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

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 [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, $\mathbf{A}$. 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 [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 $\mathbf{A}$, 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 [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 $\mathbf{v}^{(k)}$ with access to the previous query vectors $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(k-1)}$, the matrix-vector products $\mathbf{A}\mathbf{v}^{(1)}, \ldots, \mathbf{A}\mathbf{v}^{(k-1)}$, and the intermediate low-rank matrix approximations, $\mathbf{Q}^{(k)}(\mathbf{Q}^{(k)})^{\mathrm{H}}\mathbf{A}$, where $\mathbf{Q}^{(k)} \triangleq \mathrm{orth}(\mathbf{A}\mathbf{V}^{(k)})$ where $\mathbf{V}^{(k)}$ is the concatenation of vectors $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(k)}$ and $\mathbf{Q}^{(k)} \triangleq \mathrm{orth}(\mathbf{A}\mathbf{V}^{(k)})$.

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 [HP14a, 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 $\mathbf{A}$ 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.

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 $\mathbf{A}$.

2. We provide a novel theoretical analysis which utilizes subspace perturbation theory.

3. We perform extensive experiments on matrices with various spectrums and show the effec-

tiveness 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

Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ represent the target matrix. $\|\cdot\|$ represents the spectral norm, which is equivalent to the max singular value, $\sigma_{\max}(\cdot)$. The pseudoinverse is represented by $(\cdot)^+$ s.t. $\mathbf{X}^+ = (\mathbf{X}^{\mathrm{H}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{H}}$. The Projection Matrix is defined as $\mathbf{\Pi_Y} = \mathbf{Y}\mathbf{Y}^+ = \mathbf{Y}(\mathbf{Y}^{\mathrm{H}}\mathbf{Y})^{-1}\mathbf{Y}^{\mathrm{H}}$ as the projection on to the column space of $\mathbf{Y}$. If $\mathbf{Y}$ has orthogonal columns, then $\mathbf{\Pi_Y}$ is the Orthogonal Projection defined as $\mathbf{\Pi_Y} = \mathbf{Y}\mathbf{Y}^{\mathrm{H}}$. 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. $\{\mathbf{V} : \mathbf{V}^{\mathrm{H}}\mathbf{V} = \mathbf{I}_{k\times k}\}$. 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,

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

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, $y \le O(x)$, to denote $y \le Cx$ for some positive constant, C. We define $\mathbb{E}[X]$ as expection of random variable X, $\mathbb{P}\{A\}$ as probability of event A occuring, and $\mathbb{V}(X)$ as variance or a random variable X. We will denote boldface characters $\mathbf{A}, \mathbf{Q}, \mathbf{X}$ 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 $\sin\angle$ norm. Let $\mathcal{X}, \mathcal{Y}$ be subspaces, then we denote the principal angles between subspaces (PABS) $\mathcal{X}$ and $\mathcal{Y}$ as $\frac{\mathbf{\Pi}}{2} \ge \angle_1(\mathcal{X}, \mathcal{Y}) \ge \ldots \ge \angle_{m\wedge n}(\mathcal{X}, \mathcal{Y})$. We then have that for any two matrices, $\mathbf{P}$ and $\mathbf{Q}$, it follows that $\|\mathbf{\Pi_P} - \mathbf{\Pi_Q}\|_{\mathrm{F}} = \|\sin \mathbf{\Theta}(\mathbf{P}, \mathbf{Q})\|_{\mathrm{F}}$.

2.3 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 review work by [MT20] gives significant theory on the Randomized SVD. [BT22] 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. [BT23] extended upon previous work on generalizing the Randomized SVD to learning HS Operators.

The most relevant work to ours is likely [DIKMI18]. The measure of accuracy in the Krylov Subspace is measured by the $\sin\angle$ 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 [DMN22] and [Sai19].

A similar analysis of a power method is explored in [HP14b] utilizing subspace perturbation theory. In this work, they consider the Matrix-Vector products have noise. In this work, similarly to [DIKMI18], 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 [LHZ21].

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, $\mathbf{C} \triangleq \mathbf{I}$. In the generalization of the Randomized SVD, when given some prior information of the matrix, the covariance matrix is given as $\mathbf{C} \triangleq \mathbf{K}$ where $\mathbf{K}$ has some information on the right singular vectors of $\mathbf{A}$ (e.g. discretization of Green's Function of a PDE). Let $\tilde{\mathbf{V}}$ 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 $\mathbf{C}^{(k+1)} \triangleq \tilde{\mathbf{V}}_{(:,k)}\tilde{\mathbf{V}}^{\mathrm{H}}_{(:,k)}$. Throughout this paper we will only consider $\mathbf{C}^{(0)} = \mathbf{I}$, however using theory from [BT22], this can be extended to $\mathbf{C}^{(0)} = \mathbf{K}$ if one has some knowledge of the right singular vectors, e.g. if the matrix represents a Partial Differential Equation (PDE), having $\mathbf{K}$ 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 Algorithm 1. For efficient updates, we frame all operations as rank-1 updates.

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 $\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}$ where $\mathbf{W} = \mathrm{orth}(\mathbf{A}\boldsymbol{\Omega})$ where $\boldsymbol{\Omega}$ 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 $\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}$. This process is then repeated for k iterations. The final low-rank approximation, $\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}$, 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.

Algorithm 1 Bayesian Function Sampling

Input: HS Operator: $\mathscr{F}$, Rank: r, Initial Covariance: $\mathbf{C}$, Oversampling Parameter: p
Output: Rank-r Approximation, $\hat{\mathbf{A}}_r$
$\mathbf{\Omega} \leftarrow [\underbrace{\mathcal{N}(\mathbf{0}, \mathbf{C}) \overset{\text{i.i.d}}{\cdots} \mathcal{N}(\mathbf{0}, \mathbf{C})}_{p}]$ ▷ Sample Oversampling Vectors from Standard Normal Matrix
$\mathbf{Y} \leftarrow \mathbf{A\Omega}$ ▷ Matrix Vector Products
$\left[\mathbf{W}^{(0)}, \sim\right] \leftarrow \text{QR}\left(\mathbf{Y}\right)$ ▷ Find Orthonormal Basis
$\tilde{\mathbf{A}}^{(0)} \leftarrow \mathbf{0}_{m\times n}$ ▷ Initial Low-Rank Approximation
for $k \in 1, 2, \ldots, r$ **do**
  $\tilde{\mathbf{A}}^{(k)} \leftarrow \tilde{\mathbf{A}}^{(k-1)} + \mathbf{W}^{(k-1)}_{(:,k-1)} \left(\mathbf{W}^{(k-1)}_{(:,k-1)}\right)^{\mathrm{H}} \mathbf{A}$ ▷ Rank-1 update to the low-rank approximation
  $\left[\tilde{\mathbf{U}}, \tilde{\mathbf{\Sigma}}, \tilde{\mathbf{V}}\right] \leftarrow \text{SVD}(\tilde{\mathbf{A}}^{(k)})$ ▷ SVD of current low-rank approximation
  $\mathbf{C}^{(k+1)} \leftarrow \tilde{\mathbf{V}}_{(:,k)} \tilde{\mathbf{V}}^{\mathrm{H}}_{(:,k)}$ ▷ 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
  $\left[\mathbf{W}^{(k)}, \sim\right] \leftarrow \text{QR}(\mathbf{Y})$ ▷ Find Orthonormal Basis
end for
$\tilde{\mathbf{A}}^{(\rho)} \leftarrow \mathbf{W}^{(\rho)}(\mathbf{W}^{(\rho)})^{\mathrm{H}}\mathbf{A}$ ▷ Final Low-Rank Approximation
return: $\tilde{\mathbf{A}}^{(\rho)}$

4.1 Setup

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

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

Furthermore, we let $\mathbf{A}_{(k)} \triangleq \sigma_k \mathbf{u}_k \mathbf{v}_k^{\mathrm{H}}$ and $\mathbf{A}_{\perp,k} \triangleq \mathbf{A} - \mathbf{A}_{(k)}$. The low rank matrix approximation is denoted by $\tilde{\mathbf{A}} \triangleq \mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}$ for a $\mathbf{W} \in \mathbb{C}^{m\times k}$ such that $\mathbf{W}^{\mathrm{H}}\mathbf{W} = \mathbf{I}$. The matrix factorization notation described in Equation (2) holds for $\tilde{\mathbf{A}}$ with a *tilde* over the typical notation. We denote $\mathbf{\Omega} \in \mathbb{R}^{n\times \ell}$ to be a test matrix such that columns of $\mathbf{\Omega}$ are sampled i.i.d from $\mathcal{N}(\mathbf{0}_n, \mathbf{I}_{n\times n})$.

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

$$\underset{\tilde{\mathbf{v}}\in\mathbb{R}^n}{\arg\min} \|\mathbf{A} - \tilde{\mathbf{W}}\tilde{\mathbf{W}}^{\mathrm{H}}\mathbf{A}\|_{\mathrm{F}} = \underset{\substack{\mathbf{x}\in \text{Null}(\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}) \\ \|\mathbf{x}\|=1}}{\arg\max} \|\mathbf{A}^{\mathrm{H}}\mathbf{A}\mathbf{x}\| \tag{3}$$

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

Lemma 1 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 $\text{Null}(\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A})$, we note that we do not have knowledge of $\left\|s\mathbf{A}^{\mathrm{H}}\mathbf{A}\mathbf{v}\right\|$ for any $\mathbf{v} \in \text{Null}(\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A})$.

Therefore, sampling from the Null Space of $\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}$. Rather, we sample to the closest singular vector to the Null-Space, which is *known*.

Lemma 2. *Let* $\mathbf{W} \in \mathbb{C}^{m\times k}$ *be a unitary matrix such that* $\mathbf{W} \triangleq \mathrm{orth}(\mathbf{A}\boldsymbol{\Omega})$ *for a test matrix* $\boldsymbol{\Omega} \in \mathbb{C}^{n\times k}$*.*

$$\left\| \underset{\substack{\tilde{\mathbf{x}}\in\mathrm{Span}(\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A})\\ \|\mathbf{x}\|=1}}{\arg\max} \|\mathbf{A}^{\mathrm{H}}\mathbf{A}\mathbf{x}\| - \underset{\substack{\mathbf{x}\in\mathrm{Null}(\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A})\\ \|\mathbf{x}\|=1}}{\arg\max} \|\mathbf{A}^{\mathrm{H}}\mathbf{A}\mathbf{x}\| \right\| \leq \varepsilon \tag{4}$$

4.3 Query Lower Bound for Frobenius Norm

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* $(\mathbf{v}^{(\zeta_1)})^{\mathrm{H}}\mathbf{v}^{(j)} = \delta_{ij}$ *for all* $i, j \in [k-1]\times[k-1]$*, matrix-vector queries,* $\mathbf{A}\mathbf{v}^{(1)},\ldots,\mathbf{A}\mathbf{v}^{(k-1)}$*, and intermediate low-rank approximations,* $\mathbf{W}^{(1)}(\mathbf{W}^{(1)})^{\mathrm{H}}\mathbf{A},\ldots,\mathbf{W}^{(k-1)}(\mathbf{W}^{(k-1)})^{\mathrm{H}}\mathbf{A}$ *, which requires* $k = O(\Xi)$ *vector queries to obtain a rank-k matrix with orthogonal columns,* $\mathbf{W}$*, such that with probability at least* $1-\delta$ *for a* $\delta \in (0,1)$ *such that*

$$\|\mathbf{A} - \mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}\|_{\mathrm{F}} \leq (1+\varepsilon) \min_{\substack{\mathbf{U}\in\mathbb{C}^{m\times k}\\ \mathbf{U}^{\mathrm{H}}\mathbf{U}=\mathbf{I}_k}} \left\|\mathbf{A} - \mathbf{U}\mathbf{U}^{\mathrm{H}}\mathbf{A}\right\|_{\mathrm{F}} \tag{5}$$

Proof.

4.4 Analysis of Algorithm 1

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.

Lemma 4. *Let* $\mathbf{A} \in \mathbb{C}^{m\times n}$ *and* $\mathbf{W}$ *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 k-th right singular vector of* $\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}$*. Denote* $\hat{\mathbf{W}} \triangleq \mathrm{orth}([\mathbf{Y} \quad \mathbf{v}])$ *and* $\tilde{\mathbf{W}} \triangleq \mathrm{orth}([\mathbf{Y} \quad \tilde{\mathbf{v}}])$*, if* $\|\mathbf{A} - \mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}\|_{\mathrm{F}} \leq C\sigma_{k+1}$ *and* $\|\mathbf{A} - \mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}\|_2 \leq c\sigma_{k+1}$ *for positive constants* $c, C > 0$ *where* $C \geq c$*, then we have in the worst case,*

$$\left\|\mathbf{A} - \tilde{\mathbf{W}}\tilde{\mathbf{W}}^{\mathrm{H}}\mathbf{A}\right\|_{\mathrm{F}} \leq C\sigma_{k+1} - \sigma_k^2 - 2\sigma_{k+1}^2\left(c^2 + O\left(\frac{1}{\sigma_k^2}\right)\right)\left(\frac{1+\xi}{1-\xi}\right) \tag{6}$$

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

Lemma 5. *Let* $\mathbf{A} \in \mathbb{C}^{m\times n}$ *and* $\mathbf{W}$ *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 k-th right singular vector of* $\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}$*. Denote* $\hat{\mathbf{W}} \triangleq \mathrm{orth}([\mathbf{Y} \quad \mathbf{v}])$ *and* $\tilde{\mathbf{W}} \triangleq \mathrm{orth}([\mathbf{Y} \quad \tilde{\mathbf{v}}])$*, if* $\|\mathbf{A} - \mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}\|_{\mathrm{F}} \leq C\sigma_{k+1}$ *and* $\|\mathbf{A} - \mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}\|_2 \leq c\sigma_{k+1}$ *for positive constants* $c, C > 0$ *where* $C \geq c$ *and an absolute constant* $c_3 > 0$*, it then follows*

$$\mathop{\mathbb{E}}_{\hat{\mathbf{v}}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \|\mathbf{A} - \hat{\mathbf{W}}\hat{\mathbf{W}}^{\mathrm{H}}\mathbf{A}\|_{\mathrm{F}} \leq C\sigma_{k+1}\sqrt{\frac{\rho - k - \left(\frac{16}{75\sqrt{5}}\right)^2}{\rho - k}} \tag{7}$$

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

Lemma 6. *Frame the same hypothesis as Lemma 5, it then follows for a failure probability* $\delta \in (0, 1)$

$$\|\mathbf{A} - \hat{\mathbf{W}}\hat{\mathbf{W}}^{\mathrm{H}}\mathbf{A}\|_{\mathrm{F}} \leq \tag{8}$$

Proof. We have with probability almost surely 1 ■

It is clear to see in Equation (6) the strength of our near-optimal sampling described in Algorithm 1 when there is sufficient singular value gap. We will now extend our theory to capture the error bounds of Algorithm 1.

Theorem 7 (Sufficient Singular Value Gap)**.** *If* $\|\mathbf{A} - \mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}\|_{\mathrm{F}} \leq C\sigma_{k+1}$ *and* $\|\mathbf{A} - \mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}\|_2 \leq c\sigma_{k+1}$ *for positive constants* $c, C > 0$ *where* $C \geq c$ *and* $\tilde{\sigma}_k = c_4\sigma_k$*, then we have with probability* $1 - \delta$ *Bayesian near-optimal sampling described in Section 3 decreases the low-rank approximation error faster than a normal vector sample when*

$$\left(\frac{\sigma_{k+1}}{\sigma_k}\right) \leq \Xi \tag{9}$$

Proof. The proof follows from algebraic manipulations relating Equation (6) and Equation (7) from Lemma 4 and Lemma 5, respectively. ■

The constant c_4 defined in Theorem 7 is nearly zero as it is the smallest non-zero eigenvalue of the low-rank matrix approximation $\mathbf{W}^{(k)}(\mathbf{W}^{(k)})^{\mathrm{H}}\mathbf{A}$.

Lemma 8. *Let* $\mathbf{A} \in \mathbb{C}^{m\times n}$*, and* $\mathbf{W} \in \mathbb{C}^{m\times n}$ *be an orthonormal matrix representing the basis of the subspace of* $\mathbf{Y} \in \mathbb{C}^{m\times k}$*. Let* $\mathbf{v}^{(i)}$ *for* $i \in [k]$ *represent the vector-queries from Algorithm 1. It then follows,*

$$\left\|\mathbf{A} - \mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}\right\|_{\mathrm{F}}^2 \leq \left(1 + \sum_{j=1}^{k}\right)\sum_{j>k}\sigma_j^2 \tag{10}$$

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:

$$\mathcal{L}u = \frac{\partial^2 u}{\partial x^2} - 100\sin(5\pi x)\,u, \qquad x \in [0, 1] \tag{11}$$

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 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, [DH11]. In Figure 2 *(Left)*, we choose a fluid dynamics problem due to its relevance in low-rank approximation [BNK20].

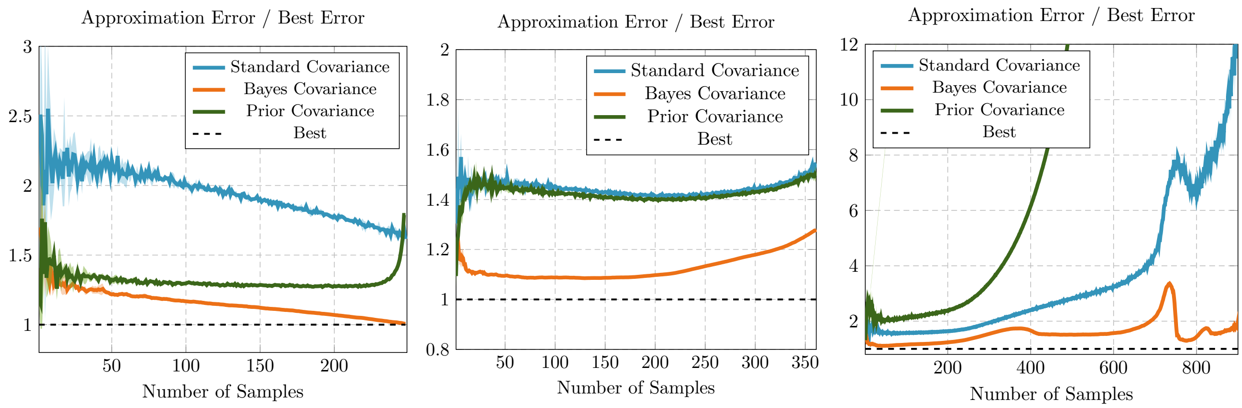

Figure 1: Low Rank Approximation for the Inverse Differential Operator given in Equation (11) *(Left)*, Differential Operator Matrix `Poisson2D` [DH11] *(Center)*, and Differential Operator Matrix `DK01R` [DH11] *(Right)*. The experiment on the left is from [BT22, Figure 2]. We use the discretized Green's Function as the prior covariance matrix.

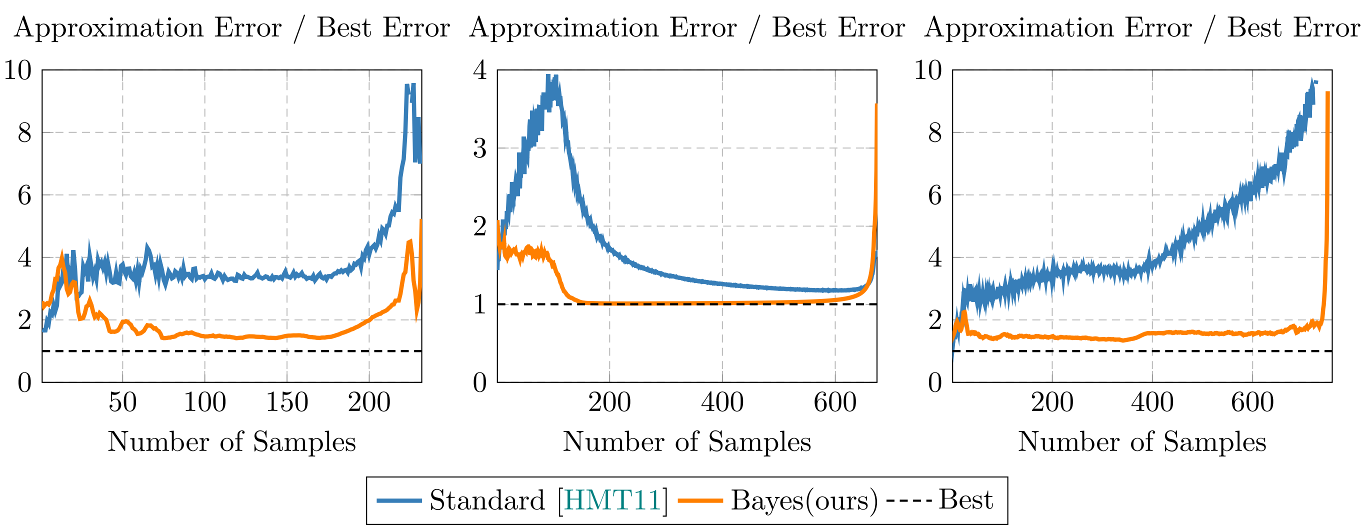

Figure 2: Low Rank Approximation for a matrix for a Computational Fluid Dynamics Problem, `saylr1` *(Left)*from [DH11]. Subsequent 2D/3D Problem `fs-680-2` *(Center)*from [DH11]. Astrophysics 2D/3D Problem `msfe` *(Right)*from [DH11].

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

$$\mathbf{A} = \sum_{i=1}^{\rho} \frac{100 i^{\ell}}{n} \mathbf{U}_{(:,i)} \mathbf{V}_{(:,i)}^{\mathrm{H}}, \quad \mathbf{U} \in \mathbb{O}_{m,k}, \mathbf{V} \in \mathbb{O}_{n,k} \tag{12}$$

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

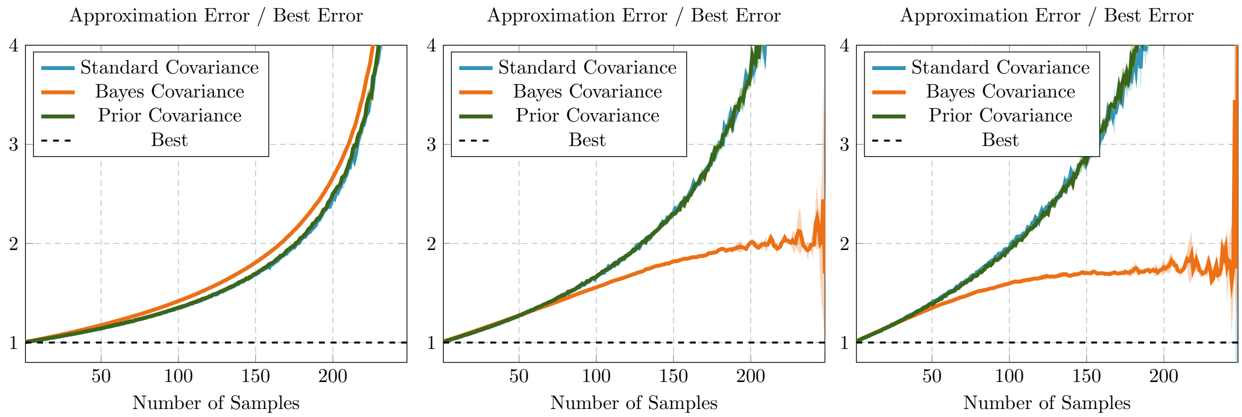

Figure 3: Low Rank Approximation for synthetic matrices with decay described in Equation (12). In *(Left)*, $\ell = 1$, in *(Center)*, $\ell = 2$, and in *(Right)*, $\ell = 3$. We use the Gram matrix with Gaussian Kernel and $\gamma = 0.01$ for the Prior Covariance Matrix.

the singular values.

6 Conclusions

We have theoretically and empirically analyzed a novel Covariance Update to iteratively construct the sampling matrix, $\mathbf{\Omega}$ in the Randomized SVD algorithm. We introduce a new adaptive sampling framework for low-rank matrix approximation when the matrix is only accessible by matrix-vector products by giving the algorithm access to intermediate low-rank matrix approximations. Our covariance update for generating sampling vectors and functions can find use various PDE learning applications, [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

[Ada15] Radoslaw Adamczak. A note on the hanson-wright inequality for random vectors with dependencies. 2015.

[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.

[BG13] 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.

[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.

[DIKMI18] 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.

[DMN22] 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.

[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.

[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.

[Gu15] Ming Gu. Subspace iteration randomization and singular value problems. *SIAM Journal on Scientific Computing*, 37(3):A1139–A1173, 2015.

[H8̈9] O. Hölder. Ueber einen mittelwerthabsatz. *Nachrichten von der Königl. Gesellschaft der Wissenschaften und der Georg-Augusts-Universität zu Göttingen*, 1889:38–47, 1889.

[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.

[HP14a] Sariel Har-Peled. Low rank matrix approximation in linear time. *arXiv preprint arXiv:1410.8802*, 2014.

[HP14b] 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.

[LHZ21] 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.

[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.

[Mar89] A. A. Markov. On a question by D. I. Mendeleev. *Zapiski Imp. Akad. Nauk*, 62(12):1–24, 1889.

[Mos21] Kamyar Moshksar. On the absolute constant in hanson-wright inequality. *arXiv preprint arXiv:2111.00557*, 2021.

[MT20] Per-Gunnar Martinsson and Joel A. Tropp. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica*, 29:403–572, 2020.

[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.

[PZ32] Raymond EAC Paley and Antoni Zygmund. A note on analytic functions in the unit circle. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 28, pages 266–272. Cambridge University Press, 1932.

[RV13] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. 2013.

[Sai19] 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.

[SgS90] G. W. Stewart and Ji guang Sun. Matrix perturbation theory. 1990.

[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.

[TB22] Lloyd N Trefethen and David Bau. *Numerical linear algebra*, volume 181. Siam, 2022.

[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.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[W+14] 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.

[Wed72] Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, Mar 1972.

[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.

A.1 Proof of Lemma 1

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

$$\begin{aligned}
\left\|\mathbf{A} - \hat{\mathbf{W}}\hat{\mathbf{W}}^{\mathrm{H}}\mathbf{A}\right\|_{\mathrm{F}}^2 &= \left\|\mathbf{A} - \mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A} - \hat{\mathbf{w}}\hat{\mathbf{w}}^{\mathrm{H}}\mathbf{A}\right\|_{\mathrm{F}}^2 && (13)\\
&= \mathrm{Tr}\left(\mathbf{A}^{\mathrm{H}}\mathbf{A} - \mathbf{A}^{\mathrm{H}}\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A} - \mathbf{A}^{\mathrm{H}}\hat{\mathbf{w}}\hat{\mathbf{w}}^{\mathrm{H}}\mathbf{A}\right) && (14)\\
&= \|\mathbf{A} - \mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}\|_{\mathrm{F}}^2 - \underbrace{\|\hat{\mathbf{w}}^{\mathrm{H}}\mathbf{A}\|_{\mathrm{F}}^2}_{c_2} && (15)
\end{aligned}$$

Similarly from Equation (15), we have

$$\|\mathbf{A} - \tilde{\mathbf{W}}\tilde{\mathbf{W}}^{\mathrm{H}}\mathbf{A}\|_{\mathrm{F}}^2 = \|\mathbf{A}\|_{\mathrm{F}}^2 - \|\mathbf{W}^{\mathrm{H}}\mathbf{A}\|_{\mathrm{F}}^2 - \underbrace{\|\tilde{\mathbf{w}}^{\mathrm{H}}\mathbf{A}\|_{\mathrm{F}}^2}_{c_1} \qquad (16)$$

Let us note for any column $\mathbf{w} \in \mathbf{W}$ and $\mathbf{v} \in \mathbb{C}^n$, we have

$$\left(\left(\mathbf{I} - \mathbf{W}\mathbf{W}^{\mathrm{H}}\right)\mathbf{A}\mathbf{v}\right)^{\mathrm{H}}\mathbf{w} = \left(\mathbf{v}^{\mathrm{H}}\mathbf{A}^{\mathrm{H}} - \mathbf{v}^{\mathrm{H}}\mathbf{A}^{\mathrm{H}}\mathbf{W}\mathbf{W}^{\mathrm{H}}\right)\mathbf{w} = (\mathbf{A}\mathbf{v})^{\mathrm{H}}\mathbf{w} - (\mathbf{A}\mathbf{v})^{\mathrm{H}}\mathbf{w} = \mathbf{0} \qquad (17)$$

Since we have $\hat{\mathbf{w}}, \tilde{\mathbf{w}} \in \mathrm{Span}(\mathbf{I} - \mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A})$, then from our formulation in Equations (15) and (16), we want to our sampled vector to be in the dominant singular space of the span of the singular vectors of $\mathbf{I} - \mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}$. We first require the following supplementary result for any matrix projector, $\mathbf{\Pi}$.

$$\begin{aligned}
\mathrm{Null}\left((\mathbf{I} - \mathbf{\Pi})\,\mathbf{A}\right) &= \{\mathbf{y} \in \mathrm{Range}(\mathbf{A}) : (\mathbf{I} - \mathbf{\Pi})\,\mathbf{y} = \mathbf{0}\} && (18)\\
&= \{\mathbf{y} \in \mathrm{Range}(\mathbf{A}) : \mathbf{\Pi}\mathbf{y} = \mathbf{y}\} && (19)\\
&= \{\mathbf{y} \in \mathrm{Range}(\mathbf{A}) \cap \mathrm{Range}(\mathbf{\Pi})\} = \{\mathbf{x} \in \mathbb{C}^n : \mathbf{A}\mathbf{x} \in \mathrm{Range}(\mathbf{\Pi})\} && (20)
\end{aligned}$$

Now we can calculate the optimal sampling vector.

$$\begin{aligned}
\mathbf{v}_{\mathrm{OPT}} &\overset{(17)}{=} \underset{\mathbf{v} \in \mathbb{C}^n}{\arg\max} \frac{\left\|\mathbf{A}^{\mathrm{H}}\left(\mathbf{I} - \mathbf{W}\mathbf{W}^{\mathrm{H}}\right)(\mathbf{A}\mathbf{v})\right\|}{\|(\mathbf{I} - \mathbf{W}\mathbf{W}^{\mathrm{H}})\,(\mathbf{A}\mathbf{v})\|} && (21)\\
&= \underset{\mathbf{x} \in \mathbb{C}^n \text{ s.t. } \mathbf{A}\mathbf{x} \in \mathrm{Range}((\mathbf{I} - \mathbf{W}\mathbf{W}^{\mathrm{H}})\mathbf{A})}{\arg\max} \frac{\left\|\mathbf{A}^{\mathrm{H}}\mathbf{A}\mathbf{x}\right\|}{\|\mathbf{x}\|} && (22)\\
&\overset{(20)}{=} \underset{\mathbf{x} \in \mathrm{Null}(\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A})}{\arg\max} \frac{\left\|\mathbf{A}^{\mathrm{H}}\mathbf{A}\mathbf{x}\right\|}{\|\mathbf{x}\|} && (23)
\end{aligned}$$

This completes the proof. ∎

A.2 Proof of Lemma 4

Let c_1 defined in Appendix A.1 in Equation (16). Let us define $\tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^{\mathrm{H}} = \mathrm{SVD}(\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A})$ where $\tilde{\mathbf{U}} \in \mathbb{C}^{m\times n}$ such that $\mathbf{U}^{\mathrm{H}}\mathbf{U} = \mathbf{I}$, $\tilde{\mathbf{\Sigma}} \in \mathbb{R}^{n\times n}$ is diagonal, and $\tilde{\mathbf{V}} \in \mathbb{C}^{n\times n}$ is unitary. Recall $\sigma_1 \geq \ldots \geq \sigma_{m\wedge n}$ are the singular values of $\mathbf{A}$ and $\tilde{\sigma}_1 \geq \ldots \geq \tilde{\sigma}_{m\wedge n}$ are the singular values of $\mathbf{W}^{(k)}(\mathbf{W}^{(k)})^{\mathrm{H}}\mathbf{A}$.

$$\left\|\tilde{\mathbf{w}}^{\mathrm{H}}\mathbf{A}\right\|_{\mathrm{F}}^2 \overset{(17)}{=} \frac{\|((\mathbf{I}-\mathbf{\Pi}_{\mathbf{W}})\,\mathbf{A}\tilde{\mathbf{v}})^{\mathrm{H}}\,\mathbf{A}\|_2^2}{\|(\mathbf{I}-\mathbf{\Pi}_{\mathbf{W}})\,\mathbf{A}\tilde{\mathbf{v}}\|_2^2} \tag{24}$$

We bound the numerator and denominator seperately. We will first upper bound the denominator.

$$\begin{aligned}
\left\|\left(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}}\right)\mathbf{A}\tilde{\mathbf{v}}_k\right\|^2 &= \tilde{\mathbf{v}}_k^{\mathrm{H}}\mathbf{A}^{\mathrm{H}}\left(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}}\right)\left(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}}\right)\mathbf{A}\tilde{\mathbf{v}}_k && (25)\\
&= \tilde{\mathbf{v}}_k^{\mathrm{H}}\mathbf{A}^{\mathrm{H}}\left(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}}\right)\mathbf{A}\tilde{\mathbf{v}}_k = \|\mathbf{A}\tilde{\mathbf{v}}_k\|^2 - \tilde{\sigma}_k^2 && (26)
\end{aligned}$$

Next, we will lower bound the numerator.

$$\begin{aligned}
\left\|\mathbf{A}^{\mathrm{H}}\left(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}}\right)\mathbf{A}\tilde{\mathbf{v}}_k\right\|^2 &= \left\|\mathbf{A}^{\mathrm{H}}\mathbf{A}\tilde{\mathbf{v}}_k\right\|^2 - 2\tilde{\mathbf{v}}_k^{\mathrm{H}}\mathbf{A}^{\mathrm{H}}\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}\mathbf{A}^{\mathrm{H}}\mathbf{A}\tilde{\mathbf{v}}_k + \left\|\tilde{\mathbf{v}}_k^{\mathrm{H}}\mathbf{A}^{\mathrm{H}}\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}\right\|^2 && (27)\\
&= \left\|\mathbf{A}^{\mathrm{H}}\mathbf{A}\tilde{\mathbf{v}}_k\right\|^2 - 2\tilde{\mathbf{v}}_k^{\mathrm{H}}\mathbf{A}^{\mathrm{H}}\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}\mathbf{A}^{\mathrm{H}}\mathbf{A}\tilde{\mathbf{v}}_k + \tilde{\sigma}_k^4 && (28)\\
&\geq \left(\left\|\mathbf{A}^{\mathrm{H}}\mathbf{A}\tilde{\mathbf{v}}_k\right\| - \sigma_k^2\right)^2 && (29)
\end{aligned}$$

In (ζ_1), we lower bound $\|\tilde{\mathbf{v}}_k^{\mathrm{H}}\mathbf{A}^{\mathrm{H}}\mathbf{A}\|$ given in Lemma 12, next we use the equality for $\|\tilde{\mathbf{u}}_k^{\mathrm{H}}\mathbf{A}\|$ with Lemma 13, finally, we combine Lemma 11 and Lemma 13 with the Cauchy-Schwarz Inequality to upper bound $\|\tilde{\mathbf{v}}_k^{\mathrm{H}}\mathbf{A}^{\mathrm{H}}\mathbf{A}\mathbf{A}^{\mathrm{H}}\tilde{\mathbf{u}}_k\|$. In (ζ_2), we upper bound $\sin^2\angle(\mathbf{v}_k,\tilde{\mathbf{v}}_k)$ with Lemma 9. In **??**, we have $\|\mathbf{W}^{\mathrm{H}}\mathbf{A}\tilde{\mathbf{v}}_k\| = \|\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}\tilde{\mathbf{v}}_k\| = \|\tilde{\sigma}_k\tilde{\mathbf{u}}_k\| = \tilde{\sigma}_k$. We then have,

$$\|\tilde{\mathbf{w}}^{\mathrm{H}}\mathbf{A}\|^2 \geq \sigma_k^2 - \frac{2\left(c^2 + O\left(\frac{1}{\sigma_k^2}\right)\right)\sigma_{k+1}^2\left(\sigma_k^2+\tilde{\sigma}_k^2\right)}{\sigma_k^2 - \tilde{\sigma}_k^2 + O\left(c^2\left(\frac{\sigma_{k+1}}{\sigma_k}\right)^2\right)} \tag{30}$$

To complete the proof, we must control $\tilde{\sigma}_k$ with relation to σ_k. First, we note for any two matrices $\mathbf{S},\mathbf{T} \in \mathbb{C}^{m\times n}$, we have $\sigma_i(\mathbf{TS}) \leq \sigma_1(\mathbf{T})\sigma_i(\mathbf{S})$ Then, using the idempotency of $\mathbf{\Pi}_{\mathbf{W}}$, we have that $\sigma_i(\mathbf{\Pi}_{\mathbf{W}}\mathbf{A}) \leq \sigma_i(\mathbf{A})$ for all $i \in [n]$. We then use Lemma 14 to lower bound $\tilde{\sigma}_k$. This gives us the following,

$$\|\tilde{\mathbf{w}}^{\mathrm{H}}\mathbf{A}\|^2 \geq \sigma_k^2 - 2\sigma_{k+1}^2\left(c^2 + O\left(\frac{1}{\sigma_k^2}\right)\right)\left(\frac{2}{1-\xi}\right) \tag{31}$$

where $\xi \leq c^2(\frac{\sigma_{k+1}}{\sigma_k})^2$. Taking the square root of both sides completes our proof. ■

A.3 Proof of Lemma 5

We will lower bound c_2 as defined in Equation (15).

$$c_2 \triangleq \|\hat{\mathbf{w}}^{\mathrm{H}}\mathbf{A}\|_{\mathrm{F}}^2 \overset{(17)}{=} \frac{\left\|\mathbf{A}^{\mathrm{H}}\left(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}}\right)\mathbf{A}\hat{\mathbf{v}}\right\|_2^2}{\|(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}})\,\mathbf{A}\hat{\mathbf{v}}\|_2^2} \tag{32}$$

For shorthand, let $\tilde{\mathbf{A}} \triangleq \mathbf{A}^{\mathrm{H}}(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}})\mathbf{A}$. Note, we have $\|(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}})\mathbf{A}\|_{\mathrm{F}}$ is given. It then follows with probability $1-\delta$,

$$\|\tilde{\mathbf{A}}\mathbf{x}\|^2 \overset{\text{lem. 15}}{\leq} \mathbb{E}\left[\|\tilde{\mathbf{A}}\mathbf{x}\|^2\right] - \|\tilde{\mathbf{A}}\|_{\mathrm{F}}^2\sqrt{\frac{1}{c_3}\log\frac{2}{\delta}} \tag{33}$$

for an absolute constant $c_3 > 0$. We further simplify this with the fact

$$\mathbb{E}\|\tilde{\mathbf{A}}\mathbf{x}\|^2 = \mathbb{E}\mathbf{x}^{\mathrm{H}}\tilde{\mathbf{A}}^{\mathrm{H}}\tilde{\mathbf{A}}\mathbf{x} = \mathbb{E}\operatorname{Tr}\left(\mathbf{x}^{\mathrm{H}}\tilde{\mathbf{A}}^{\mathrm{H}}\tilde{\mathbf{A}}\mathbf{x}\right) = \mathbb{E}\operatorname{Tr}\left(\mathbf{x}\mathbf{x}^{\mathrm{H}}\tilde{\mathbf{A}}^{\mathrm{H}}\tilde{\mathbf{A}}\right) = \operatorname{Tr}\left(\mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathrm{H}}\tilde{\mathbf{A}}^{\mathrm{H}}\tilde{\mathbf{A}}\right]\right) \tag{34}$$

$$= \operatorname{Tr}\left(\mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathrm{H}}\right]\tilde{\mathbf{A}}^{\mathrm{H}}\tilde{\mathbf{A}}\right) = \operatorname{Tr}\left(\tilde{\mathbf{A}}^{\mathrm{H}}\tilde{\mathbf{A}}\right) = \|\tilde{\mathbf{A}}\|_{\mathrm{F}}^2 \tag{35}$$

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

$$\left\|\left(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}}\right)\mathbf{A}\right\|_{\mathrm{F}}^2 = \operatorname{Tr}\left(\mathbf{A}\left(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}}\right)\left(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}}\right)\mathbf{A}^{\mathrm{H}}\right) \tag{36}$$

$$= \operatorname{Tr}\left(\mathbf{A}\left(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}}\right)\mathbf{A}^{\mathrm{H}}\right) \tag{37}$$

$$\overset{(\zeta_1)}{\leq} \sqrt{\rho-k}\left\|\mathbf{A}\left(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}}\right)\mathbf{A}^{\mathrm{H}}\right\|_{\mathrm{F}} \tag{38}$$

where (ζ_1) follows from the relation of the $\|\cdot\|_1$ and $\|\cdot\|_2$ norm and assuming $\operatorname{rank}(\mathbf{A}) = \rho$. 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

$$\mathbb{E}\left[\frac{\left\|\mathbf{A}\left(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}}\right)\mathbf{A}\hat{\mathbf{v}}\right\|_2^2}{\|(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}})\mathbf{A}\hat{\mathbf{v}}\|_2^2}\right] \overset{(\zeta_2)}{\geq} \left(\mathbb{E}\left\|\hat{\mathbf{v}}^{\mathrm{H}}\mathbf{A}^{\mathrm{H}}\left(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}}\right)\mathbf{A}\right\|_2\right)^2 \mathbb{E}\left[\left\|\left(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}}\right)\mathbf{A}\hat{\mathbf{v}}\right\|_2^2\right]^{-1}$$

$$\overset{(\zeta_3)}{\geq} \left(\frac{16}{75\sqrt{5}}\right)^2 \left\|\mathbf{A}^{\mathrm{H}}\left(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}}\right)\mathbf{A}\right\|_{\mathrm{F}}^2 \left\|\left(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}}\right)\mathbf{A}\right\|_{\mathrm{F}}^{-2} \tag{39}$$

$$\overset{(36)}{\geq} \left(\frac{16}{75\sqrt{5}}\right)^2 \frac{\left\|\left(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}}\right)\mathbf{A}\right\|_{\mathrm{F}}^4}{(\rho-k)\|(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}})\mathbf{A}\|_{\mathrm{F}}^2} \tag{40}$$

$$\gtrsim \left(\frac{1}{\rho-k}\right)\left\|\left(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}}\right)\mathbf{A}\right\|_{\mathrm{F}}^2 \tag{41}$$

(ζ_2) follows from the Reverse Hölder Inequality [H8̈9]. (ζ_3) follows from an application of Jensen's Inequality. Combining Equation (41) and Equation (35) we thus have in expectation we have

$$\mathbb{E}c_2 = \Omega\left(\frac{1}{\rho-k}\right)\left\|\left(\mathbf{I}-\mathbf{W}\mathbf{W}^{\mathrm{H}}\right)\mathbf{A}\right\|_{\mathrm{F}}^2 \tag{42}$$

This completes our proof of Claim (ii). ■

A.4 Proof of Lemma 8

Proof. We will utilize Lemma 4 for our proof.

$$\left\|\mathbf{A}-\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}\right\|_{\mathrm{F}}^2 = \|\mathbf{A}\|_{\mathrm{F}}^2 - \left\|\mathbf{W}^{\mathrm{H}}\mathbf{A}\right\|_{\mathrm{F}}^2 \tag{43}$$

$$= \sum_{i=1}^{k}\sigma_i^2 - \sum_{i=1}^{k}\|\mathbf{w}_i^{\mathrm{H}}\mathbf{A}\|_{\mathrm{F}}^2 + \|\mathbf{A}_{\perp,k}\|_{\mathrm{F}}^2 \tag{44}$$

$$\stackrel{\text{lem. 4}}{\leq} \sum_{i=1}^{k} \sigma_i^2 - \left(\sigma_i^2 - 2\sigma_{i+1}^2 \left(c_i^2 + O\left(\frac{1}{\sigma_i^2}\right) \right) \left(\frac{2}{1-\xi_i} \right) \right) + \|\mathbf{A}_{\perp,k}\|_{\mathrm{F}}^2 \tag{45}$$

$$\leq 2C \sum_{i=1}^{k} \|\mathbf{A}_{\perp,k}\|_{\mathrm{F}}^2 \left(c_i^2 + O\left(\frac{1}{\sigma_i^2}\right) \right) \left(\frac{2}{1-\xi_i} \right) + \|\mathbf{A}_{\perp,k}\|_{\mathrm{F}}^2 \tag{46}$$

Taking the square root, we have the following bound,

$$\left\| \mathbf{A} - \mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A} \right\|_{\mathrm{F}} \leq \left(1 + 2C \sum_{i=1}^{k} \left(c_i^2 + O\left(\frac{1}{\sigma_i^2}\right) \right) \left(\frac{2}{1-\xi_i} \right) \right)^{1/2} \|\mathbf{\Sigma}_{\rho-k}\|_{\mathrm{F}} \tag{47}$$

□

B Singular Subspace Perturbation Lemmas

Lemma 9 ([Wed72], [SgS90])**.** *Let $\mathbf{A}, \hat{\mathbf{A}} \in \mathbb{C}^{m\times n}$ be partitioned as follows,*

$$\mathbf{A} = \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_2 \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^{\mathrm{H}} \\ \mathbf{V}_2^{\mathrm{H}} \end{bmatrix}, \qquad \tilde{\mathbf{A}} = \begin{bmatrix} \tilde{\mathbf{U}}_1 & \tilde{\mathbf{U}}_2 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{\Sigma}}_1 & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{\Sigma}}_2 \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{V}}_1^{\mathrm{H}} \\ \tilde{\mathbf{V}}_2^{\mathrm{H}} \end{bmatrix} \tag{48}$$

Then define $\delta \triangleq \min\limits_{1\leq i\leq k, 1\leq j\leq n-k} \left\{ |\sigma_i - \tilde{\sigma}_{k+j}|, \min\limits_{1\leq i\leq k} \sigma_i \right\}$. It then follows

$$\|\sin\Theta(\mathbf{U}_1, \tilde{\mathbf{U}}_1)\|_{\mathrm{F}}^2 + \|\sin\Theta(\mathbf{V}_1, \tilde{\mathbf{V}}_1)\|_{\mathrm{F}}^2 \leq \frac{\|(\tilde{\mathbf{A}} - \mathbf{A})\mathbf{V}_1\|_{\mathrm{F}}^2 + \|\mathbf{U}_1^{\mathrm{H}}(\tilde{\mathbf{A}} - \mathbf{A})\|_{\mathrm{F}}^2}{\delta^2} \tag{49}$$

Lemma 10. *Let $\mathbf{v}$ and $\tilde{\mathbf{v}}$ be vectors s.t. $\|\mathbf{v}\| = \|\tilde{\mathbf{v}}\| = 1$ and $\mathbf{v}^{\mathrm{H}}\tilde{\mathbf{v}} \geq 0$. Then,*

$$\|\mathbf{v} - \tilde{\mathbf{v}}\| \leq \sqrt{2} \sin\angle(\mathbf{v}, \tilde{\mathbf{v}}) \tag{50}$$

Proof. Let us first note that for any two normal vectors $\mathbf{v}, \tilde{\mathbf{v}}$, we have $\cos\angle(\mathbf{v}, \tilde{\mathbf{v}}) = \mathbf{v}^{\mathrm{H}}\tilde{\mathbf{v}}$. Then, we have

$$\sin^2\angle(\mathbf{v}, \tilde{\mathbf{v}}) = 1 - \left(\mathbf{v}^{\mathrm{H}}\tilde{\mathbf{v}}\right)^2 \stackrel{(\zeta_1)}{\geq} 1 - \mathbf{v}^{\mathrm{H}}\tilde{\mathbf{v}} = 1 + \frac{1}{2}\|\mathbf{v} - \tilde{\mathbf{v}}\|^2 - \frac{1}{2}\|\mathbf{v}\|^2 - \frac{1}{2}\|\tilde{\mathbf{v}}\|^2 = \frac{1}{2}\|\mathbf{v} - \tilde{\mathbf{v}}\|^2 \tag{51}$$

(ζ_1) follows from $0 \leq \mathbf{v}^{\mathrm{H}}\tilde{\mathbf{v}} \leq 1$, therefore $\mathbf{v}^{\mathrm{H}}\tilde{\mathbf{v}} \geq (\mathbf{v}^{\mathrm{H}}\tilde{\mathbf{v}})^2$. Plugging this back into the first inequality and taking the square root gives us the desired result. □

Lemma 11. *Let $\tilde{\mathbf{v}}_k$ represent the kth right singular vector of $\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}$ where $\mathbf{W} \in \mathbb{C}^{m\times k}$ and $\mathbf{W}^{\mathrm{H}}\mathbf{W} = \mathbf{I}$. Furthermore, assume $\|(\mathbf{I} - \mathbf{W}\mathbf{W}^{\mathrm{H}})\mathbf{A}\| \leq c\sigma_{k+1}$. It then holds,*

$$\|\mathbf{A}\tilde{\mathbf{v}}_k\| \leq \sigma_k + O\left(c \left(\frac{\sigma_{k+1}}{\sigma_k} \right) \right) \tag{52}$$

Proof. Let us first describe the relation to the Rayleigh Quotient, which is defined here as

$$R(\mathbf{A}^{\mathrm{H}}\mathbf{A}, \mathbf{x}) = \frac{\mathbf{x}^{\mathrm{H}}\mathbf{A}^{\mathrm{H}}\mathbf{A}\mathbf{x}}{\mathbf{x}^{\mathrm{H}}\mathbf{x}} \tag{53}$$

Furthermore, it is well known that $\nabla_{\mathbf{x}} R(\mathbf{A}^{\mathrm{H}}\mathbf{A}, \mathbf{v}_k) = \mathbf{0}$ and $R(\mathbf{A}^{\mathrm{H}}\mathbf{A}, \mathbf{v}_k) = \sigma_k^2$ for any $k \in [n]$ [TB22]. From the Taylor Series Expansion of the Rayleigh Quotient, we have

$$R(\mathbf{A}^{\mathrm{H}}\mathbf{A}, \tilde{\mathbf{v}}_k) = R(\mathbf{A}^{\mathrm{H}}\mathbf{A}, \mathbf{v}_k) + (\tilde{\mathbf{v}}_k - \mathbf{v}_k)^{\mathrm{H}} \nabla_{\mathbf{x}} R(\mathbf{A}^{\mathrm{H}}\mathbf{A}, \mathbf{v}_k) + O\left(\|\tilde{\mathbf{v}}_k - \mathbf{v}_k\|^2\right) \tag{54}$$

$$= \sigma_k^2 + O\left(\|\tilde{\mathbf{v}}_k - \mathbf{v}_k\|^2\right) \overset{\text{lem. 10}}{\leq} \sigma_k^2 + O\left(\sin^2\angle(\mathbf{v}_k, \tilde{\mathbf{v}}_k)\right) \tag{55}$$

$$\overset{\text{lem. 9}}{\leq} \sigma_k^2 + O\left(\frac{\|\mathbf{A} - \mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}\|^2}{(\sigma_k - \tilde{\sigma}_{k+1})^2}\right) \leq \sigma_k^2 + O\left(c^2\left(\frac{\sigma_{k+1}}{\sigma_k}\right)^2\right) \tag{56}$$

Applying the triangle inequality completes the proof. □

Lemma 12. *Let $\tilde{\mathbf{v}}_k$ be the kth right singular vector of an approximation $\tilde{\mathbf{A}}$ of $\mathbf{A}$. Then,*

$$\left\|\mathbf{A}^{\mathrm{H}}\mathbf{A}\tilde{\mathbf{v}}_k\right\| \geq \sigma_k^2\left(1 - \sqrt{2}\sin\angle(\mathbf{v}_k, \tilde{\mathbf{v}}_k)\right) \tag{57}$$

Proof. We will start by applying the Matrix Pythagoras Thoerem.

$$\left\|\mathbf{A}^{\mathrm{H}}\mathbf{A}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{A}^{\mathrm{H}}\mathbf{A}\tilde{\mathbf{v}}_k\right\|_{\mathrm{F}}^2 = \left\|\mathbf{A}^{\mathrm{H}}\mathbf{A} - \tilde{\mathbf{v}}_k\tilde{\mathbf{v}}_k^{\mathrm{H}}\mathbf{A}^{\mathrm{H}}\mathbf{A}\right\|_{\mathrm{F}}^2 \tag{58}$$

$$\overset{\text{lem. 19}}{\leq} \left\|\mathbf{A}^{\mathrm{H}}\mathbf{A} - \tilde{\mathbf{v}}_k\tilde{\mathbf{v}}_k^{\mathrm{H}}\mathbf{A}_{(k)}^{\mathrm{H}}\mathbf{A}_{(k)}\right\|_{\mathrm{F}}^2 \tag{59}$$

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

$$\overset{\text{lem. 19}}{\leq} \left\|\mathbf{A}_{(k)}^{\mathrm{H}}\mathbf{A}_{(k)} - \sigma_k^2\tilde{\mathbf{v}}_k\tilde{\mathbf{v}}_k^{\mathrm{H}}\right\|_{\mathrm{F}}^2 + \left\|\mathbf{A}_{\perp,k}^{\mathrm{H}}\mathbf{A}_{\perp,k}\right\|_{\mathrm{F}}^2 \tag{61}$$

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

$$\overset{\text{lem. 10}}{=} 2\sigma_k^4\sin^2\angle(\mathbf{v}_k, \tilde{\mathbf{v}}_k) + \left\|\mathbf{A}_{\perp,k}^{\mathrm{H}}\mathbf{A}_{\perp,k}\right\|_{\mathrm{F}}^2 \tag{63}$$

(ζ_1) follows from the Matrix Pythagoras theorem [W+14]. From rearranging the inequalities and noting $\|\mathbf{A}^{\mathrm{H}}\mathbf{A}\|_{\mathrm{F}}^2 - \|\mathbf{A}_{\perp,k}^{\mathrm{H}}\mathbf{A}_{\perp,k}\|_{\mathrm{F}}^2 = \sigma_k^4$, we then obtain

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

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

Lemma 13. *Let $\tilde{\mathbf{u}}_k$ be the kth left singular vector and $\tilde{\mathbf{v}}_k$ be the kth right singular vector of $\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}$ where $\mathbf{W} \in \mathbb{C}^{m\times k}$ s.t. $\mathbf{W}^{\mathrm{H}}\mathbf{W} = \mathbf{I}$. Then, we have*

$$\|\tilde{\mathbf{u}}_k^{\mathrm{H}}\mathbf{A}\|_2 = \tilde{\mathbf{u}}_k^{\mathrm{H}}\mathbf{A}\tilde{\mathbf{v}}_k = \tilde{\sigma}_k \tag{65}$$

Proof. The second equality is simple to show.

$$\tilde{\mathbf{u}}_k^{\mathrm{H}}\mathbf{A}\tilde{\mathbf{v}}_k = \left(\frac{1}{\tilde{\sigma}_k}\right)\tilde{\mathbf{v}}_k^{\mathrm{H}}\mathbf{A}^{\mathrm{H}}\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}\tilde{\mathbf{v}}_k = \left(\frac{1}{\tilde{\sigma}_k}\right)\left\|\mathbf{W}^{\mathrm{H}}\mathbf{A}\tilde{\mathbf{v}}_k\right\|^2 \overset{(\zeta_1)}{=} \left(\frac{1}{\tilde{\sigma}_k}\right)\left\|\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}\tilde{\mathbf{v}}_k\right\|^2 = \tilde{\sigma}_k \tag{66}$$

(ζ_1) follows from noting $\mathbf{W}^{\mathrm{H}}\mathbf{W} = \mathbf{I}$. We will now show the first inequality.

$$\left\|\mathbf{A}^{\mathrm{H}}\tilde{\mathbf{u}}_k\right\| = \left(\frac{1}{\tilde{\sigma}_k}\right)\left\|\mathbf{A}^{\mathrm{H}}\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}\tilde{\mathbf{v}}_k\right\| = \left(\frac{1}{\tilde{\sigma}_k}\right)\left\|\mathbf{A}^{\mathrm{H}}\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}\tilde{\mathbf{v}}_k\right\| \tag{67}$$

Now, from the definition of the spectral norm, we have,

$$\text{Equation (67) RHS} = \left(\frac{1}{\tilde{\sigma}_k}\right) \max_{\mathbf{x}\in\mathbb{C}^n} \frac{\mathbf{x}^{\mathrm{H}}\mathbf{A}^{\mathrm{H}}\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}\tilde{\mathbf{v}}_k}{\mathbf{x}^{\mathrm{H}}\mathbf{x}} \tag{68}$$

From here, it is clear to choose $\mathbf{x} \triangleq \tilde{\mathbf{v}}_k$, as any vector in $\mathrm{Span}(\mathbf{U}_{k,\perp})$ will be orthogonal to $\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}\tilde{\mathbf{v}}_k$. Plugging this into Equation (68) gives us $\|\mathbf{A}^{\mathrm{H}}\mathbf{u}_k\| = \tilde{\sigma}_k$. We have proved both equalities and thus the proof is complete. □

Lemma 14. *Let $\tilde{\sigma}_k$ be the kth largest singular value of $\mathbf{W}\mathbf{W}^{\mathrm{H}}\mathbf{A}$, then we have*

$$\tilde{\sigma}_k \geq \sigma_k \left(1 - \sqrt{2}\sin\angle\left(\mathbf{u}_k, \tilde{\mathbf{u}}_k\right)\right) \vee 0 \tag{69}$$

Proof. From Lemma 13, we have $\|\tilde{\mathbf{u}}_k^{\mathrm{H}}\mathbf{A}\| = \tilde{\sigma}_k$. Then, we can expand this out and use the ideas discussed in the proof of Lemma 12.

$$\begin{aligned}
\|\mathbf{A}\|_{\mathrm{F}}^2 - \|\tilde{\mathbf{u}}_k^{\mathrm{H}}\mathbf{A}\|_{\mathrm{F}}^2 &= \left\|\mathbf{A} - \tilde{\mathbf{u}}_k\tilde{\mathbf{u}}_k^{\mathrm{H}}\mathbf{A}\right\|_{\mathrm{F}}^2 &(70)\\
&\overset{\text{lem. 19}}{\leq} \left\|\mathbf{A} - \tilde{\mathbf{u}}_k\tilde{\mathbf{u}}_k^{\mathrm{H}}\mathbf{A}_{(k)}\right\|_{\mathrm{F}}^2 &(71)\\
&\overset{\text{lem. 18}}{\leq} \left\|\mathbf{A}_{(k)} - \tilde{\mathbf{u}}_k\tilde{\mathbf{u}}_k^{\mathrm{H}}\mathbf{A}_{(k)}\right\|_{\mathrm{F}}^2 + \|\mathbf{A}_{\perp,k}\|_{\mathrm{F}}^2 &(72)\\
&\overset{\text{lem. 19}}{\leq} \left\|\mathbf{A}_{(k)} - \sigma_k\tilde{\mathbf{u}}_k\mathbf{v}_k\right\|_{\mathrm{F}}^2 + \|\mathbf{A}_{\perp,k}\|_{\mathrm{F}}^2 &(73)\\
&\leq \sigma_k^2 \left\|\mathbf{\Pi}_{\mathbf{u}_k} - \mathbf{\Pi}_{\tilde{\mathbf{u}}_k}\right\|_{\mathrm{F}}^2 + \|\mathbf{A}_{\perp,k}\|_{\mathrm{F}}^2 &(74)\\
&\overset{\text{lem. 10}}{=} 2\sigma_k^2 \sin^2\angle\left(\mathbf{u}_k, \tilde{\mathbf{u}}_k\right) + \|\mathbf{A}_{\perp,k}\|_{\mathrm{F}}^2 &(75)
\end{aligned}$$

Rearranging Equation (70) LHS and Equation (75) RHS and applying the triangle inequality completes the proof. □

C Concentration Inequalities

Lemma 15 (Hanson-Wright Inequality [Ada15])**.** *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 $\forall t \geq 0$,*

$$\mathbb{P}\left\{\left|\mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x} - \mathbb{E}\left[\mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x}\right]\right| \geq t\right\} \leq 2\exp\left(-c_3 \min\left\{\frac{t^2}{K^4\|\mathbf{B}\|_{\mathrm{F}}^2}, \frac{t}{K^2\|\mathbf{B}\|}\right\}\right) \tag{76}$$

where $K \triangleq \max_{i\in[n]} \|x_i\|_{\psi_2}$, where $\|\cdot\|_{\psi_2}$ represents the sub-Gaussian Norm [Ver18]. If $\mathbf{B} \in \mathbb{R}^{n\times n}$ such that $\mathbf{B} = \mathbf{B}^{\mathrm{T}}$, then $c_3 \geq 0.14$ [Mos21].

Lemma 16. *Let $X \sim \mathcal{N}(0,1)$, then it follows $\|X\|_{\psi_2} = \sqrt{\frac{8}{3}}$.*

Proof. We will first give the definition of a sub-Gaussian random variable X [RV13].

$$\|X\|_{\psi_2} \triangleq \inf_{\theta>0} \mathbb{E}\left[e^{(X/\theta)^2}\right] \leq 2 \tag{77}$$

We will expand the expectation using the density of the Gaussian.

$$\mathbb{E}\left[e^{(X/\theta)^2}\right] = \int_{-\infty}^{\infty} e^{(X/\theta)^2} \frac{1}{\sqrt{2\pi}} e^{-(X/\sqrt{2})^2} dX \tag{78}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(X^2\left(\frac{1}{\theta^2} - \frac{1}{2}\right)\right) dX \overset{(\zeta_1)}{=} \left(1 - \frac{2}{\theta^2}\right)^{-1/2} \tag{79}$$

Now we will optimize θ over the inequality.

$$\inf_{\theta>0} \left(1 - \frac{2}{\theta^2}\right)^{-1/2} \leq 2 \iff \inf_{\theta>0} \left(4 - \frac{8}{\theta^2}\right)^{1/2} \geq 1 \tag{80}$$

From here, we see $\theta = \sqrt{\frac{8}{3}}$ gives equality. This completes the proof. □

Lemma 17. *Let* $\mathbf{A} \in \mathbb{C}^{m\times n}$ *and* $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$*, then we have*

$$\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2 \geq \frac{16}{75\sqrt{5}} \left\|\mathbf{A}\right\|_{\mathrm{F}} \tag{81}$$

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

$$\frac{\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2}{\theta\sqrt{\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2^2}} \overset{(\zeta_1)}{\geq} \mathbb{P}\left\{\left\|\mathbf{A}\mathbf{x}\right\|_2 \geq \theta\sqrt{\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2^2}\right\} \tag{82}$$

$$= \mathbb{P}\left\{\left\|\mathbf{A}\mathbf{x}\right\|_2^2 \geq \theta\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2^2\right\} \tag{83}$$

$$\overset{(\zeta_2)}{\geq} \frac{(1-\theta^2)^2\left(\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2^2\right)^2}{\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2^4} \tag{84}$$

(ζ_1) follows from Markov's Inequality [Mar89]. (ζ_2) follows from the Paley-Zygmund Inequality [PZ32]. Rearranging the LHS of Equation (82) with RHS of Equation (84), we have

$$\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2 \geq \theta(1-\theta^2)^2 \frac{\left(\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2^2\right)^{5/2}}{\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2^4} \overset{(\zeta_3)}{\geq} \frac{16\left(\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2^2\right)^{5/2}}{25\sqrt{5}\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2^4} \tag{85}$$

where (ζ_3) follows from noting $\theta(1-\theta^2)^2$ is maximized at $\theta = \frac{1}{\sqrt{5}}$. Next we note $\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2^2 = \left\|\mathbf{A}\right\|_{\mathrm{F}}^2$. Furthermore, we have

$$\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2^4 = \mathbb{E}\left\|\mathbf{x}^{\mathrm{H}}\mathbf{A}^{\mathrm{H}}\mathbf{A}\mathbf{x}\right\|_2^2 = \mathrm{Tr}\left(\mathbf{A}^{\mathrm{H}}\mathbf{A}\right)^2 + 2\,\mathrm{Tr}\left(\left(\mathbf{A}^{\mathrm{H}}\mathbf{A}\right)^2\right) \leq \left\|\mathbf{A}\right\|_{\mathrm{F}}^4 + 2\left\|\mathbf{A}\right\|_{\mathrm{F}}^2\left\|\mathbf{A}\right\|_2^2 \tag{86}$$

$$= \left(\frac{\mathrm{sr}(\mathbf{A}) + 2}{\mathrm{sr}(\mathbf{A})}\right)\left\|\mathbf{A}\right\|_{\mathrm{F}}^4 \leq 3\left\|\mathbf{A}\right\|_{\mathrm{F}}^2 \tag{87}$$

Then we can substitute Equation (86) into Equation (85), and we have

$$\mathbb{E}\left\|\mathbf{A}\mathbf{x}\right\|_2 \overset{(86)}{\geq} \frac{16\left(\left\|\mathbf{A}\right\|_{\mathrm{F}}^2\right)^{5/2}}{75\sqrt{5}\left\|\mathbf{A}\right\|_{\mathrm{F}}^4} = \frac{16}{75\sqrt{5}}\left\|\mathbf{A}\right\|_{\mathrm{F}} \tag{88}$$

This concludes the proof. □

D Necessary Lemmas

Lemma 18 (Lemma 5.2 [BG13])**.** *If* $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m \times n}$ *and* $\mathbf{X}\mathbf{Y}^{\mathrm{T}} = \mathbf{0}$ *or* $\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbf{0}$*, then for both* $\xi = 2, \mathrm{F}$*, it follows*

$$\|\mathbf{X} + \mathbf{Y}\|_\xi^2 \leq \|\mathbf{X}\|_\xi^2 + \|\mathbf{Y}\|_\xi^2 \tag{89}$$

Lemma 19 (Lemma 5.3 [BG13])**.** *Given* $\mathbf{A} \in \mathbb{C}^{m \times n}$*,* $\mathbf{C} \in \mathbb{C}^{m \times r}$*, and for all* $\mathbf{X} \in \mathbb{C}^{m \times n}$ *and for both* $\xi = 2, \mathrm{F}$*, it holds*

$$\|\mathbf{A} - \mathbf{C}\mathbf{C}^{+}\mathbf{A}\|_\xi^2 \leq \|\mathbf{A} - \mathbf{C}\mathbf{X}\|_\xi^2 \tag{90}$$

Lemma 20 (Theorem 3.4 [Gu15])**.** *For any matrices,* $\mathbf{X}, \mathbf{Y} \in \mathbb{C}^{m \times n}$ *where* $\mathrm{rank}(\mathbf{Y}) \leq k$*, such that*

$$\|\mathbf{X} - \mathbf{Y}\|_{\mathrm{F}} \leq \sqrt{\eta^2 + \|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}^2} \tag{91}$$

for some $\eta \geq 0$*, then it follows*

$$\|\mathbf{X} - \mathbf{Y}\|_2 \leq \sqrt{\eta^2 + \|\mathbf{A} - \mathbf{A}_k\|_2^2} \tag{92}$$

E Additional Experiments

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

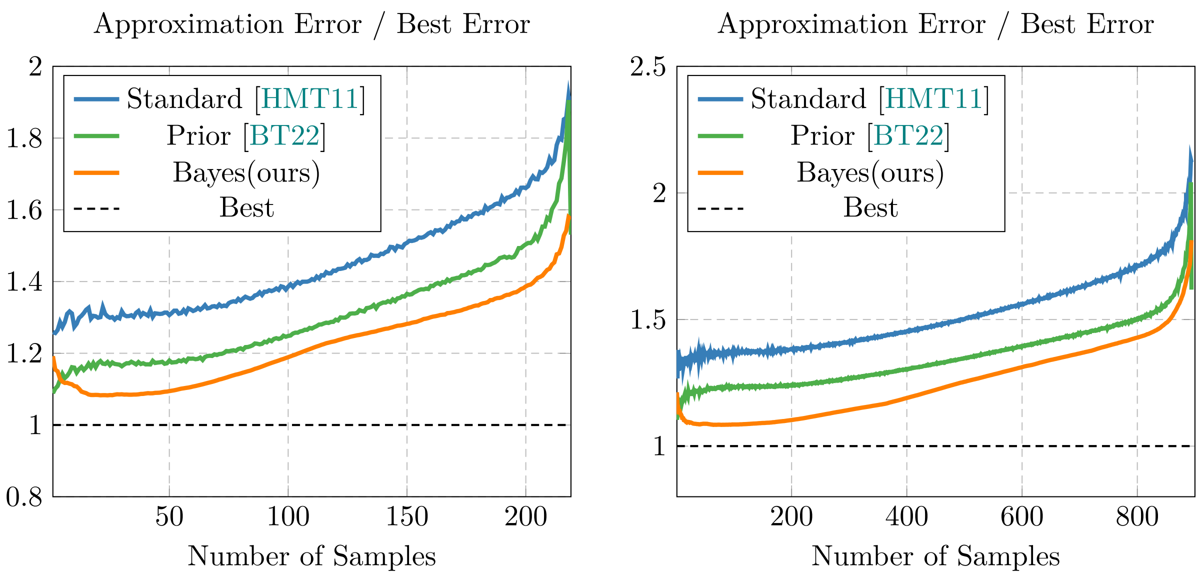

Figure 4: In *(Left)*, Matrix from TAMU Sparse Matrix Suite `pde 225`. In *(Right)*, Matrix from TAMU Sparse Matrix Suite `pde 900`. With the prior, we use the covariance matrix associated with the discrete Green's Function for the Laplacian as in Equation (11).

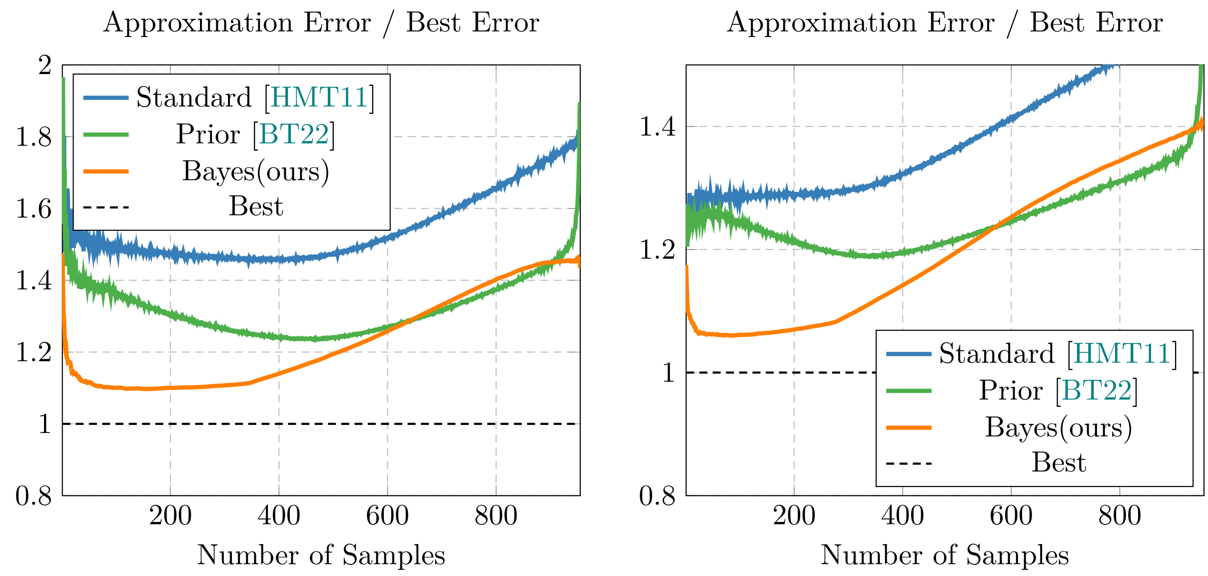

Figure 5: In these figures we look at matrices from Computational Fluid Dynamics. In *(Left)*, Matrix from TAMU Sparse Matrix Suite `cdde1`. In *(Right)*, Matrix from TAMU Sparse Matrix Suite `cdde1`. With the prior, we use the covariance matrix associated with the discrete Green's Function for the Laplacian as in Equation (11).

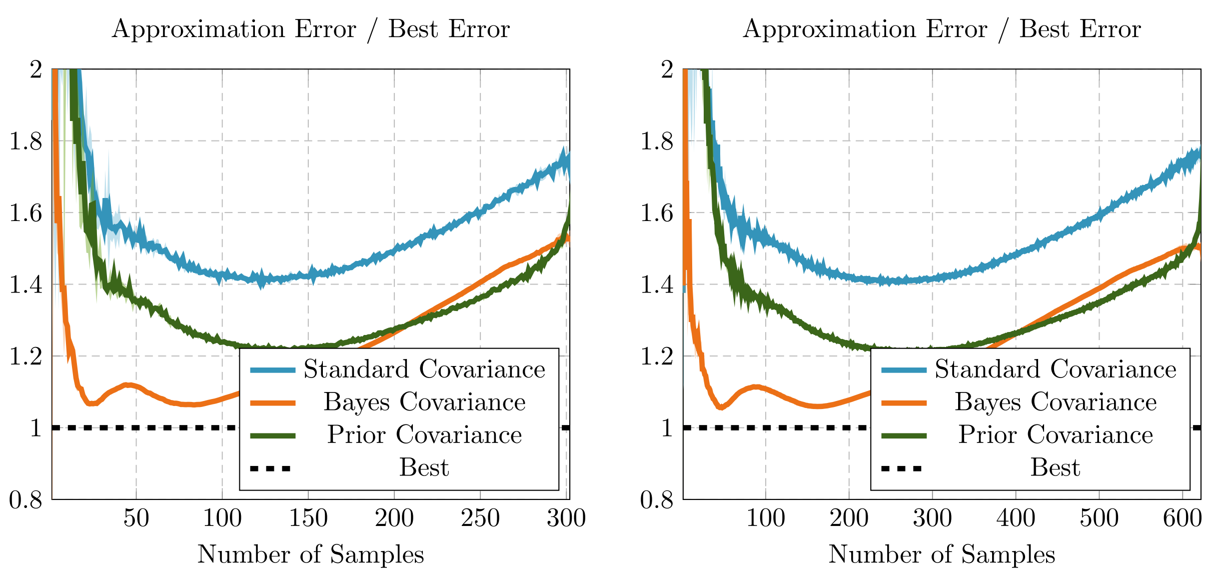

Figure 6: In these figures we look at matrices from the PDE for the Poisson Differential Operator. In *(Left)*, Matrix from TAMU Sparse Matrix Suite `cz308`. In *(Right)*, Matrix from TAMU Sparse Matrix Suite `cz628`. With the prior, we use the covariance matrix associated with the discrete Green's Function for the Laplacian as in Equation (11).