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

Arvind K. 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 in case the when the matrix A is accessible only via matrix-vector products and we are given a budget of $k + p$ matrix-vector products. This situation arises in practice when the cost of data acquisition is high, despite the Numerical Linear Algebra (NLA) costs being low. We create an adaptive sampling algorithm to optimally choose vectors to sample. The Randomized Singular Value Decomposition (rSVD) is an effective algorithm for obtaining the low rank representation of a matrix developed by [15]. Recently, [4] 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, A. In this work, we develop an adaptive sampling framework for the Matrix-Vector Product Model which does not need prior information on the matrix A. We provide a novel theoretic analysis of our algorithm with subspace perturbation theory. We extend the analysis of [22] for right singular vector approximations from the randomized SVD in the context of non-symmetric rectangular matrices. We also test our algorithm on various synthetic, real-world application, and image matrices. Furthermore, we show our theory bounds on matrices are stronger than state-of-the-art methods with the same number of matrix-vector product queries.

1 Introduction

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 [13], fluid dynamics [17]. 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, [15], 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 A, we can modify the Covariance Matrix such that the sampled vectors are within the column space of 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 [23] 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

Preprint. Under review.

model used a fixed covariance matrix structure. In this paper, we consider the adaptive setting where the algorithm $\mathcal{A}$ chooses a vector $\mathbf{x}^{(k)}$ with access to the previous query vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k-1)}$, the matrix-vector products $A\mathbf{x}^{(1)}, \ldots, A\mathbf{x}^{(k-1)}$, and the intermediate low-rank matrix approximations, $Q^{(k)}(Q^{(k)})^* A$, where $Q^{(k)} \triangleq \mathrm{orth}(AV^{(k)})$ where $V^{(k)}$ is the concatenation of vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k)}$ and $Q^{(k)} \triangleq \mathrm{orth}(AV^{(k)})$.

Adaptive Sampling techniques for Low-Rank Matrix Approximation first appeared in CUR Matrix Decomposition in [14]. Optimial column-sampling for the CUR Matrix Decomposition received much attention as can be seen in the works [16, 10, 11]. More recently, [20] gave an algorithm for sampling the rows for CUR-Matrix Factorization and proved error bounds by induction. Similar to adaptively choosing a function, in recommender systems, the company can ask users for surveys and obtain data with high probability is a better representation of the column space of A than a random sample. Choosing the right people to give an incentivized survey (e.g. gift card upon completition) can save a company significant expenses.

Adaptively sampling vectors for matrix problems has been studied in detail in [21]. The theoretical properties of adaptively sampled matrix vector queries for estimating the minimum eigenvalue of a Wishart matrix have been studied in [7]. 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 computaton of the low-rank approximation at each step to sample the next vector. Although there are are runtime limitations, both in theory under certain conditions and most real-world matrices, our algorithm gets the most value out of each sampled vector.

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 A.
2. We provide a novel theoretical analysis which utilizes subspace perturbation theory.
3. We perform extensive experiments on matrices with various spectrums and show the effectiveness of Bayes Near-Optimal Sampling comparing to State-of-the-Art Low-Rank Matrix Approximation Algorithms in the Matrix-Vector Product Query Model.

2 Adaptive Sampling

NB: We need a good literature review and background material about rSVD here, see [Martinsson, Tropp, Acta Numerica].

The Randomized SVD is a method to find an orthonormal matrix that captures the range of the the top left singular space of A by multiplying the matrix A with a Matrix with Standard Normal entries[18]. The analysis has been extended to SRTT matrices by Boutsidis and Gittens [6]. 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.

The Adaptive Range Finder Algorithm was propose in [15] as a method to guarantee a high accuracy guarantee for the low-rank approximation by sampling vectors one at a time.

Definition 1. *Given access to vector queries* $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(k-1)}$ *where w.l.o.g* $\|\mathbf{v}^{(\zeta_1)}\| = 1$ *for all* $i \in [k-1]$ *and* $(\mathbf{v}^{(\zeta_1)})^* \mathbf{v}^{(j)} = \delta_{ij}$ *for all* $i, j \in [k-1] \times [k-1]$*, matrix-vector queries,* $A\mathbf{v}^{(1)}, \ldots, A\mathbf{v}^{(k-1)}$*, and intermediate low-rank approximations,* $Q^{(1)}(Q^{(1)})^* A, \ldots, Q^{(k-1)}(Q^{(k-1)})^* A$ *, which requires* $k = O(\Xi)$ *vector queries to obtain a rank-*k *matrix with orthogonal columns,* Q*, such that with probability at least* $1 - \delta$ *for a* $\delta \in (0, 1)$ *such that* $\|A - QQ^* A\|_{\mathrm{F}} \leq (1 + \varepsilon) \min_{\substack{U \in \mathbb{C}^{m \times k} \\ U^* U = I_k}} \|A - UU^* A\|_{\mathrm{F}}$

2.1 Algorithm

NB: We are there 2 algorithms here? We need to explain them in details underneath with motivation and intuition, also discuss benefits compared to alternatives such as fewer number of queries. State that our main objective is to use as few matvecs as possible.

The Pseudo Code for the optimal function sampling is given in Algorithm 1. For efficient updates, we frame all operations as rank-1 updates.

Algorithm 1 Near-Optimal Adaptive Sampling for Low-Rank Matrix Approximation

input: Target Matrix $A \in \mathbb{C}^{m\times n}$, target rank k, and oversampling parameter p.
output: Low-Rank Approximation $\hat{A}$ of A.

$\widehat{A}^{(0)} \leftarrow \mathbf{0}$
Form $\Omega \in \mathbb{C}^{n\times p}$ with columns sampled i.i.d from $\mathcal{N}(\mathbf{0}, I)$
$Y^{(0)} \leftarrow A\Omega \in \mathbb{C}^{m\times p}$ and obtain the factorization $Y^{(0)} = QR$
for $t \in [k+p]$ **do**
    $\widehat{A}^{(t)} \leftarrow \widehat{A}^{(t)} + Q_{[:,t-1]}Q^*_{[:,t-1]}A$
    $C^{(t)} \leftarrow (\widehat{A}^{(t)})^+\widehat{A}^{(t)}$
    Sample $\boldsymbol{\omega}^{(t)} \sim \mathcal{N}(\mathbf{0}, C^{(t)})$ and compute $A\boldsymbol{\omega}^{(t)}$
    $Y^{(t)} \leftarrow [Y^{(t-1)}, A\boldsymbol{\omega}^{(t)}]$ and obtain the factorization $Y^{(t)} = QR$
return: $\widehat{A}^{(k+p)}$

Algorithm 2 Power-Method Adaptive Sampling for Low-Rank Matrix Approximation

input: Target Matrix $A \in \mathbb{C}^{m\times n}$, target rank k, and oversampling parameter p.
output: Low-Rank Approximation $\hat{A}$ of A.

$\widehat{A}^{(0)} \leftarrow 0$
Sample $\boldsymbol{\omega} \in \mathbb{C}^n$ from $\mathcal{N}(\mathbf{0}, I)$
$Y^{(0)} \leftarrow A\boldsymbol{\omega} \in \mathbb{C}^{m\times p}$ and obtain the factorization $Y^{(0)} = QR$
for $t \in [k+p]$ **do**
    $\widehat{A}^{(t)} \leftarrow \widehat{A}^{(t)} + Q_{[:,t-1]}Q^*_{[:,t-1]}A$
    $\boldsymbol{\omega}^{(t)} \leftarrow$ POWER ITERATION$\big((I - QQ^*)A, \boldsymbol{\omega}, p\big)$
    $Y^{(t)} \leftarrow [Y^{(t-1)}, A\boldsymbol{\omega}^{(t)}]$ and obtain the factorization $Y^{(t)} = QR$
return: $\widehat{A}^{(k+p)}$

In Algrithm 1, we first sample a standard normal gaussian matrix which can be considered as the oversampling vectors. These oversampling vectors are used to approximate the first singular vector. This is the first vector which is *adaptively* sampled. Next, we form the low-rank approximation QQ^*A where $Q = \mathrm{orth}(A\Omega)$ where Ω is a matrix of all our adaptively chosen vector queries. From here, we choose the kth right singular vector of the SVD of the the appoximation QQ^*A. This process is then repeated for k iterations. The final low-rank approximation, QQ^*A, is then returned.

3 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 § A.

NB: Maybe say something like "the proofs of all results in this section are deferred to the appendix". Then remove all proof statements but keep a "proof idea" paragraph detailing the main idea leading to the theorem and its main consequences. E.g.: The following result states that [...] and can be deduced from [...].

3.1 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 [4].

Theorem 2. *Let $A \in \mathbb{C}^{m\times n}$, set $k \geq 1$ an integer, an oversampling parameter $p \geq 2$. Let $\Omega \in \mathbb{R}^{n\times k+p}$ represent the test matrix with columns sampled from $\mathcal{N}(\mathbf{0}, C)$. Then, let $QR = A\Omega$ represent the economized QR decomposition of $A\Omega$, then with probability at least $1 - \delta - t^{-p}$,*

$$\|A - QQ^*A\|_{\mathrm{F}} \leq \|\Sigma_{k,\perp}\|_{\mathrm{F}} \left(1 + 3t\|K_{22}\|_2 \sqrt{\log(n(k+p)/\delta)} \sqrt{\frac{\mathrm{Tr}(\mathrm{diag}(K_{11})^{-1})}{p+1}}\right)$$

where

$$K = V^*CV = \begin{bmatrix} V_k^*CV_k & V_k^*CV_{k,\perp} \\ V_{k,\perp}^*CV_k & V_{k,\perp}^*CV_{k,\perp} \end{bmatrix}$$

The proof of Theorem 2 can be derived with the deterministic error bound of Theorem 9.1 in [15] and combining Lemma 8 with Lemma 10.

Corollary 3. *Let $A \in \mathbb{C}^{m\times n}$, set $k \geq 1$ an integer, an oversampling parameter $p \geq 2$. Let $\Omega \in \mathbb{R}^{n\times k+p}$ represent the test matrix with columns sampled from $\mathcal{N}(\mathbf{0}, C)$. Then, let $QR = A\Omega$ represent the economized QR decomposition of $A\Omega$. Then if*

$$p \geq 6\varepsilon^{-1}t^2 \log(3nk/\delta)$$

With probability exceeding $1 - \delta - t^{-p}$,

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

Corollary 3 develops a relative Frobenius norm error bound for the Generalized Randomized SVD.

3.2 Near Optimal Function Sampling

In this section, we will discuss the theoretical motivation for adaptive sampling. Adaptive sampling in works such as [11] and [20] focus on the residual matrix, $R_t = A - Q_tQ_t^*A$. Paul et al. [20] consider the residual matrix as the matrix remaining after removing known information about 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.

Lemma 4. *Let $\Omega_+ = [\Omega, \Omega_-]$, $Y = A\Omega$, $Q = \mathrm{orth}(Y)$, and $Q_+ = \mathrm{orth}([Y, A\Omega_-])$. Finally, for shorthand, define $P_\perp = I - QQ^*$, then for $\xi \in \{2, \mathrm{F}\}$,*

$$\|A - Q_+Q_+^*A\|_\xi = \|P_\perp A - Q_-Q_-^*P_\perp A\|_\xi \tag{1}$$

From Lemma 4, to minimize the RHS of Equation (1) we observe that Q_- is equal to the dominant left singular space of $P_\perp A$. From this result, we then see that at each iteration, the optimal vector query is the top right singular vector of $P_\perp A$. Moving forward, we then discuss how we can use the intermediate low rank approximations $Q_tQ_t^*A$ for $t \in [k+p]$ to obtain sampling vectors closer to the top right singular vector of $P_\perp A$.

3.3 Analysis of Algorithm 1

First we will introduce a lemma for the resultant vector of sampling from $C^{(k)}$. Since our general proof technique will be an induction. We first want to understand how well we are able to approximate the first right singular vector. To do this, we must know the singular vector perturbation from the error of the low-rank matrix approximation.

Lemma 5. *Let $A \in \mathbb{C}^{m\times n}$ and Q be an orthogonal matrix representing the basis of the subspace of $Y \in \mathbb{C}^{m\times k}$. Let $C \in \mathbb{C}^{n\times n}$ such that $C \succeq \mathbf{0}$, then suppose $\Omega \in \mathbb{C}^{n\times p}$ has columns sampled i.i.d from $\mathcal{N}(\mathbf{0}, C)$. Let $Q_+ = \mathrm{orth}([Y, A\Omega])$, it then follows,*

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

*when $p = O\left(\frac{\|\tan\Theta(\widetilde{V}_1, C^{1/2})\|}{\varepsilon}\right)$ where $\widetilde{V}$ is defined as in Equation (**??**).*

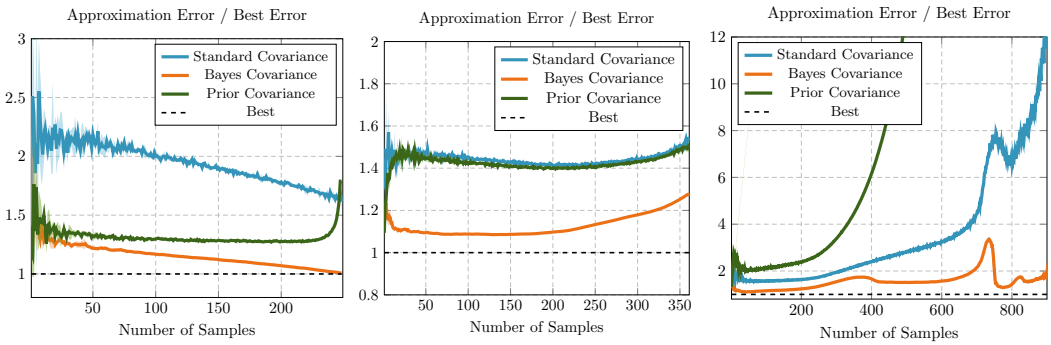

Figure 1: Low Rank Approximation for the Inverse Differential Operator given in Equation 2 Figure Left, Differential Operator Matrix `Poisson2D` [9] Figure Center, and Differential Operator Matrix `DK01R` [9] Figure Right. The experiment on the left is from [4, Figure 2]. We use the discretized Green's Function as the prior covariance matrix. In this set of experiments, we learn a low rank approximation of the inverse operator.

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

Lemma 6. *Let $A \in \mathbb{C}^{m\times n}$ and Q be an orthogonal matrix representing the basis of the subspace of $Y \in \mathbb{C}^{m\times k}$. Let $\Omega \in \mathbb{C}^{n\times p}$ such that the columns are sampled i.i.d from $\mathcal{N}(\mathbf{0}, I)$. Denote $Y_+ \triangleq \operatorname{orth}([Y, A\Omega])$ and $Q_+ \triangleq \operatorname{orth}(Y_+)$, then*

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

when $p = O(1/\varepsilon)$.

Proof. The proof follows from a direct application of Lemma 5 and setting $C = I$. ■

Theorem 7. *Let $A \in \mathbb{C}^{m\times n}$ and let $C_i \in \mathbb{C}^{n\times n}$ for $i \in [k+p]$ be PSD covariance matrices. Then suppose $\boldsymbol{\omega}_i \sim \mathcal{N}(\mathbf{0}, C_i)$ for $i \in [k+p]$. Denote $Y = \operatorname{orth}(A\Omega)$ and $Q = \operatorname{orth}(Y)$, then*

$$\|A - QQ^*A\|_{\mathrm{F}}^2 \leq \Xi$$

4 Numerical Experiments

NB: There are way too many figures here, we need to have a selection that features the main aspect of the method to discuss extensively, eventually having more experiments in Appendix. One thing we might want is comparison against Power method, computational timings. Can our method be adapted for Nystrom as well (see `https://arxiv.org/abs/2404.00960` for the general bounds)?

NB: For the tikz figures, you can generate one pdf for 3 figures by doing the following commented lines.

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. In our first experiment we attempt to learn the discretized 250×250 matrix of the inverse of the following differential operator:

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

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 [3, 5]) as the inverse Differential operator is compact and there are nice theoretical properties, such as data efficiency.

In Figure 2, note if the Covariance Matrix has eigenvectors orthogonal to the left singular vectors of A, then the randomized SVD will not perform well. Furthermore, in Figure 2, we can note even without knowledge of the Green's Function, our method achieves lower error than with the Prior Covariance. We also test our algorithm against various Sparse Matrices in the Texas A& M Sparse

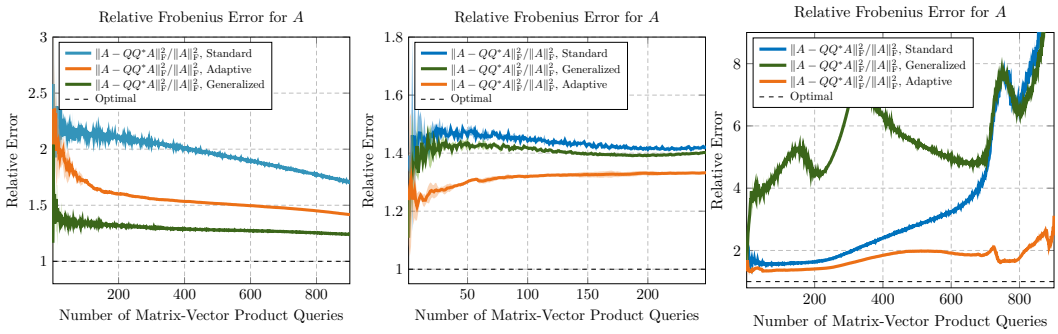

Figure 2: Results for the Power-Method based Adaptive Sampling Method described in Algorithm 2. Low Rank Approximation for the Inverse Differential Operator given in Equation 2 Figure Left, Differential Operator Matrix `Poisson2D` [9] Figure Center, and Differential Operator Matrix `DK01R` [9] Figure Right. The experiment on the left is from [4, Figure 2]. We use the discretized Green's Function as the prior covariance matrix. In this set of experiments, we learn a low rank approximation of the inverse operator.

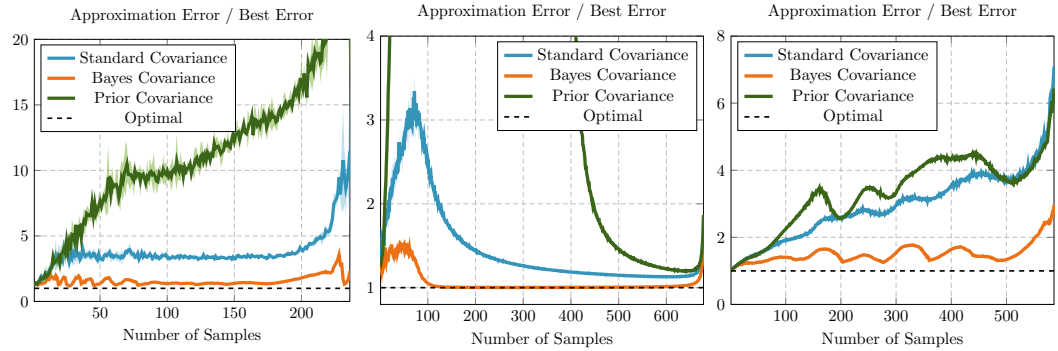

Figure 3: Low Rank Approximation for a matrix for a Computational Fluid Dynamics Problem, `saylr1` Figure Left is from [9]. Subsequent 2D/3D Problem `fs-680-2` Figure Center is from [9]. Differential Stiffness Matrix from a Nastran Buckling Problem, `bcsstk34` Figure Right is from [9].

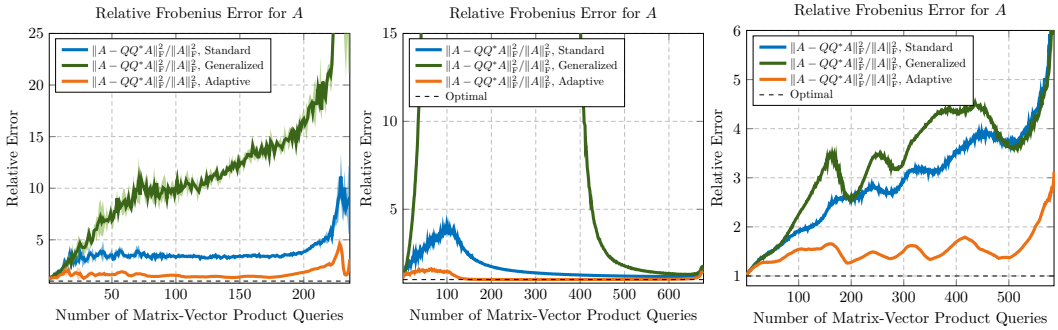

Figure 4: Results for the Power-Method based Adaptive Sampling Method described in Algorithm 2. Low Rank Approximation for a matrix for a Computational Fluid Dynamics Problem, `saylr1` is from [9]. Subsequent 2D/3D Problem `fs-680-2` Figure Center is from [9]. Differential Stiffness Matrix from a Nastran Buckling Problem, `bcsstk34` Figure Right is from [9].

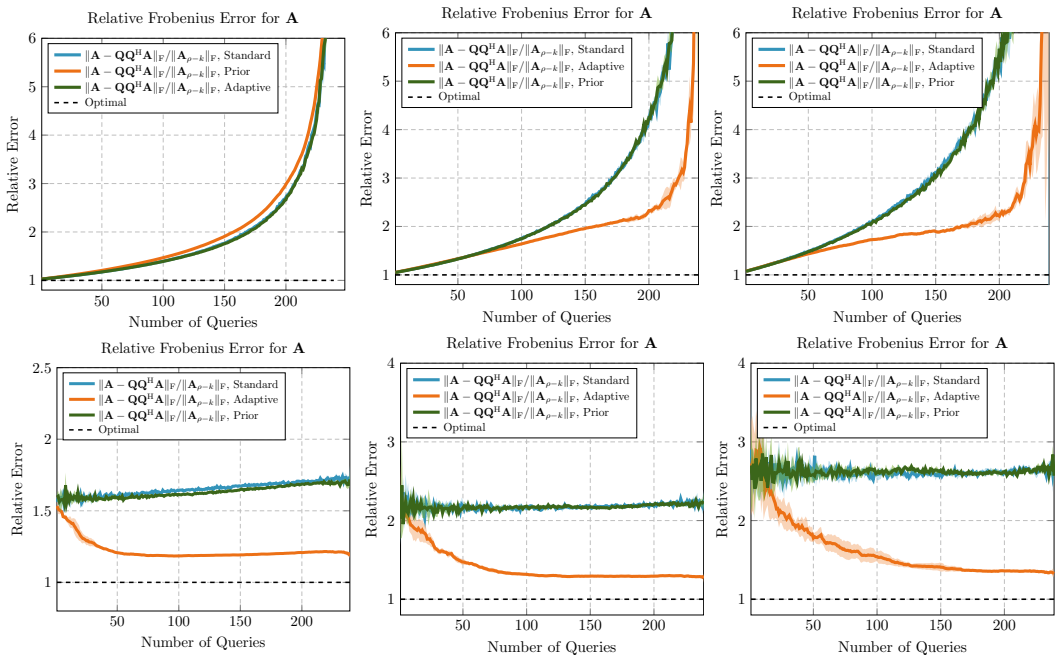

Figure 5: Low Rank Approximation for synthetic matrices with decay described in Equation 3. In *(Top Left)*, $\ell = 1$, in *(Top Center)*, $\ell = 2$, and in *(Top Right)*, $\ell = 3$. In *(Bottom Left)*, $\ell = -1$, in *(Bottom Center)*, $\ell = -2$, and in *(Bottom Right)*, $\ell = -3$. We use the Gram matrix with Gaussian Kernel and $\gamma = 0.01$ for the Prior Covariance Matrix.

Matrix Suite, [9]. In Figure 4 , we choose a fluid dynamics problem due to its relevance in low-rank approximation [8].

Singular Value Decay. Our first synthetic matrix is developed in the following scheme:

$$A = \sum_{i=1}^{\rho} \frac{100 i^{\ell}}{n} U_{(:,i)} V^*_{(:,i)}, \quad U \in \mathbb{O}_{m,k}, V \in \mathbb{O}_{n,k} \tag{3}$$

We will experiment with linear decay, $\ell = 1$, quadratic decay, $\ell = 2$, and cubic decay, $\ell = 3$. We see adaptive sampling is as good as the Randomized SVD when there is linear decay, however, when there is quadratic decay or greater, we see that there is significant improvement when using adaptive sampling. This is corroborated in our theory, where the bounds are dependent on the ratios between the singular values. Our second synthetic matrix is developed in the following scheme:

$$A = \sum_{i=1}^{\rho} (1-\delta)^i U_{(:,i)} V^*_{(:,i)}, \quad U \in \mathbb{O}_{m,k}, V \in \mathbb{O}_{n,k} \tag{4}$$

We observe when there exists exponential decay of the singular values in Figure 7, adaptive sampling in accordance without scheme

5 Conclusions

We have theoretically and empirically analyzed a novel Covariance Update to iteratively construct the sampling matrix, Ω 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, 8]. 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

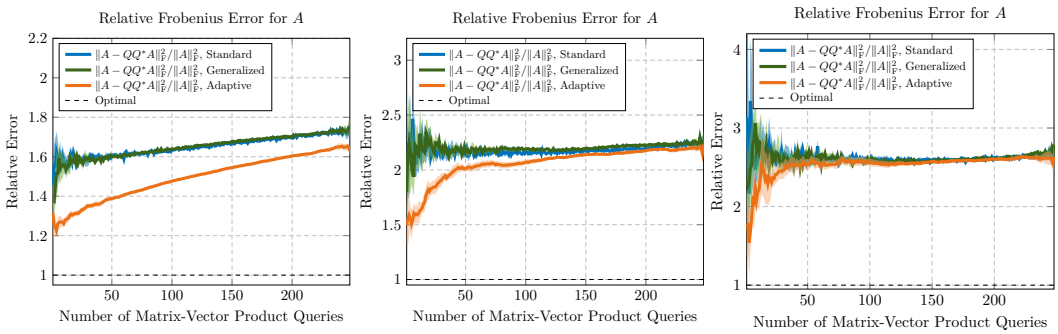

Figure 6: Results for the Power-Method based Adaptive Sampling Method desrcibed in Algorithm 2. Rank Approximation for synthetic matrices with decay described in Equation 3. In *(Top Left)*, $\ell = 1$, in *(Top Center)*, $\ell = 2$, and in *(Top Right)*, $\ell = 3$. In *(Bottom Left)*, $\ell = -1$, in *(Bottom Center)*, $\ell = -2$, and in *(Bottom Right)*, $\ell = -3$. We use the Gram matrix with Gaussian Kernel and $\gamma = 0.01$ for the Prior Covariance Matrix.

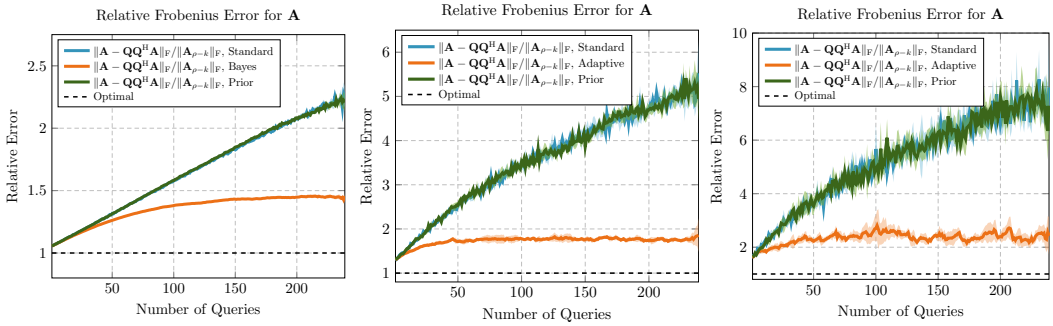

Figure 7: Low Rank Approximation for synthetic matrices with exponential decay described in Equation 4. In Figure Left, $\delta = 0.01$, in Figure Center, $\delta = 0.05$, and in Figure Right, $\delta = 0.1$. We use the Gram matrix with Gaussian Kernel and $\gamma = 0.01$ for the Prior Covariance Matrix.

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 an REU Project (give reference)... This work was supported by the Office of Naval Research (ONR), under grant N00014-23-1-2729. We thank Alex Gittens and Christopher Wang for helpful discussions.

References

[1] A. Bakshi, K. L. Clarkson, and D. 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, New York, NY, USA, 2022, Association for Computing Machinery, p. 1130–1143.

[2] N. Boullé, C. J. Earls, and A. Townsend, *Data-driven discovery of green's functions with human-understandable deep learning*, Scientific Reports, 12 (2022), p. 4824.

[3] N. Boullé, S. Kim, T. Shi, and A. Townsend, *Learning green's functions associated with time-dependent partial differential equations*, The Journal of Machine Learning Research, 23 (2022), pp. 9797–9830.

[4] N. BOULLÉ AND A. TOWNSEND, *A generalization of the randomized singular value decomposition*, in International Conference on Learning Representations, 2022.

[5] N. BOULLÉ AND A. TOWNSEND, *Learning elliptic partial differential equations with randomized linear algebra*, Foundations of Computational Mathematics, 23 (2023), pp. 709–739.

[6] C. BOUTSIDIS AND A. GITTENS, *Improved matrix algorithms via the subsampled randomized hadamard transform*, SIAM Journal on Matrix Analysis and Applications, 34 (2013), pp. 1301–1340.

[7] M. BRAVERMAN, E. HAZAN, M. SIMCHOWITZ, AND B. WOODWORTH, *The gradient complexity of linear regression*, in Conference on Learning Theory, PMLR, 2020, pp. 627–647.

[8] S. L. BRUNTON, B. R. NOACK, AND P. KOUMOUTSAKOS, *Machine learning for fluid mechanics*, Annual review of fluid mechanics, 52 (2020), pp. 477–508.

[9] T. A. DAVIS AND Y. HU, *The university of florida sparse matrix collection*, ACM Trans. Math. Softw., 38 (2011).

[10] A. DESHPANDE, L. RADEMACHER, S. S. VEMPALA, AND G. WANG, *Matrix approximation and projective clustering via volume sampling*, Theory of Computing, 2 (2006), pp. 225–247.

[11] 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.

[12] C. ECKART AND G. YOUNG, *The approximation of one matrix by another of lower rank*, Psychometrika, 1 (1936), pp. 211–218.

[13] H.-Y. FAN, G. S. DULIKRAVICH, AND Z.-X. HAN, *Aerodynamic data modeling using support vector machines*, Inverse Problems in Science and Engineering, 13 (2005), pp. 261–278.

[14] A. FRIEZE, R. KANNAN, AND S. VEMPALA, *Fast monte-carlo algorithms for finding low-rank approximations*, Journal of the ACM (JACM), 51 (2004), pp. 1025–1041.

[15] N. HALKO, P. G. MARTINSSON, AND J. A. TROPP, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Review, 53 (2011), pp. 217–288.

[16] S. HAR-PELED, *Low rank matrix approximation in linear time*, arXiv preprint arXiv:1410.8802, (2014).

[17] H. LOMAX, T. H. PULLIAM, D. W. ZINGG, T. H. PULLIAM, AND D. W. ZINGG, *Fundamentals of computational fluid dynamics*, vol. 246, Springer, 2001.

[18] P.-G. MARTINSSON AND J. A. TROPP, *Randomized numerical linear algebra: Foundations and algorithms*, Acta Numerica, 29 (2020), pp. 403–572.

[19] L. MIRSKY, *Symmetric gauge functions and unitarily invariant norms*, The Quarterly Journal of Mathematics, 11 (1960), pp. 50–59.

[20] S. PAUL, M. MAGDON-ISMAIL, AND P. DRINEAS, *Column selection via adaptive sampling*, in Advances in Neural Information Processing Systems, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds., vol. 28, Curran Associates, Inc., 2015.

[21] 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.

[22] R.-C. TZENG, P.-A. WANG, F. ADRIAENS, A. GIONIS, AND C.-J. LU, *Improved analysis of randomized svd for top-eigenvector approximation*, in Proceedings of The 25th International Conference on Artificial Intelligence and Statistics, G. Camps-Valls, F. J. R. Ruiz, and I. Valera, eds., vol. 151 of Proceedings of Machine Learning Research, PMLR, 28–30 Mar 2022, pp. 2045–2072.

[23] S. WANG, X. LI, J. SUN, AND Z. 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, 2021, pp. 184–193.

A Proofs

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

A.1 Distribution of $V^*\Omega$

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

Lemma 8. *Let $\Omega = [\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_\ell] \in \mathbb{C}^{n\times\ell}$ such that $\boldsymbol{\omega}_i \sim \mathcal{N}(\mathbf{0}, C)$ for all $i \in [\ell]$ and $C \succeq 0$ is symmetric. Then, let $V \in \mathbb{O}_{n\times k}$, it then follows the columsn of V^* are sampled from a centered multivariate Gaussian Distribution with second-moment matrix $K = \operatorname{diag}(V^*CV)$.*

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

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

Let $\mathcal{D} = \mathcal{N}(\mathbf{0}, I)$. We will first show that the entries are Gaussian. From the fact that $C \succeq 0$ and C is symmetric, we have that $C = U\Sigma U$ for a unitary U and diagonal Σ with non-negative elements on the diagonal. Then for any $i \in [k]$ and $j \in [\ell]$, we have for $\mathbf{x} \sim \mathcal{D}$,

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

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

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

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

$$\begin{aligned} &\mathop{\mathbf{E}}_{\boldsymbol{\omega}_j\sim\mathcal{N}(\mathbf{0},C)} \mathop{\mathbf{E}}_{\boldsymbol{\omega}_{j'}\sim\mathcal{N}(\mathbf{0},C_{j'})} \left[(\mathbf{v}_i^*\boldsymbol{\omega}_j - \mathop{\mathbf{E}}_{\boldsymbol{\omega}_i\sim\mathcal{N}(\mathbf{0},C)}[\mathbf{v}_i^*\boldsymbol{\omega}_j])(\mathbf{v}_{i'}^*\boldsymbol{\omega}_{i'} - \mathop{\mathbf{E}}_{\boldsymbol{\omega}_{j'}\sim\mathcal{N}(\mathbf{0},C_{j'})}[\mathbf{v}_{i'}^*\boldsymbol{\omega}_{j'}]) \right] \\ &\quad = \mathop{\mathbf{E}}_{\boldsymbol{\omega}_j\sim\mathcal{N}(\mathbf{0},C)} \mathop{\mathbf{E}}_{\boldsymbol{\omega}_{j'}\sim\mathcal{N}(\mathbf{0},C_{j'})} [(\mathbf{v}_i^*\boldsymbol{\omega}_j)(\mathbf{v}_{i'}^*\boldsymbol{\omega}_{j'})] \\ &\quad \overset{(\zeta_1)}{=} \mathop{\mathbf{E}}_{\mathbf{x}_j\sim\mathcal{D}} \left(\mathop{\mathbf{E}}_{\mathbf{x}_{j'}\sim\mathcal{D}} [(\mathbf{v}_{i'}^* C_{j'}\mathbf{x}_{j'})(\mathbf{v}_i^* C\mathbf{x}_j) \mid \mathbf{v}_i^* C\mathbf{x}_j] \right) = 0 \end{aligned}$$

In the above, (ζ_1) follows from the law of total expectation. By conditioning on $\mathbf{v}_i^*\boldsymbol{\omega}_j$, we are able to obtain zero covariance even if $C_{j'}$ is dependent on C. For the diagonal covariance elements, we have

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

We thus have for all $(i,i') \in [k]\times[k]$,

$$[K_j]_{ii'} = \begin{cases} \mathbf{v}_i^* C\mathbf{v}_i & i = i' \\ 0 & i \neq i' \end{cases}$$

Our proof is complete. ■

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

Lemma 9. *Suppose $V_k \in \mathbb{O}^{n,k}$ and $\Omega \in \mathbb{C}^{n\times k+p}$ with columns sampled from $\mathcal{N}(\mathbf{0}, K)$ and $p \geq 4$. Then with probability at least $1 - t^{-p}$,*

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

A.2 Proof of Theorem 2

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

$$K = \begin{bmatrix} V_k^* C V_k & V_k^* C V_{k,\perp} \\ V_{k,\perp}^* C V_k & V_{k,\perp}^* C V_{k,\perp} \end{bmatrix}$$

We then have the following manipulations,

$$\|A - QQ^*A\|_{\mathrm{F}} \leq \left(\|\Sigma_{k,\perp}\|_{\mathrm{F}}^2 + \|\Sigma_{k,\perp}\Omega_{k,\perp}\Omega_k^+\|_{\mathrm{F}}^2 \right)^{1/2}$$

$$= \left(\|\Sigma_{k,\perp}\|_{\mathrm{F}}^2 + \|\Sigma_{k,\perp} V_{k,\perp}^* C^{1/2} X (V_k^* C^{1/2} X)^+ \|_{\mathrm{F}}^2 \right)^{1/2}$$

$$\leq \left(\|\Sigma_{k,\perp}\|_{\mathrm{F}}^2 + 2\log(n(k+p)/\delta) \|\Sigma_{k,\perp} V_{k,\perp}^* C^{1/2}\|_{\mathrm{F}}^2 \|(V_k^* C^{1/2} X)^+\|_{\mathrm{F}}^2 \right)^{1/2}$$

$$\leq \|\Sigma_{k,\perp}\|_{\mathrm{F}} \left(1 + 6t^2 \log(n(k+p)/\delta) \lambda_1^2(V_{k,\perp}^* C^{1/2}) \frac{\mathrm{Tr}(\mathrm{diag}(K_{11})^{-1})}{p+1} \right)^{1/2}$$

In the above, the first inequality follows from the deterministic error bound in Theorem 9.1 of [15], the equality follows from noting that $\Omega = C^{1/2}X$ where the entries of X are standard normal, the second inequality follows from Lemmas 9 and 10 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 X, Y that $\|XY\|_{\mathrm{F}} \leq \|X\|_{\mathrm{F}} \|Y\|_2$. Our proof for the probabilistic bound is complete. ■

A.3 Proof of Lemma 4

Proof. Recall that $Q_+ = \mathrm{orth}([Y, A\Omega_-])$. It then follows that $Q_+ Q_+^* = QQ^* + Q_- Q_-$. We thus have that $Q_- \in \mathrm{Null}(QQ^*)$. From which we have the following,

$$A - Q_+ Q_+^* A = A - QQ^* A - Q_- Q_-^* A = (A - QQ^* A) - Q_- Q_- (A - QQ^* A)$$

Then letting, $P = QQ^*$, we have

$$A - Q_+ Q_+ A = P_\perp A - Q_- Q_-^* P_\perp A$$

Finally, we can note from the Gram-Schmidt Orthogonalization Procedure, we have $\mathrm{orth}([Y, A\Omega_-]) = \mathrm{orth}([Y, P_\perp A\Omega_-])$. Our proof is complete. ■

A.4 Proof of Theorem 7

Proof. The proof is a direct calculation. Let $R^{(t)} = (I - Q^{(t)} Q^{(t)*}) A$.

$$\begin{aligned}
\|A - Q^{k(t+1)} Q^{k(t+1)*} A\|_{\mathrm{F}}^2 &\overset{(\zeta_1)}{=} \|P_\perp^{(t)} A - Q_- Q_-^* P_\perp^{(t)} A\|_{\mathrm{F}}^2 \\
&\overset{(\zeta_2)}{\leq} (1 + \epsilon(C^{(t)})) \|P_\perp^{(t)} A - (P_\perp^{(t)} A)_k\|_{\mathrm{F}}^2 \\
&= (1 + \epsilon(C^{(t)})) \left(\|P_\perp^{(t)} A\|_{\mathrm{F}}^2 - \|(P_\perp^{(t)} A)_k\|_{\mathrm{F}}^2 \right) \\
&\overset{(\zeta_3)}{\leq} (1 + \epsilon(C^{(t)})) \left(\|P_\perp^{(t)} A\|_{\mathrm{F}}^2 - \|A - A_{k(t+1)}\|_{\mathrm{F}}^2 \right) \qquad (5)
\end{aligned}$$

In the above, (ζ_1) follows from Lemma 4, (ζ_2) follows from our probabilistic error bound for the Generalized Randomized SVD in Theorem 2, (ζ_3) follows from the following

$$\|(A - Q^{(t)} Q^{(t)} A)_k\|_{\mathrm{F}}^2 = \sum_{i \in [k]} \sigma_i^2 (A - Q^{(t)} Q^{(t)} A) \geq \inf_{\substack{B \in \mathbb{C}^{m \times n} \\ \mathrm{rank}(B) = tk}} \sum_{i \in [k]} \sigma_i^2 (A - B) = \|A_{tk} - A_{t(k+1)}\|_{\mathrm{F}}^2$$

where (ζ_4) follows from Eckart-Young-Mirsky Theorem for the Frobenius Norm [12, 19]. Then, from Theorem 2,

$$\|A - Q^k Q^{k*} A\|_{\mathrm{F}} \leq (1 + \epsilon(C^{(1)})) \|A - A_k\|_{\mathrm{F}} \qquad (6)$$

Combining Equations (5) and (6), we have

$$\|A - Q^{kt} Q^{kt} A\|_{\mathrm{F}}^2 \leq \prod_{i \in [t]} (1 + \epsilon(C^{(i)})) \cdot \|A - A_k\|_{\mathrm{F}}^2 - \sum_{i \in [t]} \prod_{j \in [i] \setminus \{1\}} (1 + \epsilon(C^{(j)})) \|A - A_{ik}\|_{\mathrm{F}}^2$$

Next, suppose we choose $\epsilon(C_i) = \epsilon(C_j)$ for all $(i, j) \in [t] \times [t]$, then we have

$$\|A - Q^{kt} Q^{kt} A\|_{\mathrm{F}}^2$$

■

B Probability Theory

Lemma 10. *Fix matrices $S \in \mathbb{R}^{k\times n}$ and $T \in \mathbb{R}^{m\times \ell}$, then for a conformal matrix G with elements sampled i.i.d from $\mathcal{N}(0,1)$, with probability exceeding $1-\delta$,*

$$\|SGT\|_{\mathrm{F}} \leq \|S\|_{\mathrm{F}}\|T\|_{\mathrm{F}}\sqrt{2\log(nm/\delta)}$$

Proof. The proof follows is a simple calculation followed by a maximal tail bound on a sample of Gaussians.

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

We now bound the maximum Gaussian over a finite sample.

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

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. Then, from some algebra, we obtain with probability exceeding $1-\delta$,

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

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