Data-Driven Sampling for Low-Rank Matrix Approximation

Arvind Rathnashyam [1] **Nicolas Boullé** [2] **Alex Townsend** [3]

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 (Halko et al., 2011b). Recently, (Boullé & Townsend, 2022) 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 (Tzeng et al., 2022) 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, (Halko et al., 2011b), has been theoretically analyzed and used in various applications. Even more recently, (Boullé et al., 2022) 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 (Wang et al., 2021) 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.

Contributions.

1. We develop a novel adaptive sampling algorithm for Low-Rank Matrix Approximation problem in the matrix-vector product model which does not utilize prior information of $\mathbf{A}$.
2. We provide a novel theoretical analysis which utilizes subspace perturbation theory.
3. We perform extensive experiments on matrices with various spectrums and compare with the state of the art methods.

2 Notation, Background Materials, and Relevant Work

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

2.1 Notation

Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ represent the target matrix. $\|\cdot\|$ represents the spectral norm, which is equivalent to the max eigenvalue

[1]Department of Mathematics, Rensselaer Polytechnic Institute, Troy, NY, USA [2]Isaac Newton Institute, Cambridge, UK [3]Cornell University, Departments of Mathematics, Ithaca, NY, USA. Correspondence to: Arvind Rathnashyam <rathna@rpi.edu>.

Lemma 4.7, Theorem 4.8, can be taken as conjectures supported by numerical experiments and the proofs are still in progress.

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,

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

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,

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

In a landmark paper by (Davis & Kahan, 1970), 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 (Yu et al., 2015).

Theorem 2.1. *(Yu et al., 2015). 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*

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

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

Theorem 2.2. *(Wedin, 1972). 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 $j \in 1, \ldots, m \vee n$,*

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

Theorem 2.3. *(O'Rourke et al., 2018). 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,*

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

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,

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

2.3 Relevant Works

The Randomized Singular Value Decomposition was developed and analyzed thoroughly in (Halko et al., 2011b); throughout this paper we will refer to this algorithm as HMT. The review work by (Martinsson & Tropp, 2020) gives significant theory on the Randomized SVD. (Boullé & Townsend, 2022) 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. (Boullé & Townsend, 2023) extended upon previous work on generalizing the Randomized SVD to learning HS Operators. (Hennig & Schuler, 2012) empirically look at Entropy Search for Probabilistic Optimization. (Park & Nakatsukasa, 2023) analyzed faster algorithms for the approximation of the null-space.

Block iterative methods have also been studied extensively in (Halko et al., 2011a). The study of block Krylov subspaces have also seen increased attention in the last few years, (Tropp & Webber, 2023). Bounds for the $(1 + \varepsilon)\|\mathbf{A} - \mathbf{A}_k\|$ approximation error with randomized block Krylov Subspace methods have been explored in (Musco & Musco, 2015; Bakshi et al., 2022).

The most relevant work to ours is likely (Drineas et al., 2018). 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 (Nakatsukasa, 2012) and improved in (Massey et al., 2020). The highly studied $\sin\Theta$ norms are studied in depth in (Cai & Zhang, 2018; O'Rourke et al., 2018).

Learning algorithms for Low-Rank Matrix Approximations have also been explored. In (Indyk et al., 2021) and (Indyk et al., 2019), the sketching matrix is learned.

A similar analysis of a power method is explored in (Hardt & Price, 2014) utilizing subspace perturbation theory. In this work, they consider the Matrix-Vector products have noise. In this work, similarly to (Drineas et al., 2018), 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 (Luo et al., 2021).

3 Data Driven Sampling

We will first give some relevant definitions. In this section we will give the update formula for the Covariance Matrix after each iteration. The update for the covariance matrix is given as follows:

$$\mathbf{C}^{(k+1)} = \widetilde{\mathbf{V}}_{(:,k)} \widetilde{\mathbf{V}}_{(:,k)}^{\mathrm{T}} \tag{7}$$

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 (Wang et al., 2021). To motivate our covariance update, we will introduce the following remark.

Remark 3.1. Let $\mathbf{U}\mathbf{\Sigma}\mathbf{V}$ be the SVD of $\mathbf{A}$, then the Covariance update described in Equation (7) is the optimal covariance update is the optimal covariance matrix for sampling vectors at iteration k.

Remark 3.1 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. 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 eigenvector.

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 Optimal Function Sampling

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

4.1 Setup.

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

$$\begin{aligned} \mathbf{A} &= \begin{matrix} k & \rho-k \end{matrix} \begin{bmatrix} \mathbf{U}_k & \mathbf{U}_{\rho-k} \end{bmatrix} \begin{matrix} k \\ \rho-k \end{matrix} \begin{bmatrix} \mathbf{\Sigma}_k & \\ & \mathbf{\Sigma}_{\rho-k} \end{bmatrix} \begin{matrix} k \\ \rho-k \end{matrix} \begin{bmatrix} \mathbf{V}_k^{\mathrm{T}} \\ \mathbf{V}_{\rho-k}^{\mathrm{T}} \end{bmatrix} \\ &= \sum_{i=1}^{\rho} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}} \end{aligned} \tag{8}$$

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 4.1. *(Halko et al., 2011b)[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 $\mathbf{Q} \triangleq \mathrm{orth}(\mathbf{A}\mathbf{\Omega})$. Then,*

$$\mathbb{E} \left\| \mathbf{A} - \left(\mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{A}\right)_k \right\|_{\mathrm{F}} \leq \left(1 + \frac{k}{p-1}\right)^{1/2} \sqrt{\sum_{j=k+1}^{n} \sigma_j^2} \tag{9}$$

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 4.2. *(Boullé & Townsend, 2022)[Theorem 2] Under the same conditions as Theorem 4.1, except assume the columns of $\mathbf{\Omega}$ are sampled from $\mathcal{N}(\mathbf{0}, \mathbf{K})$.*

$$\mathbb{E}\left\|\mathbf{A}-\left(\mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{A}\right)_k\right\|_{\mathrm{F}} \leq \left(1+\sqrt{\frac{\beta_k}{\gamma_k}\frac{(k+p)}{(p-1)}}\right)\sqrt{\sum_{j=k+1}^{n}\sigma_j^2} \tag{10}$$

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

$$\left(1+\underbrace{\left\|\mathbf{\Sigma}_{\rho-k}\left(\mathbf{V}_{\rho-k}^{\mathrm{T}}\mathbf{\Omega}\right)\left(\mathbf{V}_k^{\mathrm{T}}\mathbf{\Omega}\right)^{\dagger}\right\|}_{\psi}\right)^{1/2}\|\mathbf{\Sigma}_{\rho-k}\|_{\mathrm{F}} \tag{11}$$

See Theorems 4.1 and 4.2 and (Boutsidis & Gittens, 2013; Gittens & Mahoney, 2013). 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 (11) has been studied w.r.t to Krylov sSubspaces in (Drineas et al., 2018). In (Drineas et al., 2018)[Theorem 2.2], Drineas et al. find

$$\psi \leq \left\|\tan\mathbf{\Theta}\left(\widetilde{\mathbf{V}}, \mathbf{V}_k\right)\right\| \tag{12}$$

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 4.3. *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*

$$\mathbf{x} = \alpha\widehat{\mathbf{V}}_{(:,k)}, \quad \alpha \sim \mathcal{N}(0,1) \tag{13}$$

Lemma 4.3 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 2.1, we can now put bounds on the top right singular vector approximation by the Randomized SVD.

4.3 Deterministic Bounds

First, we introduce a necessary result.

Lemma 4.4. *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*

$$\left\|\mathbf{V}_{\rho-k}\widetilde{\mathbf{V}}_k\right\|_{\mathrm{F}} \leq \left(\sum_{i=1}^{k}\sin\Theta\left(\mathbf{v}_i, \widetilde{\mathbf{v}}_i\right)\right)^{1/2} \tag{14}$$

Now, we will come to our main results.

Lemma 4.5. *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 Algorithm 1 with oversampling parameter p. Let $\mathbf{Q} \triangleq \operatorname{orth}(\mathbf{A}\mathbf{X}) = \operatorname{orth}(\mathbf{A}\left[\widetilde{\mathbf{v}}_1 \quad \ldots \quad \widetilde{\mathbf{v}}_k\right])$. Then,*

$$\begin{aligned}&\left\|\mathbf{A}-\widetilde{\mathbf{Q}}\widetilde{\mathbf{Q}}^{\mathrm{T}}\mathbf{A}\right\|_{\mathrm{F}} \\ &\leq \left(\left\|\sin\mathbf{\Theta}(\mathbf{U}_k, \widetilde{\mathbf{U}})\mathbf{\Sigma}_k\right\|_{\mathrm{F}}^2 - (k-1)\|\mathbf{\Sigma}_k\|_{\mathrm{F}}^2\right)^{1/2} \\ &+ \sigma_{k+1}\left(\sum_{i=1}^{k}\|\sin\Theta\left(\mathbf{u}_i, \widetilde{\mathbf{u}}_i\right)\|_{\mathrm{F}}^2\right)^{1/2} + \|\mathbf{\Sigma}_{\rho-k}\|_{\mathrm{F}}\end{aligned} \tag{15}$$

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.

Lemma 4.6. *Let p be our predetermined oversampling parameter, then let $\mathbf{\Omega} \in \mathbb{R}^{n\times p}$ be a standard gaussian matrix. Define $\mathbf{Q} \triangleq \operatorname{orth}(\mathbf{A}\mathbf{\Omega})$, let $\widetilde{\mathbf{v}}_1$ be the first singular vector of the approximation $\mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{A}$, then*

$$\sin\Theta\left(\mathbf{u}_1, \widetilde{\mathbf{u}}_1\right) \leq \left(1+\left(\left(\frac{\sigma_1}{\sigma_2}\right)\cot\left(\mathbf{v}_1, \widetilde{\mathbf{v}}_1\right)\right)^2\right)^{-1/2} \tag{16}$$

The proof is deferred to § Appendix A.3. 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 4.7. *Consider the same setup in Lemma 4.6, with $j-1$ vectors sampled as described in Algorithm 1, then*

$$\sin\Theta\left(\mathbf{u}_j, \widetilde{\mathbf{u}}_j\right) \leq \left(1+\left(\left(\frac{\sigma_1}{\sigma_j}\right)\cot\left(\mathbf{v}_j, \widetilde{\mathbf{v}}_j\right)\right)^2\right)^{-1/2} \tag{17}$$

The proof is deferred to § Appendix A.4. Now we will introduce the most important theorem of our work. From Lemma 4.6, we now have an idea on the effect of sampling the eigenvector approximations.

Theorem 4.8. *Given*

We will connect this together with the error bounds of sampling k right singular vector approximations. With

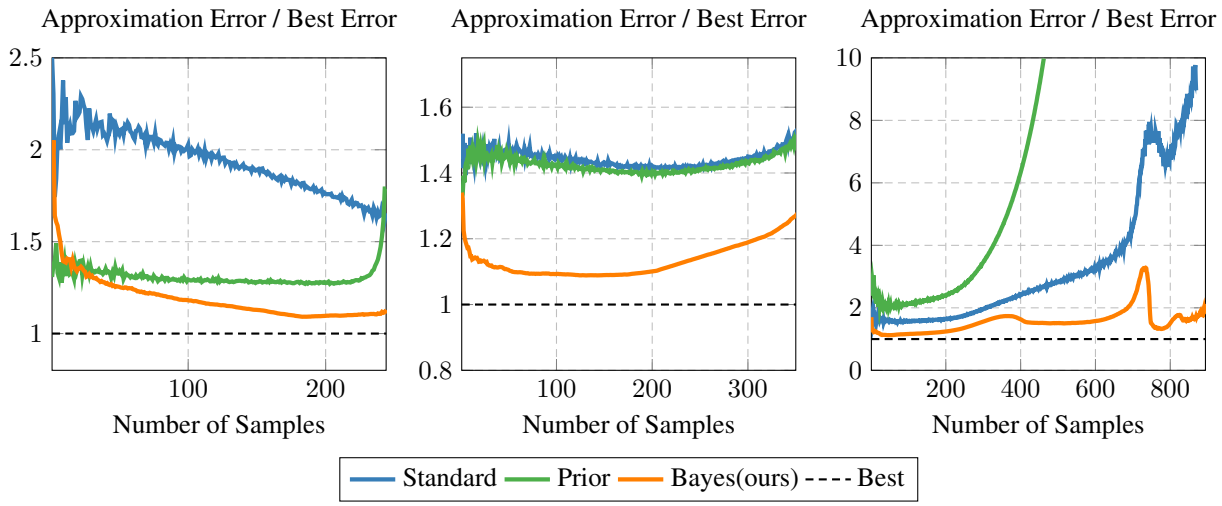

Figure 1. Low Rank Approximation for the Inverse Differential Operator given in Equation (20) *(Left)*, Differential Operator Matrix `Poisson2D` (Davis & Hu, 2011) *(Center)*, and Differential Operator Matrix `DK01R` (Davis & Hu, 2011) *(Right)*. The experiment on the left is from (Boullé & Townsend, 2022) (Figure 2).

Lemma 4.5 and Lemma 4.6, we have the following theorem.

Theorem 4.9. *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 Algorithm 1. Then,*

$$\mathbb{E}\left\|\mathbf{A}-\widetilde{\mathbf{A}}_k\right\|_{\mathrm{F}} \leq \sum_{i=1}^{k}\sigma_i\left(\frac{\sigma_{i+1}}{\sigma_i}\right) + \sigma_{k+1}\left(\sum_{i=1}^{k}\left(\frac{\sigma_{i+1}}{\sigma_i}\right)\right) + \sqrt{\sum_{j=k+1}^{n}\sigma_j} \tag{18}$$

Proof. This theorem follows from the induction formed in Lemma 4.6 and Lemma 4.7 and plugging this in to Lemma 4.5. ■

4.5 Probabilistic Bounds

Corollary 4.10. *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 Algorithm 1. Then,*

$$\left\|\mathbf{A}-\widetilde{\mathbf{A}}_k\right\|_{\mathrm{F}} \leq \sqrt{k}\sigma_{k+1} + \sqrt{\sum_{j=1}^{k}\sigma_j^2} + \sqrt{\sum_{j=k+1}^{n}\sigma_j^2} \tag{19}$$

Proof. This follows from a simple upper bound on the multiplicative term on σ_{k+1} in Lemma 4.5 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:

$$\mathcal{L}u = \frac{\partial^2 u}{\partial x^2} - 100\sin(5\pi x)\,u, \qquad x \in [0,1] \tag{20}$$

In **??***(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 **??**, 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, (Davis & Hu, 2011). The synthetic matrix is developed in the following scheme:

$$\mathbf{A} = \sum_{i=1}^{\rho}\frac{100i^{\ell}}{n}\mathbf{U}_{(:,i)}\mathbf{V}_{(i,:)}^{\mathrm{T}}, \quad \mathbf{U} \in \mathbb{O}_{m,k}, \mathbf{V} \in \mathbb{O}_{n,k} \tag{21}$$

We find our theoretical bounds stronger than Theorem 4.1.

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, (Boullé et al., 2022). 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

Bakshi, A., Clarkson, K. L., and Woodruff, D. P. 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, pp. 1130–1143, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392648. doi: 10.1145/3519935.3519988. URL https://doi.org/10.1145/3519935.3519988.

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

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

Boullé, N., Earls, C. J., and Townsend, A. Data-driven discovery of green's functions with human-understandable deep learning. *Scientific Reports*, 12 (1):4824, Mar 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-08745-5. URL https://doi.org/10.1038/s41598-022-08745-5.

Boutsidis, C. and Gittens, A. Improved matrix algorithms via the subsampled randomized hadamard transform. *SIAM Journal on Matrix Analysis and Applications*, 34 (3):1301–1340, 2013. doi: 10.1137/120874540. URL https://doi.org/10.1137/120874540.

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

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

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

Drineas, P., Ipsen, I. C. F., Kontopoulou, E.-M., and Magdon-Ismail, M. Structural convergence results for approximation of dominant subspaces from block krylov spaces. *SIAM Journal on Matrix Analysis and Applications*, 39(2):567–586, 2018. doi: 10.1137/16M1091745. URL https://doi.org/10.1137/16M1091745.

Gittens, A. and Mahoney, M. Revisiting the nystrom method for improved large-scale machine learning. In Dasgupta, S. and McAllester, D. (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 567–575, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/gittens13.html.

Halko, N., Martinsson, P.-G., Shkolnisky, Y., and Tygert, M. An algorithm for the principal component analysis of large data sets. *SIAM Journal on Scientific Computing*, 33(5):2580–2594, 2011a. doi: 10.1137/100804139. URL https://doi.org/10.1137/100804139.

Halko, N., Martinsson, P. G., and Tropp, J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011b. doi: 10.1137/090771806. URL https://doi.org/10.1137/090771806.

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

Hennig, P. and Schuler, C. J. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13(6), 2012.

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

Indyk, P., Wagner, T., and Woodruff, D. Few-shot data-driven algorithms for low rank approximation. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=6dUJPrPPUau.

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

Martinsson, P.-G. and Tropp, J. A. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica*, 29:403–572, 2020. doi: 10.1017/S0962492920000021.

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

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

Musco, C. and Musco, C. Randomized block krylov methods for stronger and faster approximate singular value decomposition. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/1efa39bcaec6f3900149160693694536-Paper.pdf.

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

Niu, F., Zhang, H., Yang, H., and Yang, D. Distribution of the smallest eigenvalue of complex central semi-correlated wishart matrices. In *2008 IEEE International Symposium on Information Theory*, pp. 1788–1792, 2008. doi: 10.1109/ISIT.2008.4595296.

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

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

Schmidt, E. Zur theorie der linearen und nichtlinearen integralgleichungen. *Mathematische Annalen*, 63(4): 433–476, Dec 1907. ISSN 1432-1807. doi: 10.1007/BF01449770. URL https://doi.org/10.1007/BF01449770.

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

Tzeng, R.-C., Wang, P.-A., Adriaens, F., Gionis, A., and Lu, C.-J. Improved analysis of randomized svd for top-eigenvector approximation. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 2045–2072. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/tzeng22a.html.

Vershynin, R. *Introduction to the non-asymptotic analysis of random matrices*, pp. 210–268. Cambridge University Press, 2012. doi: 10.1017/CBO9780511794308.006.

Wang, S., Li, X., Sun, J., and Xu, Z. Training networks in null space of feature covariance for continual learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 184–193, 2021. doi: 10.1109/CVPR46437.2021.00025. URL http://dx.doi.org/10.1109/CVPR46437.2021.00025.

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

Woolfe, F., Liberty, E., Rokhlin, V., and Tygert, M. A randomized algorithm for the approximation of matrices. 2006. URL https://api.semanticscholar.org/CorpusID:9847746.

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

A Deferred Proofs

A.1 Proof of Lemma 4.4

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

$$\zeta_1 \triangleq \left\|\mathbf{V}_{\rho-k}^{\mathrm{T}}\widetilde{\mathbf{V}}\right\|_{\mathrm{F}} = \left(\sum_{i=1}^{k}\sum_{j=k+1}^{\rho}\left\|\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{v}_j\right\|^2\right)^{1/2} = \left(\sum_{i=1}^{k}\sum_{j=1}^{n}\left\|\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{v}_j\right\|^2 - \sum_{i=1}^{k}\sum_{j=1}^{k}\left\|\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{v}_j\right\|^2\right)^{1/2} \tag{22}$$

$$= \left(\sum_{i=1}^{k}\sum_{j=1}^{n}\mathbb{1}_{i\neq j}\left\|\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{v}_j\right\|^2 - \sum_{i=1}^{k}\sum_{j=1}^{k}\left\|\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{v}_j\right\|^2 + \sum_{i=1}^{k}\left\|\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{v}_j\right\|^2\right)^{1/2} \tag{23}$$

$$= \left(\left(\sum_{i=1}^{k}\sin^2\Theta\left(\mathbf{v}_i,\widetilde{\mathbf{v}}_i\right) + \cos^2\Theta\left(\mathbf{v}_i,\widetilde{\mathbf{v}}_i\right)\right) - \sum_{i=1}^{k}\sum_{j=1}^{k}\left\|\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{v}_j\right\|^2\right)^{1/2} \tag{24}$$

$$= \left(k - \sum_{i=1}^{k}\sum_{j=1}^{k}\left\|\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{v}_j\right\|^2\right)^{1/2} = \left(k - \sum_{i=1}^{k}\left\|\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{v}_i\right\|^2 - \sum_{i=1}^{k}\sum_{j=1}^{k}\mathbb{1}_{i\neq j}\left\|\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{v}_j\right\|^2\right)^{1/2} \tag{25}$$

$$= \left(\sum_{i=1}^{k}\sin^2\Theta\left(\mathbf{v}_i,\widetilde{\mathbf{v}}_i\right) - \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{k}\mathbb{1}_{i\neq j}\left\|\widetilde{\mathbf{v}}_i^{\mathrm{T}}\mathbf{v}_j\right\|^2}_{\beta}\right)^{1/2} \leq \left(\sum_{i=1}^{k}\sin^2\Theta\left(\mathbf{v}_i,\widetilde{\mathbf{v}}_i\right)\right)^{1/2} \tag{26}$$

Here we have the desired result. ■

In Equation (26), when we use the approximation below, we lose the β term, is there any way to lower bound it?

A.2 Proof of Lemma 4.5

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 $\widetilde{\mathbf{U}} \triangleq \mathrm{orth}(\mathbf{A}\widetilde{\mathbf{V}})$.

$$\left\|\mathbf{A} - \widetilde{\mathbf{A}}_k\right\|_{\mathrm{F}} \triangleq \underbrace{\left\|\mathbf{A} - \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{A}\right\|_{\mathrm{F}}}_{\xi_1} \tag{27}$$

First we will upper bound ξ_1.

$$\xi_1^2 \triangleq \left\|\mathbf{A} - \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{A}\right\|_{\mathrm{F}}^2 \tag{28}$$

$$= \mathrm{Tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{A} - \mathbf{A}^{\mathrm{T}}\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{A} - \mathbf{A}^{\mathrm{T}}\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{A} + \mathbf{A}^{\mathrm{T}}\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^{\mathrm{T}}\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{A}\right) \tag{29}$$

$$= \mathrm{Tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{A} - \mathbf{A}^{\mathrm{T}}\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{A} - \mathbf{A}^{\mathrm{T}}\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{A} + \mathbf{A}^{\mathrm{T}}\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{A}\right) \tag{30}$$

$$= \|\mathbf{A}\|_{\mathrm{F}}^2 - \left\|\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{A}\right\|_{\mathrm{F}}^2 \tag{31}$$

$$= \|\mathbf{\Sigma}\|_{\mathrm{F}}^2 - \mathrm{Tr}\left(\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}\mathbf{V}\mathbf{\Sigma}\mathbf{U}^{\mathrm{T}}\widetilde{\mathbf{U}}\right) \tag{32}$$

$$= \|\mathbf{\Sigma}\|_{\mathrm{F}}^2 - \left\|\widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{U}\mathbf{\Sigma}\right\|_{\mathrm{F}}^2 \tag{33}$$

$$= \sum_{i=1}^{n}\sigma_i^2 - \sum_{i=1}^{n}\sigma_i^2\sum_{j=1}^{k}\cos^2\Theta\left(\mathbf{u}_i,\widetilde{\mathbf{u}}_j\right) \tag{34}$$

$$= \sum_{i=1}^{n}\sigma_i^2\left(1 - \sum_{i=1}^{k}\cos^2\Theta\left(\mathbf{u}_i,\widetilde{\mathbf{u}}_j\right)\right) \tag{35}$$

$$= \sum_{i=1}^{n} \sigma_i^2 - \sum_{i=1}^{n} \sigma_i^2 \sum_{j=1}^{k} \left(1 - \sin^2 \Theta\left(\mathbf{u}_i, \widetilde{\mathbf{u}}_j\right)\right) \tag{36}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} \sigma_i^2 \sin^2 \Theta\left(\mathbf{u}_i, \widetilde{\mathbf{u}}_j\right) - (k-1) \sum_{i=1}^{n} \sigma_i^2 \tag{37}$$

$$\leq \sum_{i=1}^{k} \sigma_i^2 \sin^2 \Theta\left(\mathbf{u}_i, \widetilde{\mathbf{u}}_i\right) + \sum_{i=1}^{k} \sum_{j=1}^{k} \mathbb{1}_{i \neq j} \sigma_i^2 - (k-1) \sum_{i=1}^{k} \sigma_i^2 + \sum_{i=k+1}^{n} \sigma_i^2 \tag{38}$$

$$= \sum_{i=1}^{k} \sigma_i^2 \sin^2 \Theta\left(\mathbf{u}_i, \widetilde{\mathbf{u}}_i\right) + \|\mathbf{\Sigma}_{\rho-k}\|_{\mathrm{F}}^2 \tag{39}$$

In Equation (38), all elements off the diagonal are upper bounded by 1, this is intuitive as are k-th left singular vector approximation is not aimed to be closed to a different vector. With this we have the bound on ξ_1,

$$\xi_1 = \left(\left\| \sin \mathbf{\Theta}(\mathbf{U}, \widetilde{\mathbf{U}}) \mathbf{\Sigma}_k \right\|_{\mathrm{F}}^2 - (k-1) \|\mathbf{\Sigma}_k\|_{\mathrm{F}}^2 \right)^{1/2} \tag{40}$$

$$\leq \left(\sum_{i=1}^{k} \sigma_i^2 \sin^2 \Theta\left(\mathbf{u}_i, \widetilde{\mathbf{u}}_i\right) \right)^{1/2} \tag{41}$$

Now we will plug this back into Equation (27).

$$\left\| \mathbf{A} - \widetilde{\mathbf{A}}_k \right\|_{\mathrm{F}} \overset{\text{eqn. (41)}}{\leq} \left(\sum_{i=1}^{k} \sigma_i^2 \sin^2 \Theta\left(\mathbf{u}_i, \widetilde{\mathbf{u}}_i\right) + \|\mathbf{A}_{\rho-k}\|_{\mathrm{F}}^2 \right)^{1/2} \tag{42}$$

In Figure 2, we see the upper bound given in Equation (42) is good. ■

A.3 Proof of Lemma 4.6

Proof. Let $\widehat{\mathbf{A}}_1$ be the approximation obtained by $(\mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{A})_1$ where $\mathbf{Q}$ is obtained by $\mathrm{orth}\left(\mathbf{A}\mathbf{\Omega}\right)$ and $\mathbf{\Omega} \in \mathbb{R}^{n \times p}$ is a standard gaussian matrix. It is important all our sine bounds are less than 1 or they will not be useful. We will first convert

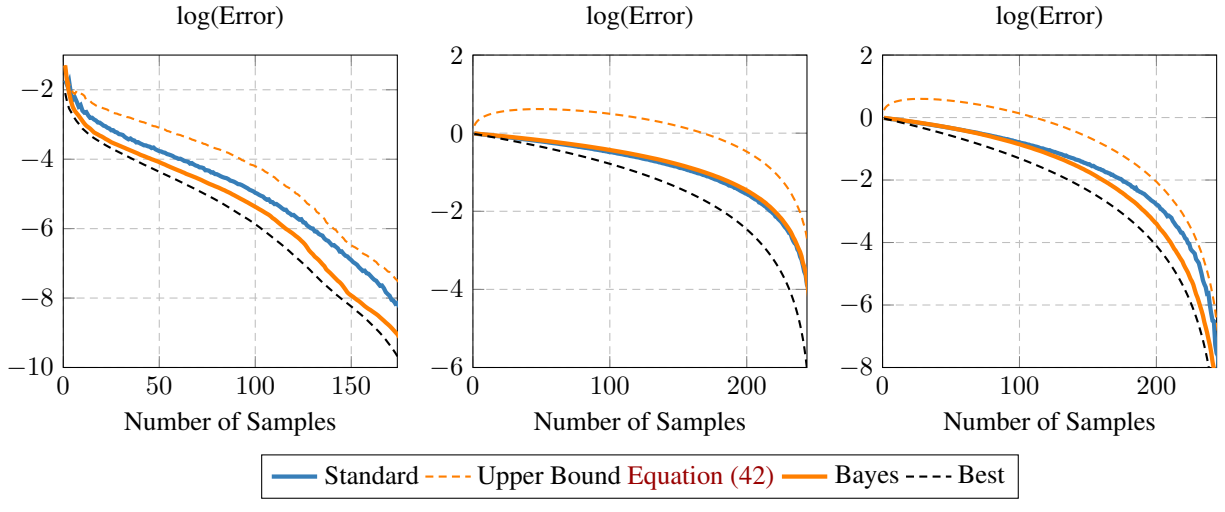

Figure 2. Low Rank Approximation Error Ratios for the Synthetic Matrix described in Equation (21) with $\ell = 1$ *(Center)*and $\ell = 2$ *(Right)*and Mona Lisa image *(Left)*. Upper bound is given in Equation (42).

this bound into a bound w.r.t. $\mathbf{v}_1$ and $\widetilde{\mathbf{v}}_1$.

$$\sin^2\Theta\left(\mathbf{u}_1,\widetilde{\mathbf{u}}_1\right) \overset{\text{eqn. (6)}}{=} \frac{\left\|\mathbf{U}_{\perp,1}^{\mathrm{T}}\mathbf{A}\widetilde{\mathbf{v}}_1\right\|_{\mathrm{F}}^2}{\|\mathbf{A}\widetilde{\mathbf{v}}_1\|_{\mathrm{F}}^2} = \frac{\left\|\mathbf{U}_{\perp,1}^{\mathrm{T}}\mathbf{A}\widetilde{\mathbf{v}}_1\right\|_{\mathrm{F}}^2}{\|\mathbf{U}^{\mathrm{T}}\mathbf{A}\widetilde{\mathbf{v}}_1\|_{\mathrm{F}}^2} \tag{43}$$

$$= \frac{\left\|\mathbf{U}_{\perp,1}^{\mathrm{T}}\mathbf{A}\widetilde{\mathbf{v}}_1\right\|_{\mathrm{F}}^2}{\left\|\mathbf{U}_{\perp,1}^{\mathrm{T}}\mathbf{A}\widetilde{\mathbf{v}}_1\right\|_{\mathrm{F}}^2 + \left\|\mathbf{U}_1^{\mathrm{T}}\mathbf{A}\widetilde{\mathbf{v}}_1\right\|_{\mathrm{F}}^2} \tag{44}$$

$$= \left(1 + \left(\frac{\left\|\mathbf{U}_1^{\mathrm{T}}\mathbf{A}\widetilde{\mathbf{v}}_1\right\|_{\mathrm{F}}}{\left\|\mathbf{U}_{\perp,1}^{\mathrm{T}}\mathbf{A}\widetilde{\mathbf{v}}_1\right\|_{\mathrm{F}}}\right)^2\right)^{-1} \tag{45}$$

$$\left\|\mathbf{U}_1^{\mathrm{T}}\mathbf{A}\widetilde{\mathbf{v}}_1\right\| = \left\|\mathbf{\Sigma}_1\mathbf{V}_1^{\mathrm{T}}\widetilde{\mathbf{v}}_1\right\| = \sigma_1\cos\Theta\left(\mathbf{v}_1,\widetilde{\mathbf{v}}_1\right) \tag{46}$$

$$\left\|\mathbf{U}_{\perp,1}^{\mathrm{T}}\mathbf{A}\widetilde{\mathbf{v}}_1\right\| = \left\|\mathbf{\Sigma}_{\perp,1}\mathbf{V}_{\perp,1}^{\mathrm{T}}\widetilde{\mathbf{v}}_1\right\| \leq \sigma_2\sin\Theta\left(\mathbf{v}_1,\widetilde{\mathbf{v}}_1\right) \tag{47}$$

The denominator in equation 44 is a result of the Matrix Pythagoras theorem.
Plugging this back into Equation (44), we have

$$\sin\Theta\left(\mathbf{u}_1,\widetilde{\mathbf{u}}_1\right) \leq \left(1 + \left(\left(\frac{\sigma_1}{\sigma_2}\right)\cot\left(\mathbf{v}_1,\widetilde{\mathbf{v}}_1\right)\right)^2\right)^{-1/2} \tag{48}$$

We will now need to upper bound $\sin(\mathbf{v}_1,\widetilde{\mathbf{v}}_1)$ and lower bound $\cos(\mathbf{v}_1,\widetilde{\mathbf{v}}_1)$. Now let us plug this back into Equation (48).

$$\sin\Theta\left(\mathbf{u}_1,\widetilde{\mathbf{u}}_1\right) \leq \frac{1}{1+\left(\frac{\sigma_1}{\sigma_2}\right)} \tag{49}$$

This completes the proof. ■

A.4 Proof of Lemma 4.7

Proof. We will perform a proof by strong induction. We have proved the base case in § Appendix A.3. Our inductive hypothesis is for $j \in \{1,\ldots,k-1\}$, it holds $\|\sin\Theta\left(\mathbf{v}_j,\widetilde{\mathbf{v}}_j\right)\| \leq \mathcal{O}\left(\frac{\sigma_{j+1}}{\sigma_j}\right)$. We will prove $\mathbb{E}\,\|\sin\Theta\left(\mathbf{v}_k,\widetilde{\mathbf{v}}_k\right)\| \leq \mathcal{O}\left(\frac{\sigma_{k+1}}{\sigma_k}\right)$. Now we must deal with the orthogonalization of $\widetilde{\mathbf{U}}$.

$$\sin^2\Theta\left(\mathbf{u}_i,\widetilde{\mathbf{u}}_i\right) \overset{\text{eqn. (6)}}{=} \frac{\|\mathbf{U}_{\perp,i}\,\mathrm{orth}\left(\mathbf{A}\widetilde{\mathbf{v}}_i\right)\|_{\mathrm{F}}^2}{\|\mathbf{U}_{\perp,i}\,\mathrm{orth}\left(\mathbf{A}\widetilde{\mathbf{v}}_i\right)\|_{\mathrm{F}}^2 + \left\|\mathbf{U}_{(i)}\,\mathrm{orth}\left(\mathbf{A}\widetilde{\mathbf{v}}_i\right)\right\|_{\mathrm{F}}^2} \tag{50}$$

$$= \left(1 + \frac{\left\|\mathbf{U}_{(i)}\,\mathrm{orth}\left(\mathbf{A}\widetilde{\mathbf{v}}_i\right)\right\|_{\mathrm{F}}^2\mu_1}{\|\mathbf{U}_{\perp,i}\,\mathrm{orth}\left(\mathbf{A}\widetilde{\mathbf{v}}_i\right)\|_{\mathrm{F}}^2\mu_2}\right)^{-1} \tag{51}$$

First, we will lower bound μ_1.

$$\sqrt{\mu_1} \triangleq \left\|\mathbf{U}_{(i)}\,\mathrm{orth}(\mathbf{A}\widetilde{\mathbf{v}}_i)\right\|_{\mathrm{F}} = \left\|\mathbf{U}_{(i)}\left(\mathbf{A}\widetilde{\mathbf{v}}_i - \sum_{j=1}^{i-1}\frac{\left(\mathbf{A}\widetilde{\mathbf{v}}_i\right)^{\mathrm{T}}\left(\mathbf{A}\widetilde{\mathbf{v}}_j\right)}{\left\|\mathbf{A}\widetilde{\mathbf{v}}_j\right\|^2}\mathbf{A}\widetilde{\mathbf{v}}_j\right)\right\|_{\mathrm{F}} \tag{52}$$

$$\geq \left\|\mathbf{\Sigma}_{(i)}\mathbf{V}_{(i)}^{\mathrm{T}}\widetilde{\mathbf{v}}_i\right\|_{\mathrm{F}} \geq \sigma_i\cos\Theta\left(\mathbf{v}_i,\widetilde{\mathbf{v}}_i\right) \tag{53}$$

We can remove the orthogonalization since the signs of $\mathrm{orth}(\mathbf{A}\widetilde{\mathbf{v}}_j)$ are arbitrary. Next we will analyze μ_2.

$$\sqrt{\mu_2} \triangleq \|\mathbf{U}_{\perp,i}\,\mathrm{orth}(\mathbf{A}\widetilde{\mathbf{v}}_i)\|^2 = \left\|\mathbf{U}_{\perp,i}\left(\mathbf{A}\widetilde{\mathbf{v}}_i - \sum_{j=1}^{i-1}\frac{\left(\mathbf{A}\widetilde{\mathbf{v}}_i\right)^{\mathrm{T}}\left(\mathbf{A}\widetilde{\mathbf{v}}_j\right)}{\left\|\mathbf{A}\widetilde{\mathbf{v}}_j\right\|^2}\mathbf{A}\widetilde{\mathbf{v}}_j\right)\right\|_{\mathrm{F}} \tag{54}$$

$$\leq \left\| \mathbf{\Sigma}_{\perp,i} \mathbf{V}_{\perp,i}^{\mathrm{T}} \widetilde{\mathbf{v}}_i \right\|_{\mathrm{F}} \leq \sigma_1 \sin \Theta \left(\mathbf{v}_i, \widetilde{\mathbf{v}}_i \right) \tag{55}$$

In our analysis of μ_2, we make the realization in the Gram-Schmidt Orthogonalization (Schmidt, 1907), the norm of the orthogonalized vector will be less than the original vector due to the subtraction of the projections and for all j we have $\|\widetilde{\mathbf{v}}_i\| = \|\widetilde{\mathbf{v}}_j\| = 1$. We thus have,

$$\sin \Theta \left(\mathbf{u}_i, \widetilde{\mathbf{u}}_i \right) \leq \left(1 + \left(\left(\frac{\sigma_i}{\sigma_1} \right) \cot \left(\mathbf{v}_i, \widetilde{\mathbf{v}}_i \right) \right)^2 \right)^{-1/2} \tag{56}$$

This completes the proof. ■

B Random Matrix Theory

Proposition B.1. *Draw a $m \times m$ matrix $\mathbf{G}$ s.t. the columns of $\mathbf{G}$ are sampled from $\mathcal{N}_m(\mathbf{0}, \mathbf{C})$ where the eigenvalues of $\mathbf{C}$ are represented as $\sigma_1 > \sigma_2 > \cdots > \sigma_m$. Then*

$$\mathbb{E}\left\|\mathbf{G}^\dagger\right\| \approx \sqrt{\pi \sum_{k=1}^{m} \frac{1}{\sigma_k}} = \sqrt{\pi \operatorname{Tr}\left(\mathbf{C}^{-1}\right)} \tag{57}$$

Proof. We will first note

$$\left\|\mathbf{G}^\dagger\right\| \overset{\text{lem. C.4}}{=} \frac{1}{\sigma_m(\mathbf{G})} = \frac{1}{\sqrt{\lambda_{\min}(\mathbf{G}\mathbf{G}^{\mathrm{T}})}} \tag{58}$$

For $\mathbf{W}$ sampled from $\mathcal{W}_m(m, \mathbf{\Sigma})$, the distribution of the minimum eigenvalue is given in (Niu et al., 2008) as

$$f_{\lambda_{\min}}(x) = \left(\sum_{k=1}^{m} \frac{1}{\sigma_k}\right) e^{-x \sum_{k=1}^{m} \frac{1}{\sigma_k}} \tag{59}$$

The Expected Value follows from a simple integration. First let us define $A \triangleq \sum_{k=1}^{m} \sigma_k^{-1}$.

$$\mathbb{E}\left\|\mathbf{G}^\dagger\right\| \overset{\text{lem. C.5}}{=} \int_0^\infty \frac{1}{\sqrt{x}} e^{-Ax} dx \tag{60}$$

$$= \sqrt{\pi A}\operatorname{erf}\left(\sqrt{\pi A}\right) \lesssim \sqrt{\pi A} \tag{61}$$

Substitute $A = \operatorname{Tr}(\mathbf{C}^{-1})$ and the proof is complete. ■

C Auxiliary Lemmas

Lemma C.1. *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,*

$$\|\mathbf{v} - \widetilde{\mathbf{v}}\| \leq \sqrt{2} \sin\Theta\left(\mathbf{v}, \widetilde{\mathbf{v}}\right) \tag{62}$$

Proof.

$$\sin^2\Theta\left(\mathbf{v}, \widetilde{\mathbf{v}}\right) = 1 - \left(\mathbf{v}^{\mathrm{T}}\widetilde{\mathbf{v}}\right)^2 \overset{(a)}{\geq} 1 - \mathbf{v}^{\mathrm{T}}\widetilde{\mathbf{v}} = 1 + \frac{1}{2}\|\mathbf{v} - \widetilde{\mathbf{v}}\|^2 - \frac{1}{2}\|\mathbf{v}\|^2 - \frac{1}{2}\|\widetilde{\mathbf{v}}\|^2 \tag{63}$$

$$= \frac{1}{2}\|\mathbf{v} - \widetilde{\mathbf{v}}\|^2 \tag{64}$$

(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 C.2. *(Mirsky, 1960). For any matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, then for any unitarily invariant norm $|||\cdot|||$, it follows*

$$|||\mathbf{ABC}||| \leq \|\mathbf{A}\|_2 \, |||\mathbf{B}||| \, \|\mathbf{C}\|_2 \tag{65}$$

Lemma C.3. *(Halko et al., 2011b). Let Π be the projection operator and $\mathbf{Q} \triangleq \operatorname{orth}(\mathbf{Y})$, it then follows*

$$\left|\left|\left|\mathbf{A} - \mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{A}\right|\right|\right| = |||\left(\mathbf{I} - \Pi_{\mathbf{Y}}\right)\mathbf{A}||| \tag{66}$$

Lemma C.4. *(Woolfe et al., 2006)[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*

$$\left\|\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathrm{T}}\right\| = \frac{1}{\sigma_m} \tag{67}$$

Lemma C.5. *It follows for a function of a random variable:*

$$\mathbb{E}g(X) = \int_{-\infty}^{\infty} g(x) f_X(x) dx \tag{68}$$

Lemma C.6. *(Boutsidis & Gittens, 2013)[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* $\xi = 2, \mathrm{F}$

$$\left\|\mathbf{A} - \mathbf{C}\mathbf{C}^{\dagger}\mathbf{A}\right\|_{\xi}^{2} \leq \|\mathbf{A} - \mathbf{C}\mathbf{X}\|_{\xi}^{2} \tag{69}$$

Lemma C.7. *(Vershynin, 2012)[Theorem 5.32]. Given a standard normal matrix* $\mathbf{G} \in \mathbb{R}^{m\times n}$, *then*

$$\mathbb{E}\left[\sigma_1(\mathbf{G})\right] \leq \left(\sqrt{m} + \sqrt{n}\right)^2 \tag{70}$$