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 rigorously studied by Halko, Martinsson, and Tropp [20]. Recently, Boullé and Townsend generalized the rSVD to Hilbert-Schmidt Operators where functions are sampled from non-standard Covariance Matrices [5]. When there is prior information on the right singular vectors within the column space of the target matrix, $\mathbf{A}$, the generalized rSVD can give stronger bounds than the rSVD. We consider the open problem posed by Boullé and Townsend in [3]. 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 $\mathbb{C}^m \to \mathbb{C}^n$. However, to run experiments, it takes hours for set up, execution, or it is expensive, e.g. aerodynamics [17], fluid dynamics [22]. 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, [20], has been theoretically analyzed and used in various applications. Even more recently, [2] 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 [33] 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 $\boldsymbol{\omega}^{(k)}$ with access to the previous query vectors $\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(k-1)}$, the matrix-vector

Preprint. Under review.

products $\mathbf{A}\boldsymbol{\omega}^{(1)}, \ldots, \mathbf{A}\boldsymbol{\omega}^{(k-1)}$, 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 $\mathbf{A}\boldsymbol{\Omega}^{(k)}$, where $\boldsymbol{\Omega}^{(k)}$ is the concatenation of vectors $\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(k)}$.

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 [21, 14, 15]. More recently, [27] 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 [31]. The theoretical properties of adaptively sampled matrix vector queries for estimating the minimum eigenvalue of a Wishart matrix have been studied in [9]. Their bounds are used in [1] to develop query lower bounds for rank-1 low-rank matrix approximation in the implicit matrix model. 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 computaton of the low-rank approximation at each round to sample the next $k + p$ vectors. Although there are 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 derive improved relative-error bounds for the Generalized Randomized SVD for both the spectral and Frobenius norms [5].
4. We perform Numerical Experiments on real-world and synthetic matrices that confirm our theoretical claims.

2 Randomized Singular Value Decomposition

The Randomized SVD is a method to find an orthonormal matrix that captures the range of the the top left singular space of $\mathbf{A}$ by multiplying the matrix $\mathbf{A}$ with a Matrix with Standard Normal entries[23]. 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. [20] 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 [8]. 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. Thus, with prior knowledge of the dominant right singular space of $\mathbf{A}$, one can incorporate this knowledge in the covariance matrix for sampling Gaussian vectors and obtain better error bounds.

3 Adaptive Sampling

The Adaptive Range Finder Algorithm was propose in [20] as a method to guarantee a high accuracy guarantee for the low-rank approximation by sampling vectors one at a time. The Adaptive Range Finder Algorithm 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 $\boldsymbol{\omega}_i$ 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 a semiorthogonal* $\mathbf{Q}$ *with* $\mathrm{rank}(\mathbf{Q}) = r(k+p)$ *that satisfies*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^*\mathbf{A}\|_{\mathrm{F}} \leq (1+\epsilon) \min_{\mathbf{U}:\mathbf{U}^*\mathbf{U}=\mathbf{I}_{r\ell}} \|\mathbf{A} - \mathbf{U}\mathbf{U}^*\mathbf{A}\|_{\mathrm{F}}$$

Our definition of *adaptive* is standard in the Theoretical Computer Science literature (see e.g. [27, 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 to minimize the number of mat-vecs necessary to obtain the same accuracy as the randomized SVD or generalized randomized SVD. We follow the convention in [27] to perform adaptive sampling in rounds.

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 $\widehat{\mathbf{A}}$ of $\mathbf{A}$ to minimize $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{\mathrm{F}}$

1: $\widehat{\mathbf{A}}^{(0)} \leftarrow \mathbf{0}$
2: **for** $t \in [r]$ **do**
3: Form $\mathbf{\Omega}^{(t)} \in \mathbb{C}^{n\times(k+p)}$ with columns sampled i.i.d from $\mathcal{N}(\mathbf{0}, \mathbf{C}^{(t)})$
4: $\mathbf{Y} \leftarrow [\mathbf{Y}, \mathbf{A}\mathbf{\Omega}^{(t)}] \in \mathbb{C}^{m\times t(k+p)}$ and obtain the economized QR decomposition $\mathbf{Y} = \mathbf{Q}\mathbf{R}$
5: Update the Low Rank Approximation

$$\widehat{\mathbf{A}}^{(t)} = \widehat{\mathbf{A}}^{(t-1)} + \mathbf{Q}_{-}^{t(k+p)}\mathbf{Q}_{-}^{t(k+p)*}\mathbf{A}$$

6: Calculate the economized SVD $\widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^* = \widehat{\mathbf{A}}^{(t)}$
7: Update the Covariance Matrix

$$\mathbf{C}_t^{1/2} \leftarrow \begin{bmatrix} \widehat{\mathbf{V}}_{t(k+p)} & \mathbf{0} \end{bmatrix}$$

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 a SVD calculation, or as from a pseudoinverse, as we have from an expansion of the SVD,

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

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 $\mathbf{A}\mathbf{\Omega}$ scales in $O(tmn(k+p))$. We perform t QR factorizations which scales in $O(tmn(k+p))$. We then must perform the SVD decomposition on the low-rank apprxoimation $\mathbf{Q}\mathbf{Q}^*\mathbf{A} \in \mathbb{C}^{m\times n}$ 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.

3.2 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 [32]. The Nyström Approximation is given as,

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

Adaptive Sampling for the Nyström Approximation is known to be difficult to study theoretically (see discussion in e.g. [19]). Recent studies on the Nyström Approximation with the columns of $\mathbf{\Omega}$ sampled from a Gaussian matrix with correlated covariance matrix [28] can likely be leveraged with our framework to develop adaptive algorithms for SPSD algorithms.

4 Theory

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 expection of random variable X, $\mathbf{Pr}\{A\}$ as probability of event A occuring, 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 $\mathbf{X}^+$. If $\mathbf{X}$ has full row-rank, then the pseudo-inverse can be given explicity as $\mathbf{X}^+ = (\mathbf{X}^*\mathbf{X})^{-1}\mathbf{X}^*$. For any matrix $\mathbf{X} = \mathbf{U\Sigma V}^*$, 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 $\mathbf{P} = \mathbf{QQ}^*$ 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,

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

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

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

The Eckart-Young-Mirsky Theorem [16, 25] tells us for the spectral and Frobenius norms this matrix is equivalent. Throughout the paper, we utilize the following norms,

$$\text{Spectral Norm: } \|\mathbf{A}\|_2^2 = \|\mathbf{A}^*\mathbf{A}\|_2 = \sigma_1^2(\mathbf{A}) \tag{1}$$

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

Our theoretical arguments rely on the following proposition given by Halko et al. [20].

Proposition 2 (Conjutation Rule)**.** *Suppose that* $\mathbf{M} \succeq \mathbf{0}$*. Then for any conformal matrix* $\mathbf{A}$*,*

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

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

Lemma 3 (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 [5].

Theorem 4. *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 $\mathbf{QR} = \mathbf{A\Omega}$ represent the economized QR decomposition of $\mathbf{A\Omega}$, then with probability at least $1 - \delta - t^{-p}$,*

$$\|\mathbf{A} - \mathbf{QQ}^*\mathbf{A}\|_{\mathrm{F}} \leq \|\mathbf{\Sigma}_{k,\perp}\|_{\mathrm{F}}\left(1 + \sqrt{6t\log(n(k+p)/\delta)}\|\mathbf{K}_{22}\|_2^{1/2}\sqrt{\mathrm{Tr}(\mathbf{K}_{11}^{-1})}\sqrt{\frac{1}{p+1}}\right)$$

$$\|\mathbf{A} - \mathbf{QQ}^*\mathbf{A}\|_{2} \leq \|\mathbf{\Sigma}_{k,\perp}\|_{2}\left(1 + \sqrt{2t\log((n-k)(k+p)/\delta)}\|\mathbf{K}_{22}\|_2^{1/2}\|\mathbf{K}_{11}^{-1}\|_2^{1/2}\frac{e\sqrt{k+p}}{p+1}\right)$$

where

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

The proof of Theorem 4 can be derived with the deterministic error bound of Theorem 9.1 in [20] and combining Lemma 9 with both relations of Proposition 16. From Theorem 4, we infer that we obtain a better theoretical bound then 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 $\|\mathbf{\Sigma}_{k,\perp}\|_{\mathrm{F}}^2$ Frobenius norm approximation error. The weakness in the bound when compared to the probabilistic bound in [20] is the $O(\log(n(k+p))$ term that is a result of Proposition 16. Our bounds presented in Theorem 4 are the strongest relative spectral and Frobenius norm error approximatiosn 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 [28] is that we do not have an additive error term, giving us the desired $(1+\epsilon)$ approximation results for low-rank approximation.

Corollary 5. *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 $\mathbf{QR} = \mathbf{A\Omega}$ represent the economized QR decomposition of $\mathbf{A\Omega}$. Then if*

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

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

$$\|\mathbf{A} - \mathbf{QQ}^*\mathbf{A}\|_{\mathrm{F}}^2 \leq (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}^2$$

Corollary 5 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 4.

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

$$\|\mathbf{A} - \mathbf{Q}_+\mathbf{Q}_+^*\mathbf{A}\|_\xi = \|\mathbf{P}_\perp\mathbf{A} - \mathbf{Q}_-\mathbf{Q}_-^*\mathbf{P}_\perp\mathbf{A}\|_\xi \tag{3}$$

From Lemma 6, to minimize the RHS of Equation (3) 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 7. *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 $1 - \delta - t^{-p}$,*

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

From Theorem 7, 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

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

Next, we consider the bound given in Theorem 7. From some simple algebraic manipulations, we obtain when

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

The adaptive bound is stronger than the classical bound given sufficient singular decay.

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 6, it is optimal to sample in the dominant right singular subspace of $\mathbf{A} - \mathbf{P}_{\mathbf{Y}}\mathbf{A}$. However, in the implicit matrix problem, we do not have access to $\mathbf{A} - \mathbf{P}_{\mathbf{Y}}\mathbf{A}$. 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 8. *Consider one round of Algorithm 1. Then,*

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

We can now note that $(1+\epsilon)^2 = 1 + O(\epsilon)$. Then, from ideas in subspace perturbation theory (see e.g. [26]), the sample complexity to obtain a $(1+\epsilon)\|\mathbf{A} - \mathbf{A}_{2k}\|_{\mathrm{F}}$ is reduced.

5 Numerical Experiments

In this section we will test various Synthetic Matrices and Differential Operators in real-world applications with our framework and compare against the 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 Figures 1 to 4 are the average over 10 randomized runs. We define

$$\mathsf{OPT} = \min_{\mathbf{Z}\in\mathbb{C}^{m\times t(k+p)}} \|\mathbf{A} - \mathbf{Z}\mathbf{Z}^*\mathbf{A}\|_{\mathrm{F}} \tag{4}$$

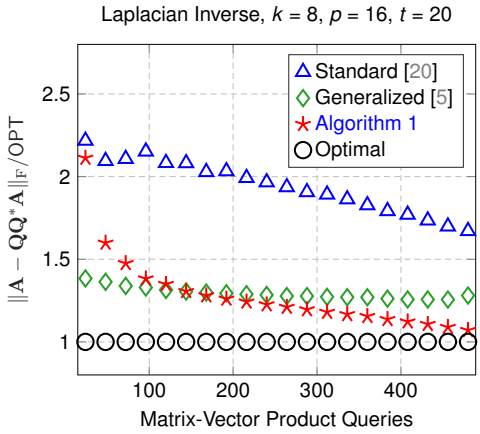
From an application of the Eckart-Young-Mirsky Theorem for the Frobenius Norm [16, 24], we find that the minimizing $\mathbf{Z} = \mathbf{U}_{k+p}$ and therefore,

$$\mathsf{OPT} = \sqrt{\sum_{i=t(k+p)+1}^{n} \sigma_i^2(\mathbf{A})} \tag{5}$$

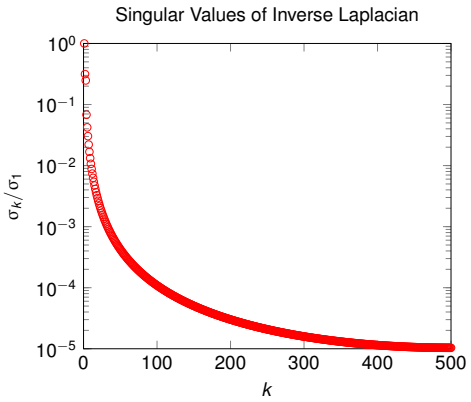
Experiment 1: Learning an Inverse Differential Operator. In our first experiment we attempt to learn the discretized $10^3 \times 10^3$ matrix of the inverse of the following differential operator:

$$\mathcal{L}u = \frac{\partial^2 u}{\partial x^2} - 100\sin(5\pi x)\,u, \qquad x \in [0,1] \tag{6}$$

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 [4, 6]) as the inverse

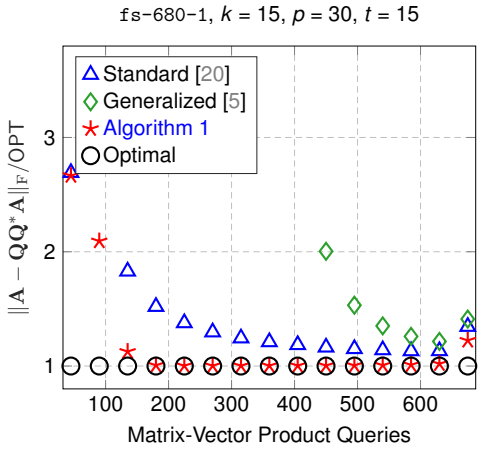


(a) Relative Frobenius Error Low-Rank Matrix Approximation

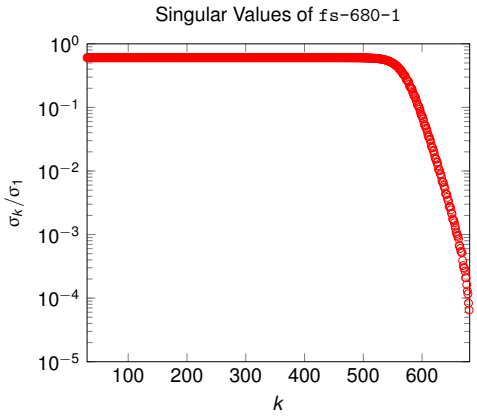


(b) Dominant Singular Value relative singular value decay

Figure 1: Low Rank Approximation for learning the inverse of a discretized Laplacian Differential Operator described in Equation (6). See Figure 2 in [5] for the same experiment.

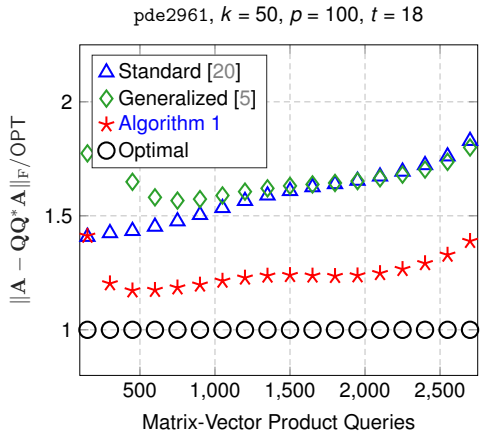


(a) Relative Frobenius Error Low-Rank Matrix Approximation

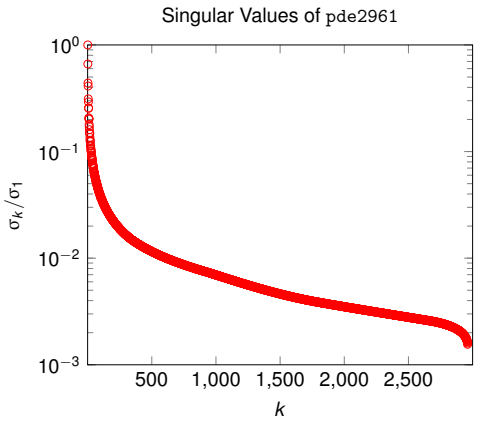


(b) Dominant Singular Value relative singular value decay

Figure 2: Low Rank Approximation for the matrix `fs-680-1` from [13] on the left. The matrix `fs-680-1` is derived from a Chemical kinetics problem. The singular values of `pde2961` is displayed on the right.

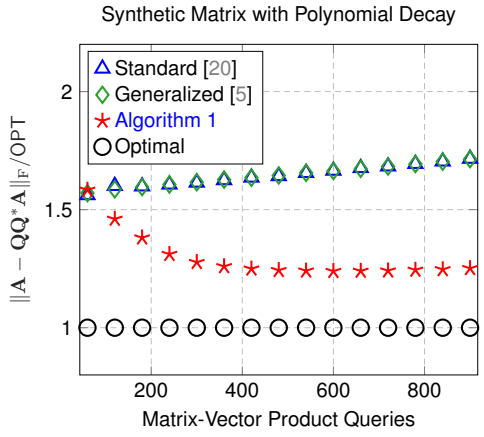


(a) Relative Frobenius Error Low-Rank Matrix Approximation

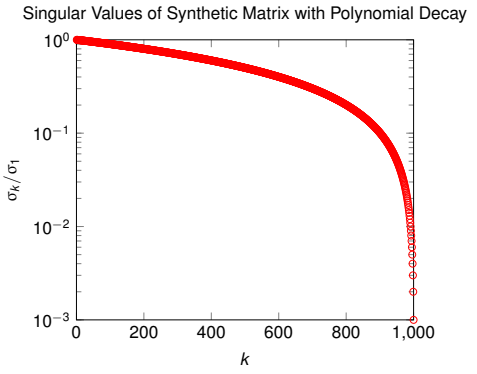


(b) Dominant Singular Value relative singular value decay

Figure 3: Low Rank Approximation for the matrix pde2961 from [13] on the left. The matrix pde2961 is derived from a finite difference discretization of a linear elliptic partial differential equation. The singular values of pde2961 is displayed on the right.



(a) Relative Frobenius Error Low-Rank Matrix Approximation



(b) Dominant Singular Value relative singular value decay

Figure 4: Low Rank Approximation for a synthetic matrix with polynomial singular value decay described in Equation (7).

Differential operator is compact. Furthermore, there are multiple works suggesting the learning of inverse differential operators is provably efficient [2, 3].

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 (7).

$$\mathbf{A} = \sum_{i \in [n]} i^{-p} \cdot \mathbf{U}_{(:,i)} \mathbf{V}^*_{(:,i)}, \quad \mathbf{U} \in \mathbb{O}_{m,m}, \mathbf{V} \in \mathbb{O}_{n,n} \tag{7}$$

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 Figure 1, 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 Figure 2, 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 Figure 3. 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 Figures 1 to 4 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 applications, [2, 10]. 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 helpful discussions.

References

[1] Ainesh Bakshi, Kenneth L. Clarkson, and David P. Woodruff. Low-rank approximation with $1/\varepsilon^3$ matrix-vector products. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2022, page 11301143, New York, NY, USA, 2022. Association for Computing Machinery.

[2] Nicolas Boullé, Christopher J. Earls, and Alex Townsend. Data-driven discovery of green's functions with human-understandable deep learning. *Scientific Reports*, 12(1):4824, Mar 2022.

[3] Nicolas Boullé, Diana Halikias, and Alex Townsend. Elliptic pde learning is provably data-efficient. *Proceedings of the National Academy of Sciences*, 120(39):e2303904120, 2023.

[4] Nicolas Boullé, Seick Kim, Tianyi Shi, and Alex Townsend. Learning green's functions associated with time-dependent partial differential equations. *The Journal of Machine Learning Research*, 23(1):9797–9830, 2022.

[5] Nicolas Boullé and Alex Townsend. A generalization of the randomized singular value decomposition. In *International Conference on Learning Representations*, 2022.

[6] Nicolas Boullé and Alex Townsend. Learning elliptic partial differential equations with randomized linear algebra. *Foundations of Computational Mathematics*, 23(2):709–739, Apr 2023.

[7] Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near-optimal column-based matrix reconstruction. *SIAM Journal on Computing*, 43(2):687–717, 2014.

[8] Christos Boutsidis and Alex Gittens. Improved matrix algorithms via the subsampled randomized hadamard transform. *SIAM Journal on Matrix Analysis and Applications*, 34(3):1301–1340, 2013.

[9] Mark Braverman, Elad Hazan, Max Simchowitz, and Blake Woodworth. The gradient complexity of linear regression. In *Conference on Learning Theory*, pages 627–647. PMLR, 2020.

[10] Steven L Brunton, Bernd R Noack, and Petros Koumoutsakos. Machine learning for fluid mechanics. *Annual review of fluid mechanics*, 52:477–508, 2020.

[11] George Casella and Roger L Berger. Statistical lnference. *Duxbury press*, 2002.

[12] Ali Civril. Column subset selection problem is ug-hard. *Journal of Computer and System Sciences*, 80(4):849–859, 2014.

[13] Timothy A. Davis and Yifan Hu. The university of florida sparse matrix collection. *ACM Trans. Math. Softw.*, 38(1), dec 2011.

[14] Amit Deshpande, Luis Rademacher, Santosh S Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. *Theory of Computing*, 2(1):225–247, 2006.

[15] Amit Deshpande and Santosh Vempala. Adaptive sampling and fast low-rank matrix approximation. In *International Workshop on Approximation Algorithms for Combinatorial Optimization*, pages 292–303. Springer, 2006.

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

[17] Hui-Yuan Fan, George S Dulikravich, and Zhen-Xue Han. Aerodynamic data modeling using support vector machines. *Inverse Problems in Science and Engineering*, 13(3):261–278, 2005.

[18] Alan Frieze, Ravi Kannan, and Santosh Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *Journal of the ACM (JACM)*, 51(6):1025–1041, 2004.

[19] Alex Gittens and Michael Mahoney. Revisiting the nystrom method for improved large-scale machine learning. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 567–575, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[20] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[21] Sariel Har-Peled. Low rank matrix approximation in linear time. *arXiv preprint arXiv:1410.8802*, 2014.

[22] Harvard Lomax, Thomas H Pulliam, David W Zingg, Thomas H Pulliam, and David W Zingg. *Fundamentals of computational fluid dynamics*, volume 246. Springer, 2001.

[23] Per-Gunnar Martinsson and Joel A Tropp. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica*, 29:403–572, 2020.

[24] L. Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 01 1960.

[25] Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *The quarterly journal of mathematics*, 11(1):50–59, 1960.

[26] Sean O'Rourke, Van Vu, and Ke Wang. Random perturbation of low rank matrices: Improving classical bounds. *Linear Algebra and its Applications*, 540:26–59, 2018.

[27] Saurabh Paul, Malik Magdon-Ismail, and Petros Drineas. Column selection via adaptive sampling. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[28] David Persson, Nicolas Boullé, and Daniel Kressner. Randomized nyström approximation of non-negative self-adjoint operators. *arXiv preprint arXiv:2404.00960*, 2024.

[29] Philippe Rigollet and Jan-Christian Hütter. High-dimensional statistics. *arXiv preprint arXiv:2310.19244*, 2023.

[30] Erhard Schmidt. Zur theorie der linearen und nichtlinearen integralgleichungen. *Mathematische Annalen*, 63(4):433–476, Dec 1907.

[31] Xiaoming Sun, David P Woodruff, Guang Yang, and Jialin Zhang. Querying a matrix through matrix-vector products. *ACM Transactions on Algorithms (TALG)*, 17(4):1–19, 2021.

[32] Joel A Tropp and Robert J Webber. Randomized algorithms for low-rank matrix approximation: Design, analysis, and applications. *arXiv preprint arXiv:2306.12418*, 2023.

[33] Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. Training networks in null space of feature covariance for continual learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 184–193, 2021.

A Proofs

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

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

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

Lemma 9. *Let $\mathbf{\Omega} = [\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_\ell] \in \mathbb{C}^{n\times\ell}$ such that $\boldsymbol{\omega}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ 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 $\mathbf{V}^*\mathbf{\Omega}$ are sampled from a centered multivariate Gaussian Distribution with second-moment matrix $\mathbf{K} = \mathbf{V}^*\mathbf{C}\mathbf{V}$.*

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

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

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

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

In the above, we have that each $[\mathbf{V}^*\mathbf{\Omega}]_{i,j}$ 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 $\mathbf{V}^*\mathbf{\Omega}$. For any $(i,j) \in [k]\times[\ell]$,

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

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

$$\begin{aligned} &\mathop{\mathbf{E}}_{\boldsymbol{\omega}_j\sim\mathcal{N}(\mathbf{0},\mathbf{C})}\left[(\mathbf{v}_i^*\boldsymbol{\omega}_j - \mathop{\mathbf{E}}_{\boldsymbol{\omega}_j\sim\mathcal{N}(\mathbf{0},\mathbf{C})}[\mathbf{v}_i^*\boldsymbol{\omega}_j])(\mathbf{v}_{i'}^*\boldsymbol{\omega}_j - \mathop{\mathbf{E}}_{\boldsymbol{\omega}_j\sim\mathcal{N}(\mathbf{0},\mathbf{C})}[\mathbf{v}_{i'}^*\boldsymbol{\omega}])\right] \\ &\quad = \mathop{\mathbf{E}}_{\boldsymbol{\omega}_j\sim\mathcal{N}(\mathbf{0},\mathbf{C})}\left[\mathbf{v}_i^*\boldsymbol{\omega}_j\boldsymbol{\omega}_j^*\mathbf{v}_{i'}\right] = \mathbf{v}_i^*\mathbf{C}\mathbf{v}_{i'} \end{aligned} \tag{8}$$

For the diagonal covariance elements, we have

$$\mathop{\mathbf{E}}_{\boldsymbol{\omega}_j\sim\mathcal{N}(\mathbf{0},\mathbf{C})}[(\mathbf{v}_i^*\boldsymbol{\omega}_j - \mathbf{E}[\mathbf{v}_i^*\boldsymbol{\omega}_j])^2] = \mathop{\mathbf{E}}_{\boldsymbol{\omega}_k\sim\mathcal{N}(\mathbf{0},\mathbf{C})}[\mathbf{v}_i^*\boldsymbol{\omega}_j\boldsymbol{\omega}_j^*\mathbf{v}_i] = \mathbf{v}_i^*\mathbf{C}\mathbf{v}_i \tag{9}$$

Then combining Equations (8) and (9), we have

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

Our proof is complete. ■

A.2 Proof of Theorem 4

We now give an improvement to the deterministic error bound given in Theorem 9.1 of [20] by Boutsidis et al. [7].

Lemma 10 (Lemma 3.2 in [7])**.** *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$, and let $\mathbf{Q}$ be an orthonormal basis of $\mathbf{A}\mathbf{\Omega}$ for $\mathbf{\Omega} \in \mathbb{R}^{n\times(k+p)}$, then*

$$\|\mathbf{A} - \mathbf{Q}(\mathbf{Q}\mathbf{A})_k\|_{\mathrm{F}}^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}^2 + \left\|\mathbf{\Sigma}_{k,\perp}\mathbf{V}_{k,\perp}\mathbf{\Omega}(\mathbf{V}_k\mathbf{\Omega})^+\right\|_{\mathrm{F}}^2$$

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

Lemma 11 (Lemma 3 in [6]). *Suppose $\mathbf{V}_k \in \mathbb{O}^{n,k}$ and $\mathbf{\Omega} \in \mathbb{C}^{n\times k+p}$ with columns sampled i.i.d from $\mathcal{N}(\mathbf{0}, \mathbf{K})$ and $p \geq 4$. Then with probability at least $1 - t^{-p}$,*

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

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

Lemma 12 (Proposition 10.4 in [20]). *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 $1 - t^{-(p+1)}$,*

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

We are now ready to prove our relative Spectral and Frobenius Error Norm Bounds for the Generalized Randomized SVD originally presented in [5].

Proof.[Proof of Theorem 4] Recall $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$, $\mathbf{\Omega}_k = \mathbf{V}_k^*\mathbf{\Omega}$ and $\mathbf{\Omega}_{k,\perp} = \mathbf{V}_{\perp,k}^*\mathbf{\Omega}$ where the columns of $\mathbf{\Omega} \in \mathbb{R}^{n\times(k+p)}$ 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,

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

In the above, the first inequality follows from the deterministic error bound in Lemma 10, the equality follows from noting that $\mathbf{\Omega} = \mathbf{C}^{1/2}\mathbf{X}$ where the entries of $\mathbf{X}$ are sampled i.i.d from $\mathcal{N}(0,1)$, the second inequality follows from Lemma 11 and Proposition 16 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 $\|\mathbf{X}\mathbf{Y}\|_{\mathrm{F}} \leq \|\mathbf{X}\|_{\mathrm{F}}\|\mathbf{Y}\|_2$. Our proof for the probabilistic bound for the Frobenius norm is complete. We now can prove the spectral norm bound,

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

The first inequality follows from the determistic error bound in Theorem 9.1 of [20]. The second relation follows from noting $\boldsymbol{\Omega} = \mathbf{C}^{1/2}\mathbf{X}$ where $\mathbf{X}$ is a Gaussian matrix. The third relation follows from Proposition 16 for the spectral norm and noting $\mathbf{X} \in \mathbb{R}^{n\times(k+p)}$. The final relation follows from noting by Cauchy-Scwharz,

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

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

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

Then, applying the bound in Lemma 12 with probability exceeding $1-t^{-p+1}$ completes the proof. ■

A.3 Proof of Lemma 6

Proof. Recall that $\mathbf{Q}_{12}$ is an orthonormal basis of $[\mathbf{A}\boldsymbol{\Omega}_1, \mathbf{A}\boldsymbol{\Omega}_2]$. Let $\mathbf{Q}_1$ be an orthnormal basis of $\mathbf{A}\boldsymbol{\Omega}_1$, it thus follows that $\mathbf{Q}_{12}$ is also an orthonormal basis of $[\mathbf{A}\boldsymbol{\Omega}_1, (\mathbf{I}-\mathbf{Q}_1\mathbf{Q}_1^*)\mathbf{A}\boldsymbol{\Omega}_2]$. This can be seen from the Classical Gram-Schmidt Procedure [30]. 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 13. *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,

$$\begin{aligned}
&\lambda_i((\mathbf{A} - \mathbf{Q}(\mathbf{Q}^*\mathbf{A})_k)^*(\mathbf{A} - \mathbf{Q}(\mathbf{Q}^*\mathbf{A})_k)) \\
&\quad= \lambda_i(\mathbf{A}^*\mathbf{A} - \mathbf{A}^*\mathbf{Q}(\mathbf{Q}^*\mathbf{A})_k - (\mathbf{Q}^*\mathbf{A})_k^*\mathbf{Q}^*\mathbf{A} + (\mathbf{Q}^*\mathbf{A})_k^*\mathbf{Q}^*\mathbf{Q}(\mathbf{Q}^*\mathbf{A})_k) \\
&\quad= \lambda_i(\mathbf{A}^*\mathbf{A} - (\mathbf{A}^*\mathbf{Q}\mathbf{Q}^*\mathbf{A})_k - (\mathbf{A}^*\mathbf{Q}\mathbf{Q}^*\mathbf{A})_k + (\mathbf{A}^*\mathbf{Q}\mathbf{Q}^*\mathbf{A})_k) \\
&\quad= \lambda_i(\mathbf{A}^*\mathbf{A} - (\mathbf{A}^*\mathbf{Q}\mathbf{Q}^*\mathbf{A})_k) \qquad (10)
\end{aligned}$$

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,

$$\begin{aligned}
\sigma_i^2(\mathbf{A} - \mathbf{Q}\mathbf{Q}^*\mathbf{A}) &= \lambda_i(\mathbf{A}^*\mathbf{A} - \mathbf{A}^*\mathbf{Q}\mathbf{Q}^*\mathbf{A}) \\
&= \lambda_i(\mathbf{A}^*\mathbf{A} - (\mathbf{A}^*\mathbf{Q}\mathbf{Q}^*\mathbf{A})_k - (\mathbf{A}^*\mathbf{Q}\mathbf{Q}^*\mathbf{A})_{k,\perp}) \\
&\leq \lambda_i(\mathbf{A}^*\mathbf{A} - (\mathbf{A}^*\mathbf{Q}\mathbf{Q}^*\mathbf{A})_k) && \text{Lemma 3} \\
&= \sigma_i^2(\mathbf{A} - \mathbf{A}(\mathbf{Q}^*\mathbf{A})_k) && \text{Equation (10)}
\end{aligned}$$

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 7

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

$$\left\|\mathbf{A} - \mathbf{Q}_+\mathbf{Q}_+^*\mathbf{A}\right\|_{\mathrm{F}} = \left\|(\mathbf{I} - \mathbf{Q}_-\mathbf{Q}_-^*)\mathbf{A} - \mathbf{Q}\mathbf{Q}^*(\mathbf{I} - \mathbf{Q}_-\mathbf{Q}_-^*)\mathbf{A}\right\|_{\mathrm{F}}$$

The above equation follows from the same argument as Lemma 6. Then from our relative-error accuracy bound in Corollary 5, we have

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

We then have from Lemma 13,

$$\begin{aligned}
\left\|(\mathbf{A} - \mathbf{Q}_-\mathbf{Q}_-^*\mathbf{A})_{k,\perp}\right\|_{\mathrm{F}}^2 &= \sum_{i\in[n]\setminus[k]} \sigma_i^2(\mathbf{A} - \mathbf{Q}_-\mathbf{Q}_-^*\mathbf{A}) && \text{Equation (2)} \\
&\leq \sum_{i\in[n]\setminus[k]} \sigma_i^2(\mathbf{A} - \mathbf{Q}_-(\mathbf{Q}_-^*\mathbf{A})_k) && \text{Lemma 13} \\
&= \left\|(\mathbf{A} - \mathbf{Q}_-(\mathbf{Q}_-^*\mathbf{A})_k)_{k,\perp}\right\|_{\mathrm{F}}^2 && \text{Equation (2)}
\end{aligned}$$

Then, from expanding out the Frobenius Norm, we have

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

We now leverage the Eckart-Young-Mirsky Theorem [16, 24], and obtain

$$\sum_{i\in[k]} \sigma_i^2(\mathbf{A} - \mathbf{Q}_-^*(\mathbf{Q}_-^*\mathbf{A})_k) \geq \inf_{\substack{\mathbf{B}\in\mathbb{C}^{m\times n} \\ \mathrm{rank}(\mathbf{B})=k}} \sum_{i\in[k]} \sigma_i^2(\mathbf{A} - \mathbf{B}) = \|\mathbf{A}_k - \mathbf{A}_{2k}\|_{\mathrm{F}}^2 \tag{11}$$

Then, from our improved Generalized Randomized SVD error bound in Theorem 4, we have with probability at least $0.99 - t^{-p}$,

$$\left\|\mathbf{A} - \mathbf{Q}_-(\mathbf{Q}_-^*\mathbf{A})_k\right\|_{\mathrm{F}}^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}^2\left(1 + 6t^2\log(100n\ell)\left\|\mathbf{V}_{k,\perp}^*\mathbf{C}^{1/2}\right\|_2^2\frac{\mathrm{Tr}((\mathbf{V}_k^*\mathbf{C}\mathbf{V}_k)^{-1})}{p+1}\right) \tag{12}$$

Then, from choosing our oversampling parameter sufficiently large in accoradance to Corollary 5,

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

We can then combine our results in Equations (11) and (12) to obtain

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

Our proof is complete. ■

A.5 Proof of Theorem 8

We first give necessary Lemmas for our proof.

Lemma 14 (Lemma 5.3 in [8]). *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{Y} \in \mathbb{R}^{m\times\ell}$, then for all $\mathbf{X} \in \mathbb{R}^{\ell\times n}$,*

$$\left\|\mathbf{A} - \mathbf{Y}\mathbf{Y}^+\mathbf{A}\right\|_{\mathrm{F}} \leq \|\mathbf{A} - \mathbf{Y}\mathbf{X}\|_{\mathrm{F}} \tag{13}$$

Lemma 15 (Matrix Pythagoras). *Let $\mathbf{X}, \mathbf{Y}$ be conformal matrices, then if $\mathbf{X}^*\mathbf{Y} = \mathbf{0}$, then*

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

Proof. The proof follows directly from Equation (2).

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

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

Proof.[Proof of Theorem 8] Our argument relies on the definition of $\mathbf{Q}$. Since we have that $\mathbf{Q}$ is an orthonormal basis of $\begin{bmatrix}\mathbf{Y}_1 & \mathbf{Y}_2\end{bmatrix}$. Then we have that $\mathbf{Q}$ is also the unique orthonormal basis of $\begin{bmatrix}\mathbf{Y}_1\mathbf{Y}_1^+\mathbf{Y}_1 & (\mathbf{I}-\mathbf{Y}_1\mathbf{Y}_1^+)\mathbf{Y}_2\end{bmatrix}$ by the definition of the Moore-Penrose Inverse. Let $\mathbf{X} \in \mathbb{R}^{2\ell\times n}$, then we have from Lemma 14,

$$\begin{aligned}\|\mathbf{A}-\mathbf{Q}\mathbf{Q}^*\mathbf{A}\|_{\mathrm{F}}^2 &\leq \left\|\mathbf{A} - \begin{bmatrix}\mathbf{P}_{\mathbf{Y}_1}\mathbf{A}\mathbf{G}_1 & (\mathbf{I}-\mathbf{P}_{\mathbf{Y}_1})\mathbf{A}\widehat{\mathbf{V}}_{k+p}\mathbf{G}_2\end{bmatrix}\begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix}\right\|_{\mathrm{F}}^2 \\ &= \underbrace{\|\mathbf{P}_{\mathbf{Y}_1}\mathbf{A}-\mathbf{P}_{\mathbf{Y}_1}\mathbf{A}\mathbf{G}_1\mathbf{X}_1\|_{\mathrm{F}}^2}_{T_1} + \underbrace{\left\|(\mathbf{A}-\mathbf{P}_{\mathbf{Y}_1}\mathbf{A}) - (\mathbf{A}-\mathbf{P}_{\mathbf{Y}_1}\mathbf{A})\widehat{\mathbf{V}}_{k+p}\mathbf{G}_2\mathbf{X}_2\right\|_{\mathrm{F}}^2}_{T_2}\end{aligned}$$

In the above, the second relation follows from Matrix Pythagoras (see Lemma 15). 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,

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

We then have for the second term,

$$\begin{aligned}T_2 &= \left\|(\mathbf{A}-\mathbf{P}_{\mathbf{Y}_1}\mathbf{A}) - (\mathbf{A}-\mathbf{P}_{\mathbf{Y}_1}\mathbf{A})\widehat{\mathbf{V}}_{k+p}\mathbf{G}_2\mathbf{X}_2\right\|_{\mathrm{F}}^2 \\ &= \left\|(\mathbf{A}-\mathbf{A}\widehat{\mathbf{V}}_{k+p}\mathbf{G}_2\mathbf{X}_2) - (\mathbf{P}_{\mathbf{Y}_1}\mathbf{A}-\mathbf{P}_{\mathbf{Y}_1}\mathbf{A}\widehat{\mathbf{V}}_{k+p}\mathbf{G}_2\mathbf{X}_2)\right\|_{\mathrm{F}}^2\end{aligned}$$

Then, setting $\mathbf{X}_2 = (\mathbf{V}_k^*\widehat{\mathbf{V}}_{k+p}\mathbf{G}_2)^+\mathbf{V}_k^*$, we obtain

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

In the above, the second inequality follows from the Conjugation Rule (see Proposition 2) with $\mathbf{P}_{\mathbf{A}_{k,\perp}\mathbf{G}_1} \preceq \mathbf{P}_{\mathbf{A}\mathbf{G}_1}$, suppose $\mathrm{range}(\mathbf{P}_1) \subset \mathrm{range}(\mathbf{P}_2)$, then $\mathbf{P}_1 \preceq \mathbf{P}_2$, and we have

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

In the above, the first and final relations follow from an alternative definition of the Frobenius Norm, the second relation follows from Equation (1), and the third relation follows from the Conjugation Rule. Then, noting that $\mathrm{range}(\mathbf{A}_{k,\perp}\mathbf{G}_1) \subset \mathrm{range}(\mathbf{A}\mathbf{G}_1)$, we obtain the desired result. The final inequality follows from the standard randomized SVD bound (see e.g. Theorem 10.6 in [20]) 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 4 with failure probability less than $\delta + t^{-(p+1)}$. ■

B Probability Theory

Proposition 16. *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 $\xi \in \{2, \mathrm{F}\}$, 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.

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

We now show an analagous 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 [29], we have the following relation for any matrix $\mathbf{B}\in\mathbb{R}^{m\times n}$,

$$\|\mathbf{B}\|_2 = \sup_{\mathbf{v}\in\mathbb{S}^{n-1}}\|\mathbf{Bv}\|_2 = \sup_{\mathbf{u}\in\mathbb{S}^{m-1}}\sup_{\mathbf{v}\in\mathbb{S}^{n-1}}|\mathbf{u}^*\mathbf{Bv}| \tag{14}$$

With the relation given in Equation (14) in hand, we have

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

We now bound the maximum Gaussian over a finite sample.

$$\begin{aligned}\mathbf{Pr}\left\{\max_{(k_1,k_2)\in[n]\times[m]}G_{k_1,k_2}^2 \geq t\right\} &= \mathbf{Pr}\left\{\max_{(k_1,k_2)\in[n]\times[m]}|G_{k_1,k_2}| \geq \sqrt{t}\right\} \\ &\leq \frac{\sqrt{2}nm}{\sqrt{\pi}}\int_{\sqrt{t}}^{\infty}e^{-x^2/2}dx \leq \frac{\sqrt{2}nm}{\sqrt{\pi}}\int_{\sqrt{t}}^{\infty}\frac{xe^{-x^2/2}}{\sqrt{t}}dx = \frac{\sqrt{2}nm}{\sqrt{\pi}}e^{-t/2} \leq \delta\end{aligned} \tag{15}$$

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 [11]. 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.

■