Data Driven Adaptive Sampling for Low-Rank Matrix Approximation

Arvind Rathnashyam* Nicolas Boullé† Alex Townsend ‡

August 29, 2023

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 [11]. Recently, [3] 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 [27] for eigenvector approximations from the randomized SVD. We also test our algorithm on various synthetic, real-world application, and image matrices. Furthermore, we show our theory bounds on matrices are stronger than state-of-the-art methods with the same number of matrix-vector product queries.

1 Introduction

In many real-world applications, it is often not possible to run experiments in parallel. 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, [11], has been theoretically analyzed and used in various applications. Even more recently, [2] discovered if we have prior information on the right singular vectors of $\mathbf{A}$, 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 [29] 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.

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.

*Math and CS, Rensselear Polytechnic Institute, rathna@rpi.edu
†Math, Cambridge, nb690@cam.cu.uk
‡Math, Cornell, ajt453@cornell.edu

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. $\mathbf{X}^\dagger = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$. The Projection Matrix is defined as $\Pi_{\mathbf{Y}} = \mathbf{Y}\mathbf{Y}^\dagger = \mathbf{Y}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1}\mathbf{Y}^{\mathrm{T}}$ 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 $\Pi_{\mathbf{Y}} = \mathbf{Y}\mathbf{Y}^{\mathrm{T}}$. 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. $\left\{\mathbf{V} : \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}_{k\times k}\right\}$. 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,

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

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.

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,

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

In a landmark paper by [7], they introduced upper bounds for $\|\sin\Theta(\mathcal{X}, \mathcal{Y})\|$ and $\|\tan\Theta(\mathcal{X}, \mathcal{Y})\|$. A generalized version of the $\sin\Theta$ theorem for rectangular matrices is given in [32].

Theorem 1. *[32]. 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*

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

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. *[30]. 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 $j \in 1, \ldots, m \vee n$,*

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

Theorem 3. *[23]. 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,*

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

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 $\cos\Theta(\mathbf{v}, \widetilde{\mathbf{v}}) = \mathbf{v}^{\mathrm{T}}\widetilde{\mathbf{v}}$. 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,

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

2.3 Relevant Works

The Randomized Singular Value Decomposition was developed and analyzed thoroughly in [11]; throughout this paper we will refer to this algorithm as HMT. The review work by [18] gives significant theory on the Randomized SVD. [3] 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. [4] extended upon previous work on generalizing the Randomized SVD to learning HS Operators. [14] empirically look at Entropy Search for Probabilistic Optimization. [24] analyzed faster algorithms for the approximation of the null-space.

Block iterative methods have also been studied extensively in [12]. The study of block Krylov subspaces have also seen increased attention in the last few years, [26]. Bounds for the $(1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|$ approximation error with randomized block Krylov Subspace methods have been explored in [21, 1].

The most relevant work to ours is likely [9]. 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.

Upper bounds on the tangent of principal angles are visited in [22] and improved in [19]. The highly studied $\sin\Theta$ norms are studied in depth in [6, 23].

Learning algorithms for Low-Rank Matrix Approximations have also been explored. In [16] and [15], the sketching matrix is learned.

A similar analysis of a power method is explored in [13] utilizing subspace perturbation theory. In this work, they consider the Matrix-Vector products have noise. In this work, similarly to [9], 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 [17].

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 [11], 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 $\mathbf{A}$. The update for the covariance matrix is given as follows:

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

Throughout this paper we will only consider $\mathbf{C}^{(0)} = \mathbf{I}$ due to simplified analysis and there is no empirical advantage in using a different initial Covariance Matrix. A similar algorithm can be found in [29]. 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.

Input: HS Operator: $\mathscr{F}$, Rank: r, Initial Covariance: $\mathbf{C}$, Oversampling Parameter: p
Output: Rank-r Approximation, $\widehat{\mathbf{A}}_r$
$\mathbf{X} \sim \underbrace{\left[\mathcal{N}(\mathbf{0}, \mathbf{C}) \overset{\text{i.i.d}}{\cdots} \mathcal{N}(\mathbf{0}, \mathbf{C})\right]}_{p}$
$\mathbf{Y} \leftarrow \mathbf{A}\mathbf{X}$
$\mathbf{Q}, \mathbf{R} \leftarrow \text{QR}(\mathbf{Y})$
for $k \in 1, 2, \ldots, r$ **do**
  $\left[\widetilde{\mathbf{U}}, \widetilde{\mathbf{\Sigma}}, \widetilde{\mathbf{V}}\right] \leftarrow \text{SVD}(\mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{A})$
  $\mathbf{C}^{(k+1)} \leftarrow \widetilde{\mathbf{V}}_{(:,k)}\widetilde{\mathbf{V}}_{(:,k)}^{\mathrm{T}}$
  $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}^{(k+1)})$
  $\mathbf{Y} \leftarrow [\mathbf{Y} \quad \mathbf{A}\mathbf{x}]$
  $\mathbf{Q}, \mathbf{R} \leftarrow \text{QR}(\mathbf{Y})$
end for
$\widetilde{\mathbf{A}}_r \leftarrow \mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{A}$
Return: $\widetilde{\mathbf{A}}_r$

Algorithm 1: Optimal Function Sampling

Remark 4. Let $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ 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 ?? 1. For efficient updates, we frame all operations as rank-1 updates.

In ?? 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 $\mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{A}$ 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 ?? 1 with the proofs in the appendix.

4.1 Setup.

We follow a similar setup as previous literature. Let rank $= \rho \leq n$, we will factorize $\mathbf{A}$ as

$$\begin{aligned}\mathbf{A} &= \begin{matrix} k & \rho-k \end{matrix} \begin{bmatrix} \mathbf{U}_k & \mathbf{U}_{\rho-k} \end{bmatrix} \begin{matrix} k \\ \rho-k \end{matrix} \begin{bmatrix} \mathbf{\Sigma}_k & \\ & \mathbf{\Sigma}_{\rho-k} \end{bmatrix} \begin{bmatrix} \mathbf{V}_k^{\mathrm{T}} \\ \mathbf{V}_{\rho-k}^{\mathrm{T}} \end{bmatrix} \begin{matrix} k \\ \rho-k \end{matrix} \\ &= \sum_{i=1}^{\rho} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}}\end{aligned} \tag{10}$$

Furthermore, we let $\mathbf{A}_{(k)} \triangleq \sigma_k \mathbf{u}_k \mathbf{v}_k^{\mathrm{T}}$. 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. *[11][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 $\mathbf{\Omega} \sim \mathcal{MN}(\mathbf{0}, \mathbf{I}_{n \times n}, \mathbf{I}_{k+p \times k+p})$, and $\mathbf{Q} \triangleq \operatorname{orth}(\mathbf{A\Omega})$. Then,*

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

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

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

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

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

$$\left(1 + \underbrace{\left\| \mathbf{\Sigma}_{\rho-k} \left(\mathbf{V}_{\rho-k}^{\mathrm{T}} \mathbf{\Omega}\right) \left(\mathbf{V}_k^{\mathrm{T}} \mathbf{\Omega}\right)^{\dagger} \right\|}_{\psi}\right)^{1/2} \left\| \mathbf{\Sigma}_{\rho-k} \right\|_{\mathrm{F}} \tag{13}$$

See Theorems 5 and 6 and [5, 10]. However, we find working with $\left\|(\mathbf{V}_k^{\mathrm{T}}\mathbf{\Omega})^{\dagger}\right\|$ is difficult since this norm can be extremely large. The 'ψ' term in Equation (13) has been studied w.r.t to Krylov sSubspaces in [9]. In [9][Theorem 2.2], Drineas et al. find

$$\psi \leq \left\| \tan \mathbf{\Theta}\left(\widetilde{\mathbf{V}}, \mathbf{V}_k\right) \right\| \tag{14}$$

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 $\mathbf{C}^{(k)}$.

Lemma 7. *Let $\widehat{\mathbf{Q}}_k \widehat{\mathbf{Q}}_k \mathbf{A}$ be the rSVD approximation for $\mathbf{A}$. Then for $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \widehat{\mathbf{V}}_{(:,k)} \widehat{\mathbf{V}}_{(:,k)}^{\mathrm{T}})$, it follows*

$$\mathbf{x} = \alpha \widehat{\mathbf{V}}_{(:,k)}, \quad \alpha \sim \mathcal{N}(0, 1) \tag{15}$$

Lemma 7 follows due to the Cholesky Factorization of the Covariance Matrix. 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.

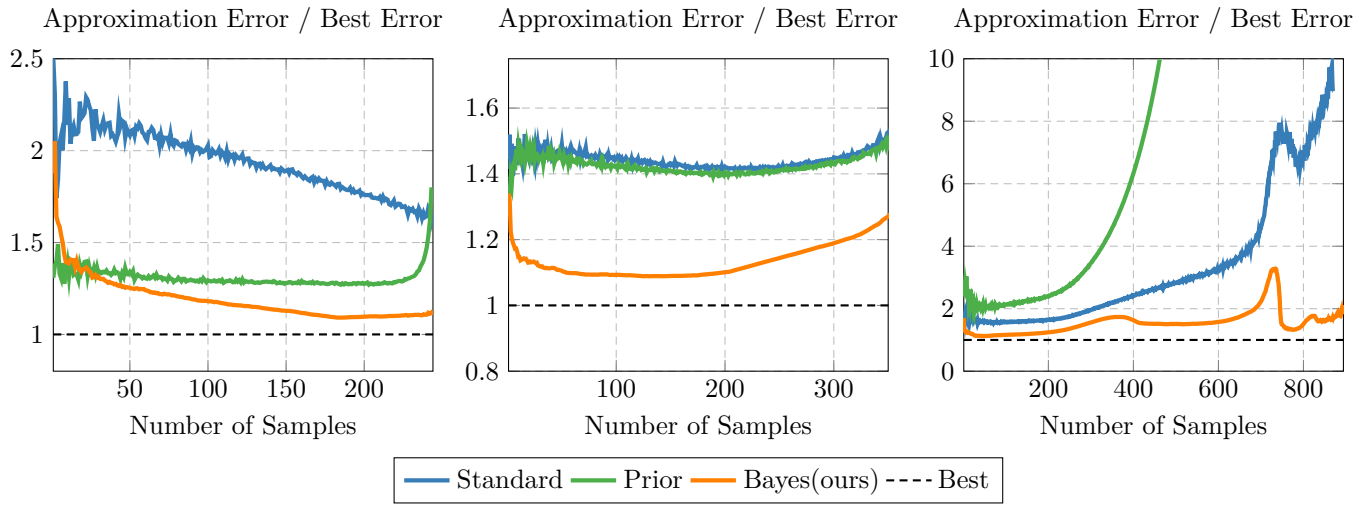

Figure 1: Low Rank Approximation for the Inverse Differential Operator given in Equation (25) *(Left)*, Differential Operator Matrix `Poisson2D` [8] *(Center)*, and Differential Operator Matrix `DK01R` [8] *(Right)*. The experiment on the left is from [3] (Figure 2).

4.3 Deterministic Bounds

First, we introduce a necessary result.

Lemma 8. *Let $\mathbf{V}_{\rho-k}$ be the last $\rho-k$ right singular vectors of $\mathbf{A}$ and let $\widetilde{\mathbf{V}}_k$ be the k orthonormal adaptively sampled vectors, then*

$$\left\|\mathbf{V}_{\rho-k}\widetilde{\mathbf{V}}_k\right\|_{\mathrm{F}} \leq \left(\sum_{i=1}^{k} \sin\Theta\left(\mathbf{v}_i, \widetilde{\mathbf{v}}_i\right)\right)^{1/2} \tag{16}$$

Now, we will come to our main results.

Lemma 9. *Let $\mathbf{A}$ have singular values $\sigma_1 \geq \ldots \geq \sigma_{m\vee n}$ and $\widetilde{\mathbf{A}}_k$ be the rank-k approximation from ?? 1 with oversampling parameter p. Let $\mathbf{Q} \triangleq \mathrm{orth}(\mathbf{AX}) = \mathrm{orth}(\mathbf{A}\left[\widetilde{\mathbf{v}}_1 \quad \ldots \quad \widetilde{\mathbf{v}}_k\right])$. Then,*

$$\begin{aligned}
&\left\|\mathbf{A} - \widetilde{\mathbf{Q}}\widetilde{\mathbf{Q}}^{\mathrm{T}}\mathbf{A}\right\|_{\mathrm{F}} \\
&\leq \left(\left\|\sin\boldsymbol{\Theta}(\mathbf{U}_k, \widetilde{\mathbf{U}})\boldsymbol{\Sigma}_k\right\|_{\mathrm{F}}^2 - (k-1)\left\|\boldsymbol{\Sigma}_k\right\|_{\mathrm{F}}^2\right)^{1/2} \\
&+ \sigma_{k+1}\left(\sum_{i=1}^{k}\left\|\sin\Theta\left(\mathbf{u}_i, \widetilde{\mathbf{u}}_i\right)\right\|_{\mathrm{F}}^2\right)^{1/2} + \left\|\boldsymbol{\Sigma}_{\rho-k}\right\|_{\mathrm{F}}
\end{aligned} \tag{17}$$

Remark 10. If we have a completely flat spectrum, then for any algorithm which produces orthonormal vectors of $\widetilde{\mathbf{V}}$ will give an optimal approximation.

Now we are interested in the sines of the angles between the sampled vectors and the right singular vectors.

4.4 Expected Bounds

Now we will analyze the expected approximation error bounds.

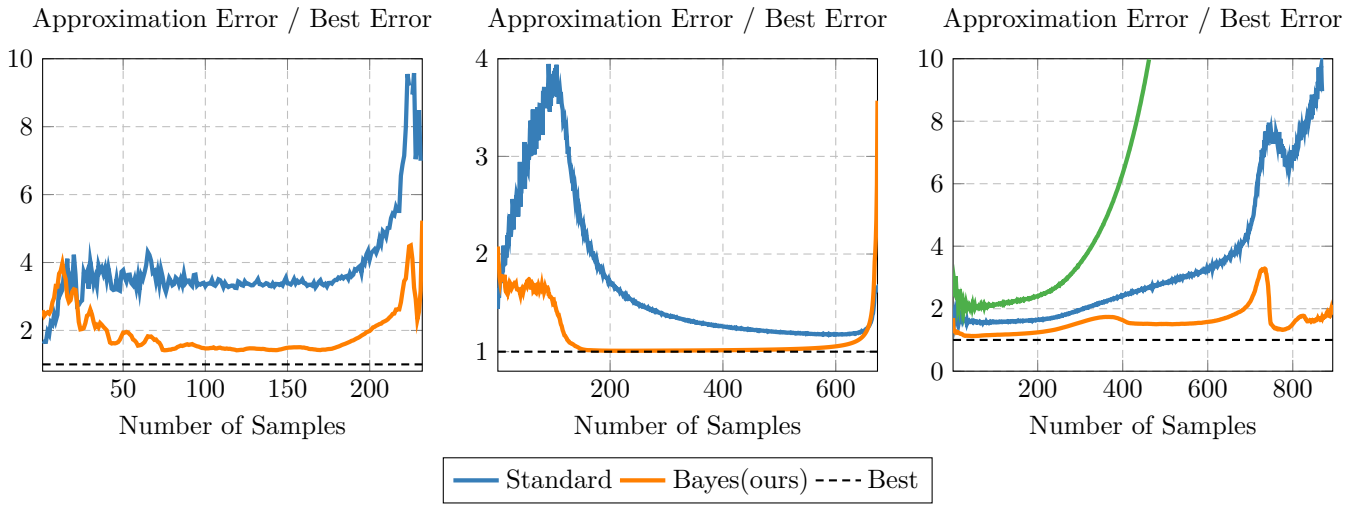

Figure 2: Low Rank Approximation for a matrix for a Computational Fluid Dynamics Problem, `saylr1` *(Left)*from [8]. Subsequent 2D/3D Problem `fs-680-2` *(Center)*from [8].

Lemma 11. *Let p be our predetermined oversampling parameter, then let $\mathbf{\Omega} \in \mathbb{R}^{n \times p}$ be a standard gaussian matrix. Define $\mathbf{Q} \triangleq \text{orth}(\mathbf{A\Omega})$, let $\widetilde{\mathbf{v}}_1$ be the first singular vector of the approximation $\mathbf{QQ}^{\mathrm{T}}\mathbf{A}$, then*

$$\sin\Theta\left(\mathbf{u}_1, \widetilde{\mathbf{u}}_1\right) \leq \left(1 + \left(\left(\frac{\sigma_1}{\sigma_2}\right)\cot\left(\mathbf{v}_1, \widetilde{\mathbf{v}}_1\right)\right)^2\right)^{-1/2} \tag{18}$$

The proof is deferred to § **??**. We now have an upper bound that depends upon the spectral gap for the dominant left singular vector. This bound is dependent on both the spectral gap and the spectral decay. When we have a flat spectrum, the denominator is minimized and our approximation of the left singular vector will be at it's worst. Now we will extend this lemma to the the j-th right singular vector.

Lemma 12. *Consider the same setup in Lemma 11, with $j-1$ vectors sampled as described in ?? 1, then*

$$\sin\Theta\left(\mathbf{u}_j, \widetilde{\mathbf{u}}_j\right) \leq \left(1 + \left(\left(\frac{\sigma_1}{\sigma_j}\right)\cot\left(\mathbf{v}_j, \widetilde{\mathbf{v}}_j\right)\right)^2\right)^{-1/2} \tag{19}$$

The proof is deferred to § **??**. Now we will introduce the most important theorem of our work. From Lemma 11, we now have an idea on the effect of sampling the eigenvector approximations.

Lemma 13. *Let p be our predetermined oversampling parameter, then let $\mathbf{\Omega} \in \mathbb{R}^{n \times p}$ be a standard gaussian matrix. Define $\mathbf{Q} \triangleq \text{orth}(\mathbf{A\Omega})$, let $\widetilde{\mathbf{v}}_1$ be the first singular vector of the approximation $\mathbf{QQ}^{\mathrm{T}}\mathbf{A}$, then*

$$\sin\Theta\left(\mathbf{v}_1, \widetilde{\mathbf{v}}_1\right) \leq \Xi \tag{20}$$

Lemma 14. *Consider the same setup in Lemma 13, with $j-1$ vectors sampled as described in ?? 1, then*

$$\sin\Theta\left(\mathbf{v}_j, \widetilde{\mathbf{v}}_j\right) \leq \Xi \tag{21}$$

Lemma 15. *Let $\widetilde{\mathbf{V}}$ be the set of vectors formed as the adaptive samples by ?? 1, it then follows*

$$\left\| \left(\widetilde{\mathbf{V}}^{\mathrm{T}}\mathbf{V}_k\right)^{\dagger} \right\| \leq \Xi \tag{22}$$

We will connect this together with the error bounds of sampling k right singular vector approximations. With Lemma 9 and Lemma 11, we have the following theorem.

Theorem 16. *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 $\widetilde{\mathbf{A}}_k$ be the matrix returned by ?? 1. Then,*

$$\begin{aligned}\mathbb{E}\left\|\mathbf{A}-\widetilde{\mathbf{A}}_k\right\|_{\mathrm{F}} &\leq \sum_{i=1}^{k}\sigma_i\left(\frac{\sigma_{i+1}}{\sigma_i}\right)\\ &+\sigma_{k+1}\left(\sum_{i=1}^{k}\left(\frac{\sigma_{i+1}}{\sigma_i}\right)\right)+\sqrt{\sum_{j=k+1}^{n}\sigma_j}\end{aligned} \tag{23}$$

Proof. This theorem follows from the induction formed in Lemma 11 and Lemma 12 and plugging this in to Lemma 9. ■

4.5 Probabilistic Bounds

Corollary 17. *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 $\widetilde{\mathbf{A}}_k$ be the matrix returned by ?? 1. Then,*

$$\left\|\mathbf{A}-\widetilde{\mathbf{A}}_k\right\|_{\mathrm{F}} \leq \sigma_{k+1}+\sqrt{\sum_{j=1}^{k}\sigma_j^2}+\sqrt{\sum_{j=k+1}^{n}\sigma_j^2} \tag{24}$$

Proof. This follows from a simple upper bound on the multiplicative term on σ_{k+1} in Lemma 9 by using the fact $\sin\Theta(\mathbf{x},\mathbf{y}) \leq 1$ for all $\mathbf{x},\mathbf{y}$. ■

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 to learn the discretized 250×250 matrix of the inverse of the following differential operator:

$$\mathcal{L}u = \frac{\partial^2 u}{\partial x^2} - 100\sin(5\pi x)\,u, \qquad x \in [0,1] \tag{25}$$

In Figure 1*(Right)*, nsote 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, [8]. The synthetic matrix is developed in the following scheme:

$$\mathbf{A} = \sum_{i=1}^{\rho}\frac{100i^{\ell}}{n}\mathbf{U}_{(:,i)}\mathbf{V}_{(i,:)}^{\mathrm{T}}, \quad \mathbf{U} \in \mathbb{O}_{m,k}, \mathbf{V} \in \mathbb{O}_{n,k} \tag{26}$$

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, $\boldsymbol{\Omega}$ in the Randomized SVD algorithm. Our covariance update for generating sampling vectors and functions can find use various PDE learning applications, [2]. 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.

References

[1] Ainesh Bakshi, Kenneth L. Clarkson, and David P. Woodruff. Low-rank approximation with $1/\varepsilon^3$ matrix-vector products. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2022, page 1130–1143, New York, NY, USA, 2022. Association for Computing Machinery.

[2] Nicolas Boullé, Christopher J. Earls, and Alex Townsend. Data-driven discovery of green's functions with human-understandable deep learning. *Scientific Reports*, 12(1):4824, Mar 2022.

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

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

[5] 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.

[6] 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.

[7] Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

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

[21] 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.

[22] Yuji Nakatsukasa. The tan θ theorem with relaxed conditions. *Linear Algebra and its Applications*, 436(5):1528–1534, 2012.

[23] 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.

[24] Taejun Park and Yuji Nakatsukasa. A fast randomized algorithm for computing an approximate null space. *BIT Numerical Mathematics*, 63(2):36, May 2023.

[25] Erhard Schmidt. Zur theorie der linearen und nichtlinearen integralgleichungen. *Mathematische Annalen*, 63(4):433–476, Dec 1907.

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

[27] 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.

[28] Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, page 210–268. Cambridge University Press, 2012.

[29] 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.

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

[31] Franco Woolfe, Edo Liberty, Vladimir Rokhlin, and Mark Tygert. A randomized algorithm for the approximation of matrices. 2006.

[32] 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. Full publication date: JUNE 2015.

A Deferred Proofs of Main Results

A.1 Proof of Lemma 8

Let $\widetilde{\mathbf{V}}_k$ be the orthonormal basis for the adaptively sampled vectors.

$$\zeta_1 \triangleq \left\|\mathbf{V}_{\rho-k}^{\mathrm{T}}\widetilde{\mathbf{V}}\right\|_{\mathrm{F}} = \left(\sum_{i=1}^{k}\sum_{j=k+1}^{\rho}\left\|\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{v}_j\right\|^2\right)^{1/2} = \left(\sum_{i=1}^{k}\sum_{j=1}^{n}\left\|\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{v}_j\right\|^2 - \sum_{i=1}^{k}\sum_{j=1}^{k}\left\|\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{v}_j\right\|^2\right)^{1/2} \tag{27}$$

$$= \left(\sum_{i=1}^{k}\sum_{j=1}^{n}\mathbb{1}_{i\neq j}\left\|\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{v}_j\right\|^2 - \sum_{i=1}^{k}\sum_{j=1}^{k}\left\|\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{v}_j\right\|^2 + \sum_{i=1}^{k}\left\|\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{v}_j\right\|^2\right)^{1/2} \tag{28}$$

$$= \left(\left(\sum_{i=1}^{k}\sin^2\Theta\left(\mathbf{v}_i,\widetilde{\mathbf{v}}_i\right) + \cos^2\Theta\left(\mathbf{v}_i,\widetilde{\mathbf{v}}_i\right)\right) - \sum_{i=1}^{k}\sum_{j=1}^{k}\left\|\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{v}_j\right\|^2\right)^{1/2} \tag{29}$$

$$= \left(k - \sum_{i=1}^{k}\sum_{j=1}^{k}\left\|\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{v}_j\right\|^2\right)^{1/2} = \left(k - \sum_{i=1}^{k}\left\|\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{v}_i\right\|^2 - \sum_{i=1}^{k}\sum_{j=1}^{k}\mathbb{1}_{i\neq j}\left\|\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{v}_j\right\|^2\right)^{1/2} \tag{30}$$

$$= \left(\sum_{i=1}^{k}\sin^2\Theta\left(\mathbf{v}_i,\widetilde{\mathbf{v}}_i\right) - \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{k}\mathbb{1}_{i\neq j}\left\|\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{v}_j\right\|^2}_{\beta}\right)^{1/2} \leq \left(\sum_{i=1}^{k}\sin^2\Theta\left(\mathbf{v}_i,\widetilde{\mathbf{v}}_i\right)\right)^{1/2} \tag{31}$$

Here we have the desired result. ■

In Equation (31), when we use the approximation below, we lose the β term, is there any way to lower bound it?

A.2 Proof of Lemma 9

Proof. We will utilize the fact that when working with projection we do not have to account for the orthogonalization of the sampled vectors from the previous eigenvector approximations. First, we define $\widetilde{\mathbf{U}} \triangleq \operatorname{orth}(\mathbf{A}\widetilde{\mathbf{V}})$.

$$\left\|\mathbf{A} - \widetilde{\mathbf{A}}_k\right\|_{\mathrm{F}} \triangleq \underbrace{\left\|\mathbf{A} - \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{A}\right\|_{\mathrm{F}}}_{\xi_1} \tag{32}$$

We will upper bound the square.

$$\left\|\mathbf{A} - \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{A}\right\|_{\mathrm{F}}^2 = \operatorname{Tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{A} - \mathbf{A}^{\mathrm{T}}\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{A} - \mathbf{A}^{\mathrm{T}}\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{A} + \mathbf{A}^{\mathrm{T}}\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^{\mathrm{T}}\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{A}\right) \tag{33}$$

$$= \operatorname{Tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{A} - \mathbf{A}^{\mathrm{T}}\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{A} - \mathbf{A}^{\mathrm{T}}\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{A} + \mathbf{A}^{\mathrm{T}}\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{A}\right) \tag{34}$$

$$= \|\mathbf{A}\|_{\mathrm{F}}^2 - \left\|\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{A}\right\|_{\mathrm{F}}^2 = \|\mathbf{\Sigma}\|_{\mathrm{F}}^2 - \operatorname{Tr}\left(\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}\mathbf{V}\mathbf{\Sigma}\mathbf{U}^{\mathrm{T}}\widetilde{\mathbf{U}}\right) \tag{35}$$

$$= \|\mathbf{\Sigma}\|_{\mathrm{F}}^2 - \left\|\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{U}\mathbf{\Sigma}\right\|_{\mathrm{F}}^2 = \sum_{i=1}^{n}\sigma_i^2 - \sum_{i=1}^{n}\sigma_i^2\sum_{j=1}^{k}\cos^2\Theta\left(\mathbf{u}_i,\widetilde{\mathbf{u}}_j\right) \tag{36}$$

$$= \sum_{i=1}^{n}\sigma_i^2\left(1 - \sum_{j=1}^{k}\cos^2\Theta\left(\mathbf{u}_i,\widetilde{\mathbf{u}}_j\right)\right) = \sum_{i=1}^{n}\sigma_i^2\left(1 - \left\|\cos\mathbf{\Theta}\left(\mathbf{U}_{(i)},\widetilde{\mathbf{U}}_k\right)\right\|_{\mathrm{F}}^2\right) \tag{37}$$

$$\overset{\text{lem. 19}}{=} \sum_{i=1}^{n}\sigma_i^2\left(\left\|\sin\mathbf{\Theta}\left(\mathbf{U}_{(i)},\widetilde{\mathbf{U}}_k\right)\right\|_{\mathrm{F}}^2 - k\right) \tag{38}$$

$$= \sum_{i=1}^{n} \sigma_i^2 \left(\frac{1}{2} \underbrace{\left\| \widetilde{\mathbf{U}}_k \widetilde{\mathbf{U}}_k^{\mathrm{T}} - \mathbf{U}_{(i)} \mathbf{U}_{(i)}^{\mathrm{T}} \right\|_{\mathrm{F}}^2}_{\xi_1} - k \right) \tag{39}$$

We will now upper bound ξ_1

$$\xi_1 \triangleq \left\| \widetilde{\mathbf{U}}_k \widetilde{\mathbf{U}}_k^{\mathrm{T}} - \mathbf{U}_{(i)} \mathbf{U}_{(i)}^{\mathrm{T}} \right\|_{\mathrm{F}}^2 = \left\| \widetilde{\mathbf{U}}_k \widetilde{\mathbf{U}}_k^{\mathrm{T}} - \mathbf{U}_i \mathbf{U}_i^{\mathrm{T}} + \mathbf{U}_i \mathbf{U}_i^{\mathrm{T}} - \mathbf{U}_{(i)} \mathbf{U}_{(i)}^{\mathrm{T}} \right\|_{\mathrm{F}}^2 \tag{40}$$

$$\leq \left\| \widetilde{\mathbf{U}}_k \widetilde{\mathbf{U}}_k^{\mathrm{T}} - \mathbf{U}_i \mathbf{U}_i^{\mathrm{T}} \right\|_{\mathrm{F}}^2 + \left\| \mathbf{U}_{i-1} \mathbf{U}_{i-1}^{\mathrm{T}} \right\|_{\mathrm{F}}^2 \tag{41}$$

$$\leq \left\| \mathbf{A}\mathbf{V}_i \left(\mathbf{A}\mathbf{V}_i \right)^{\dagger} - \mathbf{A} \left(\mathbf{V}_k - \mathbf{V}_k + \widetilde{\mathbf{V}}_k \right) \left(\mathbf{A} \left(\mathbf{V}_k - \mathbf{V}_k + \widetilde{\mathbf{V}}_k \right) \right)^{\dagger} \right\|_{\mathrm{F}}^2 + (i-1) \tag{42}$$

In Figure 3, we see the upper bound given in **??** is good.

■

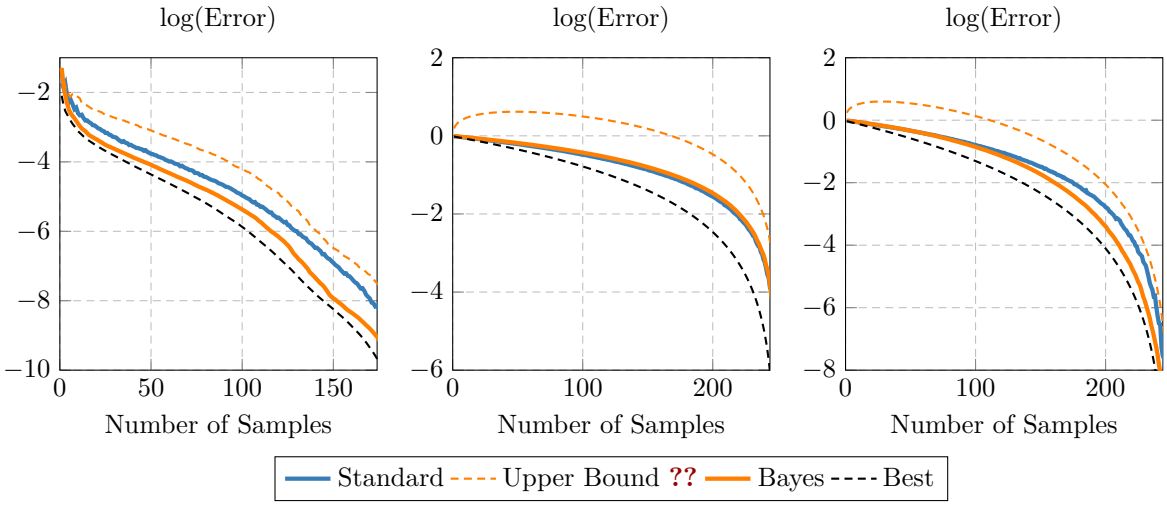

Figure 3: Low Rank Approximation Error Ratios for the Synthetic Matrix described in Equation (26) with $\ell = 1$ *(Center)*and $\ell = 2$ *(Right)*and Mona Lisa image *(Left)*. Upper bound is given in **??**.

B Singular Subspace Perturbation Lemmas

Lemma 18. *Let* $\mathbf{v}$ *and* $\widetilde{\mathbf{v}}$ *be vectors s.t.* $\|\mathbf{v}\| = \|\widetilde{\mathbf{v}}\| = 1$ *and* $\mathbf{v}^{\mathrm{T}}\widetilde{\mathbf{v}} \geq 0$*. Then,*

$$\|\mathbf{v} - \widetilde{\mathbf{v}}\| \leq \sqrt{2} \sin \Theta\left(\mathbf{v}, \widetilde{\mathbf{v}}\right) \tag{43}$$

Proof.

$$\sin^2 \Theta\left(\mathbf{v}, \widetilde{\mathbf{v}}\right) = 1 - \left(\mathbf{v}^{\mathrm{T}}\widetilde{\mathbf{v}}\right)^2 \overset{(a)}{\geq} 1 - \mathbf{v}^{\mathrm{T}}\widetilde{\mathbf{v}} = 1 + \frac{1}{2}\|\mathbf{v} - \widetilde{\mathbf{v}}\|^2 - \frac{1}{2}\|\mathbf{v}\|^2 - \frac{1}{2}\|\widetilde{\mathbf{v}}\|^2 = \frac{1}{2}\|\mathbf{v} - \widetilde{\mathbf{v}}\|^2 \tag{44}$$

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

Lemma 19. *Let* $\mathbf{P}$ *and* $\mathbf{Q}$ *be projection matrices s.t.* $\mathbf{P} \triangleq \mathbf{U}\mathbf{U}^{\dagger}$ *and* $\mathbf{Q} \triangleq \mathbf{V}\mathbf{V}^{\dagger}$*, then*

$$\|\cos \mathbf{\Theta}\left(\mathbf{U}, \mathbf{V}\right)\|_{\mathrm{F}}^2 = \frac{\mathrm{rank}(\mathbf{U}) + \mathrm{rank}(\mathbf{V})}{2} - \|\sin \mathbf{\Theta}\left(\mathbf{U}, \mathbf{V}\right)\|_{\mathrm{F}}^2 \tag{45}$$

Proof.

$$\|\sin \mathbf{\Theta}\left(\mathbf{U}, \mathbf{V}\right)\|_{\mathrm{F}}^2 = \frac{1}{2}\|\Pi_{\mathbf{U}} - \Pi_{\mathbf{V}}\|_{\mathrm{F}}^2 = \frac{1}{2}\mathrm{Tr}\left(\Pi_{\mathbf{U}}^{\mathrm{T}}\Pi_{\mathbf{U}} - \Pi_{\mathbf{U}}^{\mathrm{T}}\Pi_{\mathbf{V}} - \Pi_{\mathbf{V}}^{\mathrm{T}}\Pi_{\mathbf{U}} + \Pi_{\mathbf{V}}^{\mathrm{T}}\Pi_{\mathbf{V}}\right) \tag{46}$$

$$= \frac{1}{2}\left(\mathrm{rank}(\mathbf{U}) + \mathrm{rank}(\mathbf{V}) - 2\,\mathrm{Tr}\left(\Pi_{\mathbf{U}}^{\mathrm{T}}\Pi_{\mathbf{V}}\right)\right) \tag{47}$$

$$= \frac{1}{2}\left(\mathrm{rank}(\mathbf{U}) + \mathrm{rank}(\mathbf{V})\right) - \|\cos \mathbf{\Theta}\left(\mathbf{U}, \mathbf{V}\right)\|_{\mathrm{F}}^2 \tag{48}$$

Subtract the sine term from both sides and add the cosine terms to both sides and the proof is complete. ■

Auxiliary Lemmas

Lemma 20. *[20]. For any matrices* $\mathbf{A}$, $\mathbf{B}$, *and* $\mathbf{C}$, *then for any unitarily invariant norm* $|||\cdot|||$, *it follows*

$$|||\mathbf{ABC}||| \leq \min\left\{|||\mathbf{A}|||\,\|\mathbf{B}\|_2\,\|\mathbf{C}\|_2, \|\mathbf{A}\|_2\,|||\mathbf{B}|||\,\|\mathbf{C}\|_2, \|\mathbf{A}\|_2\,\|\mathbf{B}\|_2\,|||\mathbf{C}|||\right\} \tag{49}$$

Lemma 21. *[11]. Let* Π *be the projection operator and* $\mathbf{Q} \triangleq \operatorname{orth}(\mathbf{Y})$, *it then follows*

$$\left|||\mathbf{A} - \mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{A}|||\right| = |||\left(\mathbf{I} - \Pi_{\mathbf{Y}}\right)\mathbf{A}|||\tag{50}$$

Lemma 22. *[31][Lemma 6]. Let* $m, n \in \mathbb{N}$ *s.t.* $n \geq m$. *Suppose* $\mathbf{A} \in \mathbb{R}^{n\times m}$, *then if* $(\mathbf{A}^{\mathrm{T}}\mathbf{A})$ *is invertible*

$$\left\|\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathrm{T}}\right\| = \frac{1}{\sigma_m}\tag{51}$$

Lemma 23. *[5][Lemma 5.3]. Given* $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{C} \in \mathbb{R}^{m\times r}$, *and for all* $\mathbf{X} \in \mathbb{R}^{r\times n}$ *and for* $\xi = 2, \mathrm{F}$

$$\left\|\mathbf{A} - \mathbf{C}\mathbf{C}^{\dagger}\mathbf{A}\right\|_{\xi}^2 \leq \|\mathbf{A} - \mathbf{C}\mathbf{X}\|_{\xi}^2\tag{52}$$

Lemma 24. *[28][Theorem 5.32]. Given a standard normal matrix* $\mathbf{G} \in \mathbb{R}^{m\times n}$, *then*

$$\mathbb{E}\left[\sigma_1(\mathbf{G})\right] \leq \left(\sqrt{m} + \sqrt{n}\right)^2\tag{53}$$