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

Arvind Rathnashyam*
RPI Math, rathna@rpi.edu

Nicolas Boullé
Cambridge Math, nb690@cam.ac.uk

Alex Townsend
Cornell Math, ajt453@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 [HMT11]. Recently, [BT22] 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 [TWA+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.

*Work was partially completed at Cornell Summer 2023

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 [FDH05], fluid dynamics [LPZ+01]. 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, [HMT11], has been theoretically analyzed and used in various applications. Even more recently, [BET22] 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 [WLSX21] 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, $Q^{(k)}(Q^{(k)})^* A$, where $Q^{(k)} \triangleq \text{orth}(AV^{(k)})$ where $V^{(k)}$ is the concatenation of vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k)}$ and $Q^{(k)} \triangleq \text{orth}(AV^{(k)})$.

Adaptive Sampling techniques for Low-Rank Matrix Approximation first appeared in CUR Matrix Decomposition in [FKV04]. Optimial column-sampling for the CUR Matrix Decomposition received much attention as can be seen in the works [HP14, DRVW06, DV06]. More recently, [PMID15] 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 [SWYZ21]. The theoretical properties of adaptively sampled matrix vector queries for estimating the minimum eigenvalue of a Wishart matrix have been studied in [BHSW20]. Their bounds are used in [BCW22] 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 Notation, Background Materials, and Relevant Work

In this section we will introduce the notation we use throughout the paper, perturbations of singular spaces, as well as relevant work in the Low-Rank Matrix Approximation Literature.

2.1 Notation

We will denote boldface characters A, B, C as matrices and lower roman boldface characters $\mathbf{x}, \mathbf{y}, \mathbf{z}$ as vectors. The spectral norm of a matrix is given by $\|\cdot\|$, which is equivalent to the max singular value, $\sigma_{\max}(\cdot)$. 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 $\Pi_Y = YY^*$. Let $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. Let $\mathbb{O}_{n,k}$ be the set of all $n \times k$ matrices with orthogonal columns, i.e. $\{V : V^*V = I_{k\times k}\}$. The Frobenius Norm is denoted for a matrix A as $\|A\|_{\mathrm{F}}$. 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.

2.2 Linear Algebra

We follow a similar setup as previous literature. Let $\rho \triangleq \operatorname{rank}(A) \leq m \wedge n$, we will factorize A as

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

Furthermore, we let $A_{(k)} \triangleq \sigma_k \mathbf{u}_k \mathbf{v}_k^*$ and $A_{\perp,k} \triangleq A - A_{(k)}$. The low rank matrix approximation is denoted by $\tilde{A} \triangleq WW^*A$ for a $W \in \mathbb{C}^{m\times k}$ such that $W^*W = I$. The matrix factorization notation described in eq:A-SVD holds for $\tilde{A}$ with a *tilde* over the typical notation. 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,

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

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

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

2.3 Singular Subspace Perturbations

To represent the distance between subspaces we utilize the $\sin\Theta$ norm. Let $\mathcal{X}, \mathcal{Y}$ be subspaces, then we denote the principal angles between subspaces (PABS) $\mathcal{X}$ and $\mathcal{Y}$ as $\frac{\Pi}{2} \geq \Theta_1(\mathcal{X}, \mathcal{Y}) \geq \ldots \geq \Theta_{m\wedge n}(\mathcal{X}, \mathcal{Y})$. We then have that for any two matrices, P and Q, it follows that $\|\Pi_P - \Pi_Q\|_{\mathrm{F}} = \|\sin\Theta(P, Q)\|_{\mathrm{F}}$.

2.4 Relevant Works

The Randomized Singular Value Decomposition was developed and analyzed thoroughly in [HMT11]; throughout this paper we will refer to this algorithm as HMT. The HMT algorithm samples $k + p$ normal isotropic vectors. [BT22] extend this idea by noting that when we have knowledge of the right singular vectors, we can improve the approximation by sampling normal vectors with a chosen covariance matrix which has eigenvectors with small subspace angles to the right singular vectors of A.

3 Near Optimal Sampling

In this section, we will go over the covariance matrices proposed papers and we consider choosing the optimal covariance matrix adaptively for sampling vectors. In the seminal paper by [HMT11], the covariance matrix is given as identity matrix, $C \triangleq I$. In the generalization of the Randomized SVD, when given some prior information of the matrix, the covariance matrix is given as $C \triangleq K$ where K has some information on the right singular vectors of A (e.g. discretization of Green's Function of a PDE). Let $\tilde{V}$ be the right singular

vectors of the SVD of the low-rank approximation at iteration $k-1$, then the update for the covariance matrix is given as $C^{(k+1)} \triangleq \tilde{V}_{(:,k)}\tilde{V}^*_{(:,k)}$. Throughout this paper we will only consider $C^{(0)} = I$, however using theory from [BT22], this can be extended to $C^{(0)} = K$ if one has some knowledge of the right singular vectors, e.g. if the matrix represents a Partial Differential Equation (PDE), having K as a discretized Green's Function for the type of PDE, e.g. elliptic, parabolic, or hyperbolic [Eva22]. A similar algorithm can be found in [WLSX21]. It is intuitive that we want to continuously sample in the null space of the the matrix approximation we have already obtained. This ensures we are learning new information in each iteration as we don't want to sample vectors which will not learn significant unknown information about the matrix.

3.1 Algorithm

The Pseudo Code for the optimal function sampling is given in alg:optimal-sampling. 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 alg:optimal-sampling, 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 $Q = \text{orth}(A\Omega)$ 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.

4 Theory

In this section we will give the mathematical setup for the theoretical analysis. We will then represent theorems from relevant works on the error bounds for their low-rank approximation methods. We will then

give our error bounds and general theory of Algorithm 1 with the proofs in the appendix.

4.1 Near Optimal Function Sampling

In this section, we will describe the motivation of choosing the k-th right singular vector as our new sample. In particular, we will show why our approach is *near*-optimal.

Lemma 1. *Let $Q \in \mathbb{C}^{m\times k}$ be a unitary matrix such that $Q \triangleq \operatorname{orth}(A\Omega)$ for a test matrix $\Omega \in \mathbb{C}^{n\times k}$. Recall $\widetilde{Q} \triangleq \operatorname{orth}(A[\Omega, \widetilde{\mathbf{v}}])$ where $\widetilde{\mathbf{v}}$ is adaptively chosen. Then, we have*

$$\arg\min_{\widetilde{\mathbf{v}}\in\mathbb{R}^n} \|A - \widetilde{Q}\widetilde{Q}^* A\|_{\mathrm{F}} = \arg\max_{\substack{\mathbf{x}\in\mathrm{Null}(QQ^*A)\\ \|\mathbf{x}\|=1}} \|A^*A\mathbf{x}\|$$

The proof is deferred to § **??**. ■

lem:optimal-sample gives us the result that in the Frobenius Norm, we want to sample maximally in the Null Space of the current low-rank approximation. However, even with knowledge of the $\mathrm{Null}(QQ^*A)$, we note that we do not have knowledge of $\|A^*A\mathbf{v}\|$ for any $\mathbf{v} \in \mathrm{Null}(QQ^*A)$. If we are to take a random sample in this subspace, we can still be quite far from the optimal sampling vector, even in expectation. Rather, we sample to the closest singular vector to the Null-Space, which is *known*.

4.2 Query Lower Bound for Frobenius Norm

We will first give the error bounds for the randomized SVD using a general positive definite covariance matrix.

Theorem 2 (Theorem 2 [BT22])**.** *Let $A \in \mathbb{R}^{m\times n}$, $k \geq 2$, oversampling parameter $p \geq 2$, where $k+p \leq m\wedge n$. Let $\Omega_i \sim \mathcal{N}(\mathbf{0}, K)$ for $i \in [k+p]$, and $Q \triangleq \operatorname{orth}(A\Omega)$. Then,*

$$\|A - QQ^*A\|_{\mathrm{F}} \leq \left(1 + ut\sqrt{(k+p)\,\frac{\beta_k}{\gamma_k}\,\frac{3k}{(p+1)}}\right)\sqrt{\sum_{j=k+1}^{n}\sigma_j^2}$$

with failure probability at most $t^{-p} + [ue^{-(u^2-1)/2}]^{k+p}$, where $\gamma_k = k/(\lambda_1 \operatorname{Tr}((V_1^*KV_1)^{-1}))$ and $\beta_k = \operatorname{Tr}(\Sigma_2^2V_2^*KV_2)/(\lambda_1\|\Sigma_2\|_{\mathrm{F}}^2)$, where λ_1 represents the largest eigenvalue of K.

In this section, we give information theoretic lower bounds on query complexity. We assume the algorithm, $\mathcal{A}$ not only has access to the matrix-vector products, but also has available the SVD of the intermediate low-rank approximations.

Theorem 3. *There exists an adaptive algorithm (possibly randomized) with 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*

$$\|A - QQ^*A\|_{\mathrm{F}} \leq (1+\varepsilon)\min_{\substack{U\in\mathbb{C}^{m\times k}\\ U^*U=I_k}}\|A - UU^*A\|_{\mathrm{F}}$$

Our proof is by construction and given in 1. The rest of the section will be dedicated to the proof.

4.3 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 4. *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,*

$$\mathbf{E}\|A - Q_+Q_+^*A\|_{\mathrm{F}}^2 \leq \|A - QQ^*A\|_{\mathrm{F}}^2 + \frac{\|\widetilde{\Sigma}_2\|_2^2}{p-2} \cdot \|\tan\Theta(\widetilde{V}_1, C^{1/2})\|_{\mathrm{F}}^2 - \|A - QQ^*A\|_2^2$$

where

$$(I - QQ^*)A = \begin{bmatrix}\widetilde{U}_1 & \widetilde{U}_2\end{bmatrix}\begin{bmatrix}\widetilde{\Sigma}_1 & \\ & \widetilde{\Sigma}_2\end{bmatrix}\begin{bmatrix}\widetilde{V}_1^* \\ \widetilde{V}_2^*\end{bmatrix}$$

Proof. From noting that $Q_+ = \mathrm{orth}([Y, (I - QQ^*)A\Omega])$. Then if we consider the orthogonalization as coming from a Gram-Schmidt Procedure, we have $Q_-^*Q = 0$. We then obtain,

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

Then the problem reduces to low-rank approximation of the matrix $(I - QQ^*)A$. Then from the standard analysis, we obtain

$$\mathbf{E}\|A - Q_+Q_+^*A\|_{\mathrm{F}}^2 \leq \|\widetilde{\Sigma}_2\|_{\mathrm{F}}^2 + \mathbf{E}\|\widetilde{\Sigma}_2\widetilde{\Omega}_2\widetilde{\Omega}_1^+\|_{\mathrm{F}}^2$$

The remaining details of the proof are deferred to § A.1. ■

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 $\Omega \in \mathbb{C}^{n\times p}$ such that the columns are sampled i.i.d from $\mathcal{N}(\mathbf{0}, I)$. Denote $Y_+ \triangleq \mathrm{orth}([Y, A\Omega])$ and $Q_+ \triangleq \mathrm{orth}(Y_+)$, then*

$$\begin{aligned}\mathbf{E}\|A - Q_+Q_+^*A\|_{\mathrm{F}}^2 &\leq \|A - QQ^*A\|_{\mathrm{F}}^2 + \|\widetilde{\Sigma}_2\|_{\mathrm{F}}^2 \cdot \frac{1}{p-2} - \|A - A_k\|_2^2 \\ &\leq \|A - QQ\|\end{aligned}$$

where $\widetilde{\Sigma}_2$ is defined as in Equation (4).

Proof. The proof follows from a direct application of Lemma 4 and setting $C = I$. ■

Theorem 6. *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 $Y = \mathrm{orth}(A\Omega)$ and $Q = \mathrm{orth}(Y)$, then*

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

Proof. The proof is deferred to § A.2. ■

Lemma 7. *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 $\Omega \in \mathbb{C}^{n\times 1}$ be the k-th right singular vector of QQ^*A. Denote $Y_+ \triangleq \mathrm{orth}([Y, A\Omega])$ and $Q_+ \triangleq \mathrm{orth}(Y_+)$, then*

$$\|A - Q_+Q_+^*A\|_{\mathrm{F}}^2 \leq \|A - QQ^*A\|_{\mathrm{F}}^2 + \|\widetilde{\Sigma}_2\|_{\mathrm{F}}^2 \cdot \frac{1}{p-1} - \|A - A_k\|_2^2$$

where $\widetilde{\Sigma}_2$ is defined as in Equation (4).

Proof. The proof is deferred to § A.3. ■

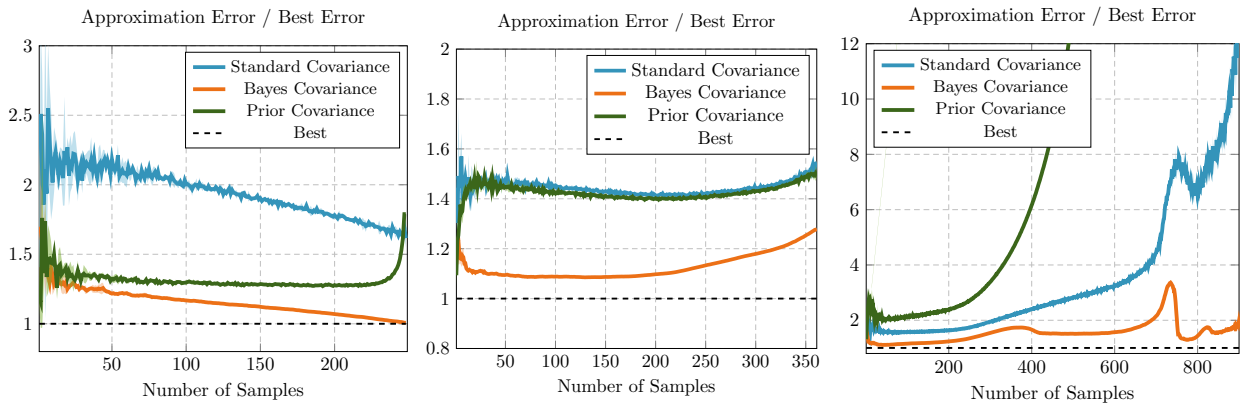

Figure 1: Low Rank Approximation for the Inverse Differential Operator given in eq:inverse-laplace Figure Left, Differential Operator Matrix `Poisson2D` [DH11] Figure Center, and Differential Operator Matrix `DK01R` [DH11] Figure Right. The experiment on the left is from [BT22, Figure 2]. We use the discretized Green's Function as the prior covariance matrix. In this set of experiments, we learn a low rank approximation of the inverse operator.

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

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

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 [BKST22, BT23]) as the inverse Differential operator is compact and there are nice theoretical properties, such as data efficiency.

In fig:inverse-matricesFigure Right, 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 fig:inverse-matricesFigure Left, 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 Matrix Suite, [DH11]. In fig:forward-matrices Figure Left, we choose a fluid dynamics problem due to its relevance in low-rank approximation [BNK20].

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

$$A = \sum_{i=1}^{\rho} \frac{100 i^{\ell}}{n} U_{(:,i)} V^*_{(:,i)}, \quad U \in \mathbb{O}_{m,k}, V \in \mathbb{O}_{n,k} \tag{2}$$

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:

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

We observe when there exists exponential decay of the singular values in fig:synthetic-singular-value-exponential-decay, adaptive sampling in accordance without scheme

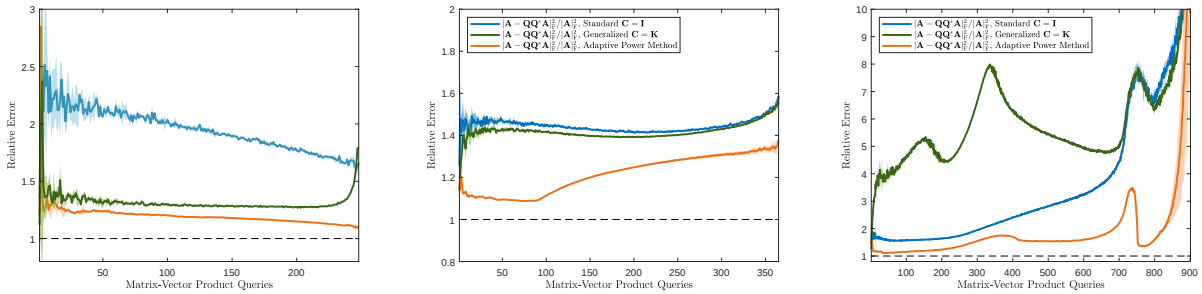

Figure 2: Results for the Power-Method based Adaptive Sampling Method described in Algorithm 2. Low Rank Approximation for the Inverse Differential Operator given in eq:inverse-laplace Figure Left, Differential Operator Matrix `Poisson2D` [DH11] Figure Center, and Differential Operator Matrix `DK01R` [DH11] Figure Right. The experiment on the left is from [BT22, 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.

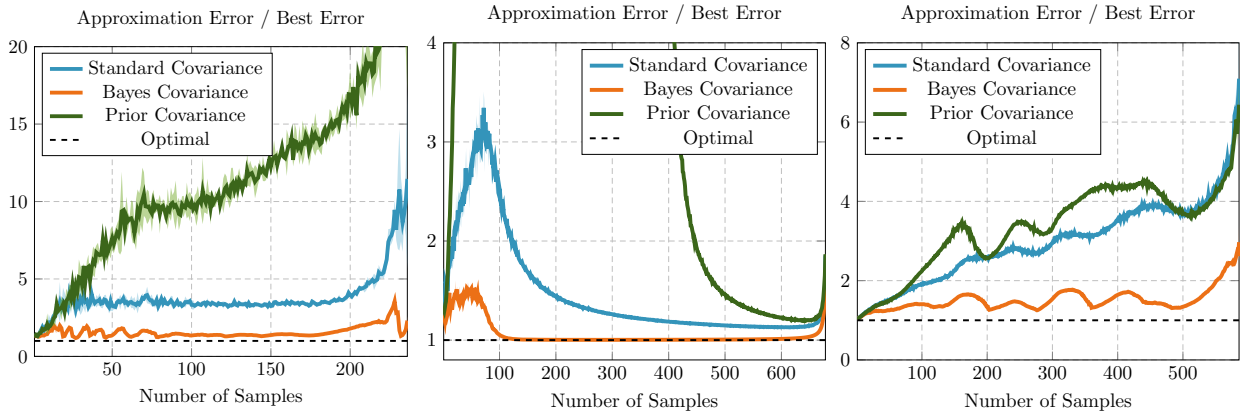

Figure 3: Low Rank Approximation for a matrix for a Computational Fluid Dynamics Problem, `saylr1` Figure Left is from [DH11]. Subsequent 2D/3D Problem `fs-680-2` Figure Center is from [DH11]. Differential Stiffness Matrix from a Nastran Buckling Problem, `bcsstk34` Figure Right is from [DH11].

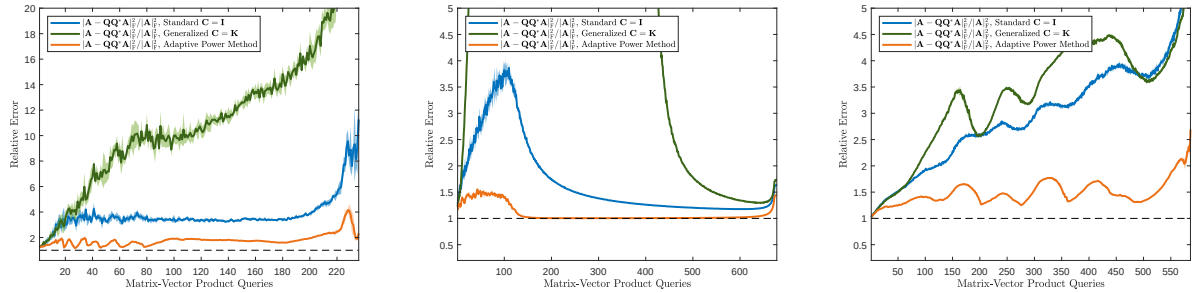

Figure 4: Results for the Power-Method based Adaptive Sampling Method described in Algorithm 2. Low Rank Approximation for a matrix for a Computational Fluid Dynamics Problem, `saylr1` Figure Left is from [DH11]. Subsequent 2D/3D Problem `fs-680-2` Figure Center is from [DH11]. Differential Stiffness Matrix from a Nastran Buckling Problem, `bcsstk34` Figure Right is from [DH11].

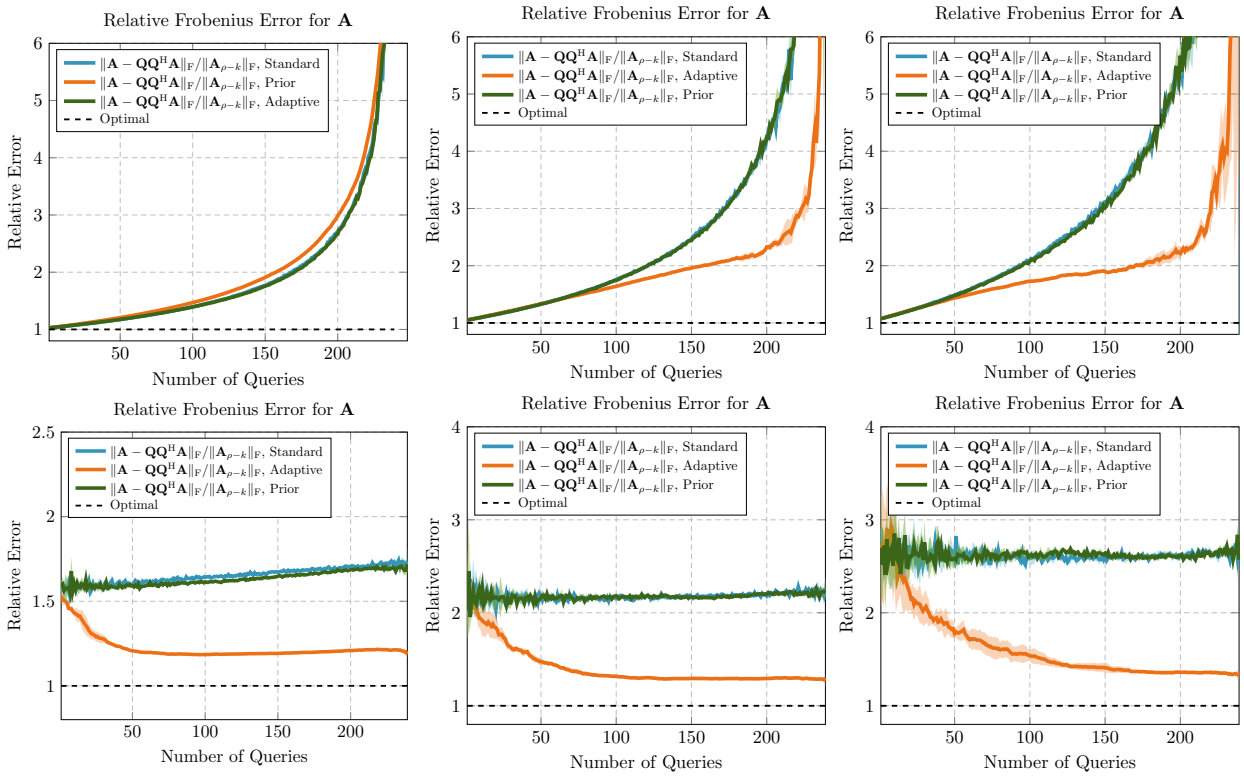

Figure 5: Low Rank Approximation for synthetic matrices with decay described in eq:synthetic-matrix. In *(Top Left)*, $\ell = 1$, in *(Top Center)*, $\ell = 2$, and in *(Top Right)*, $\ell = 3$. In *(Bottom Left)*, $\ell = -1$, in *(Bottom Center)*, $\ell = -2$, and in *(Bottom Right)*, $\ell = -3$.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

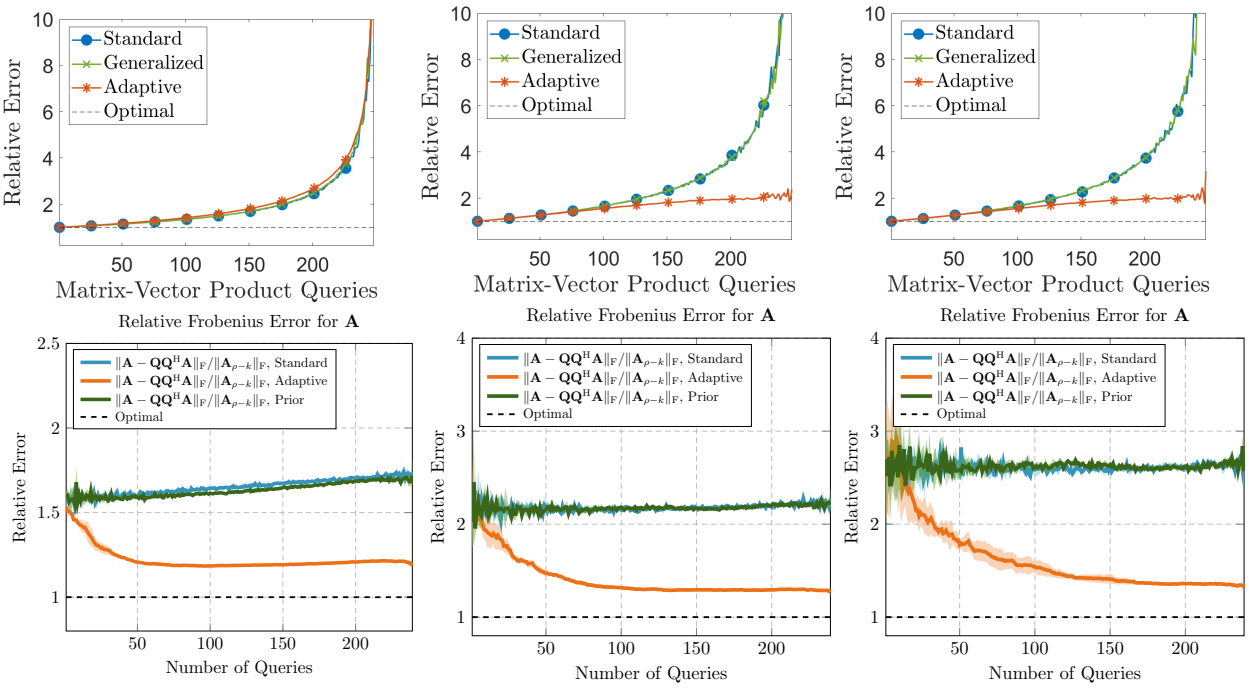

Figure 6: Results for the Power-Method based Adaptive Sampling Method desrcibed in Algorithm 2. Rank Approximation for synthetic matrices with decay described in eq:synthetic-matrix. In *(Top Left)*, $\ell = 1$, in *(Top Center)*, $\ell = 2$, and in *(Top Right)*, $\ell = 3$. In *(Bottom Left)*, $\ell = -1$, 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.

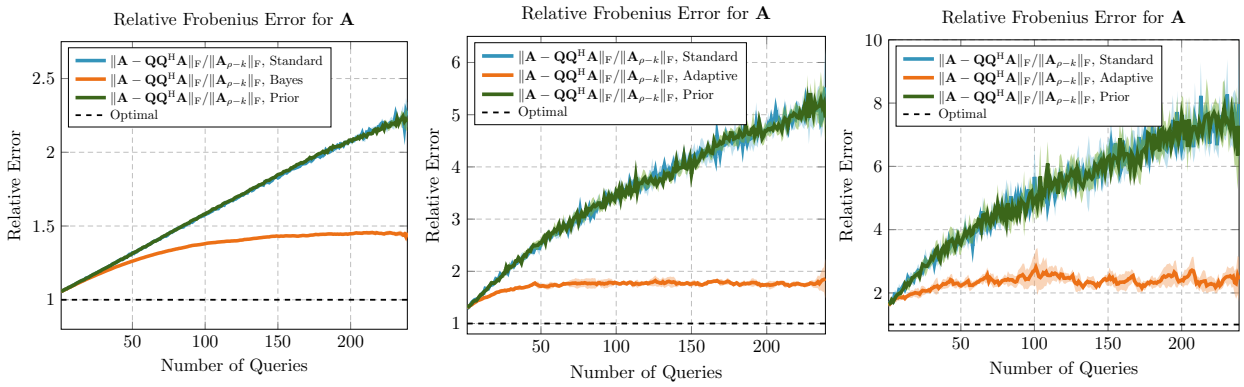

Figure 7: Low Rank Approximation for synthetic matrices with exponential decay described in eq:exponential-decay. In Figure Left, $\delta = 0.01$, in Figure Center, $\delta = 0.05$, and in Figure Right, $\delta = 0.1$. We use the Gram matrix with Gaussian Kernel and $\gamma = 0.01$ for the Prior Covariance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

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, [BET22, BNK20]. 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

We thank mentors Christopher Wang and Nicolas Boullé and supervisor Alex Townsend for the idea of extending Adaptive Sampling for the Matrix-Vector Product Model and the numerous helpful discussions leading to the formulation of the algorithm and the development of the theory. We also would like to thank Alex Gittens for his helpful discussions and encouragement.

References

[BCW22] 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.

[BET22] 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.

[BHSW20] 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.

[BKST22] 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.

[BNK20] Steven L Brunton, Bernd R Noack, and Petros Koumoutsakos. Machine learning for fluid mechanics. *Annual review of fluid mechanics*, 52:477–508, 2020.

[BT22] Nicolas Boullé and Alex Townsend. A generalization of the randomized singular value decomposition. In *International Conference on Learning Representations*, 2022.

[BT23] Nicolas Boullé and Alex Townsend. Learning elliptic partial differential equations with randomized linear algebra. *Foundations of Computational Mathematics*, 23(2):709–739, Apr 2023.

[DH11] Timothy A. Davis and Yifan Hu. The university of florida sparse matrix collection. *ACM Trans. Math. Softw.*, 38(1), dec 2011.

[DRVW06] 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.

[DV06] 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.

[Eva22] Lawrence C Evans. *Partial differential equations*, volume 19. American Mathematical Society, 2022.

[EY36] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[FDH05] 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.

[FKV04] 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.

[HMT11] 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.

[HP14] Sariel Har-Peled. Low rank matrix approximation in linear time. *arXiv preprint arXiv:1410.8802*, 2014.

[LPZ+01] Harvard Lomax, Thomas H Pulliam, David W Zingg, Thomas H Pulliam, and David W Zingg. *Fundamentals of computational fluid dynamics*, volume 246. Springer, 2001.

[Mir60] L. Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 01 1960.

[PMID15] 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.

[SWYZ21] 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.

[TWA+22] Ruo-Chun Tzeng, Po-An Wang, Florian Adriaens, Aristides Gionis, and Chi-Jen Lu. Improved analysis of randomized svd for top-eigenvector approximation. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 2045–2072. PMLR, 28–30 Mar 2022.

[WLSX21] 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.

Lemma 8. *Let $\Omega \in \mathbb{C}^{n\times \ell}$ s.t. the columns of Ω are sampled i.i.d from $\mathcal{N}(\mathbf{0}, C)$ where $C = U\Lambda U^*$ for $U \in \mathbb{O}_{n\times n}$ and $\Lambda \in \mathbb{R}^{n\times n}$ is a PSD diagonal matrix. Then, let $V \in \mathbb{O}_{n\times k}$, it then follows the columns of $V^*\Omega$ are sampled from $\mathcal{N}(\mathbf{0}, K)$ where $K = \operatorname{diag}(V^*CV)$.*

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

$$V^*\Omega = \begin{bmatrix} \mathbf{v}_1 \cdot \boldsymbol{\omega}_1 & \cdots & \mathbf{v}_k \cdot \boldsymbol{\omega}_1 \\ \vdots & \ddots & \vdots \\ \mathbf{v}_1 \cdot \boldsymbol{\omega}_\ell & \cdots & \mathbf{v}_k \cdot \boldsymbol{\omega}_\ell \end{bmatrix} \tag{4}$$

We will first show that the entries are Gaussian, for any $i \in [k]$ and $j \in [\ell]$, we have

$$\mathbf{v}_i \cdot \boldsymbol{\omega}_j = \mathbf{v}_i \cdot C^{1/2}\mathbf{g} = \sum_{k\in[n]} \mathbf{v}_i \cdot \mathbf{u}_k \sqrt{\lambda_1} g_k \tag{5}$$

In the above, we have that each $\left[V^*\Omega\right]_{i,j}$ is Gaussian for $(i,j) \in [k]\times[\ell]$. We will calculate the mean and covariance. We then have for any $(i,j) \in [k]\times[\ell]$,

$$\mathop{\mathbf{E}}_{\boldsymbol{\omega}_j\sim\mathcal{N}(\mathbf{0},C)} [\mathbf{v}_i \cdot \boldsymbol{\omega}_j] = \mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(\mathbf{0},I)} \left[\mathbf{v}_i \cdot C^{1/2}\mathbf{g}\right] = \mathbf{v}_i \cdot C^{1/2} \mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(\mathbf{0},I)} [\mathbf{g}] = \sum_{j=1}^{n} \mathbf{v}_i \cdot \mathbf{u}_j \sqrt{\lambda_j} \mathop{\mathbf{E}}_{g_i\sim\mathcal{N}(0,1)} [g_i] = 0 \tag{6}$$

Then, to calculate the covariance matrix, we have for the non-diagonal elements

$$\mathop{\mathbf{E}}_{\boldsymbol{\omega}_j,\boldsymbol{\omega}_\ell\sim\mathcal{N}(\mathbf{0},C)} \left[\left(\mathbf{v}_i \cdot \boldsymbol{\omega}_j - \mathbf{E}\left[\mathbf{v}_i \cdot \boldsymbol{\omega}_j\right]\right)\left(\mathbf{v}_k \cdot \boldsymbol{\omega}_\ell - \mathbf{E}\left[\mathbf{v}_k \cdot \boldsymbol{\omega}_\ell\right]\right)\right] = \mathop{\mathbf{E}}_{\boldsymbol{\omega}_j,\boldsymbol{\omega}_\ell\sim\mathcal{N}(\mathbf{0},C)} \left[\left(\mathbf{v}_i \cdot \boldsymbol{\omega}_j\right)\left(\mathbf{v}_k \cdot \boldsymbol{\omega}_\ell\right)\right] \tag{7}$$

$$= \mathop{\mathbf{E}}_{\boldsymbol{\omega}_j\sim\mathcal{N}(\mathbf{0},C)} \left[\mathbf{v}_i \cdot \boldsymbol{\omega}_j\right] \mathop{\mathbf{E}}_{\boldsymbol{\omega}_\ell\sim\mathcal{N}(\mathbf{0},C)} \left[\mathbf{v}_k \cdot \boldsymbol{\omega}_\ell\right] = 0 \tag{8}$$

For the diagonal covariance elements, we have

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

We thus have,

$$K_{ij} = \begin{cases} \mathbf{v}_i^* C \mathbf{v}_i & i = j \\ 0 & i \neq j \end{cases} \tag{10}$$

Our proof is complete. ■

A.1 Proof of Lemma 4

Proof. Our first key note is that adaptive low rank matrix approximation is equivalent to performing one sample of the generalized rSVD to the residual $A - QQ^*A$. Let $Q = \operatorname{orth}(Y)$ and $Q_+ = \operatorname{orth}([Y, A\Omega])$ for a sample matrix Ω, and $Q_-Q_-^* = Q_+Q_+^* - QQ^*$, then we have

$$(A - QQ^*A) - Q_-Q_-^*(A - QQ^*A) = A - QQ^*A - Q_-Q_-^*A = A - Q_+Q_+^*A \tag{11}$$

We first give the factorization $A \triangleq (I - QQ^*)A + QQ^*A \triangleq P_\perp A + PA$, it then follows $Q_- = \operatorname{orth}((I - QQ^*)A\Omega) \overset{\text{def}}{=} \operatorname{orth}(P_\perp A\Omega) \triangleq \operatorname{orth}(\widetilde{Y}_-)$. Then, we factorize the residual term as follows,

$$P_\perp A \triangleq (I - QQ^*)A \triangleq \begin{matrix} k & n-k \end{matrix} \begin{bmatrix} \widetilde{U}_k & \widetilde{U}_{k,\perp} \end{bmatrix} \begin{matrix} k \\ n-k \end{matrix} \begin{bmatrix} \widetilde{\Sigma}_k & \\ & \widetilde{\Sigma}_{k,\perp} \end{bmatrix} \begin{matrix} k \\ n-k \end{matrix} \begin{bmatrix} \widetilde{V}_k^* \\ \widetilde{V}_{k,\perp}^* \end{bmatrix} \tag{12}$$

Defining $\widetilde{\Omega}_k \triangleq \widetilde{V}_k^* \Omega$ and $\widetilde{\Omega}_{k,\perp} \triangleq \widetilde{V}_{k,\perp}^* \Omega$. We then have,

$$\mathbf{E}\|A - Q_+Q_+^*A\|_{\mathrm{F}} \overset{(11)}{=} \mathbf{E}\|P_\perp A - Q_-Q_-^*P_\perp A\|_{\mathrm{F}} \tag{13}$$

$$\overset{(\zeta_1)}{\leq} \left(\mathbf{E}\|P_\perp A - Q_-Q_-^*P_\perp A\|_{\mathrm{F}}^2\right)^{1/2} \tag{14}$$

$$\overset{(\zeta_2)}{\leq} \left(\|\widetilde{\Sigma}_{k,\perp}\|_{\mathrm{F}}^2 + \mathbf{E}\|\widetilde{\Sigma}_{k,\perp}\widetilde{\Omega}_{k,\perp}\widetilde{\Omega}_k^+\|_{\mathrm{F}}^2\right)^{1/2} \tag{15}$$

where (ζ_1) follows from Jensen's Inequality and (ζ_2) follows from [HMT11, Theorem 9.1], which holds from noting that $Q_- = \mathrm{orth}(P_\perp A\Omega_-)$. Term by term, we have

$$\|\widetilde{\Sigma}_{k,\perp}\|_{\mathrm{F}}^2 = \|P_\perp A\|_{\mathrm{F}}^2 - \|P_\perp A\|_2^2 = \|A - QQ^*A\|_{\mathrm{F}}^2 - \|A - QQ^*A\|_2^2 \overset{(\zeta_3)}{\leq} \|A - QQ^*A\|_{\mathrm{F}}^2 - \|A - A_k\|_2^2 \tag{16}$$

where in (ζ_3) we used the following,

$$\|A - QQ^*A\|_2^2 \geq \min_{\substack{P\in\mathbb{C}^{m\times m}\\ \mathrm{rank}(P)=k\\ P^2=P}} \|A - PA\|_2^2 \geq \min_{\substack{B\in\mathbb{C}^{m\times n}\\ \mathrm{rank}(B)=k}} \|A - B\|_2^2 \overset{(\zeta_4)}{=} \|A - A_k\|_2^2 \tag{17}$$

where (ζ_4) follows from the Eckart-Young-Mirsky Theorem [EY36, Mir60]. An immediate consequence of Equation (17), is that,

$$\|A - Q_+Q_+^*A\|_{\mathrm{F}}^2 \leq \|A - QQ^*A\|_{\mathrm{F}}^2 + \|\widetilde{\Sigma}_{k,\perp}\widetilde{\Omega}_{k,\perp}\widetilde{\Omega}_k^+\|_{\mathrm{F}}^2 - \|A - A_k\|_2^2 \tag{18}$$

We thus have, for any algorithm we can always improve the approximation by $\|A - A_k\|_2^2$ with some approximation error under the condition that $\mathrm{rank}(Q) = k$ and $\mathrm{rank}(Q_+) = k + p$. Consider the case when $Q = \mathbf{0}$, then we recover the HMT bounds. Expanding out the error term, we have

$$\mathbf{E}\|\widetilde{\Sigma}_{k,\perp}\widetilde{\Omega}_{k,\perp}\widetilde{\Omega}_k^+\|_{\mathrm{F}}^2 = \mathbf{E}\|\widetilde{\Sigma}_{k,\perp}\widetilde{V}_{k,\perp}^*C^{1/2}X(\widetilde{V}_k^*C^{1/2}X)^+\|_{\mathrm{F}}^2 \tag{19}$$

$$= \mathbf{E}\left(\mathbf{E}\left[\|\widetilde{\Sigma}_{k,\perp}\widetilde{V}_{k,\perp}^*C^{1/2}X(\widetilde{V}_k^*C^{1/2}X)^+\|_{\mathrm{F}}^2 \mid \widetilde{V}_k^*C^{1/2}X\right]\right) \tag{20}$$

$$\overset{(\zeta_5)}{=} \|\widetilde{\Sigma}_{k,\perp}\widetilde{V}_{k,\perp}^*C^{1/2}\|_{\mathrm{F}}^2 \cdot \mathbf{E}\|(\widetilde{V}_k^*C^{1/2}X)^+\|_{\mathrm{F}}^2 \tag{21}$$

where (ζ_5) follows from Lemma 11 as X is a standard Gaussian matrix. We then have from [BT23],

$$\mathbf{E}\|\Omega_k^+\|_{\mathrm{F}}^2 = \mathbf{E}\,\mathrm{Tr}((\Omega_k^+)^*\Omega_k^+) = \mathbf{E}\,\mathrm{Tr}((\Omega_k\Omega_k^*)^{-1}) = \mathrm{Tr}(\mathbf{E}[(\Omega_k\Omega_k^*)^{-1}]) = \frac{\mathrm{Tr}(K^{-1})}{p-2} \tag{22}$$

where from Lemma 8, we have $K = \mathrm{diag}(\widetilde{V}_k^*C\widetilde{V}_k)$. From which we obtain the claimed bound,

$$\mathbf{E}\|A - Q_+Q_+^*A\|_{\mathrm{F}}^2 \leq \|A - QQ^*A\|_{\mathrm{F}}^2 + \frac{1}{p-2}\cdot\|\widetilde{\Sigma}_{k,\perp}\widetilde{V}_{k,\perp}^*C^{1/2}\|_{\mathrm{F}}^2 \cdot \mathrm{Tr}([\mathrm{diag}(\widetilde{V}_k^*C\widetilde{V}_k)]^{-1}) - \|A - A_k\|_2^2 \tag{23}$$

$$\leq \|A - QQ^*A\|_{\mathrm{F}}^2 + \frac{1}{p-2}\cdot\|\widetilde{\Sigma}_{k,\perp}\|_{\mathrm{F}}^2\|\tan\Theta(\widetilde{V}_k, C^{1/2})\|_2^2 - \|A - A_k\|_2^2 \tag{24}$$

Our proof is complete. ■

A.2 Proof of Theorem 6

Proof. We have from the definition of the Frobenius Norm,

$$\|A - QQ^*A\|_{\mathrm{F}}^2 = \mathrm{Tr}(A^*A - A^*QQ^*A) = \|A\|_{\mathrm{F}}^2 - \|Q^*A\|_{\mathrm{F}}^2 \tag{25}$$

$$= \|A\|_{\mathrm{F}}^2 - \sum_{i\in[k+p]} \|P_\perp A\boldsymbol{\omega}\|^{-2}\|A^*P_\perp A\boldsymbol{\omega}\|^2 \tag{26}$$

$$\leq \|A\|_{\mathrm{F}}^2 - \sum_{i\in[k+p]} (\delta/2)\|P_\perp A\|_{\mathrm{F}}^2 + \delta\|P_\perp A\|_2^2 \tag{27}$$

■

A.3 Proof of Lemma 7

Proof. Recall for a vector $\boldsymbol{\omega}$, we have

$$\|A - Q_+Q_+^*A\|_{\mathrm{F}}^2 \leq \|A - QQ^*A\|_{\mathrm{F}}^2 + \|\widetilde{\Sigma}_{k,\perp}\|_2^2 \tan^2\Theta(\widetilde{\mathbf{v}}_k, \boldsymbol{\omega}) - \|A - QQ^*A\|_2^2 \tag{28}$$

It becomes clear we want to adaptively choose $\boldsymbol{\omega}$, that reduces $\tan\Theta(\widetilde{\mathbf{v}}_k, \boldsymbol{\omega})$. This inequality is sharp when $Q = U_k$, then choosing $\boldsymbol{\omega} = \mathbf{v}_{k+1}$ is optimal. The natural approach is power iteration. Now, consider $p = 2$, we have,

$$\boldsymbol{\omega}_{k,\perp} = ((I - QQ^*)A)^*(I - QQ^*)A\mathbf{v} = A^*(I - QQ^*)A\mathbf{v} = A^*A\mathbf{v} - A^*QQ^*QQ^*A\mathbf{v} \tag{29}$$

Recall at any iteration t, we have QQ^*A, and thus the second term is known. The first term on the other hand requires two matrix-vector multiplications. Let $\mathbf{v}_i$ for $i \in [n]$ represent the right singular vectors of $(I - QQ^*)A$. Then, we have for some coefficients, α_i for $i \in [n]$ such that $\boldsymbol{\omega} = \sum_{i\in[n]} \alpha_i\widetilde{\mathbf{v}}_i$. We then, have $\boldsymbol{\omega}_{k,\perp} = A^*(I - QQ^*)A\boldsymbol{\omega} = \sum_{i\in[n]} \alpha_i\lambda_i\big(A^*(I - QQ^*)A\big)\widetilde{\mathbf{v}}_i$. Then we have,

$$\cos^2\Theta(\widetilde{\mathbf{v}}_k, \boldsymbol{\omega}_p) = \alpha_k^2\lambda_1^p(A^*(I - QQ^*)A) \tag{30}$$

and for sin we have

$$\sin^2\Theta(\widetilde{\mathbf{v}}_k, \boldsymbol{\omega}_p) = \sum_{i=2}^{n} \alpha_i^2\lambda_i(A^*(I - QQ^*)A) \leq \left(\sum_{i=2}^{n} \alpha_i^2\right)\lambda_2(A^*(I - QQ^*)A) \tag{31}$$

Combining our upper bound for sin and lower bound for cos, we have

$$\tan^2\Theta(\widetilde{\mathbf{v}}_k, \boldsymbol{\omega}_p) \leq \left(\frac{\lambda_2\big(A^*(I - QQ^*)A\big)}{\lambda_1\big(A^*(I - QQ^*)A\big)}\right)^p \tan^2\Theta(\widetilde{\mathbf{v}}_k, \boldsymbol{\omega}) \tag{32}$$

We now bound the top eigenvalue ratio of $A^*(I - QQ^*)A$,

$$\begin{aligned}
\lambda_1(A^*(I - QQ^*)A) &= \lambda_1(A^*(I - U_kU_k^* + U_kU_k^* - QQ^*)A) && (33)\\
&\geq \lambda_1(A^*A - A_k^*A_k) + \lambda_n(A^*(U_kU_k^* - QQ^*)A) && (34)\\
&\geq \lambda_1(A^*A - A_k^*A_k) - \lambda_1(A^*A)\|U_kU_k^* - QQ^*\|_2 && (35)
\end{aligned}$$

and from the similar idea, we obtain the following upper bound from Weyl's Inequality,

$$\lambda_2(A^*(I - QQ^*)A) \leq \lambda_2(A^*A - A_k^*A_k) + \lambda_1(A^*A)\|U_kU_k - QQ^*\|_2 \tag{36}$$

Combining our previous two estimates, we obtain,

$$\begin{aligned}
\tan^2\Theta(\widetilde{\mathbf{v}}_k, \boldsymbol{\omega}_p) &\leq \left(\frac{\lambda_{k+1}(A^*A) + \lambda_1(A^*A)\|U_kU_k - QQ^*\|_2}{\lambda_1(A^*A) - \lambda_1(A^*A)\|U_kU_k - QQ^*\|_2}\right)^p \tan^2\Theta(\widetilde{\mathbf{v}}_k, \boldsymbol{\omega}) && (37)\\
&\leq \left(\frac{\lambda_{k+1}(A^*A)}{\lambda_1(A^*A)}\right)^p \left(\frac{1+\epsilon}{1-\epsilon}\right)^p \tan^2\Theta(\widetilde{\mathbf{v}}_k, \boldsymbol{\omega}) && (38)
\end{aligned}$$

Our proof is complete.

■

B Probability Theory

Lemma 9. *Let $\mathbf{v} \in \mathbb{R}^n$ have unitary ℓ_2 norm and $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, I)$. Then with probability at least $1 - \delta$,*

$$\tan\Theta(\mathbf{v}, \mathbf{x}) \leq \tag{39}$$

Proof. We have the following decomposition,

$$\mathbf{Pr}\big\{\tan^2\Theta(\mathbf{v},\mathbf{x}) \geq t\big\} = \mathbf{Pr}\big\{\sin^2\Theta \geq t\cos^2\Theta\big\} = \mathbf{Pr}\bigg\{\cos^2\Theta \leq \frac{1}{t+1}\bigg\} \leq \mathbf{Pr}\bigg\{|\mathbf{x}\cdot\mathbf{v}| \leq \|\mathbf{x}\|\sqrt{\frac{1}{t}}\bigg\} \tag{40}$$

$$= \mathbf{Pr}\bigg\{|\mathbf{x}\cdot\mathbf{v}| \leq \|\mathbf{x}\|\sqrt{\frac{1}{t}} \mid \|\mathbf{x}\| \leq C\bigg\} + \mathbf{Pr}\bigg\{|\mathbf{x}\cdot\mathbf{v}| \leq \|\mathbf{x}\|\sqrt{\frac{1}{t}} \mid \|\mathbf{x}\| \geq C\bigg\} \tag{41}$$

$$\leq \mathbf{Pr}\bigg\{|\mathbf{x}\cdot\mathbf{v}| \leq C\sqrt{\frac{1}{t}}\bigg\} + \mathbf{Pr}\{\|\mathbf{x}\| \geq C\} \tag{42}$$

We now make a ε-net argument. Let $\mathcal{N}$ be a ε-net of $\mathbb{S}^{n-1}$. Then let $\mathbf{v}^* = \arg\max_{\|\mathbf{v}\|=1}|\mathbf{x}\cdot\mathbf{v}|$ and let $\mathbf{u}\in\mathcal{N}$ such that $\|\mathbf{u}-\mathbf{v}^*\| \leq \varepsilon$, then we have

$$|\mathbf{x}\cdot\mathbf{u}| \geq |\mathbf{x}\cdot\mathbf{v}^*| - |\mathbf{x}\cdot\mathbf{v}^* - \mathbf{x}\cdot\mathbf{u}| \geq \|\mathbf{x}\|(1-\varepsilon) \tag{43}$$

From which we obtain, $\|\mathbf{x}\| \leq (1-\varepsilon)^{-1}\arg\max_{\mathbf{u}\in\mathcal{N}}|\mathbf{x}\cdot\mathbf{u}|$. Then, from a union bound, we have,

$$\mathbf{Pr}\{\|\mathbf{x}\| \geq C\} \leq \left(\frac{3}{\varepsilon}\right)^n \mathbf{Pr}\{|\mathbf{x}\cdot\mathbf{v}| \geq C(1-\varepsilon)\} \leq 12^n \exp\left(-\frac{C^2}{2}\right) \leq \frac{\delta}{2} \tag{44}$$

Suppose we choose $\varepsilon = \frac{1}{4}$, then the above probabilistic condition is satisfied when $C^2 \geq 2n\log(12)+2\log(1/\delta)$. We then have,

$$\mathbf{Pr}\bigg\{|\mathbf{x}\cdot\mathbf{v}| \leq C\sqrt{\frac{1}{t}}\bigg\} \leq \tag{45}$$

The above probabilistic condition holds when $t \leq \delta^2 C^2$. We then obtain with probability $1-\delta$,

$$\tan\Theta(\mathbf{v},\mathbf{x}) \leq \delta^2(2n\log(12)) \tag{46}$$

■

Lemma 10. *Let $A \in \mathbb{C}^{m\times n}$, then sample $\mathbf{x}\sim\mathcal{N}(\mathbf{0},I)$, then with probability at least $1-\delta$,*

$$\frac{\|A^*A\mathbf{x}\|}{\|A\mathbf{x}\|} \geq (\delta/2)\|A\|_{\mathrm{F}}^2 - \delta\|A\|_2^2 = (\delta/2)\big(\|A_{1,\perp}\|_{\mathrm{F}}^2 - \|A\|_2^2\big) \tag{47}$$

Proof. Let $\mathcal{D}$ be the distribution $\mathcal{N}(\mathbf{0},I)$ for shorthand,

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}}\big\{\|A\mathbf{x}\|^2 \geq t^{-1}\|A^*A\mathbf{x}\|^2\big\}$$

$$= \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}}\big\{\|A\mathbf{x}\|^2 \geq t^{-1}\|A^*A\mathbf{x}\|^2, \|A^*A\mathbf{x}\|^2 \leq C\big\} + \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}}\big\{\|A^*A\mathbf{x}\|^2 \geq t^{-1}\|A^*A\mathbf{x}\|^2, \|A^*A\mathbf{x}\|^2 \geq C\big\} \tag{48}$$

$$\leq \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}}\big\{\|A\mathbf{x}\|^2 \geq Ct^{-1}\big\} + \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}}\Big\{\|A^*A\mathbf{x}\|^2 \leq C\Big\} \tag{49}$$

We can note from standard analysis in trace estimation,

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}\|A^*A\mathbf{x}\|^2 = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}\mathrm{Tr}(\mathbf{x}^*A^*AA^*A\mathbf{x}) = \mathrm{Tr}(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^*]A^*AA^*A) = \|A^*A\|_{\mathrm{F}}^2 \tag{50}$$

We can also note that,

$$\mathop{\mathbf{Var}}_{\mathbf{x}\sim\mathcal{D}}(\|A^*A\mathbf{x}\|^2) = 2\|A^*AA^*A\|_{\mathrm{F}}^2 \tag{51}$$

Then set $C = \varepsilon\|A^*A\|_{\mathrm{F}}^2$ for $\varepsilon\in(0,1)$. We then have,

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}}\Big\{\|A^*A\mathbf{x}\|^2 \geq C\Big\} \leq \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}}\Big\{|\|A^*A\mathbf{x}\|^2 - \|A^*A\|_{\mathrm{F}}^2| \geq (1-\varepsilon)\|A^*A\|_{\mathrm{F}}^2\Big\} \tag{52}$$

$$\leq \frac{2\|A^*AA^*A\|_{\mathrm{F}}^2}{(1-\varepsilon)^2\|A^*A\|_{\mathrm{F}}^4} \leq \frac{2\|A^*A\|_2^2}{(1-\varepsilon)^2\|A^*A\|_{\mathrm{F}}^2} \leq \frac{\delta}{2} \tag{53}$$

where the above condition holds when,

$$\varepsilon \leq 1 - \frac{2\|A^*A\|_2}{\|A^*A\|_{\mathrm{F}}} \tag{54}$$

Then from Markov's Inequality, we have

$$\Pr_{\mathbf{x}\sim\mathcal{D}}\Big\{\|A\mathbf{x}\|^2 \geq \varepsilon t^{-1}\|A^*A\|_{\mathrm{F}}^2\Big\} \leq \frac{t}{\varepsilon\|A^*A\|_{\mathrm{F}}} \leq \frac{\delta}{2} \tag{55}$$

Where the final inequality holds when

$$t \leq \frac{\varepsilon\delta\|A^*A\|_{\mathrm{F}}}{2} \leq (\delta/2)\|A^*A\|_{\mathrm{F}} - \delta\|A^*A\|_2 \tag{56}$$

Combining our results, we have with probability exceeding $1-\delta$,

$$\frac{\|A^*A\mathbf{x}\|}{\|A\mathbf{x}\|} \geq (\delta/2)\|A\|_{\mathrm{F}}^2 - \delta\|A\|_2^2 \tag{57}$$

Our proof is complete. ■

Lemma 11 (Proposition 10.1 [HMT11])**.** *Fix matrices $S \in \mathbb{R}^{K\times N}$ and $T \in \mathbb{R}^{N\times L}$, then for a matrix G such that elements of G are sampled i.i.d from $\mathcal{N}(0,1)$, then*

$$\big(\mathbf{E}\|SGT\|_{\mathrm{F}}^2\big)^{1/2} = \|S\|_{\mathrm{F}}\|T\|_{\mathrm{F}} \tag{58}$$

Proof. The proof is a simple calculation.

$$\mathbf{E}\|SGT\|_{\mathrm{F}}^2 = \sum_{i\in[K]}\sum_{j\in[L]}\sum_{(k_1,k_2)\in[N]\times[N]} S_{i,k_1}^2 T_{k_2,j}^2\,\mathbf{E}\big[G_{k_1,k_2}^2\big] \tag{59}$$

$$= \sum_{i\in[K]}\sum_{j\in[L]}\sum_{(k_1,k_2)\in[N]\times[N]} S_{i,k_1}^2 T_{k_2,j}^2 \tag{60}$$

$$= \|S\|_{\mathrm{F}}^2\|T\|_{\mathrm{F}}^2 \tag{61}$$

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