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 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 [16]. 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.

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 [14], fluid dynamics [18]. 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, [16], 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 [24] 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 $\mathbf{x}^{(k)}$ 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,

Preprint. Under review.

$Q^{(k)}(Q^{(k)})^* A$, 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 [15]. Optimal column-sampling for the CUR Matrix Decomposition received much attention as can be seen in the works [17, 11, 12]. More recently, [21] 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 [23]. 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 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 A.
2. We provide a novel theoretical analysis for adaptive sampling in the matrix-vector product query model.
3. We perform Numerical Experiments on real-world and synthetic matrices that confirm our theoretical claims.

2 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 A by multiplying the matrix A with a Matrix with Standard Normal entries[19]. One first samples $k+p$ gaussian vectors $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, I)$ 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 $QR = AX$. It is then proved by Halko et al. [16] that Q is a good approximation of the range of the top right singular space of A and thus the approximation QQ^*A is close to A in both the spectral and Frobenius Norms. 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.

3 Adaptive Sampling

The Adaptive Range Finder Algorithm was propose in [16] 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 $Q^i Q^{i*} A$. Obtain an approximation to A, that minimizes*

$$\|A - Q^{r(k+p)} Q^{r(k+p)*} A\|_{\mathrm{F}}$$

3.1 Continued Sampling

We will

3.2 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 [21] to perform adaptive sampling in rounds.

Algorithm 1 Adaptive Sampling for Low-Rank Matrix Approximation

input: Target Matrix $A \in \mathbb{C}^{m\times n}$, target rank k, oversampling parameter p, number of rounds r
output: Low-Rank Approximation $\hat{A}$ of A to minimize $\|\widehat{A} - A\|_{\mathrm{F}}$

1: $\widehat{A}^{(0)} \leftarrow \mathbf{0}$
2: **for** $t \in [r]$ **do**
3: Form $\Omega^{(t)} \in \mathbb{C}^{n\times k+p}$ with columns sampled i.i.d from $\mathcal{N}(\mathbf{0}, C^{(t)})$
4: $Y \leftarrow [Y, A\Omega^{(t)}] \in \mathbb{C}^{m\times t(k+p)}$ and obtain the factorization $Y = QR$
5: Update the Low Rank Approximation $\widehat{A}^{(t)} = \widehat{A}^{(t-1)} + Q_{-}^{t\ell} Q_{-}^{t\ell *} A$
6: Calculate the economized SVD $\widehat{U}\widehat{\Sigma}\widehat{V}^{*} = \widehat{A}^{(t)}$
7: Update the Covariance Matrix $C^{(t)} \leftarrow \widehat{V}_{t(k+p)}\widehat{V}^{*}_{t(k+p)}$

return: $Q^{\ell r} Q^{\ell r *} A$

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 Line 7. 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,

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

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 $A\Omega$ 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 $QQ^{*}A \in \mathbb{C}^{m\times n}$ a total of t times which can be done on the order of $O(tmn^2)$ as well. Thus the total complexity can be observed as $O(tmn^2)$. The dominating runtime is the economized SVD calculation.

4 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.

4.1 Notation

We define $[t]$ as the set of integers $\{1, \ldots, t\}$. Let $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$ for any two real numbers a, b. Let $\mathbb{O}_{n,k}$ be the set of all $n \times k$ matrices with orthogonal columns, i.e. $\{V : V^{*}V = I_{k\times k}\}$. We use Big-O notation, $y \leq O(x)$, to denote $y \leq Cx$ for some positive constant, C. 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 is represented by $(\cdot)^{+}$ s.t. $X^{+} = (X^{*}X)^{-1}X^{*}$. The Projection Matrix is defined as $\Pi_{Y} = YY^{+} = Y(Y^{*}Y)^{-1}Y^{*}$ as the projection on to the column space of Y. If Y has orthogonal

columns, then Π_Y is the Orthogonal Projection defined as $\Pi_Y = YY^*$. We follow a similar setup as previous literature. Let $\rho \triangleq \text{rank}(A) \leq m \wedge n$, we will factorize A as

$$A = \begin{matrix} k & n-k \\ [U_k & U_{k,\perp}] \end{matrix} \begin{bmatrix} \Sigma_k & \\ & \Sigma_{k,\perp} \end{bmatrix} \begin{bmatrix} V_k^* \\ V_{k,\perp}^* \end{bmatrix} \begin{matrix} k \\ n-k \end{matrix} = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^*$$

Furthermore, we let $A_{(k)} = \sigma_k \mathbf{u}_k \mathbf{v}_k^*$ and $A_{\perp,k} \triangleq A - A_{(k)}$. We denote $\Omega \in \mathbb{R}^{n \times \ell}$ to be a test matrix such that columns of Ω are sampled i.i.d from $\mathcal{N}(\mathbf{0}_n, I_{n \times n})$. Throughout the paper, we utilize the following norms,

$$\text{Spectral Norm: } \|A\|_2 = \max_{\mathbf{v} \in \mathbb{S}^{n-1}} \|A\mathbf{v}\| = \sigma_1$$

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

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 [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 $1 - \delta - t^{-p}$,*

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

where

$$K = V^* C V = \begin{bmatrix} V_k^* C V_k & V_k^* C V_{k,\perp} \\ V_{k,\perp}^* C V_k & V_{k,\perp}^* C V_{k,\perp} \end{bmatrix} = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}$$

The proof of Theorem 2 can be derived with the deterministic error bound of Theorem 9.1 in [16] and combining Lemma 6 with Lemma 8. From Theorem 2, we infer that we obtain a better theoretical bound then the randomized SVD when C has better alignment with the top right singular space of A than the identity. As example, choosing $C = V_k V_k^*$, we obtain $K_{22} = 0$ and we then obtain the optimal $\|\Sigma_{k,\perp}\|$ Frobenius norm approximation error. The weakness in the bound when compared to the probabilistic bound in [16] is the $O(\log(n(k+p))$ term that is a result of Lemma 8. The difference between our bound and the bound in Theorem 2.4 of [22] is we have a relative error bound whereas they also have an additive term.

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*

$$p \geq (6t^2/\varepsilon) \log(n^2/\delta)$$

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

$$\|A - QQ^* A\|_{\text{F}}^2 \leq (1 + \varepsilon) \|A - A_k\|_{\text{F}}^2$$

Corollary 3 develops a relative Frobenius norm error bound for the Generalized Randomized SVD and can be derived from elementary algebraic manipulations after an application of Theorem 2.

4.4 Near Optimal Adaptive Sampling

In this section, we will discuss the theoretical motivation for adaptive sampling. Adaptive sampling in works such as [12] and [21] for the column subselection problem (see e.g. [9] § 1 for a problem definition) sample columns from the residual matrix, $A - Q^{(t)} Q^{(t)*} A$, at each iteration t instead of directly from 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 $P_\perp A$.

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

Theorem 5. *Let* $A \in \mathbb{C}^{m\times n}$ *and let* $C \in \mathbb{C}^{n\times n}$ *be a PSD covariance matrix. Let the sampling matrix be decomposed as* $\Omega = [\Omega, \Omega_-]$, *the matrix-matrix products* $Y = A\Omega$, $Y_- = A\Omega_-$, *then let* Q_+ *be the orthonormal matrix from an economized QR decomposition of* $[A\Omega_-, Y]$. *Let* $p \geq 4$ *be the oversampling parameter, then with probability exceeding* $1 - \delta - t^{-p}$,

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

4.5 Analysis of Algorithm 1

In this subsection, we will study the covariance update in Line 7 of Algorithm 1.

5 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)?

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

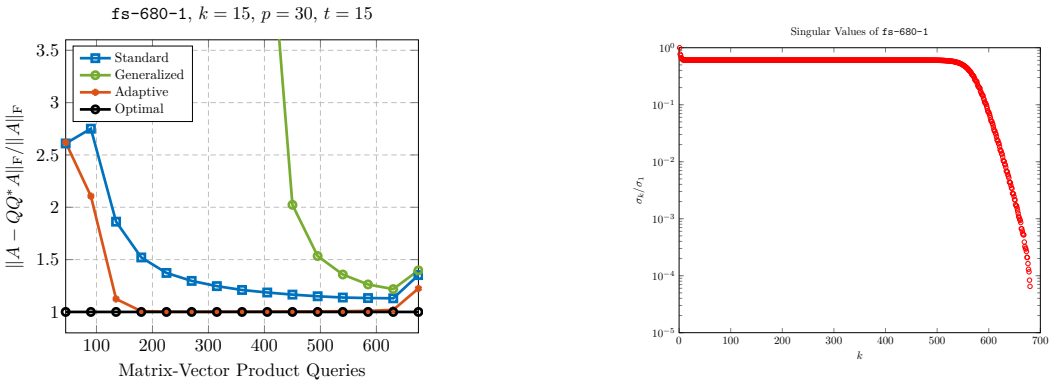

Figure 1: Low Rank Approximation for the matrix `fs-680-1` from [10] on the left. The matrix `fs-680-1` is derived from a Chemical kinetics problem. The singular values of `fs-680-1` is displayed on the right.

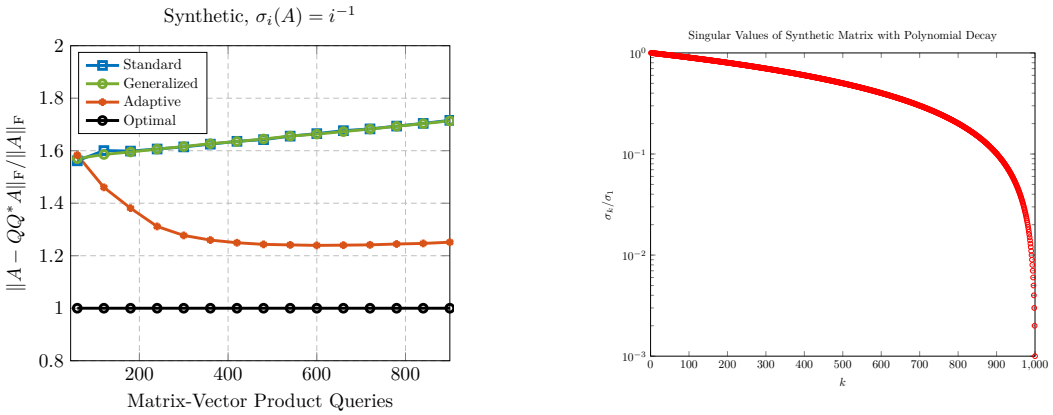

Figure 2: Low Rank Approximation for a synthetic matrix with polynomial singular value decay described in Equation (3).

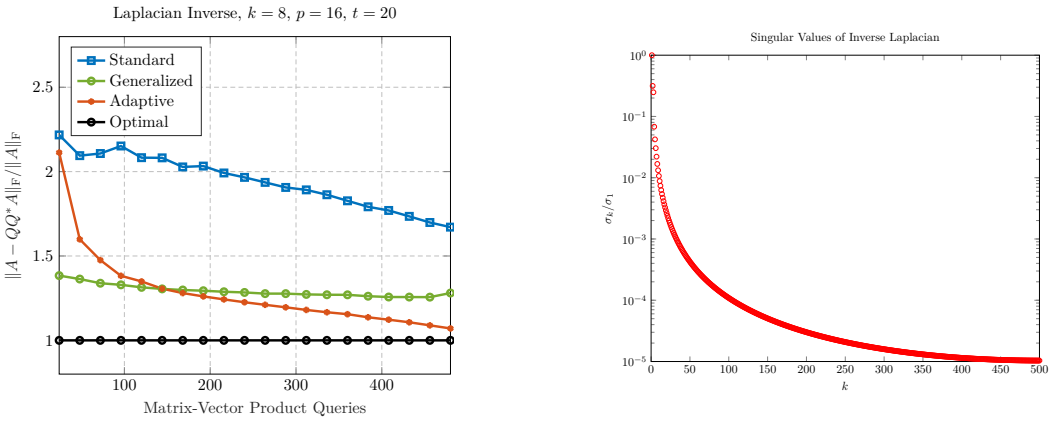

Figure 3: Low Rank Approximation for a synthetic matrix with polynomial singular value decay described in Equation (2). Also see Figure 2 in [4] for the same experiment.

Figure 2, we find that Algorithm 1 obtains nearly-optimal Frobenius norm error.

We observe in Figure 3, even without prior knowledge of the dominant right singular space of A, after approximately 150 matrix vector products adaptive sampling learns a better low-rank approximation with respect to the Frobenius norm.

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

$$A = \sum_{i \in [n]} i^{-p} \cdot U_{(:,i)} V^*_{(:,i)}, \quad U \in \mathbb{O}_{m,k}, V \in \mathbb{O}_{n,k} \tag{3}$$

In Figure 3, we experiment on the matrix with singular values described in Equation (3) and setting $p = 1$. We observe adaptive sampling obtains a better low-rank matrix approximation after the first round of adaptive sampling.

6 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 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 N.B. 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). We 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 1130–1143, 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, 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.

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

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

[6] 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.

[7] 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.

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

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

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

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

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

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

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

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

[21] 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.

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

[23] 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.

[24] 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.

[25] Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, Mar 1972.

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 6. *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 columns of $V^*\Omega$ are sampled from a centered multivariate Gaussian Distribution with second-moment matrix $K = V^*CV$.*

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 of each column of $V^*\Omega$ are Gaussian. From the fact that 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 first calculate the mean of a column of $V^*\Omega$. 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 for any $(i,i',j) \in [k]\times[k]\times[\ell]$, and $i \neq i'$, we have

$$\begin{aligned} &\mathop{\mathbf{E}}_{\boldsymbol{\omega}_j\sim\mathcal{N}(\mathbf{0},C)}\left[(\mathbf{v}_i^*\boldsymbol{\omega}_j - \mathop{\mathbf{E}}_{\boldsymbol{\omega}_j\sim\mathcal{N}(\mathbf{0},C)}[\mathbf{v}_i^*\boldsymbol{\omega}_j])(\mathbf{v}_{i'}^*\boldsymbol{\omega}_j - \mathop{\mathbf{E}}_{\boldsymbol{\omega}_j\sim\mathcal{N}(\mathbf{0},C)}[\mathbf{v}_{i'}^*\boldsymbol{\omega}])\right] \\ &\quad = \mathop{\mathbf{E}}_{\boldsymbol{\omega}_j\sim\mathcal{N}(\mathbf{0},C)}\left[\mathbf{v}_i^*\boldsymbol{\omega}_j\boldsymbol{\omega}_j^*\mathbf{v}_{i'}\right] = \mathbf{v}_i^*C\mathbf{v}_{i'} \end{aligned} \tag{4}$$

For the diagonal covariance elements, we have

$$\mathop{\mathbf{E}}_{\boldsymbol{\omega}_j\sim\mathcal{N}(\mathbf{0},C)}[(\mathbf{v}_i^*\boldsymbol{\omega}_j - \mathbf{E}[\mathbf{v}_i^*\boldsymbol{\omega}_j])^2] = \mathop{\mathbf{E}}_{\boldsymbol{\omega}_k\sim\mathcal{N}(\mathbf{0},C)}[\mathbf{v}_i^*\boldsymbol{\omega}_j\boldsymbol{\omega}_j^*\mathbf{v}_i] = \mathbf{v}_i^*C\mathbf{v}_i \tag{5}$$

Then combining Equations (4) and (5), we have

$$K = V^* C V$$

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 7 (Lemma 3 in [5]). *Suppose $V_k \in \mathbb{O}^{n,k}$ and $\Omega \in \mathbb{C}^{n\times k+p}$ with columns sampled i.i.d from $\mathcal{N}(\mathbf{0}, K)$ and $p \geq 4$. Then with probability at least $1 - t^{-p}$,*

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

We are now ready to prove our relative Frobenius Error Norm Bound for the Generalized Randomized SVD originally presented in [4].

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

$$K = \begin{bmatrix} V_k^* C V_k & V_k^* C V_{k,\perp} \\ V_{k,\perp}^* C V_k & V_{k,\perp}^* C V_{k,\perp} \end{bmatrix} = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}$$

We then have the following manipulations,

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

In the above, the first inequality follows from the deterministic error bound in Theorem 9.1 of [16], 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 7 and 8 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_+ is an orthonormal basis of $[Y, A\Omega_-]$. 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 Classical Gram-Schmidt Orthogonalization Procedure, we have

$$\mathrm{orth}([Y, A\Omega_-]) = \mathrm{orth}([Y, (I - QQ^*)A\Omega_-])$$

Our proof is complete. ■

A.4 Proof of Theorem 5

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

$$\|A - Q_+Q_+^*A\|_{\mathrm{F}} = \|(I - Q_-Q_-^*)A - QQ^*(I - Q_-Q_-^*)A\|_{\mathrm{F}}$$

The above equation follows from the same argument as Lemma 4. Then from our relative-error accuracy bound in Corollary 3, we have

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

Then, from expanding out the Frobenius Norm, we have

$$\|(A - Q_-Q_-^*A)_k\|_{\mathrm{F}}^2 = \|A - Q_-Q_-^*A\|_{\mathrm{F}}^2 - \sum_{i\in[k]} \sigma_i^2(A - Q_-Q_-^*A)$$

We now leverage the Eckart-Young-Mirsky Theorem [13, 20], and obtain

$$\sum_{i\in[k]} \sigma_i^2(A - Q_-^*Q_-^*A) \geq \inf_{\substack{B\in\mathbb{C}^{m\times n}\\ \mathrm{rank}(B)=k+p}} \sum_{i\in[k]} \sigma_i^2(A-B) = \|A_{k+p} - A_{2k+p}\|_{\mathrm{F}}^2 \tag{6}$$

Then, from our improved Generalized Randomized SVD error bound in Theorem 2, we have with probability at least $0.99 - t^{-p}$,

$$\|A - Q_-Q_-^*A\|_{\mathrm{F}}^2 \leq \|A - A_k\|_{\mathrm{F}}^2 \left(1 + 6t^2\log(100n\ell)\|V_{k,\perp}^*C^{1/2}\|_2^2 \frac{\mathrm{Tr}((V_k^*CV_k)^{-1})}{p+1}\right) \tag{7}$$

Then, from choosing our oversampling parameter sufficiently large,

$$p \geq (6t^2/\epsilon)\|V_{k,\perp}^*C^{1/2}\|\,\mathrm{Tr}((V_k^*CV_k)^{-1})$$

We can then combine our results in Equations (6) and (7) to obtain

$$\begin{aligned}
&\|A - Q_+Q_+^*A\|_{\mathrm{F}}^2 = (1+\epsilon)^2\|A - A_k\|_{\mathrm{F}}^2 - (1+\epsilon)\|A_{k+p} - A_{2k+p}\|_{\mathrm{F}}^2\\
&= (1+\epsilon)^2\|A_k - A_{k+p}\|_{\mathrm{F}}^2 + \epsilon(1+\epsilon)\|A_{k+p} - A_{2k+p}\|_{\mathrm{F}}^2 + (1+\epsilon^2)\|A - A_{2k+p}\|_{\mathrm{F}}^2\\
&\leq (1+3\epsilon)\|A_k - A_{k+p}\|_{\mathrm{F}}^2 + 2\epsilon\|A_{k+p} - A_{2k+p}\|_{\mathrm{F}}^2 + (1+3\epsilon)\|A - A_{2k+p}\|_{\mathrm{F}}^2
\end{aligned}$$

Our proof is complete. ■

A.5 Proof of Theorem

Proof. We can now understand why choosing $C^{(t)} = \widehat{V}_k\widehat{V}_k^*$ is a strong choice, as when $\widehat{V}_k$ is close to V_k, we have both $\|V_{k,\perp}^*\widehat{V}_k\|_2$ and $\mathrm{Tr}((V_k^*\widehat{V}_k\widehat{V}_k^*V_k)^{-1})$ are small. To formalize this, we have from Wedin's Theorem [25],

$$\|V_{k,\perp}^*\widehat{V}_k\|_2 \leq \frac{\sqrt{2}\|A - QQ^*A\|_2}{\sigma_k} \wedge 1 \leq \sqrt{2}(1+\epsilon)\frac{\sigma_{k+1}}{\sigma_k} \wedge 1$$

The $\sqrt{2}$ results from the lack of sharpness in Wedin's Theorem. The ϵ can be derived from the probabilistic spectral error bound for the rSVD given in Theorem 10.8 of [16].

Extension to k steps. We have,

$$\begin{aligned}
\|A - Q_{\ell(t+1)}Q_{\ell(t+1)}^*A\|_{\mathrm{F}}^2 &= \|(I - Q_-Q_-^*)A - Q_{\ell t}Q_{\ell t}^*(I - Q_-Q_-^*)A\|_{\mathrm{F}}^2\\
&\leq (1+\epsilon)\left(\|A - Q_-Q_-A\|_{\mathrm{F}}^2 - \sum_{i\in[kt]}\sigma_i^2(A - Q_-Q_-^*A)\right)
\end{aligned}$$

Leveraging once again the Eckart-Young-Mirsky Theorem [13, 20], we have

$$\sum_{i\in[kt]} \sigma_i^2(A - Q_{\ell t}Q_{\ell t}^*A) \geq \min_{\substack{B\in\mathbb{C}^{m\times n}\\ \mathrm{rank}(B)=\ell t}} \|(A-B)_{kt}\|_{\mathrm{F}}^2$$

■

B Probability Theory

Lemma 8. *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 brute-force 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 elementary algebraic manipulations, 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. ■