}_i\|_2^2 = \sum_{i \in [n]} \mathbf{e}_i^*\mathbf{A}^*(\mathbf{I} - \mathbf{P}_2)\mathbf{A}\mathbf{e}_i \\ &\leq \sum_{i \in [n]} \mathbf{e}_i^*\mathbf{A}^*(\mathbf{I} - \mathbf{P}_1)\mathbf{A}\mathbf{e}_i = \|(\mathbf{I} - \mathbf{P}_1)\mathbf{A}\|_{\text{F}}^2 \end{aligned}$$

In the above, the first and final relations follow from an alternative definition of the Frobenius Norm, the second relation follows from Equation (3), and the third relation follows from the Conjugation Rule. Then, noting that $\text{range}(\mathbf{A}_{k,\perp} \mathbf{G}_1) \subset \text{range}(\mathbf{A} \mathbf{G}_1)$, we obtain the desired result. The final inequality follows from the standard randomized SVD bound (see e.g. Theorem 10.6 in [19]) for $k + p$ samples on the matrix $\mathbf{A} - \mathbf{A}_k$, and the second term follows from the manipulations in the Spectral Norm bound proof given in Theorem 5 with failure probability less than $\delta + t^{-(p+1)}$. ■

Theorem 18. *Consider Algorithm 1 for t steps, then*

$$\|\mathbf{A} - \mathbf{Q} \mathbf{Q}^* \mathbf{A}\|_F \leq \Xi$$

Proof. Proof follows similarly to the above theorem.

$$\begin{aligned} \|\mathbf{A} - \mathbf{Q} \mathbf{Q}^* \mathbf{A}\|_F^2 &\leq \left\| \mathbf{A} - \begin{bmatrix} \mathbf{P}_{\mathbf{Y}_1} \mathbf{A} \mathbf{C}_1^{1/2} \mathbf{G}_1 & (\mathbf{I} - \mathbf{P}_{\mathbf{Y}_1}) \mathbf{P}_{\mathbf{Y}_2} \mathbf{A} \mathbf{C}_2^{1/2} \mathbf{G}_2 & \cdots & (\mathbf{I} - \mathbf{P}_{\mathbf{Y}_t}) \mathbf{A} \mathbf{C}_t^{1/2} \mathbf{G}_t \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_t \end{bmatrix} \right\|_F^2 \\ &= \sum_{i \in [t-1]} \left\| \mathbf{P}_{\mathbf{Y}_i} \mathbf{A} - \mathbf{P}_{\mathbf{Y}_i} \mathbf{A} \mathbf{C}_i^{1/2} \mathbf{G}_i \mathbf{X}_i \right\|_F^2 + \underbrace{\left\| (\mathbf{A} - \mathbf{P}_{\mathbf{Y}} \mathbf{A}) - (\mathbf{A} - \mathbf{P}_{\mathbf{Y}} \mathbf{A}) \mathbf{C}^{1/2} \mathbf{G}_2 \mathbf{X}_2 \right\|_F^2}_{T_2} \end{aligned}$$

We can then note that by choosing $\mathbf{C}_i^{1/2} = (\mathbf{P}_{\mathbf{Y}_i} \mathbf{A})^+ (\mathbf{P}_{\mathbf{Y}_i} \mathbf{A})$, we have that $\mathbf{P}_{\mathbf{Y}_i} \mathbf{A} \mathbf{C}_i^{1/2} = \mathbf{P}_{\mathbf{Y}_i} \mathbf{A}$ and then choosing $\mathbf{X}_i = (\widehat{\mathbf{V}}_{k+p}^* \mathbf{G})^+ \widehat{\mathbf{V}}_{k+p}^*$, with probability almost surely unitary, we obtain

$$\sum_{i \in [t-1]} \left\| \mathbf{P}_{\mathbf{Y}_i} \mathbf{A} - \mathbf{P}_{\mathbf{Y}_i} \mathbf{A} \mathbf{C}_i^{1/2} \mathbf{G}_i \mathbf{X}_i \right\|_F^2 = \sum_{i \in [t-1]} \left\| \mathbf{P}_{\mathbf{Y}_i} \mathbf{A} - \mathbf{P}_{\mathbf{Y}_i} \mathbf{A} \right\|_F^2 = 0$$

Then, considering the second term, we have

$$\begin{aligned} T_2 &= \left\| (\mathbf{A} - \mathbf{P}_{\mathbf{Y}} \mathbf{A}) - (\mathbf{A} - \mathbf{P}_{\mathbf{Y}} \mathbf{A}) \widehat{\mathbf{V}}_{k+p} \mathbf{G}_2 \mathbf{X}_2 \right\|_F^2 \\ &= \end{aligned}$$

■

B Probability Theory

Proposition 19. *Fix matrices $\mathbf{S} \in \mathbb{R}^{k \times n}$ and $\mathbf{T} \in \mathbb{R}^{m \times \ell}$, then for a conformal matrix \mathbf{G} with elements sampled i.i.d from $\mathcal{N}(0, 1)$, then for $\xi \in \{2, F\}$, with probability exceeding $1 - \delta$,*

$$\|\mathbf{S} \mathbf{G} \mathbf{T}\|_\xi \leq \|\mathbf{S}\|_\xi \|\mathbf{T}\|_\xi \sqrt{2 \log(nm/\delta)}$$

Proof. The proof for the Frobenius norm follows from a brute-force calculation followed by a maximal tail bound on a sample of Gaussians.

$$\begin{aligned} \|\mathbf{S} \mathbf{G} \mathbf{T}\|_F^2 &= \sum_{i \in [k]} \sum_{j \in [\ell]} \sum_{(k_1, k_2) \in [n] \times [m]} \mathbf{S}_{i, k_1}^2 \mathbf{T}_{k_2, j}^2 \mathbf{G}_{k_1, k_2}^2 \\ &\leq \sum_{i \in [k]} \sum_{j \in [\ell]} \sum_{(k_1, k_2) \in [n] \times [m]} \mathbf{S}_{i, k_1}^2 \mathbf{T}_{k_2, j}^2 \max_{(k_1, k_2) \in [n] \times [m]} \mathbf{G}_{k_1, k_2}^2 \\ &= \|\mathbf{S}\|_F^2 \|\mathbf{T}\|_F^2 \max_{(k_1, k_2) \in [n] \times [m]} \mathbf{G}_{k_1, k_2}^2 \end{aligned}$$

We now show an analogous result for the spectral norm with just a few more steps to recover the result. We will use classical techniques in probability theory to obtain a similar proof structure as the Frobenius norm result. From [27], we have the following relation for any matrix $\mathbf{B} \in \mathbb{R}^{m \times n}$,

$$\|\mathbf{B}\|_2 = \sup_{\mathbf{v} \in \mathbb{S}^{n-1}} \|\mathbf{B} \mathbf{v}\|_2 = \sup_{\mathbf{u} \in \mathbb{S}^{m-1}} \sup_{\mathbf{v} \in \mathbb{S}^{n-1}} |\mathbf{u}^* \mathbf{B} \mathbf{v}| \quad (16)$$

With the relation given in Equation (16) in hand, we have

$$\begin{aligned}
\|\mathbf{SGT}\|_2^2 &= \sup_{\mathbf{u} \in \mathbb{S}^{k-1}} \sup_{\mathbf{v} \in \mathbb{S}^{\ell-1}} (\mathbf{u}^* \mathbf{SGT} \mathbf{v})^2 \\
&= \sup_{\mathbf{u} \in \mathbb{S}^{k-1}} \sup_{\mathbf{v} \in \mathbb{S}^{\ell-1}} \sum_{(i,j) \in [n] \times [m]} (\mathbf{u}^* \mathbf{S})_i^2 \mathbf{G}_{i,j}^2 (\mathbf{T} \mathbf{v})_j^2 \\
&\leq \sup_{\mathbf{u} \in \mathbb{S}^{k-1}} \sup_{\mathbf{v} \in \mathbb{S}^{\ell-1}} \sum_{(i,j) \in [n] \times [m]} (\mathbf{u}^* \mathbf{S})_i^2 (\mathbf{T} \mathbf{v})_j^2 \max_{(i,j) \in [n] \times [m]} \mathbf{G}_{i,j}^2 \\
&\leq \sup_{\mathbf{u} \in \mathbb{S}^{k-1}} \sup_{\mathbf{v} \in \mathbb{S}^{\ell-1}} \|\mathbf{S}^* \mathbf{u}\|_2^2 \|\mathbf{T} \mathbf{v}\|_2^2 \max_{(i,j) \in [n] \times [m]} \mathbf{G}_{i,j}^2 \\
&= \|\mathbf{S}\|_2^2 \|\mathbf{T}\|_2^2 \max_{(i,j) \in [n] \times [m]} \mathbf{G}_{i,j}^2
\end{aligned}$$

We now bound the maximum Gaussian over a finite sample.

$$\begin{aligned}
\Pr \left\{ \max_{(k_1, k_2) \in [n] \times [m]} G_{k_1, k_2}^2 \geq t \right\} &= \Pr \left\{ \max_{(k_1, k_2) \in [n] \times [m]} |G_{k_1, k_2}| \geq \sqrt{t} \right\} \\
&\leq \frac{\sqrt{2nm}}{\sqrt{\pi}} \int_{\sqrt{t}}^{\infty} e^{-x^2/2} dx \leq \frac{\sqrt{2nm}}{\sqrt{\pi}} \int_{\sqrt{t}}^{\infty} \frac{x e^{-x^2/2}}{\sqrt{t}} dx = \frac{\sqrt{2nm}}{\sqrt{\pi}} e^{-t/2} \leq \delta
\end{aligned}$$

In the above, the first inequality follows from a union bound over $[n] \times [m]$ and then integrating over the PDF of a standard normal Gaussian [2]. Then, from elementary algebraic manipulations, we obtain with probability exceeding $1 - \delta$,

$$\|\mathbf{SGT}\|_{\xi}^2 \leq 2 \log(nm/\delta) \|\mathbf{S}\|_{\xi}^2 \|\mathbf{T}\|_{\xi}^2$$

Taking the square root of both sides completes the proof. ■

C Additional details on the numerical experiments

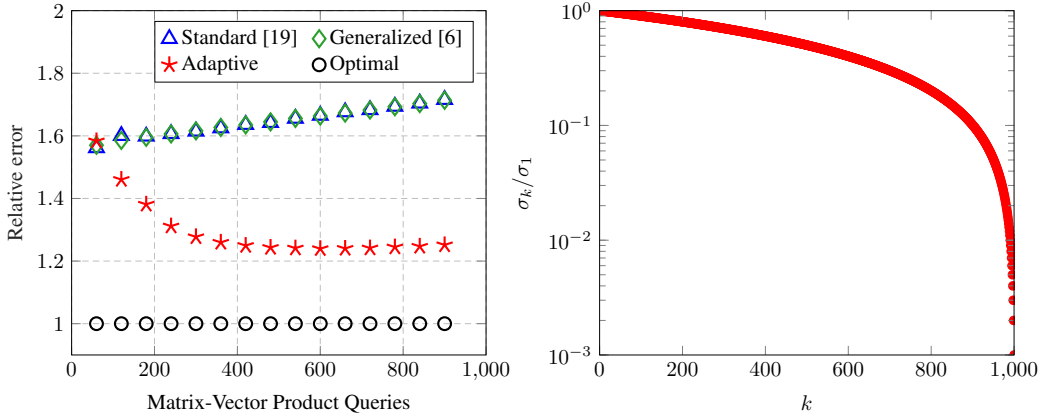


Figure 2: Low-rank approximation for a synthetic matrix with polynomial singular value decay described in Equation (9).