Data Driven Adaptive Sampling for Low-Rank Matrix Approximation

Arvind Rathnashyam*
RPI Math and CS, rathna@rpi.edu

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

Alex Townsend
Cornell Math, ajt453@cornell.edu

Abstract

We consider the problem of low-rank matrix approximation in case the when the matrix $\mathbf{A}$ 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 [24]. Recently, [2] generalized the rSVD to Hilbert-Schmidt Operators where functions are sampled from non-standard Covariance Matrices when there is already prior information on the right singular vectors within the column space of the target matrix, $\mathbf{A}$. In this work, we develop an adaptive sampling framework for the Matrix-Vector Product Model which does not need prior information on the matrix $\mathbf{A}$. We provide a novel theoretic analysis of our algorithm with subspace perturbation theory. We extend the analysis of [45] 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 [17], fluid dynamics [32]. 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, [24], has been theoretically analyzed and used in various applications. Even more recently, [4] discovered if we have prior information on the right singular vectors of $\mathbf{A}$, we can modify the Covariance Matrix such that the sampled vectors are within the column space of $\mathbf{A}$. They extended the theory for Randomized SVD where the covariance matrix is now a general PSD matrix. The basis of our analysis is the idea of sampling vectors in the Null-Space of the Low-Rank Approximation. This idea has been introduced recently in Machine Learning in [47] 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.

Adaptive Sampling techniques for Low-Rank Matrix Approximation first appeared in CUR Matrix Decomposition in [19]. Optimial column-sampling for the CUR Matrix Decomposition received much attention as can be seen in the works [26, 13, 12]. More recently, [42] 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 $\mathbf{A}$ 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.

The theoretical properties of adaptively sampled matrix vector queries 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 $\mathbf{A}$.
2. We provide a novel theoretical analysis which utilizes subspace perturbation theory.
3. We perform extensive experiments on matrices with various spectrums and compare with the state of the art methods.

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

Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ represent the target matrix. $\|\cdot\|$ represents the spectral norm, which is equivalent to the max eigenvalue of the argument, $\sigma_{\max}(\cdot)$. Quasimatrices (matrices with infinite rows and finite columns) will be denoted as a variation of the symbol, $\mathbf{\Omega}$. The pseudoinverse is represented by $(\cdot)^\dagger$ s.t. $\mathbf{X}^\dagger = (\mathbf{X}^*\mathbf{X})^{-1}\mathbf{X}^*$. The Projection Matrix is defined as $\Pi_{\mathbf{Y}} = \mathbf{Y}\mathbf{Y}^\dagger = \mathbf{Y}(\mathbf{Y}^*\mathbf{Y})^{-1}\mathbf{Y}^*$ as the projection on to the column space of $\mathbf{Y}$. If $\mathbf{Y}$ has orthogonal columns, then $\Pi_{\mathbf{Y}}$ is the Orthogonal Projection defined as $\Pi_{\mathbf{Y}} = \mathbf{Y}\mathbf{Y}^*$. 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. $\{\mathbf{V} : \mathbf{V}^*\mathbf{V} = \mathbf{I}_{k\times k}\}$. We also denote $\mathcal{MN}(\mathbf{0}, \mathbf{I}_{n\times n}, \mathbf{I}_{m\times m})$, denote the distribution of $m \times n$ standard gaussian matrices. The Frobenius norm for a matrix is defined as,

$$\|\mathbf{A}\|_{\mathrm{F}} = \left(\sum_{i\in[m]}\sum_{j\in[n]} \mathbf{A}_{i,j}^2\right)^{1/2} = \sqrt{\operatorname{Tr}\left(\mathbf{A}^*\mathbf{A}\right)} = \sqrt{\operatorname{Tr}\left(\mathbf{A}\mathbf{A}^*\right)} \tag{1}$$

We define $[\![\mathbf{A}]\!]_r$ as the best rank-r approximation to $\mathbf{A}$ w.r.t the Frobenius norm. We use Big-O notation, $y \leq \mathcal{O}(x)$, to denote $y \leq Cx$ for some positive constant, C. We define $\mathbb{E}$ as expection, $\mathbb{P}$ as probability, and $\mathbb{V}$ as variance. We will denote normal text characters A, X, Y as matrices and lower roman boldface characters $\mathbf{x}, \mathbf{y}, \mathbf{z}$ as vectors.

2.2 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 $\frac{\pi}{2} \geq \Theta_1(\mathcal{X}, \mathcal{Y}) \geq \ldots \geq \Theta_{m\wedge n}(\mathcal{X}, \mathcal{Y})$. Typically, the norm for distance between subspaces $\mathcal{X}$ and $\mathcal{Y}$ is defined as,

$$\|\sin\Theta\left(\mathcal{X}, \mathcal{Y}\right)\|_{\mathrm{F}} = \frac{1}{2}\|\Pi_{\mathcal{X}} - \Pi_{\mathcal{Y}}\|_{\mathrm{F}} \tag{2}$$

In a landmark paper by [10], they introduced upper bounds for $\|\sin\Theta\left(\mathcal{X}, \mathcal{Y}\right)\|$ and $\|\tan\Theta\left(\mathcal{X}, \mathcal{Y}\right)\|$. A generalized version of the $\sin\Theta$ theorem for rectangular matrices is given in [50].

Theorem 1. *[50]. Let $\mathbf{A}, \widehat{\mathbf{A}} \in \mathbb{R}^{m\times n}$ have singular values $\sigma_1 \geq \ldots \geq \sigma_{m\vee n}$ and $\widehat{\sigma}_1 \geq \ldots \geq \widehat{\sigma}_{m\vee n}$, respectively. Given $j \in 1, \ldots, m \vee n$, it follows*

$$\|\sin\Theta\left(\widehat{\mathbf{v}}_j, \mathbf{v}_j\right)\|_{\mathrm{F}} \leq \frac{2\left(2\sigma_1 + \left\|\widehat{\mathbf{A}} - \mathbf{A}\right\|\right)\left\|\widehat{\mathbf{A}} - \mathbf{A}\right\|_{\mathrm{F}}}{\sigma_j^2 - \sigma_{j+1}^2} \wedge 1 \tag{3}$$

However, we would like to note this theorem tends to not be sharp enough for theoretical use. Instead, we introduce Wedin's Theorem,

Theorem 2. *[48]. Let $\mathbf{A}, \widehat{\mathbf{A}} \in \mathbb{R}^{m\times n}$ have singular values $\sigma_1 \geq \ldots \geq \sigma_{m\vee n}$ and $\widehat{\sigma}_1 \geq \ldots \geq \widehat{\sigma}_{m\vee n}$, respectively. Given $j \in 1, \ldots, m \vee n$,*

$$\sin\Theta\left(\mathbf{v}_1, \widetilde{\mathbf{v}}_1\right) \leq \frac{\left\|\mathbf{A} - \widehat{\mathbf{A}}\right\|}{\sigma_1 - \widehat{\sigma}_2} \tag{4}$$

Theorem 3. *[40]. Let $\mathbf{A}, \widehat{\mathbf{A}} \in \mathbb{R}^{m\times n}$ have singular values $\sigma_1 \geq \ldots \geq \sigma_{m\vee n}$ and $\widehat{\sigma}_1 \geq \ldots \geq \widehat{\sigma}_{m\vee n}$, respectively. Then,*

$$\sin\Theta\left(\mathbf{v}_1, \widetilde{\mathbf{v}}_1\right) \leq \frac{2\left\|\mathbf{A} - \widehat{\mathbf{A}}\right\|}{\sigma_1 - \sigma_2} \tag{5}$$

Now we give some introduction to singular vector perturbation theory. Given two vectors, $\mathbf{v}, \widetilde{\mathbf{v}} \in \mathbb{R}^n$ s.t. $\|\mathbf{v}\| = \|\widetilde{\mathbf{v}}\| = 1$, it follows $\cos\Theta\left(\mathbf{v}, \widetilde{\mathbf{v}}\right) = \mathbf{v}^*\widetilde{\mathbf{v}}$. Let $\mathbf{V}$ be the matrix representing an orthonormal basis of vectors in $\mathbb{R}^n$: $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. Using basic ideas in trigonometry we find that,

$$\sin^2\Theta\left(\mathbf{v}_j, \widetilde{\mathbf{v}}\right) = 1 - \cos^2\Theta\left(\mathbf{v}_j, \widetilde{\mathbf{v}}\right) = \|\mathbf{V}^*\widetilde{\mathbf{v}}\|^2 - \left\|\mathbf{v}_j^*\widetilde{\mathbf{v}}\right\|^2 = \sum_{i=1}^{n} \mathbb{I}\left\{i \neq j\right\}\|\mathbf{v}_i^*\widetilde{\mathbf{v}}\|^2 \triangleq \left\|\mathbf{V}_{\perp,j}^*\widetilde{\mathbf{v}}\right\|^2 \tag{6}$$

2.3 Relevant Works

The Randomized Singular Value Decomposition was developed and analyzed thoroughly in [24]; throughout this paper we will refer to this algorithm as HMT. The review work by [34] gives significant theory on the Randomized SVD. [2] proposed learning the Hilbert-Schmidt Operators associated with the Green's Functions with the randomized SVD algorithm. One of their key findings is they can better approximate the HS Operator when they use functions drawn from $\mathcal{GP}(\mathbf{0}, \mathbf{K})$ where $\mathbf{K}$ is not the identity. [3] extended upon previous work on generalizing the Randomized SVD to learning HS Operators. [28] empirically look at Entropy Search for Probabilistic Optimization. [41] analyzed faster algorithms for the approximation of the null-space.

Block iterative methods have also been studied extensively in [25]. The study of block Krylov subspaces have also seen increased attention in the last few years, [44]. Bounds for the $(1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|$ approximation error with randomized block Krylov Subspace methods have been explored in [38, 1].

The most relevant work to ours is likely [15]. The measure of accuracy in the Krylov Subspace is measured by the $\sin\Theta$ norm. We would like to note the Krylov Subspace method takes q times more matrix-vector products and thus is not a suitable method for our problem. Work similar to ours with regards to upper bounding the sine of the principal angles between subspaces in the context of the Randomized SVD is explored in [14] and [43].

Upper bounds on the tangent of principal angles are visited in [39] and improved in [35]. The highly studied $\sin\Theta$ norms are studied in depth in [9, 40].

Learning algorithms for Low-Rank Matrix Approximations have also been explored. In [30] and [29], the sketching matrix is learned.

A similar analysis of a power method is explored in [27] utilizing subspace perturbation theory. In this work, they consider the Matrix-Vector products have noise. In this work, similarly to [15], it takes d times more matrix-vector products to recover the right singular space. Furthermore, a similar projection-based analysis based on the sines of the singular vector perturbations is done in [33].

The theoretical analysis of greedy algorithms for low-rank matrix approximation tend to be difficult as noted in the discussion of [21]. Such algorithms have been developed for Nyström Approximation in [18, 31].

3 Data Driven 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 [24], the covariance matrix is given as:

$$\mathbf{C} = \mathbf{I} \tag{7}$$

In the generalization of the Randomized SVD, when given some prior information of the matrix, the covariance matrix is given as:

$$\mathbf{C} = \mathbf{K} \tag{8}$$

where $\mathbf{K}$ should be close to the right singular vectors of $\mathbf{A}$. Let $\widetilde{\mathbf{V}}$ 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 follows:

$$\mathbf{C}^{(k+1)} = \widetilde{\mathbf{V}}_{(:,k)} \widetilde{\mathbf{V}}^*_{(:,k)} \tag{9}$$

Throughout this paper we will only consider $\mathbf{C}^{(0)} = \mathbf{I}$ due to simplified analysis, however using theory from [2], this can be extended to $\mathbf{C}^{(0)} = \mathbf{K}$ if one has some knowledge of the right singular vectors. A similar algorithm can be found in [47]. Naturally, 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 'waste' samples which do not learn any new information about the matrix. To further motivate our covariance update, we will introduce the following remark.

Remark 4. Let $\mathbf{U\Sigma V}^*$ be the SVD of $\mathbf{A}$, then the Covariance update described in Equation (9) is the optimal covariance update is the optimal covariance matrix for sampling vectors at iteration k.

Remark 4 is an intuitive result, in that when we are learning a matrix $\mathbf{A}$, we would optimally want to sample the right singular vectors, so the resultant matrix product is the left singular vectors.

3.1 Algorithm

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 Bayesian Function Sampling

Input: HS Operator: $\mathscr{F}$, Rank: r, Initial Covariance: $\mathbf{C}$, Oversampling Parameter: p
Output: Rank-r Approximation, $\widehat{\mathbf{A}}_r$
$\mathbf{\Omega} \leftarrow [\underbrace{\mathcal{N}(\mathbf{0}, \mathbf{C}) \quad \overset{\text{i.i.d}}{\cdots} \quad \mathcal{N}(\mathbf{0}, \mathbf{C})}_{p}]$ ▷ Sample Oversampling Vectors from Standard Normal Matrix
$\mathbf{Y} \leftarrow \mathbf{A\Omega}$ ▷ Matrix Vector Products
$[\mathbf{Q}_0, \sim] \leftarrow \text{QR}(\mathbf{Y})$ ▷ Find Orthonormal Basis
$\widetilde{\mathbf{A}} \leftarrow \mathbf{0}_{m\times n}$ ▷ Initial Low-Rank Approximation
for $k \in 1, 2, \dots, r$ **do**
  $\widetilde{\mathbf{A}}_k \leftarrow \widetilde{\mathbf{A}}_{k-1} + \mathbf{Q}_{(k-1)}\mathbf{Q}^*_{(k-1)}\mathbf{A}$ ▷ Rank-1 update to the low-rank approximation
  $\left[\widetilde{\mathbf{U}}, \widetilde{\mathbf{\Sigma}}, \widetilde{\mathbf{V}}\right] \leftarrow \text{SVD}(\widetilde{\mathbf{A}}_k)$ ▷ SVD of current low-rank approximation
  $\mathbf{C}^{(k+1)} \leftarrow \widetilde{\mathbf{V}}_{(:,k)}\widetilde{\mathbf{V}}^*_{(:,k)}$ ▷ Form new Covariance Matrix
  $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}^{(k+1)})$ ▷ Adaptive Sampling of Vector
  $\mathbf{Y} \leftarrow [\mathbf{Y} \quad \mathbf{Ax}]$ ▷ Matrix-Vector Product
  $[\mathbf{Q}_k, \sim] \leftarrow \text{QR}(\mathbf{Y})$ ▷ Find Orthonormal Basis
end for
$\widetilde{\mathbf{A}}_r \leftarrow \mathbf{Q}_r\mathbf{Q}_r^*\mathbf{A}$ ▷ Final Low-Rank Approximation
Return: $\widetilde{\mathbf{A}}_r$

In Algorithm 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 $\mathbf{QQ}^*\mathbf{A}$ with the adaptive matrix vector query. From here, we adaptively query with the kth right singular value of the SVD of the the low rank approximation at iteration $k-1$. We believe this algorithm to be the closest to replicate the idea given in Remark 4.

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 Setup.

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

$$\mathbf{A} = \begin{array}{cc} k & \rho-k \\ \multicolumn{2}{c}{\begin{bmatrix}\mathbf{U}_k & \mathbf{U}_{\rho-k}\end{bmatrix}} \end{array} \begin{array}{cc} k & \rho-k \\ \multicolumn{2}{c}{\begin{bmatrix}\mathbf{\Sigma}_k & \\ & \mathbf{\Sigma}_{\rho-k}\end{bmatrix}} \end{array} \begin{array}{c} \begin{bmatrix}\mathbf{V}_k^* \\ \mathbf{V}_{\rho-k}^*\end{bmatrix} \end{array} \begin{array}{c} k \\ \rho-k \end{array} = \sum_{i=1}^{\rho} \sigma_i \mathbf{u}_i \mathbf{v}_i^* = \sum_{i=1}^{\rho} \mathbf{U}_{(i)}\mathbf{\Sigma}_{(i)}\mathbf{V}^*_{(i)} \tag{10}$$

Furthermore, we let $\mathbf{A}_{(k)} \triangleq \sigma_k \mathbf{u}_k \mathbf{v}_k^*$. Let $\mathbf{\Omega} \in \mathbb{R}^{n\times \ell}$ be a test matrix where $\ell = k + p$ denotes the number of samples and p is the oversampling parameter.

4.2 Previous Literature

We first restate the expected Frobenius error in the Low-Rank Approximation obtained by the Randomized SVD with data sampling from a central and uncorrelated Normal Distribution.

Theorem 5. *[24][Theorem 10.5] Let $\mathbf{A} \in \mathbb{R}^{m\times n}$, $k \geq 2$, oversampling parameter $p \geq 2$, where $k+p \leq m \wedge n$. Let $\mathbf{\Omega} \sim \mathcal{MN}\left(\mathbf{0}, \mathbf{I}_{n\times n}, \mathbf{I}_{k+p\times k+p}\right)$, and $\mathbf{Q} \triangleq \operatorname{orth}\left(\mathbf{A\Omega}\right)$. Then,*

$$\mathbb{E}\left\|\mathbf{A} - \left(\mathbf{Q}\mathbf{Q}^*\mathbf{A}\right)_k\right\|_{\mathrm{F}} \leq \left(1 + \frac{k}{p-1}\right)^{1/2} \sqrt{\sum_{j=k+1}^{n} \sigma_j^2} \tag{11}$$

We will now restate the expected frobenius norm error in the Low-Rank Approximation obtained by the Randomized SVD with vector sampling from a central and correlated Normal Distribution.

Theorem 6. *[2][Theorem 2] Under the same conditions as Theorem 5, except assume the columns of $\mathbf{\Omega}$ are sampled from $\mathcal{N}(\mathbf{0}, \mathbf{K})$.*

$$\mathbb{E}\left\|\mathbf{A} - \left(\mathbf{Q}\mathbf{Q}^*\mathbf{A}\right)_k\right\|_{\mathrm{F}} \leq \left(1 + \sqrt{\frac{\beta_k}{\gamma_k}\frac{(k+p)}{(p-1)}}\right) \sqrt{\sum_{j=k+1}^{n} \sigma_j^2} \tag{12}$$

where

$$\gamma_k = \frac{k}{\lambda_1 \operatorname{Tr}\left(\left(\mathbf{V}_1^*\mathbf{K}\mathbf{V}_1\right)^{-1}\right)} \quad \textit{and} \quad \beta_k = \frac{\operatorname{Tr}\left(\mathbf{\Sigma}_2^2\mathbf{V}_2^*\mathbf{K}\mathbf{V}_2\right)}{\lambda_1\left\|\mathbf{\Sigma}_2\right\|_{\mathrm{F}}^2} \tag{13}$$

In the literature, approximation error bounds on $\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^*\mathbf{A}\|$ typically are of the form

$$\left(1 + \underbrace{\left\|\mathbf{\Sigma}_{\rho-k}\left(\mathbf{V}_{\rho-k}^*\mathbf{\Omega}\right)\left(\mathbf{V}_k^*\mathbf{\Omega}\right)^{\dagger}\right\|}_{\psi}\right)^{1/2} \left\|\mathbf{\Sigma}_{\rho-k}\right\|_{\mathrm{F}} \tag{14}$$

See Theorems 5 and 6 and [6, 20]. However, we find working with $\left\|\left(\mathbf{V}_k^*\mathbf{\Omega}\right)^{\dagger}\right\|$ is difficult since this norm can be extremely large. The 'ψ' term in Equation (14) has been studied w.r.t to Krylov sSubspaces in [15]. In [15][Theorem 2.2], Drineas et al. find

$$\psi \leq \left\|\tan \mathbf{\Theta}\left(\widetilde{\mathbf{V}}, \mathbf{V}_k\right)\right\| \tag{15}$$

Since the tan function is unbounded, upper bounding ψ is difficult and may not lead to strong bounds. First we will introduce a lemma for the resultant vector of sampling from $\mathbf{C}^{(k)}$. 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. With Theorem 1, we can now put bounds on the top right singular vector approximation by the Randomized SVD.

Lemma 7. *Let $\mathbf{A} \in \mathbb{C}^{m\times n}$ and $\mathbf{Q}$ be an orthogonal matrix representing the basis of the subspace of $\mathbf{Y} \in \mathbb{C}^{m\times k}$. Let $\mathbf{v} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I})$ and $\tilde{\mathbf{v}}$ represent the right singular vector of $\mathbf{Q}\mathbf{Q}^*\mathbf{A}$. Denote $\widehat{\mathbf{Q}} \triangleq \operatorname{orth}([\mathbf{Y} \quad \mathbf{v}])$ and $\widetilde{\mathbf{Q}} \triangleq \operatorname{orth}([\mathbf{Y} \quad \tilde{\mathbf{v}}])$, then we have with probability $1-\delta$*
Claim (i):

$$\left\|\mathbf{A} - \widetilde{\mathbf{Q}}\widetilde{\mathbf{Q}}^*\mathbf{A}\right\|_{\mathrm{F}} \leq \left\|\left(\mathbf{I} - \mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}\right\|_{\mathrm{F}} \sqrt{1 - \frac{\left(\sigma_k - 1\right)^2}{2\left(1 + \frac{1}{\sigma_k^2}\right)\left\|\left(\mathbf{I} - \mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}\right\|_{\mathrm{F}}^2} + \frac{\sigma_{k+1}^2 + \eta^2}{\left\|\left(\mathbf{I} - \mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}\right\|_{\mathrm{F}}}} \tag{16}$$

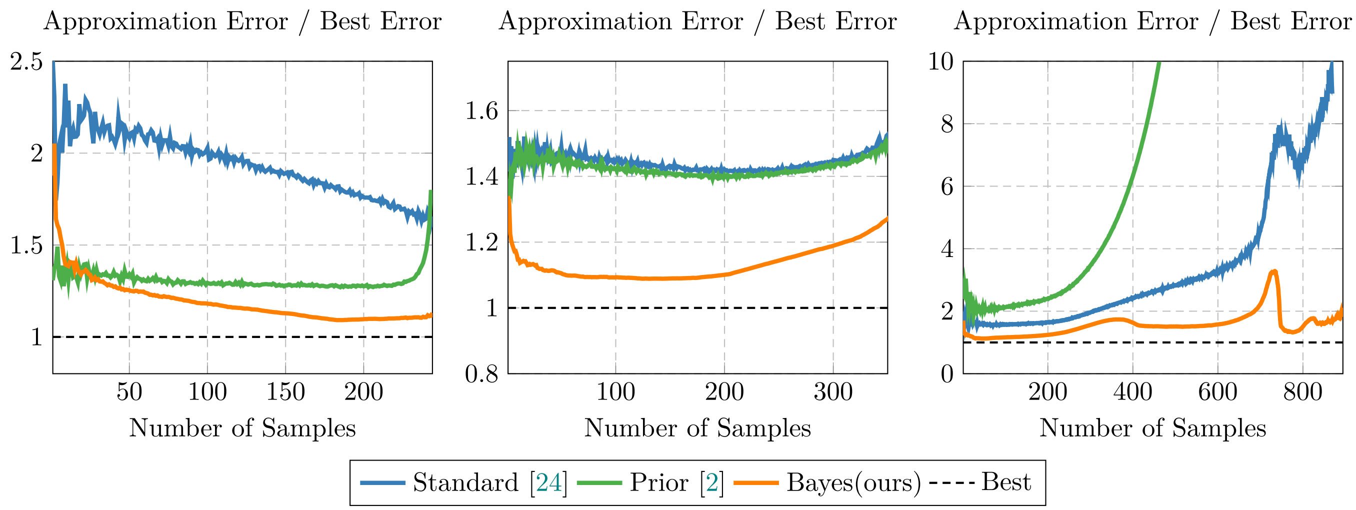

Figure 1: Low Rank Approximation for the Inverse Differential Operator given in Equation (19) *(Left)*, Differential Operator Matrix `Poisson2D` [11] *(Center)*, and Differential Operator Matrix `DK01R` [11] *(Right)*. The experiment on the left is from [2, Figure 2].

Claim (ii):

$$\left\|\mathbf{A}-\widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^{*}\mathbf{A}\right\|_{\mathrm{F}} \leq \left\|\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^{*}\right)\mathbf{A}\right\|_{\mathrm{F}} \sqrt{1-\mathcal{O}\left(\frac{1}{rn}\right)\left(1-\sqrt{\frac{1}{c}\log\left(\frac{2}{\delta}\right)}\right)} \tag{17}$$

Proof. The proof is deferred to Appendix A.1. ■

Theorem 8 (Sufficient Singular Value Gap)**.** *Let* $\gamma \triangleq \frac{\sigma_k}{\sigma_{k+1}}$ *and* $\eta^2 \leq \beta\sigma_{k+1}^2$. *Bayesian sampling described in Equation* (9) *decrease the low-rank approximation error faster than a normal vector sample when*

$$\left(\sigma_k-\sigma_{k+1}\right)^2 \leq 2\left\|\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^{*}\right)\mathbf{A}\right\|_{\mathrm{F}}^2+\frac{2rn\left(\sigma_k^4+\sigma_k^2\right)}{\left\|\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^{*}\right)\mathbf{A}\right\|_2^2} \tag{18}$$

Proof. The proof is deferred to **??**

5 Numerical Experiments

In this section we will test various Synthetic Matrices, Differential Operators, Images, and real-world applications, with our framework compared to fixed covariance matrices. In our first experiment we attempt to learn the discretized 250×250 matrix of the inverse of the following differential operator:

$$\mathcal{L}u = \frac{\partial^2 u}{\partial x^2} - 100\sin\left(5\pi x\right)u, \qquad x \in [0,1] \tag{19}$$

Learning the inverse operator of a PDE is equivalent to learning the Green's Function of a PDE. This has been theoretically proven for certain classes of PDEs (Linear Parabolic [5, 3]) as the inverse Differential operator is compact and there are nice theoretical properties, such as data efficiency.

In Figure 1*(Right)*, note if the Covariance Matrix has eigenvectors orthogonal to the left singular vectors of $\mathbf{A}$, then the randomized SVD will not perform well. Furthermore, in Figure 1*(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, [11]. In

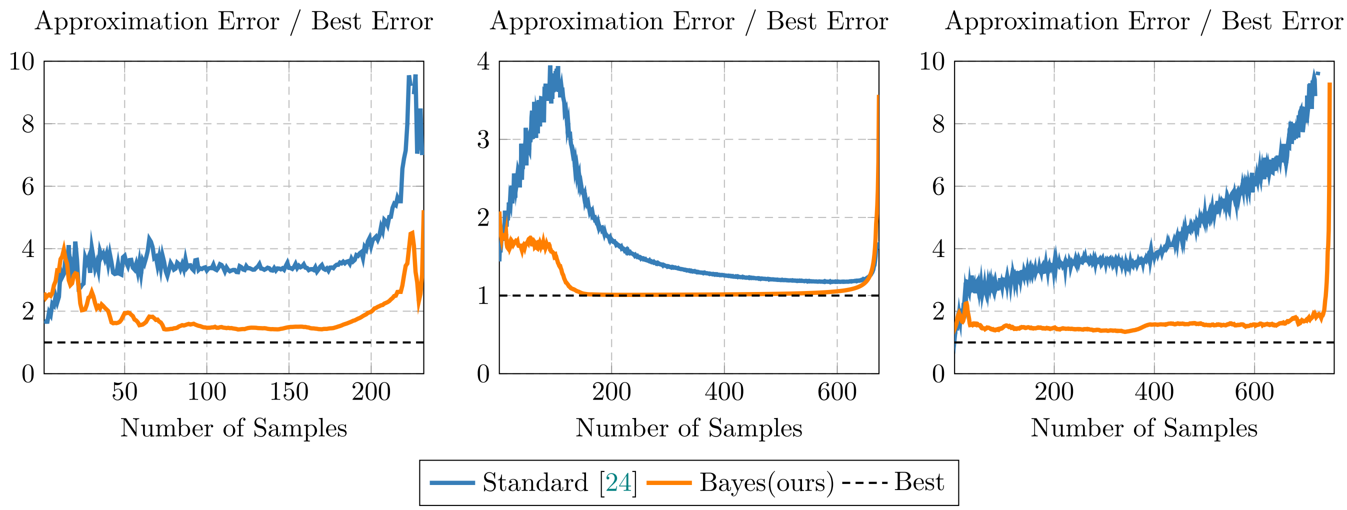

Figure 2: Low Rank Approximation for a matrix for a Computational Fluid Dynamics Problem, `saylr1` *(Left)*from [11]. Subsequent 2D/3D Problem `fs-680-2` *(Center)*from [11]. Astrophysics 2D/3D Problem `msfe` *(Right)*from [11].

Figure 2 *(Left)*, we choose a fluid dynamics problem due to its relevance in low-rank approximation [8]. The synthetic matrix is developed in the following scheme:

$$\mathbf{A} = \sum_{i=1}^{\rho} \frac{100 i^{\ell}}{n} \mathbf{U}_{(:,i)} \mathbf{V}^{*}_{(i,:)}, \quad \mathbf{U} \in \mathbb{O}_{m,k}, \mathbf{V} \in \mathbb{O}_{n,k} \tag{20}$$

We find our theoretical bounds stronger than Theorem 5.

6 Conclusions

We have theoretically and empirically analyzed a novel Covariance Update to iteratively construct the sampling matrix, $\mathbf{\Omega}$ in the Randomized SVD algorithm. Our covariance update for generating sampling vectors and functions can find use various PDE learning applications, [4, 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

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

[1] Ainesh Bakshi, Kenneth L. Clarkson, and David P. Woodruff. Low-rank approximation with $1/\varepsilon^3$ matrix-vector products. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2022, page 11301143, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392648. doi: 10.1145/3519935.3519988. URL https://doi.org/10.1145/3519935.3519988.

[2] Nicolas Boullé and Alex Townsend. A generalization of the randomized singular value decomposition. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=hgKtwSb4S2.

[3] Nicolas Boullé and Alex Townsend. Learning elliptic partial differential equations with randomized linear algebra. *Foundations of Computational Mathematics*, 23(2):709–739, Apr 2023. ISSN 1615-3383. doi: 10.1007/s10208-022-09556-w. URL https://doi.org/10.1007/s10208-022-09556-w.

[4] 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. ISSN 2045-2322. doi: 10.1038/s41598-022-08745-5. URL https://doi.org/10.1038/s41598-022-08745-5.

[5] 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.

[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. doi: 10.1137/120874540. URL https://doi.org/10.1137/120874540.

[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] T. Tony Cai and Anru Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60 – 89, 2018. doi: 10.1214/17-AOS1541. URL https://doi.org/10.1214/17-AOS1541.

[10] Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. ISSN 00361429. URL http://www.jstor.org/stable/2949580.

[11] Timothy A. Davis and Yifan Hu. The university of florida sparse matrix collection. *ACM Trans. Math. Softw.*, 38(1), dec 2011. ISSN 0098-3500. doi: 10.1145/2049662.2049663. URL https://doi.org/10.1145/2049662.2049663.

[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] 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.

[14] Yijun Dong, Per-Gunnar Martinsson, and Yuji Nakatsukasa. Efficient bounds and estimates for canonical angles in randomized subspace approximations. *arXiv preprint arXiv:2211.04676*, 2022.

[15] Petros Drineas, Ilse C. F. Ipsen, Eugenia-Maria Kontopoulou, and Malik Magdon-Ismail. Structural convergence results for approximation of dominant subspaces from block krylov spaces. *SIAM Journal on Matrix Analysis and Applications*, 39(2):567–586, 2018. doi: 10.1137/16M1091745. URL https://doi.org/10.1137/16M1091745.

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

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

[18] Ahmed Farahat, Ali Ghodsi, and Mohamed Kamel. A novel greedy algorithm for Nyström approximation. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 269–277, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/farahat11a.html.

[19] 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.

[20] Alex Gittens and Michael Mahoney. Revisiting the nystrom method for improved large-scale machine learning. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 567–575, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/gittens13.html.

[21] Alex Gittens and Michael W. Mahoney. Revisiting the nyström method for improved large-scale machine learning. *J. Mach. Learn. Res.*, 17(1):39774041, jan 2016. ISSN 1532-4435.

[22] Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

[23] Ming Gu. Subspace iteration randomization and singular value problems. *SIAM Journal on Scientific Computing*, 37(3):A1139–A1173, 2015.

[24] 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. doi: 10.1137/090771806. URL https://doi.org/10.1137/090771806.

[25] Nathan Halko, Per-Gunnar Martinsson, Yoel Shkolnisky, and Mark Tygert. An algorithm for the principal component analysis of large data sets. *SIAM Journal on Scientific Computing*, 33(5):2580–2594, 2011. doi: 10.1137/100804139. URL https://doi.org/10.1137/100804139.

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

[27] Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/729c68884bd359ade15d5f163166738a-Paper.pdf.

[28] Philipp Hennig and Christian J Schuler. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13(6), 2012.

[29] Piotr Indyk, Ali Vakilian, and Yang Yuan. Learning-based low-rank approximations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/1625abb8e458a79765c62009235e9d5b-Paper.pdf.

[30] Piotr Indyk, Tal Wagner, and David Woodruff. Few-shot data-driven algorithms for low rank approximation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=6dUJPrPPUau.

[31] Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling methods for the nyström method. *The Journal of Machine Learning Research*, 13(1):981–1006, 2012.

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

[33] Yuetian Luo, Rungang Han, and Anru R Zhang. A schatten-q low-rank matrix perturbation analysis via perturbation projection error bound. *Linear Algebra and its Applications*, 630:225–240, 2021.

[34] Per-Gunnar Martinsson and Joel A. Tropp. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica*, 29:403572, 2020. doi: 10.1017/S0962492920000021.

[35] Pedro G. Massey, Demetrio Stojanoff, and Sebastian Zarate. Majorization bounds for ritz values of self-adjoint matrices. *SIAM Journal on Matrix Analysis and Applications*, 41(2):554–572, 2020. doi: 10.1137/19M1263996. URL https://doi.org/10.1137/19M1263996.

[36] L. Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 01 1960. ISSN 0033-5606. doi: 10.1093/qmath/11.1.50. URL https://doi.org/10.1093/qmath/11.1.50.

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

[38] Cameron Musco and Christopher Musco. Randomized block krylov methods for stronger and faster approximate singular value decomposition. 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. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/1efa39bcaec6f3900149160693694536-Paper.pdf.

[39] Yuji Nakatsukasa. The tan θ theorem with relaxed conditions. *Linear Algebra and its Applications*, 436(5):1528–1534, 2012. ISSN 0024-3795. doi: https://doi.org/10.1016/j.laa.2011.08.038. URL https://www.sciencedirect.com/science/article/pii/S0024379511006227.

[40] Sean O'Rourke, Van Vu, and Ke Wang. Random perturbation of low rank matrices: Improving classical bounds. *Linear Algebra and its Applications*, 540:26–59, 2018. doi: 10.1016/j.laa.2017.11.014. URL https://doi.org/10.1016/j.laa.2017.11.014.

[41] Taejun Park and Yuji Nakatsukasa. A fast randomized algorithm for computing an approximate null space. *BIT Numerical Mathematics*, 63(2):36, May 2023. ISSN 1572-9125. doi: 10.1007/s10543-023-00979-7. URL https://doi.org/10.1007/s10543-023-00979-7.

[42] 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. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/d395771085aab05244a4fb8fd91bf4ee-Paper.pdf.

[43] Arvind K Saibaba. Randomized subspace iteration: Analysis of canonical angles and unitarily invariant norms. *SIAM Journal on Matrix Analysis and Applications*, 40(1):23–48, 2019.

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

[45] 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. URL https://proceedings.mlr.press/v151/tzeng22a.html.

[46] Roman Vershynin. High-dimensional probability. *University of California, Irvine*, 2020.

[47] 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. doi: 10.1109/CVPR46437.2021.00025. URL http://dx.doi.org/10.1109/CVPR46437.2021.00025.

[48] Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, Mar 1972. ISSN 1572-9125. doi: 10.1007/BF01932678. URL https://doi.org/10.1007/BF01932678.

[49] David P Woodruff et al. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

[50] Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the davis-kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2023/07/23/ 2015. ISSN 00063444. URL http://www.jstor.org/stable/43908537. Full publication date: JUNE 2015.

A Deferred Proofs of Main Results

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

A.1 Proof of Lemma 7

We have $\mathbf{Q}$ is an orthonormal basis of $\mathbf{Y} \in \mathbb{C}^{m\times k}$ where $\mathbf{Y} \triangleq \mathbf{A\Omega}$ for $\mathbf{\Omega} \in \mathbb{C}^{n\times k}$ is an arbitrary test matrix. Then let us denote $\widehat{\mathbf{Q}} \triangleq \mathrm{orth}([\mathbf{Y} \quad \mathbf{Av}])$ where $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\widetilde{\mathbf{Q}} \triangleq \mathrm{orth}([\mathbf{Y} \quad \mathbf{A}\widetilde{\mathbf{v}}])$ where $\widetilde{\mathbf{v}}$ is the kth right singular vector of $\mathbf{QQ}^*\mathbf{A}$. Note $\widehat{\mathbf{q}} \in \mathrm{Span}(\mathbf{I} - \mathbf{QQ}^*\mathbf{A})$.

$$\left\|\mathbf{A} - \widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^*\mathbf{A}\right\|_{\mathrm{F}}^2 = \|\mathbf{A} - \mathbf{QQ}^*\mathbf{A} - \widehat{\mathbf{q}}\widehat{\mathbf{q}}^*\mathbf{A}\| \tag{21}$$

$$\begin{aligned} &= \mathrm{Tr}\Big(\mathbf{A}^*\mathbf{A} - \mathbf{A}^*\mathbf{QQ}^*\mathbf{A} - \mathbf{A}^*\widehat{\mathbf{q}}\widehat{\mathbf{q}}^*\mathbf{A} - \mathbf{A}^*\mathbf{QQ}^*\mathbf{A} + \mathbf{A}^*\mathbf{QQ}^*\mathbf{QQ}^*\mathbf{A} \\ &\quad + \mathbf{A}^*\mathbf{QQ}^*\widehat{\mathbf{q}}\widehat{\mathbf{q}}^*\mathbf{A} - \mathbf{A}^*\widehat{\mathbf{q}}\widehat{\mathbf{q}}^*\mathbf{A} + \mathbf{A}^*\widehat{\mathbf{q}}\widehat{\mathbf{q}}^*\mathbf{QQ}^*\mathbf{A} + \mathbf{A}^*\widehat{\mathbf{q}}\widehat{\mathbf{q}}^*\widehat{\mathbf{q}}\widehat{\mathbf{q}}^*\mathbf{A}\Big) \end{aligned} \tag{22}$$

$$= \mathrm{Tr}\left(\mathbf{A}^*\mathbf{A} - \mathbf{A}^*\mathbf{QQ}^*\mathbf{A} - \mathbf{A}^*\widehat{\mathbf{q}}\widehat{\mathbf{q}}^*\mathbf{A}\right) \tag{23}$$

$$= \|\mathbf{A} - \mathbf{QQ}^*\mathbf{A}\|_{\mathrm{F}}^2 - \underbrace{\|\widehat{\mathbf{q}}^*\mathbf{A}\|_{\mathrm{F}}^2}_{c_2} \tag{24}$$

Similarly from Equation (24), we have

$$\left\|\mathbf{A} - \widetilde{\mathbf{Q}}\widetilde{\mathbf{Q}}^*\mathbf{A}\right\|_{\mathrm{F}}^2 = \|\mathbf{A}\|_{\mathrm{F}}^2 - \|\mathbf{Q}^*\mathbf{A}\|_{\mathrm{F}}^2 - \underbrace{\|\widetilde{\mathbf{q}}^*\mathbf{A}\|_{\mathrm{F}}^2}_{c_1} \tag{25}$$

Since we have $\widehat{\mathbf{q}}, \widetilde{\mathbf{q}} \in \mathrm{Span}(\mathbf{I} - \mathbf{QQ}^*\mathbf{A})$, then from our formulation in Equations (24) and (25), we want to our sampled vector to be in the dominant singular space of the span of the singular vectors of $\mathbf{I} - \mathbf{QQ}^*\mathbf{A}$.

$$\mathbf{q}_{\mathrm{OPT}} = \operatorname*{arg\,max}_{\mathbf{q}\in\mathrm{Span}(\mathbf{I}-\mathbf{QQ}^*):\|\mathbf{q}\|=1} \|\mathbf{q}^*\mathbf{A}\| \tag{26}$$

$$= \operatorname*{arg\,max}_{\mathbf{v}\in\mathbb{R}^n} \frac{\mathbf{A}^*\left(\mathbf{I} - \mathbf{QQ}^*\right)\left(\mathbf{Av}\right)}{\|\left(\mathbf{I} - \mathbf{QQ}^*\right)\left(\mathbf{Av}\right)\|} \tag{27}$$

Let us note for any column $\mathbf{q} \in \mathbf{Q}$, we have

$$\left(\left(\mathbf{I} - \mathbf{QQ}^*\right)\mathbf{v}\right)^*\mathbf{q} = \left(\mathbf{v}^* - \mathbf{v}^*\mathbf{QQ}^*\right)\mathbf{q} = \mathbf{v}^*\mathbf{q} - \mathbf{v}^*\mathbf{q} = 0 \tag{28}$$

Claim (i):
Let us define $\widetilde{\mathbf{U}}\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{V}}^* = \mathrm{SVD}(\mathbf{QQ}^*\mathbf{A})$.

$$c_1 \triangleq \|\widetilde{\mathbf{q}}^*\mathbf{A}\|_{\mathrm{F}}^2 \overset{(28)}{=} \frac{\left\|\left(\left(\mathbf{I} - \mathbf{QQ}^*\right)\mathbf{A}\widetilde{\mathbf{v}}\right)^*\mathbf{A}\right\|_2^2}{\|\left(\mathbf{I} - \mathbf{QQ}^*\right)\mathbf{A}\widetilde{\mathbf{v}}\|_2^2} \overset{(i)}{\geq} \frac{\|\widetilde{\mathbf{v}}^*\mathbf{A}^*\left(\mathbf{I} - \mathbf{QQ}^*\right)\mathbf{A}\|_2^2}{\|\mathbf{A}\widetilde{\mathbf{v}}_k\|_2^2} \tag{29}$$

$$\overset{\text{lem. 12}}{\geq} \frac{\|\widetilde{\mathbf{v}}^*\mathbf{A}^*\mathbf{A} - \widetilde{\mathbf{u}}_k^*\mathbf{A}\|^2}{\sigma_k^2} \overset{\text{lem. 12}}{\geq} \frac{\|\widetilde{\mathbf{v}}_k^*\mathbf{A}^*\mathbf{A}\|^2 + \|\widetilde{\mathbf{u}}_k^*\mathbf{A}\|^2 - 2\|\widetilde{\mathbf{v}}_k^*\mathbf{A}^*\mathbf{A}\|\,\|\widetilde{\mathbf{u}}_k\mathbf{A}\|}{\sigma_k^2} \tag{30}$$

$$\overset{\text{lem. 12,lem. 11}}{\geq} \frac{\sigma_k^4\left(1 - 2\sin^2\Theta\left(\mathbf{v}_k, \widetilde{\mathbf{v}}_k\right)\right) + \sigma_k^2\left(1 - 2\sin^2\Theta\left(\mathbf{u}_k, \widetilde{\mathbf{u}}_k\right)\right) - 2\sigma_k^3}{\sigma_k^2} \tag{31}$$

$$\overset{\text{thm. 2}}{\geq} \sigma_k^2 - 2\sigma_k + 1 - 2\sigma_k^2\frac{\|\left(\mathbf{I} - \mathbf{QQ}^*\right)\mathbf{A}\widetilde{\mathbf{v}}_k\|_2^2}{\sigma_k^2} - 2\frac{\|\left(\mathbf{I} - \mathbf{QQ}^*\right)\mathbf{A}^*\widetilde{\mathbf{u}}_k\|_2^2}{\sigma_k^2} \tag{32}$$

$$\overset{(ii)}{\geq} \left(\sigma_k - 1\right)^2 - 2\left(1 + \frac{1}{\sigma_k^2}\right)\left(\sigma_{k+1}^2 + \eta^2\right) \tag{33}$$

(i) follows from noting that Π is a contraction. (ii) follows from Lemma 10 which has been seen in [15]. We thus have

$$\left\|\mathbf{A}-\widetilde{\mathbf{Q}}\widetilde{\mathbf{Q}}^*\mathbf{A}\right\|_{\mathrm{F}}^2 \overset{(33)}{\leq} \left\|\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}\right\|_{\mathrm{F}}^2 - \left(\sigma_k-1\right)^2 + 2\left(1+\frac{1}{\sigma_k^2}\right)\left(\sigma_{k+1}^2+\eta^2\right) \tag{34}$$

$$= \left\|\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}\right\|_{\mathrm{F}}^2\left(1-\frac{\left(\sigma_k-1\right)^2-2\left(1+\frac{1}{\sigma_k^2}\right)\left(\sigma_{k+1}^2+\eta^2\right)}{\left\|\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}\right\|_{\mathrm{F}}^2}\right) \tag{35}$$

Taking the square root of both sides completes our proof for Claim (i). ■

Claim (ii):

We will lower bound c_2.

$$c_2 \triangleq \left\|\widehat{\mathbf{q}}^*\mathbf{A}\right\|_{\mathrm{F}}^2 \overset{(28)}{=} \frac{\left\|\left(\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}\widehat{\mathbf{v}}\right)^*\mathbf{A}\right\|_2^2}{\left\|\left(\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}\widehat{\mathbf{v}}\right)\right\|_2^2} \tag{36}$$

$$= \frac{\left\|\widehat{\mathbf{v}}^*\mathbf{A}^*\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}\right\|_2^2}{\left\|\left(\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}\widehat{\mathbf{v}}\right)\right\|_2^2} \geq \frac{\left\|\widehat{\mathbf{v}}^*\mathbf{A}^*\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}\right\|_2^2}{\left\|\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}\right\|_2^2\left\|\widehat{\mathbf{v}}\right\|_2^2} \tag{37}$$

For shorthand, let $\widetilde{\mathbf{A}} \triangleq \mathbf{A}^*(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*)\mathbf{A}$. Note, we have $\|(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*)\mathbf{A}\|$ is given. From Lemma 14, we have

$$\mathbb{P}\left\{\left|\mathbf{x}^*\widetilde{\mathbf{A}}^*\widetilde{\mathbf{A}}\mathbf{x}-\mathbb{E}\left[\mathbf{x}^*\widetilde{\mathbf{A}}^*\widetilde{\mathbf{A}}\mathbf{x}\right]\right|\geq t\right\} \leq 2\exp\left(-c\frac{t^2}{\left\|\widetilde{\mathbf{A}}^*\widetilde{\mathbf{A}}\right\|_{\mathrm{F}}^2}\right) \tag{38}$$

It then follows with probability $1-\delta$

$$\left\|\widetilde{\mathbf{A}}\mathbf{x}\right\|^2 \geq \mathbb{E}\left\|\widetilde{\mathbf{A}}\mathbf{x}\right\|^2 - \left\|\widetilde{\mathbf{A}}\right\|_{\mathrm{F}}^2\sqrt{\frac{1}{c}\log\left(\frac{2}{\delta}\right)} \tag{39}$$

We further simplify this with the fact

$$\mathbb{E}\left\|\widetilde{\mathbf{A}}\mathbf{x}\right\|^2 = \mathbb{E}\mathbf{x}^*\widetilde{\mathbf{A}}^*\widetilde{\mathbf{A}}\mathbf{x} = \mathbb{E}\operatorname{Tr}\left(\mathbf{x}^*\widetilde{\mathbf{A}}^*\widetilde{\mathbf{A}}\mathbf{x}\right) = \mathbb{E}\operatorname{Tr}\left(\mathbf{x}\mathbf{x}^*\widetilde{\mathbf{A}}^*\widetilde{\mathbf{A}}\right) = \operatorname{Tr}\left(\mathbb{E}\left[\mathbf{x}\mathbf{x}^*\widetilde{\mathbf{A}}^*\widetilde{\mathbf{A}}\right]\right) \tag{40}$$

$$= \operatorname{Tr}\left(\mathbb{E}\left[\mathbf{x}\mathbf{x}^*\right]\widetilde{\mathbf{A}}^*\widetilde{\mathbf{A}}\right) = \operatorname{Tr}\left(\widetilde{\mathbf{A}}^*\widetilde{\mathbf{A}}\right) = \left\|\widetilde{\mathbf{A}}\right\|_{\mathrm{F}}^2 \tag{41}$$

Let us also note the following fact for a rank-k orthonormal matrix, $\mathbf{Q}$, and rank-r matrix $\mathbf{A}$.

$$\left\|\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}\right\|_{\mathrm{F}}^2 = \operatorname{Tr}\left(\mathbf{A}\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}^*\right) \tag{42}$$

$$= \operatorname{Tr}\left(\mathbf{A}\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}^*\right) \tag{43}$$

$$\overset{\zeta_1}{\leq} \sqrt{r}\left\|\mathbf{A}\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}^*\right\|_{\mathrm{F}} \tag{44}$$

where (ζ_1) follows from the relation of the $\|\cdot\|_1$ and $\|\cdot\|_2$ norm. Now we can calculate the expectation. We then note $\mathbb{E}\|\mathbf{v}\|_2^2$ is the expectation of n-degree of freedom Chi-squared variable and thus is equal to n, then

$$\mathbb{E}c_2 = \left[\mathbb{E}\frac{\left\|\widehat{\mathbf{v}}^*\mathbf{A}^*\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}\right\|_2^2}{\left\|\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}\widehat{\mathbf{v}}\right\|_2^2}\right] \overset{(\zeta_1)}{\geq} \left(\mathbb{E}\left\|\widehat{\mathbf{v}}^*\mathbf{A}^*\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}\right\|_2\right)^2\mathbb{E}\left[\left\|\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}\widehat{\mathbf{v}}\right\|_2^2\right]^{-1} \tag{45}$$

$$\overset{(\zeta_2)}{\geq} \left(\frac{16}{75\sqrt{5}}\right)^2\left\|\mathbf{A}^*\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}\right\|_{\mathrm{F}}^2\left\|\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}\right\|_{\mathrm{F}}^{-2} \tag{46}$$

$$\overset{(42)}{\geq} \left(\frac{16}{75\sqrt{5}}\right)^2\frac{\left\|\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}\right\|_{\mathrm{F}}^4}{r\left\|\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}\right\|_{\mathrm{F}}^2} \gtrsim \left(\frac{1}{r}\right)\left\|\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*\right)\mathbf{A}\right\|_{\mathrm{F}}^2 \tag{47}$$

(ζ_1) follows from the Reverse Hölder Inequality. (ζ_2) follows from an application of Jensen's Inequality. Combining Equation (47) and Equation (41) we thus have with probability $1-\delta$

$$c_2 \geq \|(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*)\,\mathbf{A}\|_{\mathrm{F}}^2 \left(\Omega\left(\frac{1}{r}\right)\left(1-\sqrt{\frac{1}{c_3}\log\left(\frac{2}{\delta}\right)}\right)\right) \tag{48}$$

Furthermore, in expectation we have

$$\mathbb{E}c_2 \geq \Omega\left(\frac{1}{r}\right)\|(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*)\,\mathbf{A}\|_{\mathrm{F}}^2 \tag{49}$$

This completes our proof of Claim (ii). ■

A.2 Proof of Theorem 8

Proof. We will find γ s.t.

$$\left(\frac{16}{75\sqrt{5}}\right)^2 \|(\mathbf{I}-\mathbf{Q}\mathbf{Q}^*)\,\mathbf{A}\|_{\mathrm{F}}^2 \leq (\sigma_k-1)^2 + 2\left(1+\frac{1}{\sigma_k^2}\right)\left(\sigma_{k+1}^2+\eta^2\right) \tag{50}$$

■

B Singular Subspace Perturbation Lemmas

Lemma 9. *Let* $\mathbf{v}$ *and* $\widetilde{\mathbf{v}}$ *be vectors s.t.* $\|\mathbf{v}\| = \|\widetilde{\mathbf{v}}\| = 1$ *and* $\mathbf{v}^*\widetilde{\mathbf{v}} \geq 0$. *Then,*

$$\|\mathbf{v}-\widetilde{\mathbf{v}}\| \leq \sqrt{2}\sin\Theta\left(\mathbf{v},\widetilde{\mathbf{v}}\right) \tag{51}$$

Proof.

$$\sin^2\Theta\left(\mathbf{v},\widetilde{\mathbf{v}}\right) = 1-\left(\mathbf{v}^*\widetilde{\mathbf{v}}\right)^2 \overset{(a)}{\geq} 1-\mathbf{v}^*\widetilde{\mathbf{v}} = 1+\frac{1}{2}\|\mathbf{v}-\widetilde{\mathbf{v}}\|^2 - \frac{1}{2}\|\mathbf{v}\|^2 - \frac{1}{2}\|\widetilde{\mathbf{v}}\|^2 = \frac{1}{2}\|\mathbf{v}-\widetilde{\mathbf{v}}\|^2 \tag{52}$$

(a) follows from $0 \leq \mathbf{v}^*\widetilde{\mathbf{v}} \leq 1$, therefore $\mathbf{v}^*\widetilde{\mathbf{v}} \geq \left(\mathbf{v}^*\widetilde{\mathbf{v}}\right)^2$.
Plugging this back into the first inequality and taking the square root gives us the desired result. ■

Lemma 10. *Let* $\widetilde{\mathbf{u}}_k$ *represent the* k*th left singular vector of a low-rank approximation* $\widetilde{\mathbf{A}}$ *of* $\mathbf{A}$. *Then,*

$$\|\mathbf{A}^*\widetilde{\mathbf{u}}_k\| \leq \sigma_k(\mathbf{A}) \tag{53}$$

Proof. We provide the proof for completeness. From the minimax Courant-Fischer Theorem [22] we have,

$$\sigma_k(\mathbf{A}) = \max_{\dim(S)=k}\ \min_{\mathbf{x}\in S\setminus\{\mathbf{0}\}} \frac{\|\mathbf{A}^*\mathbf{x}\|}{\|\mathbf{x}\|} \tag{54}$$

$$\overset{\zeta_1}{=} \min_{\mathbf{x}\in\mathrm{Span}(\{\mathbf{u}_1,\ldots,\mathbf{u}_k\})\setminus\{\mathbf{0}\}} \frac{\|\mathbf{A}^*\mathbf{x}\|}{\|\mathbf{x}\|} \tag{55}$$

$$\geq \min_{\mathbf{x}\in\mathrm{Span}\left(\left\{\widetilde{\mathbf{u}}_1,\ldots,\widetilde{\mathbf{u}}_k\right\}\right)\setminus\{\mathbf{0}\}} \frac{\|\mathbf{A}^*\mathbf{x}\|}{\|\mathbf{x}\|} \tag{56}$$

$$= \|\mathbf{A}^*\widetilde{\mathbf{u}}_k\| \tag{57}$$

where (ζ_1) follows from the optimality of the SVD given by the Eckart-Young-Mirsky Theorem [16, 37].
For $k=1$, we have

$$\|\mathbf{A}^*\widetilde{\mathbf{u}}_1\| = \|\mathbf{V}\mathbf{\Sigma}\mathbf{U}^*\mathbf{u}_1\| = \|\mathbf{\Sigma}\mathbf{U}^*\mathbf{u}_1\| \leq \sigma_1\cos\left(vu_1,\mathbf{u}_1\right) + \sigma_2\sin\Theta\left(\mathbf{u}_1,\widetilde{\mathbf{u}}_1\right) \tag{58}$$

■

Lemma 11. *Let $\widetilde{\mathbf{v}}_k$ be the kth right singular vector of an approximation $\widetilde{\mathbf{A}}$ of $\mathbf{A}$. Then,*

$$\|\mathbf{A}^*\mathbf{A}\widetilde{\mathbf{v}}_k\| \geq \sigma_k^2\left(1-\sqrt{2}\sin\Theta\left(\mathbf{v}_k,\widetilde{\mathbf{v}}_k\right)\right) \tag{59}$$

Proof.

$$\|\mathbf{A}^*\mathbf{A}\|_\mathrm{F}^2 - \|\mathbf{A}^*\mathbf{A}\widetilde{\mathbf{v}}_k\|_\mathrm{F}^2 = \|\mathbf{A}^*\mathbf{A} - \widetilde{\mathbf{v}}_k\widetilde{\mathbf{v}}_k^*\mathbf{A}^*\mathbf{A}\|_\mathrm{F}^2 \overset{\zeta_1}{=} \left\|\mathbf{A}_{(k)}^*\mathbf{A}_{(k)} - \widetilde{\mathbf{v}}_k\widetilde{\mathbf{v}}_k^*\mathbf{A}^*\mathbf{A}\right\|_\mathrm{F}^2 + \left\|\mathbf{A}_{\perp,k}^*\mathbf{A}_{\perp,k}\right\|_\mathrm{F}^2 \tag{60}$$

$$\leq \left\|\mathbf{A}_{(k)}^*\mathbf{A}_{(k)} - \sigma_k^2\widetilde{\mathbf{v}}_k\widetilde{\mathbf{v}}_k^*\right\|_\mathrm{F}^2 + \left\|\mathbf{A}_{\perp,k}^*\mathbf{A}_{\perp,k}\right\|_\mathrm{F}^2 \tag{61}$$

$$= \sigma_k^4\left\|\Pi_{\mathbf{v}_k} - \Pi_{\widetilde{\mathbf{v}}_k}\right\|_\mathrm{F}^2 + \left\|\mathbf{A}_{\perp,k}^*\mathbf{A}_{\perp,k}\right\|_\mathrm{F}^2 \tag{62}$$

$$\overset{\text{lem. 9}}{=} 2\sigma_k^4\sin^2\Theta\left(\mathbf{v}_k,\widetilde{\mathbf{v}}_k\right) + \left\|\mathbf{A}_{\perp,k}^*\mathbf{A}_{\perp,k}\right\|_\mathrm{F}^2 \tag{63}$$

(ζ_1) follows from the Matrix Pythagoras theorem [49]. From rearranging the inequalities and noting $\|\mathbf{A}^*\mathbf{A}\|_\mathrm{F}^2 - \left\|\mathbf{A}_{\perp,k}^*\mathbf{A}_{\perp,k}\right\|_\mathrm{F}^2 = \sigma_k^4$, we then obtain

$$\|\mathbf{A}^*\mathbf{A}\widetilde{\mathbf{v}}_k\|_\mathrm{F}^2 \geq \sigma_k^4\left(1-2\sin^2\Theta\left(\mathbf{v}_k,\widetilde{\mathbf{v}}_k\right)\right) \tag{64}$$

Taking the square root and reverse triangle inequality gives us the desired result. ■

Lemma 12. *Let $\widetilde{\mathbf{u}}_k$ be the kth left singular vector of approximation $\widetilde{\mathbf{A}}$ of $\mathbf{A}$, then we have*

$$\|\mathbf{A}^*\widetilde{\mathbf{u}}_k\| \geq \sigma_k\left(1-\sqrt{2}\sin\Theta\left(\mathbf{u}_k,\widetilde{\mathbf{u}}_k\right)\right) \tag{65}$$

Proof.

$$\|\mathbf{A}\|_\mathrm{F}^2 - \|\widetilde{\mathbf{u}}_k^*\mathbf{A}\|_\mathrm{F}^2 = \|\mathbf{A} - \widetilde{\mathbf{u}}_k\widetilde{\mathbf{u}}_k^*\mathbf{A}\|_\mathrm{F}^2 \tag{66}$$

$$\leq \left\|\mathbf{A} - \widetilde{\mathbf{u}}_k\widetilde{\mathbf{u}}_k\mathbf{A}_{(k)}\right\|_\mathrm{F}^2 \tag{67}$$

$$= \left\|\mathbf{A}_{(k)} - \widetilde{\mathbf{u}}_k\widetilde{\mathbf{u}}_k^*\mathbf{A}_{(k)}\right\|_\mathrm{F}^2 + \left\|\mathbf{A}_{\perp,k}\right\|_\mathrm{F}^2 \tag{68}$$

$$\leq \left\|\mathbf{A}_{(k)} - \sigma_k\widetilde{\mathbf{u}}_k\mathbf{v}_k^*\right\|_\mathrm{F}^2 + \left\|\mathbf{A}_{\perp,k}\right\|_\mathrm{F}^2 \tag{69}$$

$$\leq 2\sigma_k^2\sin^2\Theta\left(\mathbf{u}_k,\widetilde{\mathbf{u}}_k\right) + \left\|\mathbf{A}_{\perp,k}\right\|_\mathrm{F}^2 \tag{70}$$

Rearranging we obtain the desired inequality. ■

C Concentration Inequalities

Lemma 13 ([46], Exercise 6.3.5)**.** *Let $\mathbf{B} \in \mathbb{C}^{m\times n}$ and $\mathbf{x} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$, then we have*

$$\mathbb{P}\left\{\|\mathbf{B}\mathbf{x}\|_2 \geq CK\|\mathbf{B}\|_\mathrm{F} + t\right\} \leq \exp\left(-\frac{ct^2}{K\|\mathbf{B}\|^2}\right) \tag{71}$$

Lemma 14 (Hanson-Wright Inequality)**.** *Let $\mathbf{x} \in \mathbb{R}^n$ be a vector with sub-Gaussian random vector synmmetric about $\mathbf{0}$. Let $\mathbf{B}$ be a symmetric $n\times n$ matrix, then $\forall t \geq 0$,*

$$\mathbb{P}\left\{\left|\mathbf{x}^\top\mathbf{B}\mathbf{x} - \mathbb{E}\left[\mathbf{x}^\top\mathbf{B}\mathbf{x}\right]\right| \geq t\right\} \leq 2\exp\left(-c\min\left\{\frac{t^2}{K^4\|\mathbf{B}\|_\mathrm{F}^2}, \frac{t}{K^2\|\mathbf{B}\|}\right\}\right) \tag{72}$$

Lemma 15. *[36]. For any matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, then for any unitarily invariant norm $|||\cdot|||$, it follows*

$$|||\mathbf{ABC}||| \leq \min\left\{|||\mathbf{A}|||\,\|\mathbf{B}\|_2\,\|\mathbf{C}\|_2, \|\mathbf{A}\|_2\,|||\mathbf{B}|||\,\|\mathbf{C}\|_2, \|\mathbf{A}\|_2\,\|\mathbf{B}\|_2\,|||\mathbf{C}|||\right\} \tag{73}$$

Lemma 16. *Let* $\mathbf{A} \in \mathbb{C}^{m\times n}$ *and* $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$*, then we have*

$$\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2 \geq \frac{16}{75\sqrt{5}}\left\|\mathbf{A}\right\|_{\mathrm{F}} \tag{74}$$

Proof. For a $\theta \in (0,1)$, we have

$$\frac{\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2}{\theta\sqrt{\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2^2}} \overset{(\zeta_1)}{\geq} \mathbb{P}\left\{\left\|\mathbf{A}\mathbf{x}\right\|_2 \geq \theta\sqrt{\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2^2}\right\} = \mathbb{P}\left\{\left\|\mathbf{A}\mathbf{x}\right\|_2^2 \geq \theta\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2^2\right\} \tag{75}$$

$$\overset{(\zeta_2)}{\geq} \frac{(1-\theta^2)^2\left(\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2^2\right)^2}{\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2^4} \tag{76}$$

(ζ_1) follows from Markov's Inequality. (ζ_2) follows from the Paley-Zygmund Inequality. Rearranging the first term in Equation (75) with Equation (76), we have

$$\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2 \geq \theta(1-\theta^2)^2\frac{\left(\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2^2\right)^{5/2}}{\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2^4} \tag{77}$$

The term, $\theta(1-\theta^2)^2$ is maximized at $\theta = \frac{1}{\sqrt{5}}$, we then have

$$\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2 \geq \frac{16\left(\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2^2\right)^{5/2}}{25\sqrt{5}\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2^4} \tag{78}$$

Next we note $\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2^2 = \left\|\mathbf{A}\right\|_{\mathrm{F}}^2$. Furthermore, we have

$$\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2^4 = \mathbb{E}\left\|\mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{x}\right\|_2^2 = \mathrm{Tr}\left(\mathbf{A}^*\mathbf{A}\right)^2 + 2\,\mathrm{Tr}\left(\left(\mathbf{A}^*\mathbf{A}\right)^2\right) \leq \left\|\mathbf{A}\right\|_{\mathrm{F}}^4 + 2\left\|\mathbf{A}\right\|_{\mathrm{F}}^2\left\|\mathbf{A}\right\|_2^2 \leq 3\left\|\mathbf{A}\right\|_{\mathrm{F}}^4 \tag{79}$$

Then from using the result in Equation (79), we have

$$\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2 \geq \frac{16}{75\sqrt{5}}\left\|\mathbf{A}\right\|_{\mathrm{F}} \tag{80}$$

This concludes the proof. ■

D Additional Experiments

In this section we perform more experiments on learning the inverse operator for PDE matrices with State of the Art Matrix Experiments.

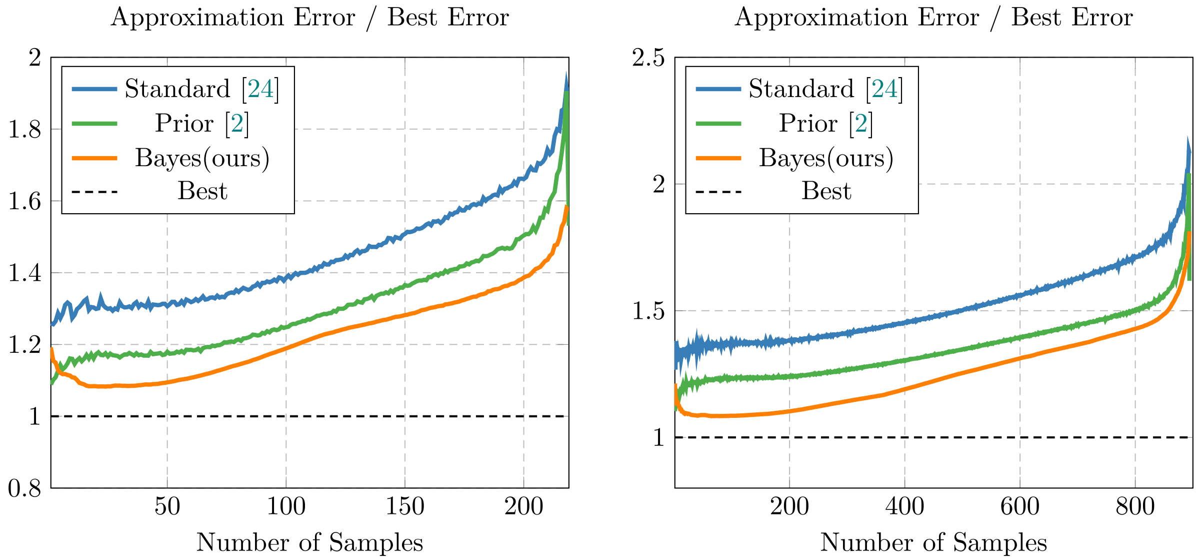

Figure 3: In *(Left)*, Matrix from TAMU Sparse Matrix Suite `pde 225`. In *(Right)*, Matrix from TAMU Sparse Matrix Suite `pde 900`. With the prior, we use the covariance matrix associated with the discrete Green's Function for the Laplacian as in Equation (19).

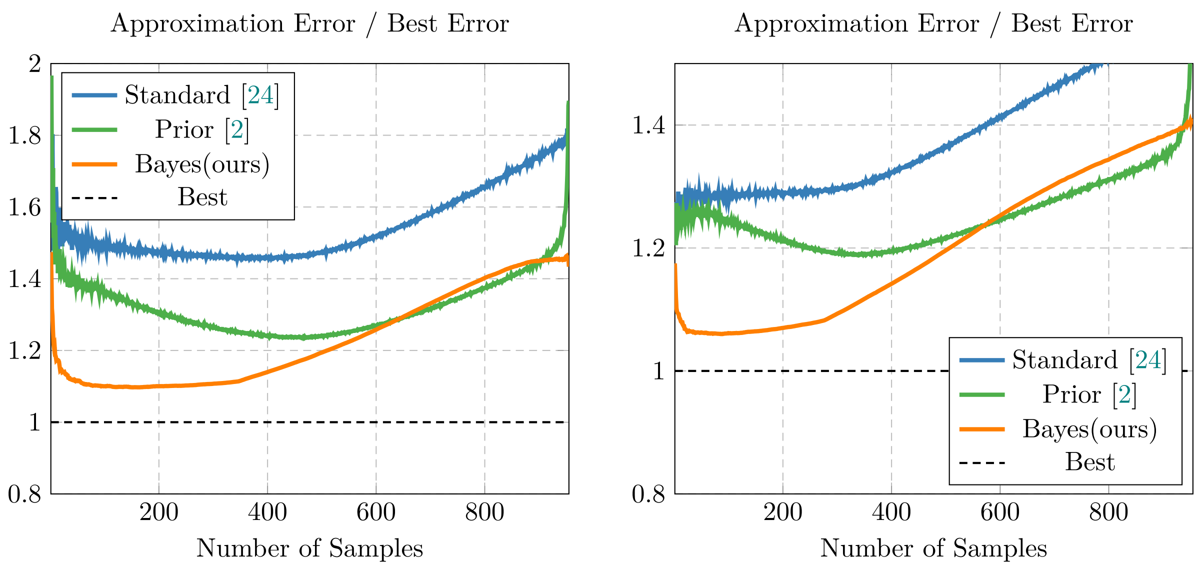

Figure 4: In these figures we look at matrices from Computational Fluid Dynamics. In *(Left)*, Matrix from TAMU Sparse Matrix Suite `cdde1`. In *(Right)*, Matrix from TAMU Sparse Matrix Suite `cdde1`. With the prior, we use the covariance matrix associated with the discrete Green's Function for the Laplacian as in Equation (19).

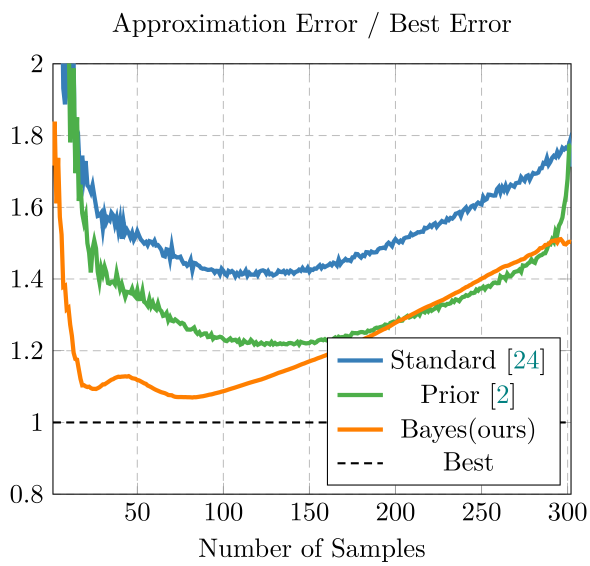

Figure 5: In these figures we look at matrices from the PDE for the Poisson Differential Operator. In *(Left)*, Matrix from TAMU Sparse Matrix Suite `cz308`. In *(Right)*, Matrix from TAMU Sparse Matrix Suite `cz628`. With the prior, we use the covariance matrix associated with the discrete Green's Function for the Laplacian as in Equation (19).