RANDOMIZED LOW-RANK APPROXIMATION WITH ADAPTIVE SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

The randomized singular value decomposition (rSVD) is an efficient algorithm for computing low-rank approximations to matrices through matrix-vector products with standard Gaussian vectors. Recently, Boullé and Townsend (ICLR 2022) generalized the rSVD to leverage information on the right singular vectors of the target matrix A using Gaussian sketches with correlated entries. One open problem is whether this approach allows for the adaptive selection of the covariance matrix. We develop an adaptive sampling randomized singular value decomposition algorithm (AS-rSVD), which iteratively updates the covariance matrix after batches of rSVD applications. After each batch, the distribution of the random sampling vectors is updated to incorporate knowledge of the dominant right singular vectors of A obtained from the intermediate low-rank approximations. Our approach is demonstrated by a theoretical analysis and numerical experiments, and yields lower approximation errors than the rSVD and generalized rSVD while using the same number of matrix-vector products.

NB: Check and cite papers by Chris and Cameron Musco on randomized SVD, such as (Musco & Musco, 2015). Make the distinction clear between our approach and the power method.

1 INTRODUCTION

Computing low-rank matrix approximations through sketching has been a problem of interest at the intersection of computational linear algebra and machine learning for the past two decades (Drineas et al., 2006; Halko et al., 2011; Liberty et al., 2007; Martinsson et al., 2011; Musco & Musco, 2015; Nakatsukasa, 2020; Woodruff, 2014). In many real-world applications, it is often not possible to run experiments in parallel. Consider the following setting, there are a set of n inputs and m outputs, and there exists a PDE such it maps any set of inputs in $\mathbb{C}^m \rightarrow \mathbb{C}^n$. However, PDE experiments have intensive time and monetary requirements, e.g. aerodynamics (Fan et al., 2005), fluid dynamics (Lomax et al., 2001). 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. The randomized SVD is a popular algorithm for low rank matrix approximation in the matrix vector product query model (Halko et al., 2011). The rSVD utilizes Gaussian sketching to approximate the range of A and then constructs the matrix QQ^*A where Q is an orthonormal basis of the range of the sketch AG for a standard normal matrix G. This method has been generalized recently by Boulle & Townsend (2022) to allow for Gaussian input vectors with a general positive semi-definite covariance matrix. The current state-of-the-art algorithms for low-rank matrix approximation in the matrix-vector product model used a fixed covariance matrix structure.

Adaptively sampling vectors for matrix problems have been studied in detail by Sun et al. (2021). The theoretical properties of adaptively sampled matrix-vector queries for estimating the minimum eigenvalue of a Wishart matrix have been studied by Braverman et al. (2020). Their results are used by Bakshi et al. (2022) to develop lower bounds for rank-one low-rank matrix approximation with adaptive matrix-vector product queries. Adaptive sampling techniques for low-rank matrix approximation first appeared in finding a CUR matrix approximation (Frieze et al., 2004). Optimal column sampling for a CUR matrix decomposition received much attention in the literature (Deshpande et al., 2006; Deshpande & Vempala, 2006; Har-Peled, 2014). More recently, Paul et al. (2015) gave an adaptive algorithm for the column subset selection problem and proved it is possible to improve upon any relative-error column subset selection algorithm by adaptive sampling.

In this paper, we consider the adaptive setting where the algorithm $\mathcal{A}$ chooses a vector $\boldsymbol{\omega}^{(k)}$ with access to the previous query vectors $\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(k-1)}$, the matrix-vector products $A\boldsymbol{\omega}^{(1)}, \ldots, A\boldsymbol{\omega}^{(k-1)}$, and the intermediate low-rank matrix approximations, $Q_kQ_k^*A$, where Q_kR_k is the economized QR decomposition of $A\Omega^{(k)}$, where $\Omega^{(k)}$ is the concatenation of vectors $\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(k)}$. 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. Our algorithm utilizes the SVD computation of the low-rank approximation at each step to sample the next vector. Although there are runtime limitations, both in theory under certain conditions and many real-world matrices, our algorithm obtains a closer estimate of the optimal in the Frobenius norm.

Main Contributions.

1. We develop a novel adaptive sampling algorithm for the low-rank matrix approximation problem in the matrix-vector product model which does not utilize prior information of A.
2. We provide a novel theoretical analysis for adaptive sampling in the matrix-vector product query model.
3. We derive improved relative-error bounds for the generalized randomized SVD for both the spectral and Frobenius norms (Boulle & Townsend, 2022, Thm. 2) and show the spectral-norm bound is also an improvement for the rSVD (Halko et al., 2011, Thm. 10.8).
4. We perform numerical experiments on real-world and synthetic matrices that confirm our theoretical claims.

2 Background and related works

The randomized SVD is a method to find an orthonormal matrix that captures the range of the top left singular space of an $m \times n$ matrix A by multiplying it with a matrix with standard normal entries (Martinsson & Tropp, 2020). One first samples A at $k + p$ Gaussian vectors $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, I)$ for $1 \leq i \leq k + p$, where k is the target rank, p is the oversampling parameter, and I denotes the $n \times n$ identity matrix. One then calculates an orthonormal basis for the range by the economized QR decomposition and obtains $QR = AX$. Halko et al. (2011) that Q is a good approximation to the range of the top right singular space of A and thus the approximation QQ^*A is close to A in both the spectral and Frobenius norms. The analysis has also been extended to subsampled randomized trigonometric transform (SRTT) matrices (Boutsidis & Gittens, 2013). More recently, a generalized of the randomized SVD was proposed by Boulle & Townsend (2022) to allow for the columns of the Gaussian matrix to be sampled from a multivariate normal distribution with correlated covariance. Their results indicate that there exist covariance matrices that are able to obtain better approximation bounds than the standard rSVD. Correlated Gaussian sketching has also been studied for performing Nyström approximations (Persson et al., 2024).

3 Randomized SVD with adaptive samples

The adaptive range finder algorithm was proposed by Halko et al. (2011) as a method to guarantee a high accuracy low-rank approximation by sampling vectors one at a time until a stop condition is reached, which is not fixed as the sampled vectors are Gaussian, hence the algorithm is adaptive. It does not modify the distribution of the sampling vectors to reduce sample complexity or error. To clarify the distinction in the meaning of the term *adaptive*, we provide a formal definition of the problem we study.

Definition 1 (Adaptive sampling)**.** *With access to $r\ell$ right matrix-vector product queries. Sample $A\boldsymbol{\omega}_i$ for $1 \leq i \leq r\ell$ such that, after each matrix-vector query, one has access to the low-rank approximation $Q_tQ_t^*A$ for $t \in [r]$. Obtain a matrix $Q \in \mathbb{O}_{m,r\ell}$ using maximally $r\ell$ left and right matrix-vector product queries to A to find Q such that*

$$\|A - QQ^*A\|_{\mathrm{F}} \leq (1 + \epsilon) \min_{U:U^*U=I} \|A - UU^*A\|_{\mathrm{F}}.$$

Our definition of *adaptive* is standard in the theoretical computer science literature (see e.g., Deshpande & Vempala (2006); Paul et al. (2015)), in that we are using our previous samples to inform how we sample our future matrix vector product queries. We define *continued* sampling as sampling all vectors at once, i.e. not using the information in earlies matrix vector product queries to inform future choices.

3.1 Description of the algorithm

NB: Rename the number of samples per round as batch size throughout the paper, and rename round as iteration. Define the two terms in this section.

NB: We need a more extensive description of the algorithm before the pseudo-code.

We follow the convention in Paul et al. (2015) to perform adaptive sampling in rounds. Algorithm 1 runs in multiple rounds. In each round, we use the information we have learned from the matrix to update the covariance matrix.

Covariance Update in Algorithm 1 . In each round, we update our covariance matrix to be the projection matrix of the singular space of the low-rank approximation. It can be calculated either from an SVD calculation or from a pseudoinverse, as we have from an expansion of the SVD,

$$\widehat{V}_{t(k+p)}\widehat{V}^*_{t(k+p)} = \widehat{V}^{(t)}\widehat{\Sigma}^{(t)+}\widehat{U}^{(t)*}\widehat{U}^{(t)}\widehat{\Sigma}^{(t)}\widehat{V}^{(t)*} = \widehat{A}^{(t)+}\widehat{A}^{(t)} \tag{1}$$

Therefore, one can call a pseudoinverse procedure of SVD procedure to update the covariance matrix.

The pseudo-code for the optimal function sampling is given in Algorithm 1. Algorithm 1 is developed with the goal of minimizing the number of matrix-vector products necessary to obtain the same accuracy as the randomized SVD or generalized randomized SVD.

Algorithm 1 AS-rSVD, Adaptive sampling for low-rank matrix approximation

```
input: Target matrix $A \in \mathbb{C}^{m\times n}$, batch size $\ell$, number of iterations $r$
output: Low-rank approximation $\widehat{A}$ of $A$ to minimize $\|\widehat{A} - A\|_{\mathrm{F}}$
$\widetilde{A}_0 \leftarrow 0$
for $t \in [r]$ do
    Form $\Omega_t \in \mathbb{C}^{n\times\ell}$ with columns sampled i.i.d from $\mathcal{N}(\mathbf{0}, C_t)$
    $Y \leftarrow [Y, A\Omega_t] \in \mathbb{C}^{m\times t\ell}$ and obtain the economized QR decomposition $Y = QR$
    Update the low-rank approximation: $\widehat{A}^{(t)} = \widehat{A}^{t-1} + Q^{t\ell}_{-}Q^{t\ell*}_{-}A$
    Calculate the SVD: $\widehat{U}\widehat{\Sigma}\widehat{V}^* = Q^*A$
    Update the covariance matrix: $C_t \leftarrow \widehat{V}_{k(t-1),\ell t}\widehat{V}^*_{k(t-1),\ell t}$
return: $Q_tQ_t^*A$
```

Complexity analysis. In total, we sample $k+p$ Gaussian vectors from t different Gaussian Distributions and therefore the matrix-matrix product $A\Omega$ scales in $O(tmn\ell)$. We perform t QR factorizations which scale in $O(tmn\ell)$. We then must perform the SVD decomposition on the $Q^*A \in \mathbb{C}^{m\times n}$ a total of t times which can be done on the order of $O(tmn\ell)$. Thus the total complexity can be observed as $O(tmn\ell)$. We note that the right singular vectors of QQ^*A are equivalent to the right singular vectors of Q^*A as Q is semi-orthogonal.

4 Theoretical analysis

In this section, we will first give the mathematical setup for the theoretical analysis. We next derive improvements to the Generalized Randomized SVD Approximation Relative Error Bounds. We utilize our improved approximation bounds to derive relative error bounds for adaptive sampling. The proofs of all results presented in this section are deferred to Appendix A.

For any integer t, we define $[t]$ as the set of integers $\{1, \dots, t\}$. Let $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$ for any two real numbers a, b. We define $\mathbb{O}_{n,k}$ as the set of all $n \times k$ matrices with orthonormal columns, i.e. $\{V \in \mathbb{C}^{n\times k} : V^*V = I_{k\times k}\}$. For any vector $\mathbf{x} \in \mathbb{C}^d$, we denote the ℓ_2 norm as $\|\mathbf{x}\|_2^2 = \sum_{i=1}^d |x_i|^2$. We define $\mathbb{S}^{d-1} = \{\mathbf{x} \in \mathbb{C}^d : \|\mathbf{x}\|_2 = 1\}$. We use Big-O notation, $y = O(x)$, to denote there exists a positive constant C such that $y \le Cx$. We define $\mathbb{E}[X]$ as the expectation of the random variable X and $\mathbf{Pr}\{A\}$ as the probability of event A occurring.

4.1 Linear algebra preliminaries

NB: This should be much shorter.

AKR: Removed all preliminaries not necessary in the main body.

The Moore–Penrose Inverse (Moore, 1920; Penrose, 1955) of $X \in \mathbb{C}^{m\times n}$ is denoted as $X^+ \in \mathbb{C}^{n\times m}$. The matrix X^+ satisfies the following four conditions,

$$XX^+X = X, \quad X^+XX^+ = X^+, \quad (XX^+)^* = XX^+, \quad (X^+X)^* = X^+X.$$

If X has full row-rank, then the pseudo-inverse can be given explicitly as $X^+ = (X^*X)^{-1}X^*$. For any matrix X with economized singular value decomposition $X = U\Sigma V^*$, the pseudoinverse is given as $X^+ = V\Sigma^{-1}U^*$. An orthogonal projector matrix is a Hermitian Matrix and satisfies $P^2 = P$. This property of the orthogonal projector matrix implies $\mathrm{O} \preceq P \preceq I$. Suppose Q represents an orthonormal basis of the column space of a matrix Y, then $P_Y = YY^+$ represents the unique projection matrix on to the range of Y. One way to construct P_Y is to find the economized QR decomposition of Y, denoted as the matrix product QR, where Q is semi-orthogonal and R is upper-right triangular, then $P_Y = QQ^*$. We factorize A as follows,

$$A = \begin{array}{c} \begin{array}{cc} k & n-k \end{array} \\ \left[\begin{array}{cc} U_k & U_{k,\perp} \end{array}\right] \end{array} \begin{array}{c} \begin{array}{cc} k & n-k \end{array} \\ \begin{bmatrix} \Sigma_k & \\ & \Sigma_{k,\perp} \end{bmatrix} \end{array} \begin{bmatrix} V_k^* \\ V_{k,\perp}^* \end{bmatrix} \begin{array}{c} k \\ n-k \end{array} = \sum_{i=1}^n \sigma_i \mathbf{u}_i \mathbf{v}_i^*.$$

The trace of a matrix is given as $\mathrm{Tr}(A) = \sum_{i=1}^n \sigma_i(A)$. The spectral norm for a matrix $A \in \mathbb{C}^{m\times n}$ and $m \ge n$ is given as $\|A\|_2 = \sigma_1(A)$ and the Frobenius norm is given as $\|A\|_\mathrm{F}^2 = \sum_{i=1}^n \sigma_i^2(A)$. We define for $\xi \in \{2, \mathrm{F}\}$,

$$A_k = \underset{\mathrm{rank}(B)\le k}{\arg\min} \|A - B\|_\xi.$$

The Eckart-Young-Mirsky Theorem (Eckart & Young, 1936; Mirsky, 1960) tells us for the spectral and Frobenius norms that $A_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^*$.

4.2 Generalized Randomized SVD

In this subsection we present our result for the Frobenius norm approximation bounds for the Generalized Randomized SVD presented by Boulle & Townsend (2022). To our knowledge these are the first relative-error upper bounds for the spectral and Frobenius norms for the generalized rSVD. NB: We need to remove the dependence on the dimension here using (Persson et al., 2024, Lem. 2.2). This gives you a tighter bound than Prop.. 19 that you can replace simply later on in the proof. To control the expectation term in the Lem. 2.2, you can use (Persson et al., 2024, Lem. A.2).

Theorem 2. *Let $A \in \mathbb{C}^{m\times n}$, set $k \ge 2$ an integer, an oversampling parameter $p \ge 4$. Let $\Omega \in \mathbb{R}^{n\times k+p}$ represent the test matrix with columns sampled from $\mathcal{N}(\mathbf{0}, C)$. Let*

$$K = V^*CV = \begin{bmatrix} V_k^*CV_k & V_k^*CV_{k,\perp} \\ V_{k,\perp}^*CV_k & V_{k,\perp}^*CV_{k,\perp} \end{bmatrix} = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}.$$

Suppose $QR = A\Omega$ is the economized QR decomposition of $A\Omega$, then for all $u \ge 1$,

$$\|A - QQ^*A\|_\mathrm{F} \le \|\Sigma_{k,\perp}\|_\mathrm{F} \left(1 + \sqrt{6ut\log((n-k)(k+p))}\|K_{22}\|_2^{1/2}\sqrt{\mathrm{Tr}(K_{11}^{-1})}\sqrt{\frac{1}{p+1}}\right),$$

with probability at least $1 - \sqrt{2/\pi}(nm)^{1-u} - t^{-p}$.

Remark 1. *A similar spectral norm bound can be developed similarly to the Frobenius norm in Theorem 2 and is derived in Theorem 13 in Appendix A.*

The proof of Theorem 2 can be derived with the deterministic error bound of (Halko et al., 2011, Thm. 9.1) and combining Lemma 9 with both relations of Proposition 19. From Theorem 2, we infer that we obtain a better theoretical bound then the randomized SVD when C has better alignment with the top right singular space of A than the identity. As example, choosing $C = V_k V_k^*$, we obtain $K_{22} = \mathrm{O}$ and we then obtain the optimal $\|\Sigma_{k,\perp}\|_\mathrm{F}^2$ Frobenius norm approximation error. The weakness in the bound when compared to the probabilistic bound in Halko et al. (2011) is the $O(\log((n-k)(k+p)))$ term that is a result of Proposition 19. Our bounds presented in Theorem 2 are the strongest relative spectral and Frobenius norm error approximations to the generalized Randomized SVD presented in the literature to our knowledge. We formalize our relative error bound in the following corollary.

Corollary 3. *Let $A \in \mathbb{C}^{m\times n}$, set $k \geq 1$ an integer, an oversampling parameter $p \geq 4$. Let $\Omega \in \mathbb{R}^{n\times(k+p)}$ represent the test matrix with columns sampled from $\mathcal{N}(\mathbf{0}, C)$. Then, let $QR = A\Omega$ represent the economized QR decomposition of $A\Omega$. Then if*

$$p \geq (24/\epsilon)\log((n-k)(k+p))\|K_{22}\|_2 \operatorname{Tr}(K_{11}^{-1}),$$

with probability exceeding $1 - \sqrt{2/\pi}(nm)^{-1} - 0.0625$,

$$\|A - QQ^*A\|_\mathrm{F}^2 \leq (1+\epsilon)\|A - A_k\|_\mathrm{F}^2.$$

Corollary 3 develops a relative Frobenius norm error bound for the Generalized Randomized SVD and can be derived from elementary algebraic manipulations after an application of the first relation in Theorem 2. Since (Persson et al., 2024, Rem. 2.5) does not obtain a $(1+\epsilon)$ relative error bound, there exists a sufficiently large oversampling parameter p such that our bound gives a stronger error bound. Furthermore, our relative-error bound for the spectral norm is stronger than the relative-error spectral norm bound given by Halko et al. (Halko et al., 2011, Equation (1.9)). We are able to reduce the $O(\sqrt{n-k})$ upper bound to $O(\sqrt{\log((n-k)(k+p))})$. We formalize our improvement and give numerical illustrations in Appendix C.

4.3 Residual adaptive sampling theoretical analysis

In this section, we make theoretical connections from adaptive sampling in the implicit matrix model to the more well-studied problem of adaptive sampling in the column subset selection problem in a framework we refer to as *residual adaptive sampling*. Adaptive sampling in works such as Deshpande & Vempala (2006) and Paul et al. (2015) for the column subset selection problem (see e.g. (Civril, 2014, § 1) for a problem definition) sample columns from the residual matrix, $A - Q_tQ_t^*A$, at each iteration t instead of directly from A. Our first result is to formalize this statement and prove that indeed, adaptive low-rank matrix approximation in the implicit matrix model performs low-rank matrix approximation at each round for the residual matrix, $A - Q_tQ_t^*A$. We require the following structural lemma, which forms the basis for residual adaptive sampling.

Lemma 4. *Let $\Omega = [\Omega_1, \Omega_2]$, $Y_\psi = A\Omega_\psi$ for all $\psi \in \{1, 2\}$, and $Y = A\Omega$. Then for all $\xi \in \{2, \mathrm{F}\}$:*

$$\|A - P_YA\|_\xi = \|(I - P_{Y_1})A - P_{(I-P_{Y_1})Y_2}(I - P_{Y_1}A)\|_\xi. \tag{2}$$

From Lemma 4, to minimize the RHS of Equation (2) we observe that Q_- is equal to the dominant left singular space of $P_\perp A$. From this result, we then see that at each iteration, the optimal matrix vector queries are in the dominant right singular vector of $(I - P_{Y_1})A$. Thus, if one has knowledge of the matrix $(I - P_Y)A$ at any iteration, then performing a power iteration by sampling X as a Gaussian matrix, then calculate

$$X_q = (I - P_{Y_t})\big((I - P_{Y_t})^*(I - P_{Y_t})\big)^q X$$

Then, sample X_q as the sampling vectors for the next round.

Theorem 5. *Let $A \in \mathbb{C}^{m\times n}$ and let $C \in \mathbb{C}^{n\times n}$ be a PSD covariance matrix. Let the sampling matrix be decomposed as $\Omega = [\Omega, \Omega_-]$, the matrix-matrix products $Y = A\Omega$, $Y_- = A\Omega_-$, then let Q_+ be the orthonormal matrix from an economized QR decomposition of $[A\Omega_-, Y]$. Let $p \geq 4$ be the oversampling parameter, then with probability exceeding $1 - \delta - t^{-p}$,*

$$\left\|A - Q_+Q_+^*A\right\|_\mathrm{F}^2 \leq \epsilon(1+\epsilon)\|A - A_k\|_\mathrm{F}^2 + (1+\epsilon)\|A - A_{2k}\|_\mathrm{F}^2$$

From Theorem 5, we observe that our choice of covariance matrix reduces the number of matrix-vector products required to obtain the same accuracy as the Randomized SVD or even generalized randomized SVD.

Comparison with continued sampling. Consider the case where $t = 2$. In the continued sampling model, we sample all the $2k$ vectors initially. Continued sampling incurs Frobenius norm error,

$$\|A - Q_+Q_+A\|_{\mathrm{F}}^2 \leq \left(1 + \frac{\epsilon}{2}\right)\|A - A_k\|_{\mathrm{F}}^2.$$

Next, we consider the bound given in Theorem 5. From some simple algebraic manipulations, we obtain when

$$\|A - A_{2k}\|_{\mathrm{F}}^2 \leq \left(\frac{1 - \epsilon/2 - \epsilon^2}{1 + \epsilon}\right)\|A - A_k\|_{\mathrm{F}}^2 \leq (1 - \epsilon)\|A - A_k\|_{\mathrm{F}}^2$$

The adaptive bound is stronger than the classical bound given sufficient singular decay when the goal is a $(1+\epsilon)\|A - A_k\|_{\mathrm{F}}^2$ approximation. Even stronger, is we can note there exists an ϵ for which the adaptive bound is always stronger than the classical bound,

$$\sum_{i=2k+1}^{n} \sigma_i^2 \leq (1 - \epsilon) \sum_{i=k+1}^{n} \sigma_i^2 \iff \epsilon \sum_{i=k+1}^{n} \sigma_i^2 \leq \sum_{i=k+1}^{2k} \sigma_i^2$$

Therefore, we see there exists a sufficiently small ϵ that gives a stronger $(1 + \epsilon)\|A - A_k\|_{\mathrm{F}}$ approximation. We now consider the analysis for T steps.

Theorem 6. *Let $A \in \mathbb{C}^{m\times n}$ and let $C_i \in \mathbb{C}^{n\times n}$ for all $i \in [T]$ be PSD covariance matrices. The sampling matrix is given as $\Omega = [C_1^{1/2}G_1, \ldots, C_t^{1/2}G_t]$. Let QR be the economized QR decomposition of $A\Omega$, then*

$$\|A - QQ^*A\|_{\mathrm{F}}^2 \leq \|A - A_{tk}\|_{\mathrm{F}}^2 + \epsilon \sum_{i=1}^{T-1} (1 + \epsilon)^{T-i}\|A - A_{ik}\|_{\mathrm{F}}^2,$$

with probability exceeding $1 - T(\delta + t^{-p})$.

The theorem can be proved by induction, see Theorem 1 in Paul et al. (2015). Interestingly, without any algorithmic modifications to the randomized SVD, the fact that Q is an orthonormal basis of $A\Omega$ gives us the same result as in Paul et al. (2015) without explicit knowledge of the matrix $A - YY^+A$.

4.4 Analysis of AS-rSVD

NB: I don't understand this sentence: Our covariance update in Algorithm 1 of Algorithm 1 does not align with the theory we have developed in Section 4.3.

AKR: There are two ways to do adaptive sampling from what I can see, one is our way with AS-rSVD, where you sample in the dominant subspace, the other is residual adaptive sampling, where you sample in the dominant subspace of the residual $A - QQ^*A$. I was trying to point out this difference.

Consider Lemma 4, it is optimal to sample in the dominant right singular subspace of $A - P_YA$. However, in the implicit matrix problem, we do not have access to $A - P_YA$. We first show for one round of our adaptive procedure in Algorithm 1, it is sufficient to sample in our estimation of the dominant right singular subspace of A.

Theorem 7. *Consider one round of Algorithm 1 with a covariance matrix $C \in \mathbb{C}^{n\times n}$. Suppose $G_1, G_2 \in \mathbb{C}^{n\times(k+p)}$ are Gaussian Matrices and $\Omega = [G_1 \quad G_2]$. Suppose SVD of $P_{AG_1}A = \widehat{U}\widehat{\Sigma}\widehat{V}^*$. If $\widehat{V}_{k+p}^*G_1$ and $V_k^*C^{1/2}G_2$ have full-row rank, then for all $\xi \in \{2, \mathrm{F}\}$,*

$$\|A - P_{A\Omega}A\|_\xi \lesssim \left(1 + \sqrt{8\log(2\ell(n - 2k))}\|V_{k,\perp}^*C^{1/2}\|_2\|(V_k^*C^{1/2})^+\|_2 \frac{\mathrm{e}\sqrt{k+p}}{p+1}\right)\|A - A_{2k}\|_\xi,$$

with probability at least 0.95.

We can now note that $(1 + \epsilon)^2 = 1 + O(\epsilon)$. Then, from ideas in subspace perturbation theory (see e.g. Wedin (1972)), the sample complexity to obtain a $(1 + \epsilon)\|A - A_{2k}\|_{\mathrm{F}}$ is reduced. We will now extend Theorem 7 to multiple rounds.

Theorem 8. *Consider t rounds of Algorithm 1 with covariance matrices $C_i \in \mathbb{C}^{n\times n}$ for $i \in [t]$. Suppose $\Psi_i \in \mathbb{C}^{n\times(k+p)}$ for $i \in [t]$ are Gaussian Matrices. If $\widehat{V}_{k+p}^* G_1$ and $V_k^* C^{1/2} G_2$ have full-row rank, then for $\xi \in \{2, \mathrm{F}\}$, with probability at least* 0.95*,*

$$\|A - QQ^*A\|_\xi \lesssim \prod_{i=1}^{t}\left(1 + \sqrt{8\log(10n)}\|V_{\mathcal{I}_i,\perp}^* C_i^{1/2}\|_2 \|(V_{\mathcal{I}_t}^* C_i^{1/2})^+\|_2 \frac{\mathrm{e}\sqrt{k+p}}{p+1}\right)\|A - A_{tk}\|_\xi$$

Theorem 8 implies we can first sample $k + p$ standard Gaussian vectors. Then, at the ith iteration, we can sample vectors that align with $V_{\mathcal{I}_i}$. When given the intermediate low-rank matrix approximations, from obtaining the SVD, we can then sample our approximation of $V_{\mathcal{I}_i}$. We have nearly arrived at the covariance update given in Algorithm 1. We next note in practice minimizing $\|A - QQ^*A\|_\xi$ is not equivalent to minimizing ϵ for $(1+\epsilon)\|A - A_k\|_\xi$. Therefore, it is often the case even if we know the optimal covariance matrix $V_k V_k^*$ it might not be optimal since $\|A - A_k\|_\xi \geq \Omega(\|A - A_{k+s}\|_\xi)$ where s is an integer. Hence at each iteration it is not optimal to choose a rank-k covariance matrix.

Remark 2 (Randomized Nyström approximation with adaptive samples)**.** *When the matrix $A \in \mathbb{R}^{n\times n}$ is symmetric positive semi-definite (SPSD), the randomized Nyström approximation is more suitable than the randomized SVD (Tropp & Webber, 2023). It is given as $\widehat{A} := A\Omega(\Omega^* A\Omega)^+\Omega^* A \approx A$. While analyzing adaptive sampling for computing Nyström approximations is notoriously challenging (Gittens & Mahoney, 2016), Persson et al. (2024) recently generalized the randomized Nyström approximation to allow for a general covariance matrix. Then, our algorithm and analysis should extend naturally to improve Nyström approximations with adaptive samples.*

5 Numerical experiments

In this section, we evaluate the performance of our adaptive sampling algorithm on performing low-rank approximations to matrices (issued from the discretization of differential operators or standard benchmarks) and compare it against the non-adaptive rSVD and generalized rSVD. Each point represents one round of sampling. We define

$$\mathsf{OPT} = \min_{Z\in\mathbb{C}^{m\times t\ell}} \|A - ZZ^*A\|_{\mathrm{F}}.$$

From an application of the Eckart-Young-Mirsky Theorem for the Frobenius norm (Eckart & Young, 1936; Mirsky, 1960), we find that the minimizing $Z = U_{t\ell}$ and obtain $\mathsf{OPT} = \sqrt{\sum_{i=t\ell+1}^{n}\sigma_i^2(A)}$.

Experiment 1: Learning an Inverse Differential Operator.[1] In our first experiment we attempt to learn the discretized $10^3 \times 10^3$ matrix of the inverse of the following differential operator:

$$\mathcal{L}u = \frac{\partial^2 u}{\partial x^2} - 100\sin(5\pi x)\,u, \qquad x \in [0,1] \tag{3}$$

Learning the inverse operator of a PDE is equivalent to learning the Green's Function of a PDE. This has been theoretically proven for certain classes of PDEs (Linear Parabolic (Boullé et al., 2022b; Boullé & Townsend, 2023)) as the inverse Differential operator is compact. Furthermore, there are multiple works suggesting the learning of inverse differential operators is provably efficient (Boulle & Townsend, 2022; Boullé et al., 2023). We observe in Fig. 1(a), even without prior knowledge of the dominant right singular space of A, after approximately 280 matrix-vector products, AS-rSVD learns a better low-rank approximation with respect to the Frobenius norm than the generalized rSVD which utilizes prior information on the Green's function for inverse Laplacian PDE.

Experiment 2: Learning an Inverse Differential Operator in the Real World. We attempt to learn the inverse of the discretized differential operator given in matrix `pde2961`, sourced from the TAMU Matrix Suite Davis & Hu (2011). `pde2961` is the matrix associated with a Model Partial Differential Equations Problem. For the generalized rSVD, use the following matrix as the covariance matrix,

$$K_{i,j} = \exp\bigl(-|x-y|^2/(2\ell^2)\bigr), \quad x, y \in D, \tag{4}$$

[1] All experiments are run in MATLAB R2022b on a 3.6GHz CPU with 16 GB of RAM.

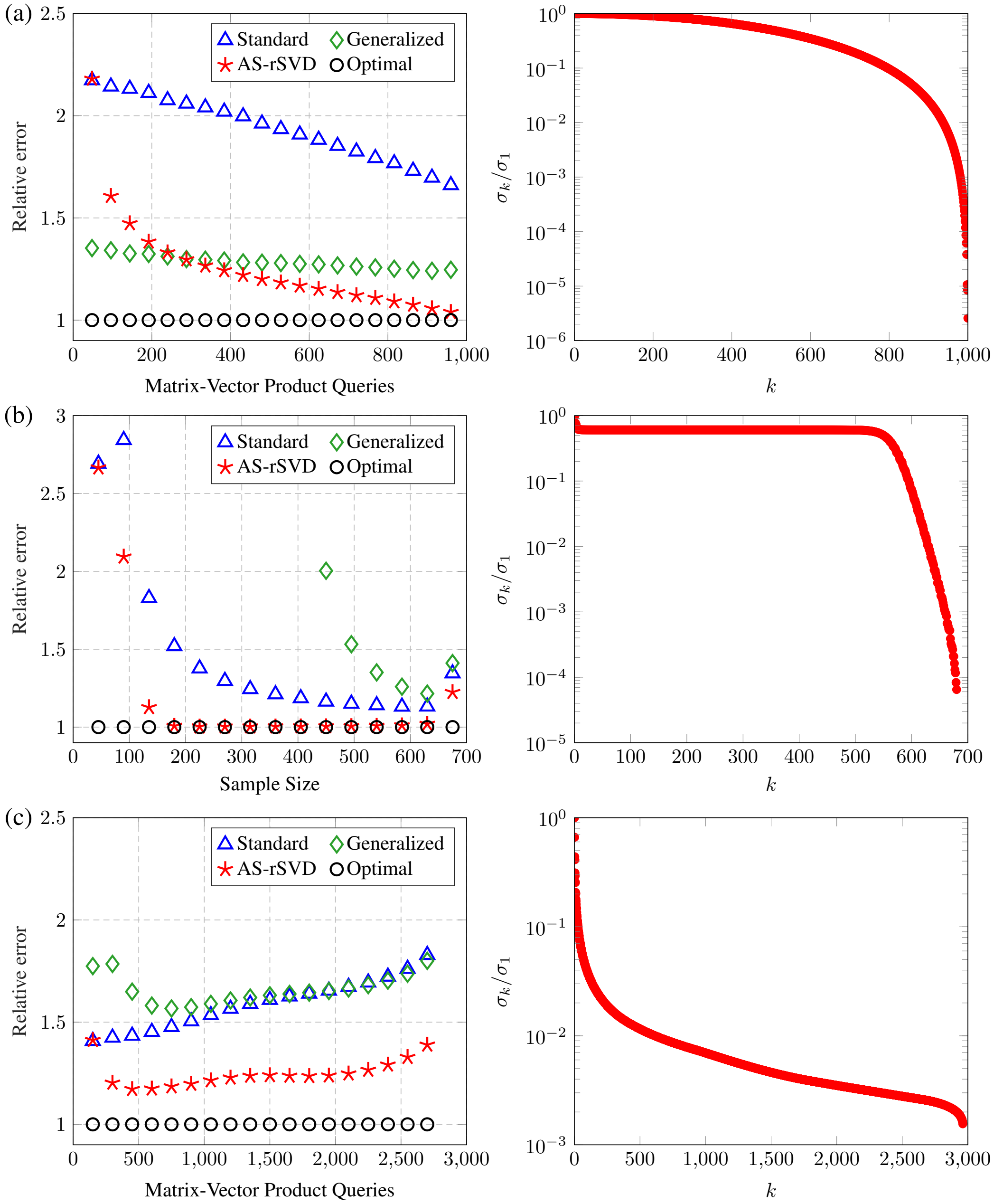

Figure 1: Left panels: Relative error (in the Frobenius norm) of the low-rank matrix approximation with respect to the number of matrix-vector products. Standard denotes the rSVD with i.i.d. Gaussian input vectors (Halko et al., 2011) while generalized uses a prior covariance matrix (Boulle & Townsend, 2022). Right panels: Singular value decay of the matrix. (a) Inverse of a discretized Laplacian differential operator (see Eq. (3) and (Boulle & Townsend, 2022, Fig. 2)), $k = 8, p = 16, t = 20$. (b) Matrix `fs-680-1` from Davis & Hu (2011), $k = 16, p = 32, t = 20$. (c) Matrix `pde2961` from Davis & Hu (2011), $k = 50, p = 100, t = 18$. All points in the plots for Fig. 1 are the average over 10 randomized runs.

where we select $\ell = 0.01$ in accordance with (Boulle & Townsend, 2022, Fig. 4). In Fig. 1(b), we find that Algorithm 1 obtains nearly-optimal Frobenius norm error from 175 to 650 matrix-vector product queries. We find from Fig. 1(a) and Fig. 1(b) that As-RSVD obtains promising results for learning inverse differential operators in synthetic and real world experiments.

Experiment 3: Learning a forward matrix in the Real World. We learn the a low rank matrix approximation of `fs-680-1`, also sourced from the TAMU Sparse Matrix Suite Davis & Hu (2011). `fs-680-1` is the associated matrix for a chemical kinetics problem. For the generalized rSVD we use the covariance matrix in Equation (4). Algorithm 1 performs well in large matrices after considering the results in Fig. 1(c). The `pde2961` matrix is of size 2961 × 2961 and Algorithm 1 outperforms rSVD and the generalized rSVD.

We find that in general, Algorithm 1 works well in a wide variety of singular value spectra. This can be concluded from observing that Fig. 1 display a diverse range of singular value spectra and cover both real and synthetic scenarios. Our experiments on real-world matrices are promising and indicate that our algorithm and implementation can be used in real-world applications of learning low-rank approximations of matrices that are only accessible via matrix-vector products.

NB: Need a timing comparison in the Appendix (with a comment in the main text) and the following sentence:

6 Conclusions

We have theoretically and empirically analyzed a novel Covariance Update to iteratively construct the sampling matrix, Ω in the Randomized SVD algorithm. We introduce a new adaptive sampling framework for low-rank matrix approximation when the matrix is only accessible by matrix-vector products by giving the algorithm access to intermediate low-rank matrix approximations. Our covariance update for generating sampling vectors and functions find applications in PDE learning (Boullé et al., 2022a; Brunton et al., 2020). Numerical Experiments indicate without prior knowledge of the matrix, we are able to obtain superior performance to the rSVD, and the generalized rSVD with covariance matrix utilizing prior information of the right singular vectors of A. Theoretically, we provide an analysis of our update extended to k-steps and show under certain singular value decay conditions, we obtain better performance expectation.

References

Ainesh Bakshi, Kenneth L Clarkson, and David P Woodruff. Low-rank approximation with $1/\epsilon^{1/3}$ matrix-vector products. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1130–1143, 2022.

Roger L Berger and George Casella. *Statistical inference*. Duxbury, 2001.

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

Nicolas Boullé and Alex Townsend. Learning elliptic partial differential equations with randomized linear algebra. *Found. Comput. Math.*, 23(2):709–739, Apr 2023.

Nicolas Boullé, Christopher J. Earls, and Alex Townsend. Data-driven discovery of Green's functions with human-understandable deep learning. *Sci. Rep.*, 12(1):4824, 2022a.

Nicolas Boullé, Seick Kim, Tianyi Shi, and Alex Townsend. Learning Green's functions associated with time-dependent partial differential equations. *J. Mach. Learn. Res.*, 23(1):9797–9830, 2022b.

Nicolas Boullé, Diana Halikias, and Alex Townsend. Elliptic PDE learning is provably data-efficient. *Proc. Natl. Acad. Sci. U.S.A.*, 120(39):e2303904120, 2023.

Christos Boutsidis and Alex Gittens. Improved matrix algorithms via the subsampled randomized Hadamard transform. *SIAM J. Matrix Anal. Appl.*, 34(3):1301–1340, 2013.

Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near-optimal column-based matrix reconstruction. *SIAM J. Comput.*, 43(2):687–717, 2014.

Mark Braverman, Elad Hazan, Max Simchowitz, and Blake Woodworth. The gradient complexity of linear regression. In *Conference on Learning Theory*, pp. 627–647, 2020.

Steven L Brunton, Bernd R Noack, and Petros Koumoutsakos. Machine learning for fluid mechanics. *Annu. Rev. Fluid Mech.*, 52:477–508, 2020.

Ali Civril. Column subset selection problem is ug-hard. *J. Comput. Syst. Sci.*, 80(4):849–859, 2014.

Timothy A Davis and Yifan Hu. The University of Florida sparse matrix collection. *ACM Trans. Math. Softw.*, 38(1):1–25, 2011.

Amit Deshpande and Santosh Vempala. Adaptive sampling and fast low-rank matrix approximation. In *International Workshop on Approximation Algorithms for Combinatorial Optimization*, pp. 292–303. Springer, 2006.

Amit Deshpande, Luis Rademacher, Santosh S Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. *Theory Comput.*, 2(1):225–247, 2006.

Petros Drineas, Ravi Kannan, and Michael W Mahoney. Fast monte carlo algorithms for matrices ii: Computing a low-rank approximation to a matrix. *SIAM Journal on computing*, 36(1):158–183, 2006.

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

Hui-Yuan Fan, George S Dulikravich, and Zhen-Xue Han. Aerodynamic data modeling using support vector machines. *Inverse Probl. Sci. En.*, 13(3):261–278, 2005.

Alan Frieze, Ravi Kannan, and Santosh Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *J. ACM*, 51(6):1025–1041, 2004.

Alex Gittens and Michael W Mahoney. Revisiting the nyström method for improved large-scale machine learning. *J. Mach. Learn. Res.*, 17(1):3977–4041, 2016.

Yehoram Gordon. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50:265–289, 1985.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2): 217–288, 2011.

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

Edo Liberty, Franco Woolfe, Per-Gunnar Martinsson, Vladimir Rokhlin, and Mark Tygert. Randomized algorithms for the low-rank approximation of matrices. *Proceedings of the National Academy of Sciences*, 104(51):20167–20172, 2007.

Harvard Lomax, Thomas H Pulliam, David W Zingg, Thomas H Pulliam, and David W Zingg. *Fundamentals of computational fluid dynamics*. Springer, 2001.

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

Per-Gunnar Martinsson, Vladimir Rokhlin, and Mark Tygert. A randomized algorithm for the decomposition of matrices. *Applied and Computational Harmonic Analysis*, 30(1):47–68, 2011.

Leonid Mirsky. Symmetric gauge functions and unitarily invariant norms. *Q. J. Math.*, 11(1):50–59, 1960.

Eliakim H Moore. On the reciprocal of the general algebraic matrix. *Bulletin of the american mathematical society*, 26:294–295, 1920.

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 (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.

Yuji Nakatsukasa. Fast and stable randomized low-rank matrix approximation, 2020.

Saurabh Paul, Malik Magdon-Ismail, and Petros Drineas. Column selection via adaptive sampling. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Roger Penrose. A generalized inverse for matrices. *Mathematical proceedings of the Cambridge philosophical society*, 51(3):406–413, 1955.

David Persson, Nicolas Boullé, and Daniel Kressner. Randomized Nyström approximation of non-negative self-adjoint operators. *arXiv preprint arXiv:2404.00960*, 2024.

Philippe Rigollet and Jan-Christian Hütter. High-dimensional statistics. *arXiv preprint arXiv:2310.19244*, 2023.

Erhard Schmidt. Zur theorie der linearen und nichtlinearen integralgleichungen. *Math. Ann.*, 63(4): 433–476, 1907.

Xiaoming Sun, David P Woodruff, Guang Yang, and Jialin Zhang. Querying a matrix through matrix-vector products. *ACM Transactions on Algorithms (TALG)*, 17(4):1–19, 2021.

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

Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. Training networks in null space of feature covariance for continual learning. In *Conference on Computer Vision and Pattern Recognition*, pp. 184–193, 2021.

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

David P. Woodruff. Sketching as a tool for numerical linear algebra. In *Foundations and Trends in Theoretical Computer Science*, volume 10, pp. 1–157. Now Publishers, Inc., 2014.

A Proofs

In this section, we will prove all results deferred from the main text. We also discuss some of the theoretical implications of our work to the rSVD and generalized rSVD. Finally, we perform additional experiments on synthetic matrices.

A.1 Distribution of $V^*\Omega$

In this section we will derive the distribution of $V^*\Omega$.

Lemma 9. *Let $\Omega = [\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_\ell] \in \mathbb{C}^{n\times\ell}$ such that $\boldsymbol{\omega}_i \sim \mathcal{N}(\mathbf{0}, C)$ for all $i \in [\ell]$ and $C \succeq 0$ is symmetric. For $V \in \mathbb{O}_{n,k}$, it follows that the columns of $V^*\Omega$ are sampled from a centered multivariate Gaussian distribution with second-moment matrix $K = V^*CV$.*

Proof of Lemma 9. The matrix $V^*\Omega$ can be decomposed as follows,

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

Let $\mathcal{D} = \mathcal{N}(\mathbf{0}, I)$ and $\mathcal{D}_C = \mathcal{N}(\mathbf{0}, C)$. We will first show that the entries of each column of $V^*\Omega$ are Gaussian. From the fact that C is symmetric, we have that $C = U\Sigma U$ for a unitary U and diagonal $\Sigma \succeq 0$. Then for any $i \in [k]$ and $j \in [\ell]$, we have for $\mathbf{x} \sim \mathcal{D}$,

$$\mathbf{v}_i^*\boldsymbol{\omega}_j = \mathbf{v}_i^* C^{1/2}\mathbf{x} = \mathbf{v}_i^* U\Sigma^{1/2}\mathbf{x} = \sum_{k\in[n]} \mathbf{v}_i^*\mathbf{u}_k \sqrt{\lambda_k(C)}\, x_k.$$

In the above, we have that each $[V^*\Omega]_{i,j}$ is Gaussian for $(i,j) \in [k]\times[\ell]$ as a linear combination of Gaussians is Gaussian. We will calculate the mean and covariance. We first calculate the mean of a column of $V^*\Omega$. For any $(i,j) \in [k]\times[\ell]$,

$$\begin{aligned} \underset{\boldsymbol{\omega}_j\sim\mathcal{D}_C}{\mathbb{E}}[\mathbf{v}_i^*\boldsymbol{\omega}_j] &= \underset{\mathbf{x}\sim\mathcal{D}}{\mathbb{E}}[\mathbf{v}_i^* C^{1/2}\mathbf{x}] = \mathbf{v}_i^* C^{1/2} \underset{\mathbf{x}\sim\mathcal{D}}{\mathbb{E}}[\mathbf{x}] \\ &= \sum_{p\in[n]} \mathbf{v}_i^*\mathbf{u}_p \sqrt{\lambda_p(C)} \underset{\mathbf{x}\sim\mathcal{D}}{\mathbb{E}}[x_p] = 0. \end{aligned}$$

Now we calculate the covariance matrix. Let $\mathbf{v} \in \mathbb{S}^{n-1}$, then for any $(i, i', j) \in [k]\times[k]\times[\ell]$, and $i \neq i'$, we have

$$\begin{aligned} &\underset{\boldsymbol{\omega}_j\sim\mathcal{D}_C}{\mathbb{E}}\left[\left(\mathbf{v}_i^*\boldsymbol{\omega}_j - \underset{\boldsymbol{\omega}_j\sim\mathcal{D}_C}{\mathbb{E}}[\mathbf{v}_i^*\boldsymbol{\omega}_j]\right)\left(\mathbf{v}_{i'}^*\boldsymbol{\omega}_j - \underset{\boldsymbol{\omega}_j\sim\mathcal{D}_C}{\mathbb{E}}[\mathbf{v}_{i'}^*\boldsymbol{\omega}]\right)\right] \\ &\qquad = \underset{\boldsymbol{\omega}_j\sim\mathcal{D}_C}{\mathbb{E}}\left[\mathbf{v}_i^*\boldsymbol{\omega}_j\boldsymbol{\omega}_j^*\mathbf{v}_{i'}\right] = \mathbf{v}_i^* C\mathbf{v}_{i'}. \end{aligned} \tag{5}$$

For the diagonal covariance elements, we have

$$\underset{\boldsymbol{\omega}_j\sim\mathcal{D}_C}{\mathbb{E}}[(\mathbf{v}_i^*\boldsymbol{\omega}_j - \mathbb{E}[\mathbf{v}_i^*\boldsymbol{\omega}_j])^2] = \underset{\boldsymbol{\omega}_k\sim\mathcal{D}_C}{\mathbb{E}}[\mathbf{v}_i^*\boldsymbol{\omega}_j\boldsymbol{\omega}_j^*\mathbf{v}_i] = \mathbf{v}_i^* C\mathbf{v}_i. \tag{6}$$

Then combining Equations (5) and (6), we have $K = V^*CV$, and our proof is complete. ■

A.2 Improvements to the Generalized rSVD

In this section, we will derived our improved bounds for the generalized rSVD. These bounds will be used in our proofs for general adaptive sampling and the analysis of AS-rSVD. We will first restate the deterministic error bound derived by Boutsidis et al. (2014).

Lemma 10 (Boutsidis et al. 2014, Lem. 3.2)**.** *Let $A \in \mathbb{R}^{m\times n}$, and let Q be an orthonormal basis of $A\Omega$ for $\Omega \in \mathbb{R}^{n\times(k+p)}$, then*

$$\|A - Q(Q^*A)_k\|_{\mathrm{F}}^2 \leq \|A - A_k\|_{\mathrm{F}}^2 + \left\|\Sigma_{k,\perp}V_{k,\perp}\Omega(V_k\Omega)^+\right\|_{\mathrm{F}}^2.$$

Before we give our improvement to the Generalized Randomized SVD, we present the following necessary lemma on the concentration of $\left\|(V^*\Omega)^+\right\|_{\mathrm{F}}$ for Ω with columns sampled from $\mathcal{N}(\mathbf{0}, C)$.

Lemma 11 (Boullé & Townsend 2023, Lem. 3)**.** *Suppose $V_k \in \mathbb{O}^{n,k}$ and $\Omega \in \mathbb{C}^{n\times k+p}$ with columns sampled i.i.d from $\mathcal{N}(\mathbf{0}, K)$ and $p \geq 4$. Then,*

$$\left\|(V_k^*\Omega)^+\right\|_{\mathrm{F}} \leq \sqrt{\frac{3\,\mathrm{Tr}(K^{-1})}{p+1}} \cdot t^2,$$

with probability at least $1 - t^{-p}$.

For our improvement to the Generalized Randomized SVD for the spectral norm, we present the following necessary lemma.

Lemma 12 (Halko et al. 2011, Prop. 10.4)**.** *Let $G \in \mathbb{R}^{k\times(k+p)}$ have elements sampled i.i.d from $\mathcal{N}(0,1)$ for $p \geq 4$. Then for $t \geq 1$, with probability exceeding $1 - t^{-(p+1)}$,*

$$\left\|G^+\right\|_2 \leq \frac{\mathrm{e}\sqrt{k+p}}{p+1} \cdot t.$$

We are now ready to prove our relative Frobenius error norm bound for the generalized rSVD originally presented in Boulle & Townsend (2022).

NB: explain the difference with the bound from this paper

AKR: should I restate theorem 2 in the generalized rSVD paper? otherwise I was going to say we reduce from $O(k^2)$ to $O(k\log((n-k)\ell))$.

Proof of Theorem 2. Recall $A = U\Sigma V^*$, $\Omega_k = V_k^*\Omega$ and $\Omega_{k,\perp} = V_{\perp,k}^*\Omega$ where the columns of $\Omega \in \mathbb{R}^{n\times(k+p)}$ are sampled from $\mathcal{N}(\mathbf{0}, C)$. Let

$$K = \begin{bmatrix} V_k^*CV_k & V_k^*CV_{k,\perp} \\ V_{k,\perp}^*CV_k & V_{k,\perp}^*CV_{k,\perp} \end{bmatrix} = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}.$$

We then have the following manipulations, from Lemma 10,

$$\begin{aligned}
\|A - Q(Q^*A)_k\|_{\mathrm{F}} &\leq \left(\|\Sigma_{k,\perp}\|_{\mathrm{F}}^2 + \|\Sigma_{k,\perp}\Omega_{k,\perp}\Omega_k^+\|_{\mathrm{F}}^2\right)^{1/2} \\
&= \left(\|\Sigma_{k,\perp}\|_{\mathrm{F}}^2 + \|\Sigma_{k,\perp}V_{k,\perp}^*C^{1/2}X(V_k^*C^{1/2}X)^+\|_{\mathrm{F}}^2\right)^{1/2} \\
&= \left(\|\Sigma_{k,\perp}\|_{\mathrm{F}}^2 + \|\Sigma_{k,\perp}V_{k,\perp}^*K_{22}^{1/2}X_2(K_{11}^{1/2}X_1)^+\|_{\mathrm{F}}^2\right)^{1/2} \\
&\leq \left(\|\Sigma_{k,\perp}\|_{\mathrm{F}}^2 + 2u\log((n-k)(k+p))\|\Sigma_{k,\perp}K_{22}^{1/2}\|_{\mathrm{F}}^2\|(K_{11}^{1/2}X_1)^+\|_{\mathrm{F}}^2\right)^{1/2} \\
&\leq \|\Sigma_{k,\perp}\|_{\mathrm{F}}\left(1 + t^2\sqrt{6u\log((n-k)(k+p))}\|K_{22}\|_2^{1/2}\sqrt{\mathrm{Tr}(K_{11}^{-1})}\sqrt{\frac{1}{p+1}}\right).
\end{aligned}$$

In the above, the first relation follows from the deterministic error bound in Lemma 10. The second relation follows from noting that the columns of Ω are sampled from $\mathcal{N}(\mathbf{0}, C)$, thus $\Omega = C^{1/2}X$ where $X \in \mathbb{R}^{n\times(k+p)}$ is a standard Gaussian matrix. In the third relation, we apply Lemma 9 to note that $V_{k,\perp}^*C^{1/2}X = K_{22}^{1/2}X_2$ and $V_k^*C^{1/2}X = K_{11}^{1/2}X_1$ for standard Gaussian matrices $X_1 \in \mathbb{R}^{k\times(k+p)}$ and $X_2 \in \mathbb{R}^{(n-k)\times(k+p)}$. The fourth relation follows from Proposition 19 for the Frobenius norm with failure probability at most $\sqrt{2/\pi}(nm)^{1-u}$, and the final inequality follows from noting by the sub-multiplicativity of the Frobenius Norm, which gives us

$$\|\Sigma_{k,\perp}V_{k,\perp}^*C^{1/2}\|_{\mathrm{F}}^2 \leq \|\Sigma_{k,\perp}\|_{\mathrm{F}}^2\|V_{k,\perp}^*C^{1/2}\|_2^2 = \|\Sigma_{k,\perp}\|_{\mathrm{F}}^2\|K_{22}\|_2$$

and Lemma 11 which holds with failure probability at most t^{-p}. We then apply the elementary inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, to complete our proof for the Frobenius norm tail bound. ■

Theorem 13. *Let $A \in \mathbb{C}^{m\times n}$, set $k \geq 2$ an integer, an oversampling parameter $p \geq 4$. Let $\Omega \in \mathbb{R}^{n\times(k+p)}$ represent the test matrix with columns sampled from $\mathcal{N}(\mathbf{0}, C)$. Suppose $QR = A\Omega$ is the economized QR decomposition of $A\Omega$, then for all $u \geq 1$,*

$$\|A - QQ^*A\|_2 \leq \|A - A_k\|_2 \left(1 + \sqrt{2ut\log((n-k)(k+p))}\|K_{22}\|_2^{1/2}\|K_{11}^{-1}\|_2 \frac{\mathrm{e}\sqrt{k+p}}{p+1}\right).$$

Proof of Theorem 13. We follow the same notation as in the proof of Theorem 2. From the deterministic error bound (Halko et al., 2011, Thm. 9.1),

$$\begin{aligned}
\|A - QQ^*A\|_2 &\leq (\|\Sigma_{k,\perp}\|_2^2 + \|\Sigma_{k,\perp}\Omega_{k,\perp}\Omega_k^+\|_2^2)^{1/2} \\
&= \left(\|\Sigma_{k,\perp}\|_2^2 + \|\Sigma_{k,\perp}V_{k,\perp}^*C^{1/2}X(V_k^*C^{1/2}X)^+\|_2^2\right)^{1/2} \\
&= \left(\|\Sigma_{k,\perp}\|_2^2 + \|\Sigma_{k,\perp}K_{22}^{1/2}X_2(K_{11}^{1/2}X_1)^+\|_2^2\right)^{1/2} \\
&\leq \left(\|\Sigma_{k,\perp}\|_2^2 + 2\log((n-k)(k+p)/\delta)\|\Sigma_{k,\perp}K_{22}^{1/2}\|_2^2\|(K_{11}^{1/2}X_1)^+\|_2^2\right)^{1/2} \\
&\leq \|\Sigma_{k,\perp}\|_2\left(1 + \sqrt{2t\log((n-k)(k+p)/\delta)}\|K_{22}\|_2^{1/2}\|K_{11}^{-1}\|_2^{1/2}\frac{\mathrm{e}\sqrt{k+p}}{p+1}\right).
\end{aligned}$$

The first inequality follows from the deterministic error bound in Theorem 9.1 of Halko et al. (2011). The second relation follows from noting $\Omega = C^{1/2}X$ where X is a Gaussian matrix. The third relation follows from Lemma 9 which gives Gaussian matrices $X_1 \in \mathbb{R}^{k\times(k+p)}$ and $X_2 \in \mathbb{R}^{(n-k)\times(k+p)}$. The final relation follows from noting by Cauchy–Schwarz,

$$\|\Sigma_{k,\perp}V_{k,\perp}^*C^{1/2}\|_2^2 \leq \|\Sigma_{k,\perp}\|_2^2\|V_{k,\perp}^*C^{1/2}\|_2^2 = \|\Sigma_{k,\perp}\|_2^2\|K_{22}\|_2.$$

Furthermore, from noting that for any rank-k matrix $\|B^+\|_2 = \sigma_k^{-1}(B)$, we have

$$\|(V_k^*C^{1/2}X)^+\|_2^2 = \sigma_k^{-2}(V_k^*C^{1/2}X_1) \leq \lambda_k^{-1}(K_{11})\sigma_k^{-2}(X_1) = \|K_{11}^{-1}\|_2\|X_1^+\|_2^2.$$

Then, applying the bound in Lemma 12 with failure probability at most t^{-p+1} completes the proof. ■

A.3 Residual Adaptive Sampling

In this section, we will prove deferred results on residual adaptive sampling. We first will prove Lemma 4, then we present and prove Lemma 14, which we then use to prove our main result of this section, Theorem 5.

Proof of Lemma 4. Recall that Q_{12} is an orthonormal basis of $[A\Omega_1, A\Omega_2]$. Let Q_1 be an orthnormal basis of $A\Omega_1$, it thus follows that Q_{12} is also an orthonormal basis of $[A\Omega_1, (I - Q_1Q_1^*)A\Omega_2]$. This can be seen from the Classical Gram-Schmidt Procedure (Schmidt, 1907). Thus, we have

$$Q_{12}Q_{12}^* = Q_1Q_1^* + Q_{2|1}Q_{2|1}^*.$$

Then, expanding out the residual matrix, we have

$$A - Q_{12}Q_{12}^*A = A - Q_1Q_1^*A - Q_{2|1}Q_{2|1}^*A.$$

Then from noting, $Q_1^*Q_{2|1} = 0$, we have

$$A - Q_1Q_1^*A - Q_{2|1}Q_{2|1}^*A = (A - Q_1Q_1^*A) - Q_{2|1}Q_{2|1}^*(A - Q_1Q_1^*A),$$

which completes the proof. ■

We now require the following linear algebraic lemma.

Lemma 14. *Let $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{m\times n}$ and $Q \in \mathbb{O}_{m,\ell}$ for $\ell > k$. Then for all $i \in [n]$:*

$$\sigma_i^2(A - QQ^*A) \leq \sigma_i^2(A - Q(Q^*A)_k).$$

For the proof of this lemma we will utilize the following corollary of Weyl's Inequality, a classical result in Perturbation Theory.

Lemma 15 (Weyl's Inequality)**.** *Suppose A, B have real eigenvalues (e.g. they are Hermitian), then for all $i \in [n]$:*

$$\lambda_i(A+B) \le \lambda_i(A) + \lambda_1(B)$$

Proof of Lemma 14. We can note that $\sigma_i(QQ^*A) \le \sigma_i(A)$ for all $i \in [n]$ by noting $QQ^* \preceq I$ as $Q \in \mathbb{O}_{m,k}$. We first note that for any matrix $B \in \mathbb{C}^{m\times n}$, the singular values and eigenvalues are related by $\sigma_i^2(B) = \lambda_i(B^*B)$ for any $i \in [n]$. Then let us consider any $i \in [n]$.

$$\lambda_i((A - QQ^*A)^*(A - QQ^*A)) = \lambda_i(A^*A - A^*QQ^*A).$$

Let us now consider the matrix in the RHS of the lemma statement,

$$\begin{aligned}\sigma_i^2(A - Q(Q^*A)_k) &= \lambda_i((A - Q(Q^*A)_k)^*(A - Q(Q^*A)_k)) \\ &= \lambda_i(A^*A - A^*Q(Q^*A)_k - (Q^*A)_k^*Q^*A + (Q^*A)_k^*Q^*Q(Q^*A)_k) \\ &= \lambda_i(A^*A - (A^*QQ^*A)_k - (A^*QQ^*A)_k + (A^*QQ^*A)_k) \\ &= \lambda_i(A^*A - (A^*QQ^*A)_k). \end{aligned} \tag{7}$$

In the above, the final relation follows noting $Q^*Q = I$ because $Q \in \mathbb{O}_{m,\ell}$, and for any matrix B, we have that $B_k^*B = (B^*B)_k$ by an expansion of the SVD. We will now complete the proof. From noting $\sigma_i^2(B) = \lambda_i(B^*B)$, we have

$$\sigma_i^2(A - QQ^*A) = \lambda_i(A^*A - A^*QQ^*A)$$

Then, from the SVD,

$$\lambda_i(A^*A - A^*QQ^*A) = \lambda_i(A^*A - (A^*QQ^*A)_k - (A^*QQ^*A)_{k,\perp})$$

We then have from Weyl's Inequality (see Lemma 15), and noting $(A^*QQ^*A)_{k,\perp} \succeq \mathrm{O}$,

$$\lambda_i(A^*A - (A^*QQ^*A)_k - (A^*QQ^*A)_{k,\perp}) \le \lambda_i(A^*A - (A^*QQ^*A)_k).$$

Finally, from Equation (7),

$$\lambda_i(A^*A - (A^*QQ^*A)_k) = \sigma_i^2(A - Q(Q^*A)_k).$$

We then find that for all $i \in [n]$,

$$\sigma_i^2(A - QQ^*A) \le \sigma_i^2(A - Q(Q^*A)_k).$$

This completes the proof.

NB: These comments should be after the equation, like: "where the first inequality is due to Lemma 15..."

AKR: Updated formatting. Not sure if this is preferable over the presentation of Equation (7). ■

We are now ready prove our residual adaptive sampling approximation bound.

Proof of Theorem 5. For any $t \in [T]$, from the Gram-Schmidt orthogonalization procedure (Schmidt, 1907), we can decompose $Q_{t+1}Q_{t+1}^*$ as a sum of two parts,

$$Q_{t+1}Q_{t+1}^* = Q_tQ_t^* + \widetilde{Q}_{t+1}\widetilde{Q}_{t+1}^*.$$

From Lemma 4, we obtain

$$\|A - Q_{t+1}Q_{t+1}^*A\|_\mathrm{F} = \|(I - Q_tQ_t^*)A - \widetilde{Q}_{t+1}\widetilde{Q}_{t+1}^*(I - Q_tQ_t^*)A\|_\mathrm{F}.$$

We now note that

$$\widetilde{Q}_{t+1}\widetilde{R}_{t+1} = (I - Q_tQ_t^*)A\Omega_{t+1},$$

then from our relative-error accuracy bound in Corollary 3, we have

$$\|(I - Q_tQ_t^*)A - \widetilde{Q}_{t+1}\widetilde{Q}_{t+1}^*(I - Q_tQ_t^*)A\|_\mathrm{F}^2 \le (1+\epsilon)\|(A - Q_tQ_t^*A)_{k,\perp}\|_\mathrm{F}^2.$$

We then have from Lemma 14,

$$\begin{aligned}
\|(A - Q_tQ_t^*A)_{k,\perp}\|_{\mathrm{F}}^2 &= \sum_{i\in[n]\setminus[k]} \sigma_i^2(A - Q_tQ_t^*A) && \text{Frobenius Norm Defn.}\\
&\le \sum_{i\in[n]\setminus[k]} \sigma_i^2(A - Q_t(Q_t^*A)_{kt}) && \text{Lemma 14}\\
&= \|(A - Q_t(Q_t^*A)_{kt})_{k,\perp}\|_{\mathrm{F}}^2. && \text{Frobenius Norm Defn.}
\end{aligned}$$

Then, from expanding out the Frobenius Norm, we have

$$\|(A - Q_t(Q_t^*A)_{kt})_{k,\perp}\|_{\mathrm{F}}^2 = \|A - Q_t(Q_t^*A)_{kt}\|_{\mathrm{F}}^2 - \sum_{i\in[k]} \sigma_i^2(A - Q_t(Q_t^*A)_{kt}).$$

We now leverage the Eckart-Young-Mirsky Theorem (Eckart & Young, 1936; Mirsky, 1960), and obtain

$$\sum_{i\in[k]} \sigma_i^2(A - Q_t^*(Q_t^*A)_{kt}) \ge \inf_{\substack{B\in\mathbb{C}^{m\times n}\\ \operatorname{rank}(B)=kt}} \sum_{i\in[k]} \sigma_i^2(A - B) = \|A_{kt} - A_{k(t+1)}\|_{\mathrm{F}}^2. \tag{8}$$

Suppose $A - Q_tQ_t^*A = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^*$, then we define the constants

$$\beta_t = \|\widetilde{V}_{k,\perp}^* C_t \widetilde{V}_{k,\perp}\|_2, \quad \gamma_t = \operatorname{Tr}((\widetilde{V}_k^* C_t \widetilde{V}_k^*)^{-1}).$$

Then, from our improved generalized rSVD error bound in Theorem 2, we have with probability at least $1 - \sqrt{2/\pi}(nm)^{-1} - 2^{-p}$,

$$\|A - Q_{t+1}(Q_{t+1}^*A)_{kt}\|_{\mathrm{F}}^2 \le \left(\|A - Q_tQ_t^*A\|_{\mathrm{F}}^2 - \|A_{kt} - A_{k(t+1)}\|_{\mathrm{F}}^2\right)\left(1 + 24\log((n-kt)\ell t)\beta_t\frac{\gamma_t}{p+1}\right)$$

From Corollary 3, we obtain,

$$\|A - Q_{t+1}Q_{t+1}^*A\|_{\mathrm{F}}^2 \le (1+\epsilon)\left(\|A - Q_tQ_t^*A\|_{\mathrm{F}}^2 - \|A_{kt} - A_{k(t+1)}\|_{\mathrm{F}}^2\right), \tag{9}$$

when the oversampling parameter is chosen sufficiently large,

$$p \ge (24/\epsilon)\log((n-k)(k+p))\beta_t\gamma_t,$$

with failure probability at most $\sqrt{2/\pi}(nm)^{-1} + 2^{-p}$. Next, we can note for the base case when we choose $C_0 = I$,

$$\|A - Q_1Q_1^*A\|_{\mathrm{F}}^2 \le (1+\epsilon)\|A - A_k\|_{\mathrm{F}}^2,$$

when the oversampling parameter p_1 is chosen sufficiently large,

$$p_1 \ge (24k/\epsilon)\log((n-k)(k+p)),$$

with probability at least $\sqrt{2/\pi}(nm)^{-1} + 0.0325$. The remainder of this proof will follow similarly to (Paul et al., 2015, Thm. 1). We now consider the following inductive hypothesis,

$$\|A - Q_tQ_t^*A\|_{\mathrm{F}}^2 \le (1+\epsilon)\|A - A_{kt}\|_{\mathrm{F}}^2 + \sum_{i=2}^{t-1} \epsilon(1+\epsilon)^{t-i+1}\|A - A_{ki}\|_{\mathrm{F}}^2,$$

for any $t \in [T]$. Then, from Equation (9), we have

$$\begin{aligned}
\|A - Q_{t+1}Q_{t+1}^*A\|_{\mathrm{F}}^2 &\le (1+\epsilon)^2\|A - A_{kt}\|_{\mathrm{F}}^2 + \sum_{i=2}^{t-1} \epsilon(1+\epsilon)^{t-i+2}\|A - A_{ki}\|_{\mathrm{F}}^2\\
&\quad - (1+\epsilon)\|A_{kt} - A_{k(t+1)}\|_{\mathrm{F}}^2\\
&= (1+\epsilon)\|A - A_{k(t+1)}\|_{\mathrm{F}}^2 + \sum_{i=2}^{t} \epsilon(1+\epsilon)^{t-i+2}\|A - A_{ki}\|_{\mathrm{F}}^2,
\end{aligned}$$

where the second relation follows from noting, $\|A_{kt} - A_{k(t+1)}\|_{\mathrm{F}}^2 = \|A - A_{kt}\|_{\mathrm{F}}^2 - \|A - A_{k(t+1)}\|_{\mathrm{F}}^2$. The proof is complete. ■

A.4 Analysis of AS-rSVD

We first give the necessary lemmata for our proof of Theorem 7.

Proposition 16 (Conjutation Rule)**.** *Suppose that $M \succeq \mathrm{O}$. Then for any conformal matrix A,*

$$M \preceq N \implies A^*MA \preceq A^*NA$$

Lemma 17 (Matrix Pythagoras)**.** *If $X, Y \in \mathbb{C}^{m\times n}$ and $X^*Y = \mathrm{O}$ or $Y^*X = \mathrm{O}$, then*

$$\|X+Y\|_2^2 \leq \|X\|_2^2 + \|Y\|_2^2 \quad \textit{and} \quad \|X+Y\|_{\mathrm{F}}^2 = \|X\|_{\mathrm{F}}^2 + \|Y\|_{\mathrm{F}}^2.$$

Proof of Lemma 17. The proof for the Spectral norm follows directly from the definition.

$$\|X+Y\|_2^2 = \|X^*X + X^*Y + Y^*X + Y^*Y\|_2 = \|X^*X + Y^*Y\|_2 \leq \|X\|_2^2 + \|Y\|_2^2$$

In the above, the final relation follows from Weyl's inequality (see Lemma 15) and noting for any matrix $B \in \mathbb{C}^{m\times n}$, it holds $\|B^*B\|_2 = \|B\|_2^2$. The proof for the Frobenius norm follows directly from the definition of the Frobenius Norm.

$$\|X+Y\|_{\mathrm{F}}^2 = \mathrm{Tr}(X^*X + X^*Y + Y^*X + Y^*Y) = \mathrm{Tr}(X^*X) + \mathrm{Tr}(Y^*Y) = \|X\|_{\mathrm{F}}^2 + \|Y\|_{\mathrm{F}}^2$$

In the above, we use the fact that $\mathrm{Tr}(A+B) = \mathrm{Tr}(A) + \mathrm{Tr}(B)$ for conformal matrices, A, B. ■

Lemma 18 (Boutsidis & Gittens 2013, Lem. 5.3)**.** *Let $A \in \mathbb{C}^{m\times n}$ and $Y \in \mathbb{C}^{m\times \ell}$, then for all $X \in \mathbb{C}^{\ell\times n}$ and $\xi \in \{2, \mathrm{F}\}$:*

$$\|A - YY^+A\|_\xi \leq \|A - YX\|_\xi.$$

We are now ready to prove Theorem 7.

Proof of Theorem 7. Our argument relies on an equivalence for the range of Q. Since we have that Q is an orthonormal basis of $[Y_1 \quad Y_2]$. Then we have that Q is also the unique orthonormal basis of $\begin{bmatrix} Y_1Y_1^+Y_1 & (I - Y_1Y_1^+)Y_2 \end{bmatrix}$ by the definition of the Moore-Penrose Inverse (see the first property of the pseudoinverse in Section 4.1) and the classical Gram-Schmidt orthonormalization procedure (Schmidt, 1907). Let $X \in \mathbb{R}^{2\ell\times n}$, then we have from Lemma 18,

$$\begin{aligned}
\|A - YY^+A\|_\xi^2 &\leq \left\| A - [P_{Y_1}A\Omega_1 \quad (I - P_{Y_1})A\Omega_2] \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \right\|_\xi^2 \\
&\leq \underbrace{\|P_{Y_1}A - P_{Y_1}A\Psi_1X_1\|_\xi^2}_{\mathcal{A}} + \underbrace{\|(A - P_{Y_1}A) - (A - P_{Y_1}A)C^{1/2}\Psi_2X_2\|_\xi^2}_{\mathcal{B}}.
\end{aligned}$$

In the above, the second relation follows from Matrix Pythagoras (see Lemma 17). Let the SVD of $P_{Y_1}A$ be $\widehat{U}\widehat{\Sigma}\widehat{V}^*$, we then choose $X_1 = (\widehat{V}_{k+p}^*\Psi_1)^+\widehat{V}_{k+p}^*$, then we have $\widehat{V}_{k+p}^*\Psi_1(\widehat{V}_{k+p}^*\Psi_1)^+\widehat{V}_{k+p}^* = \widehat{V}_{k+p}^*$ from our assumption that $\widehat{V}_{k+p}^*\Psi_1$ is full-rank. Then we can note that $\widehat{V}_{k+p}^*\Psi_1$ is invertible almost surely with probability 1. We then obtain,

$$\mathcal{A} = \|P_{Y_1}A - P_{Y_1}A\Psi_1X_1\|_\xi^2 = \|P_{Y_1}A - P_{Y_1}A\|_\xi^2 = 0.$$

We then have for the second term,

$$\begin{aligned}
\mathcal{B} &= \|(A - P_{Y_1}A) - (A - P_{Y_1}A)C^{1/2}\Psi_2X_2\|_\xi^2 \\
&= \|(A - AC^{1/2}\Psi_2X_2) - (P_{Y_1}A - P_{Y_1}AC^{1/2}\Psi_2X_2)\|_\xi^2.
\end{aligned}$$

Then, setting $X_2 = (V_k^*C^{1/2}\Psi_2)^+V_k^*$, we obtain

$$\begin{aligned}
\mathcal{B}^{1/2} &\leq \|A_{k,\perp} - P_{Y_1}A_{k,\perp}\|_\xi + \|(A_{k,\perp} - P_{Y_1}A_{k,\perp})C^{1/2}\Psi_2(V_k^*C^{1/2}\Psi_2)^+\|_\xi \\
&\leq \|A_{k,\perp} - P_{A_{k,\perp}\Psi_1}A_{k,\perp}\|_\xi + \|(A_{k,\perp} - P_{A_{k,\perp}\Psi_1}A_{k,\perp})C^{1/2}\Psi_2(V_k^*C^{1/2}\Psi_2)^+\|_\xi \\
&\leq \|A_{k,\perp} - P_{A_{k,\perp}\Psi_1}A_{k,\perp}\|_\xi \left(1 + \|V_{k,\perp}^*C^{1/2}\Psi_2(V_k^*C^{1/2}\Psi_2)^+\|_2\right) \\
&\leq (1+\epsilon)\|A - A_{2k}\|_\xi \left(1 + \sqrt{8\log((n-k)(k+p))}\|V_{k,\perp}^*C^{1/2}\|_2\|(V_k^*C^{1/2})^+\|_2 \frac{\mathrm{e}\sqrt{k+p}}{p+1}\right).
\end{aligned}$$

In the above, the second relation follows from the Conjugation Rule (see Proposition 16) with $P_{A_{k,\perp}\Psi_1} \preceq P_{A\Psi_1}$. Suppose for two conformal matrices, Z_1, Z_2 such that $\mathrm{range}(Z_1) \subset \mathrm{range}(Z_2)$, it then follows that $P_{Z_1} \preceq P_{Z_2}$, and we have for the Frobenius norm,

$$\begin{aligned}\|(I - P_{Z_2})A\|_{\mathrm{F}}^2 &= \sum_{i\in[n]} \|(I - P_{Z_2})A\mathbf{e}_i\|_2^2 = \sum_{i\in[n]} \mathbf{e}_i^* A^*(I - P_{Z_2})A\mathbf{e}_i \\ &\leq \sum_{i\in[n]} \mathbf{e}_i^* A^*(I - P_{Z_1})A\mathbf{e}_i = \|(I - P_{Z_1})A\|_{\mathrm{F}}^2.\end{aligned}$$

In the above, the first and final relations follow from an alternative definition of the Frobenius Norm, the second relation follows from the definition of the spectral norm, and the third relation follows from the Conjugation Rule. Then, noting that $\mathrm{range}(A_{k,\perp}\Psi_1) \subset \mathrm{range}(A\Psi_1)$, we obtain the desired result. A similar result for the spectral norm follows from a simpler argument. The final inequality follows from the standard rSVD Frobenius norm error bound (see e.g. Halko et al. 2011, Thm. 10.6) for $k + p$ samples on the matrix $A - A_k$, we thus obtain

$$\|A_{k,\perp} - P_{A_{k,\perp}\Psi_1} A_{k,\perp}\|_\xi \leq (1 + \epsilon)\|A - A_{2k}\|_\xi,$$

and the second term follows from the manipulations in the spectral norm bound proof given in Theorem 13 with failure probability less than $\delta + t^{-p+1} \leq 0.01 + 0.0325 \leq 0.05$. ■

With the ideas in the proof of Theorem 7, we can extend our proof to multiple adaptive sampling iterations.

Proof of Theorem 8. The proof follows similarly to the above theorem. Let $\widetilde{\Omega} = [\Omega \quad \Omega_t]$ be all the sampled vectors. We furthermore decompose $\Omega = [\Omega_1 \quad \cdots \quad \Omega_{t-1}]$. Note $\Omega_i \in \mathbb{R}^{n\times(k+p)}$ and let $Y = A\Omega$ and $Y_i = A\Omega_i$ for all $i \in [t]$. We then obtain,

$$\begin{aligned}\|A - QQ^*A\|_{\mathrm{F}}^2 &\leq \left\| A - [P_Y A\Omega \quad (I - P_Y)A\Omega_t] \begin{bmatrix} X \\ X_t \end{bmatrix} \right\|_\xi^2 \\ &= \sum_{i=1}^{t-1} \underbrace{\|P_{Y_i}A - P_{Y_i}AC_i^{1/2}\Psi_i X_i\|_\xi^2}_{\mathcal{A}_i} + \underbrace{\|(A - P_Y A) - (A - P_Y A)C^{1/2}\Psi_t X_t\|_\xi^2}_{\mathcal{B}}.\end{aligned}$$

We can then note that by choosing $C_i^{1/2} = (P_{Y_i}A)^+(P_{Y_i}A)$, we have that $P_{Y_i}AC_i^{1/2} = P_{Y_i}A$ and then choosing $X_i = (\widehat{V}_{k+p}^* C_i^{1/2}\Psi)^+\widehat{V}_{k+p}^*$, with probability almost surely unitary, we obtain

$$\sum_{i=1}^{t-1} \mathcal{A}_i \overset{\text{def}}{=} \sum_{i=1}^{t-1} \|P_{Y_i}A - P_{Y_i}AC_i^{1/2}\Psi_i X_i\|_\xi^2 = \sum_{i=1}^{t-1} \|P_{Y_i}A - P_{Y_i}A\|_\xi^2 = 0$$

Then, considering the second term, we have

$$\begin{aligned}\mathcal{B} &= \|(A - P_Y A) - (A - P_Y A)C_t^{1/2}\Psi_t X_t\|_\xi^2 \\ &= \|(A - AC_t^{1/2}\Psi_t X_t) - (P_Y A - P_Y AC_t^{1/2}\Psi_t X_t)\|_\xi^2\end{aligned}$$

Let $\mathcal{I} = \{kt, \ldots, k(t+1)\}$. We now set $X_t = (V_\mathcal{I}^* C_t^{1/2}\Psi_t)^+ V_\mathcal{I}^*$. We then obtain,

$$\begin{aligned}\mathcal{B}^{1/2} &\leq \|A_\mathcal{I} - P_Y A_\mathcal{I}\|_\xi \left(1 + \|V_{\mathcal{I},\perp}^* C_t^{1/2}\Psi_t (V_\mathcal{I}^* C_t^{1/2}\Psi_t)^+\|_2\right) \\ &\leq \|A_{\mathcal{I},\perp} - P_{A_{\mathcal{I},\perp}\Omega} A_{\mathcal{I},\perp}\|_\xi \left(1 + \sqrt{8\log(n/\delta)}\|V_{\mathcal{I},\perp}^* C_t^{1/2}\|_2 \|(V_\mathcal{I}^* C_t^{1/2})^+\|_2 \frac{\mathrm{e}\sqrt{k+p}}{p+1}\right)\end{aligned}$$

In the above, we observe when C_t is a good approximation of $V_\mathcal{I}$ we obtain an improvement. Let us now consider the term $\|A_\mathcal{I} - P_{A_\mathcal{I}} A_\mathcal{I}\|_\xi$.

$$\|A_\mathcal{I} - P_{A_\mathcal{I}\Omega} A_\mathcal{I}\|_\xi^2 \leq \left\| A_\mathcal{I} - [P_{A_\mathcal{I}\Omega} A_\mathcal{I}\Omega \quad (I - P_{A_\mathcal{I}\Omega})A\Omega_{t-1}] \begin{bmatrix} X \\ X_{t-1} \end{bmatrix} \right\|_\xi^2$$

$$\leq \underbrace{\|P_{A_{\mathcal{I}}\Omega}A_{\mathcal{I}} - P_{A_{\mathcal{I}}\Omega}A_{\mathcal{I}}\Omega X\|_{\xi}^{2}}_{I} + \underbrace{\|(A_{\mathcal{I}} - P_{A_{\mathcal{I}}\Omega}A) - (A_{\mathcal{I}} - P_{A_{\mathcal{I}}\Omega}A)C_{t-1}^{1/2}\Psi_{t-1}X_{t-1}\|_{\xi}^{2}}_{II}$$

We will now consider I. Let the SVD of $P_{A_{\mathcal{I}}\Omega_i}A_{\mathcal{I}} = U_{\mathcal{I},i}\Sigma_{\mathcal{I},i}V_{\mathcal{I},i}^*$. Then we choose

$$X_i = \left(\left(V_{\mathcal{I},i}^*\right)_{k+p}C_i^{1/2}\Psi_i\right)^+\left(V_{\mathcal{I},i}^*\right)_{k+p}$$

From which we obtain,

$$I = \sum_{i=1}^{t-1}\|P_{A_{\mathcal{I}}\Omega}A_{\mathcal{I}} - P_{A_{\mathcal{I}}\Omega}A_{\mathcal{I}}C_i^{1/2}\Psi_i X_i\|_{\xi}^2 = \sum_{i=1}^{t-1}\|P_{A_{\mathcal{I}}\Omega_i}A_{\mathcal{I}} - P_{A_{\mathcal{I}}\Omega_i}A_{\mathcal{I}}\|_{\xi}^2 = 0$$

Let $\mathcal{F} = \{k(t-1), \dots, k(t+1)\}$. Next, we set $X_{t-1} = (V_{\mathcal{I}}^*C_{t-1}^{1/2}\Psi_{t-1})^+V_{\mathcal{I}}^*$, we then obtain

$$\begin{aligned}\sqrt{II} &\leq \|A_{\mathcal{F},\perp} - P_{A_{\mathcal{F},\perp}\Omega}A_{\mathcal{F},\perp}\|_{\xi}\left(1 + \|V_{\mathcal{I},\perp}^*C_t^{1/2}\Psi_{t-1}(V_{\mathcal{I}}^*C_{t-1}^{1/2}\Psi_{t-1})^+V_{\mathcal{I}}^*\|_2\right)\\ &\leq \|A_{\mathcal{F},\perp} - P_{A_{\mathcal{F},\perp}\Omega}A_{\mathcal{F},\perp}\|_{\xi}\left(1 + \sqrt{8\log(n/\delta)}\|V_{\mathcal{I},\perp}^*C_{t-1}^{1/2}\|_2\|(V_{\mathcal{I}}^*C_{t-1}^{1/2})^+\|_2\frac{\mathrm{e}\sqrt{k+p}}{p+1}\right)\end{aligned}$$

We can then complete the proof with induction, which gives us the desired result. ■

B Probability Theory

In this section, we present and prove an interesting property of Gaussian matrices that gives us the improved bounds for the generalized rSVD.

Proposition 19. *Fix matrices $S \in \mathbb{R}^{k\times n}$ and $T \in \mathbb{R}^{m\times \ell}$, then for a conformal matrix G with elements sampled i.i.d from $\mathcal{N}(0,1)$, then for all $\xi \in \{2, \mathrm{F}\}$ and $u \geq 1$,*

$$\|SGT\|_{\xi} \leq \|S\|_{\xi}\|T\|_{\xi}\sqrt{2u\log(nm)},$$

with failure probability at most $\sqrt{2/\pi}(nm)^{1-u}$.

Proof. The proof for the Frobenius norm follows from a brute-force calculation followed by a maximal tail bound on a sample of Gaussians.

$$\begin{aligned}\|SGT\|_{\mathrm{F}}^2 &= \sum_{i\in[k]}\sum_{j\in[\ell]}\sum_{(k_1,k_2)\in[n]\times[m]} S_{i,k_1}^2T_{k_2,j}^2G_{k_1,k_2}^2\\ &\leq \sum_{i\in[k]}\sum_{j\in[\ell]}\sum_{(k_1,k_2)\in[n]\times[m]} S_{i,k_1}^2T_{k_2,j}^2\max_{(k_1,k_2)\in[n]\times[m]}G_{k_1,k_2}^2 \qquad (10)\\ &= \|S\|_{\mathrm{F}}^2\|T\|_{\mathrm{F}}^2\max_{(k_1,k_2)\in[n]\times[m]}G_{k_1,k_2}^2.\end{aligned}$$

We now show an analogous result for the spectral norm with just a few more steps to recover the result. We will use classical techniques in probability theory to obtain a similar proof structure as the Frobenius norm result. From Rigollet & Hütter (2023), we have the following relation for any matrix $B \in \mathbb{R}^{m\times n}$,

$$\|B\|_2 = \sup_{\mathbf{v}\in\mathbb{S}^{n-1}}\|B\mathbf{v}\|_2 = \sup_{\mathbf{u},\mathbf{v}\in\mathbb{S}^{k-1}}|\mathbf{u}^*B\mathbf{v}|. \qquad (11)$$

With the relation given in Equation (11) in hand, we have

$$\begin{aligned}\|SGT\|_2^2 = \sup_{\mathbf{u},\mathbf{v}\in\mathbb{S}^{k-1}}(\mathbf{u}^*SGT\mathbf{v})^2 &= \sup_{\mathbf{u},\mathbf{v}\in\mathbb{S}^{k-1}}\sum_{(i,j)\in[n]\times[m]}(\mathbf{u}^*S)_i^2G_{i,j}^2(T\mathbf{v})_j^2\\ &\leq \sup_{\mathbf{u},\mathbf{v}\in\mathbb{S}^{k-1}}\sum_{(i,j)\in[n]\times[m]}(\mathbf{u}^*S)_i^2(T\mathbf{v})_j^2\max_{(i,j)\in[n]\times[m]}G_{i,j}^2 \qquad (12)\\ &\leq \sup_{\mathbf{u},\mathbf{v}\in\mathbb{S}^{k-1}}\|S^*\mathbf{u}\|_2^2\|T\mathbf{v}\|_2^2\max_{(i,j)\in[n]\times[m]}G_{i,j}^2 = \|S\|_2^2\|T\|_2^2\max_{(i,j)\in[n]\times[m]}G_{i,j}^2.\end{aligned}$$

Combining our results from Equations (10) and (12), we obtain for all $\xi \in \{2, \mathrm{F}\}$,

$$\|SGT\|_\xi^2 \leq \|S\|_\xi^2 \|T\|_\xi^2 \max_{(i,j)\in[n]\times[m]} G_{i,j}^2.$$

We now bound the maximum Gaussian over a finite sample.

$$\mathbf{Pr}\left\{\max_{(i,j)\in[n]\times[m]} G_{k_1,k_2}^2 \geq ut\right\} = \mathbf{Pr}\left\{\max_{(i,j)\in[n]\times[m]} |G_{k_1,k_2}| \geq \sqrt{ut}\right\}$$
$$\leq \frac{\sqrt{2}nm}{\sqrt{\pi}} \int_{\sqrt{ut}}^{\infty} e^{-x^2/2} dx \leq \frac{\sqrt{2}nm}{\sqrt{\pi}} \int_{\sqrt{ut}}^{\infty} \frac{xe^{-x^2/2}}{\sqrt{ut}} dx = \frac{\sqrt{2}nm}{\sqrt{\pi}} e^{-ut/2}.$$

In the above, the first inequality follows from a union bound over $[n] \times [m]$, and integrating over the PDF of a standard normal Gaussian (Berger & Casella, 2001). Then, from a change of variables $t = 2\log(nm)$,

$$\|SGT\|_\xi^2 \leq 2u\log(nm)\|S\|_\xi^2 \|T\|_\xi^2,$$

with failure probability at most $\sqrt{2/\pi}(nm)^{1-u}$. Taking the square root of both sides completes the proof. ■

C THEORETICAL IMPROVEMENTS TO RSVD

In this section, we take a closer look at our spectral norm bound given in Theorem 13. We show that our bound implies a stronger relative spectral norm error bound than (Halko et al., 2011, Thm. 10.8). We next will show the conditions where our Frobenius norm error bound is stronger than the bound derived by Persson et al. (2024). We will first restate the implication of (Halko et al., 2011, Thm. 10.8).

Theorem 20 (Implication of HMT Theorem 10.8)**.** *Let $A \in \mathbb{C}^{m\times n}$, choose target rank $k \geq 2$ and oversampling parameter $p \geq 4$. Then sample $\Omega \in \mathbb{R}^{m\times n}$ with standard normal entries. Construct the sample matrix $Y = A\Omega$ and let QR be the economized QR decomposition of Y. Then,*

$$\|A - QQ^*A\|_2 \leq \|A - A_k\|_2 \left(1 + t\sqrt{12k/p} + t \cdot \frac{\mathrm{e}\sqrt{k+p}}{p+1}\sqrt{n-k} + ut \cdot \frac{\sqrt{k+p}}{p+1}\right),$$

with failure probability at most $5t^{-p} + \mathrm{e}^{-u^2/2}$.

The dependence on the $O(\sqrt{n-k})$ factor arises from Chevet's inequality (Gordon, 1985).

Lemma 21 (Chevet's inequality)**.** *Fix matrices $S \in \mathbb{R}^{k\times n}$ and $\mathbb{T} \in \mathbb{R}^{m\times \ell}$, then for a Gaussian matrix $G \in \mathbb{R}^{n\times \ell}$ with elements sampled i.i.d from $\mathcal{N}(0,1)$,*

$$\mathbb{E}\|SGT\|_2 \leq \|S\|_2\|T\|_\mathrm{F} + \|S\|_\mathrm{F}\|T\|_2.$$

In the rSVD analysis (see e.g. (Halko et al., 2011, Thm. 10.7) or (Boutsidis & Gittens, 2013, Lem. 5.5)), $S = A - A_k$, and thus to obtain a relative error upper bound, we note $A - A_k$ is maximally rank $n - k$ and use the following inequality,

$$\|A - A_k\|_\mathrm{F} \leq \sqrt{n-k}\|A - A_k\|_2.$$

Although this upper bound is loose with strong singular value decay, we show we can give a stronger bound without any assumptions on the singular value spectra. We now will formalize our comparison by giving a corollary where we set $C = I$ and use our result from Theorem 13.

Corollary 22. *Consider the same hypothesis as Theorem 20, then for all $u \geq 1$,*

$$\|A - QQ^*A\|_2 \leq \|A - A_k\|_2 \left(1 + \sqrt{2ut\log((n-k)(k+p))}\frac{\mathrm{e}\sqrt{k+p}}{p+1}\right),$$

with failure probability at most $\sqrt{2/\pi}(nm)^{1-u} + t^{-p}$.

Proof. Consider the spectral norm result of Theorem 13 and set $C = I$. ■

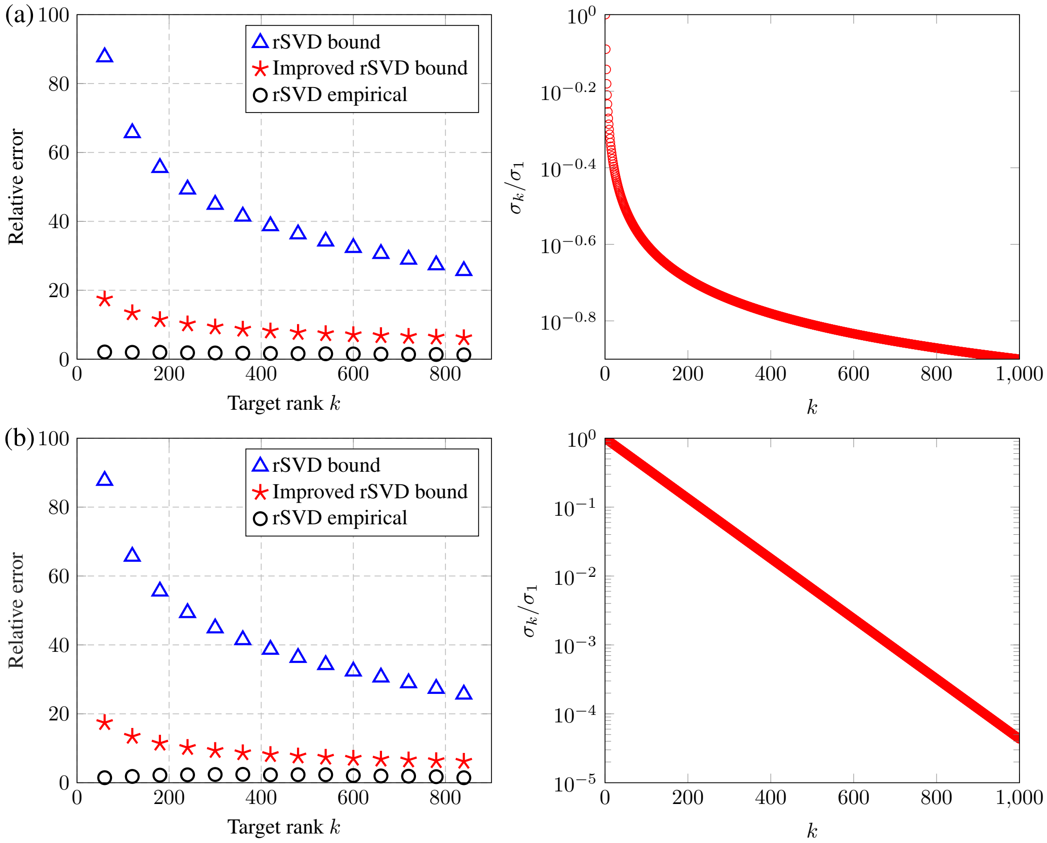

Figure 2: Left panels: Relative error (in the spectral norm) of the low-rank matrix approximation with respect to the number of matrix-vector products. At each target rank k, we set the oversampling parameter $p = \lceil k/4 \rceil$. Right panels: Singular value decay of the matrix. (a) Polynomial singular value decay (see Equation (13) with $p = 0.3$). (b) Exponential singular value decay (see Equation (14) with $\delta = 0.05$). Empirical Mean Error points in Figure 2 are averaged over 10 randomized runs.

Our analysis allows us to remove the $O(\sqrt{n-k})$ term and reduce it to a $O(\log((n-k)\ell))$ term, giving us a stronger bound. We showcase the improved relative error bound in the following synthetic experiment. We observe our bounds offer a significant improvement for relative-error spectral norm $\|A - Q(Q^*A)_k\|_2 / \|A_{k,\perp}\|_2$ compared to Theorem 20 and also are bounds are a proper upper bound.

We now give the Frobenius norm error bound with tools derived in Persson et al. (2024).

Theorem 23. *Let $A \in \mathbb{C}^{m\times n}$, let $k \geq 2$ be a target rank, and $p \geq 4$ be an oversampling parameter. Let Ω be a random sketch matrix with columns sampled i.i.d from $\mathcal{N}(0, C)$, then*

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

Proof. Following (Persson et al., 2024, Lem. 2.2) and Jensen's inequality, for $u \geq 1$, we have

$$\|SGT\|_{\mathrm{F}} = \|SGT\|_{(2)} \leq (\mathbb{E}[\|SGT\|_{(2)}^2])^{1/2} + u\|S\|_2\|T\|_2,$$

with probability $\geq 1 - e^{-u^2}$. Then, by (Persson et al., 2024, Lem. A.2),

$$\|SGT\|_{\mathrm{F}} \leq \|S\|_{\mathrm{F}}\|T\|_{\mathrm{F}} + u\|S\|_2\|T\|_2.$$

If we do the change of variables $\delta = e^{-u^2}$ and bound $\|\cdot\|_2 \leq \|\cdot\|_{\mathrm{F}}$, we obtain

$$\|SGT\|_{\mathrm{F}} \leq (1 + \sqrt{\log(1/\delta)})\|S\|_{\mathrm{F}}\|T\|_{\mathrm{F}}.$$

■

D Additional details on the numerical experiments

NB: This requires a lot more details and discussion. In this section, we perform additional experiments on synthetic matrices with polynomial and exponential singular value decay.

Experiment 4: Matrices with Polynomial and Exponential Singular Value Decay. We sample U randomly according to the Haar Measure over $\mathbb{O}_{m,m}$ and V randomly according to the Haar Measure over $\mathbb{O}_{n,n}$. We then form the matrices A in the scheme described in Equation (13) and Equation (14), respectively.

$$A = \sum_{i\in[n]} i^{-p} \cdot \mathbf{u}_i \mathbf{v}_i^*, \quad U \in \mathbb{O}_{m,m}, V \in \mathbb{O}_{n,n} \tag{13}$$

$$A = \sum_{i\in[n]} (1-\delta)^i \cdot \mathbf{u}_i \mathbf{v}_i^*, \quad U \in \mathbb{O}_{m,m}, V \in \mathbb{O}_{n,n}. \tag{14}$$

We observe Algorithm 1 is stronger than the rSVD and generalized rSVD in both experiments. Furthermore, we are able to observe that the variance of AS-rSVD is less than rSVD and generalized rSVD. AS-rSVD has comparable runtime to the rSVD and grSVD, where the dominating cost is the computation of the covariance matrix in each iteration.

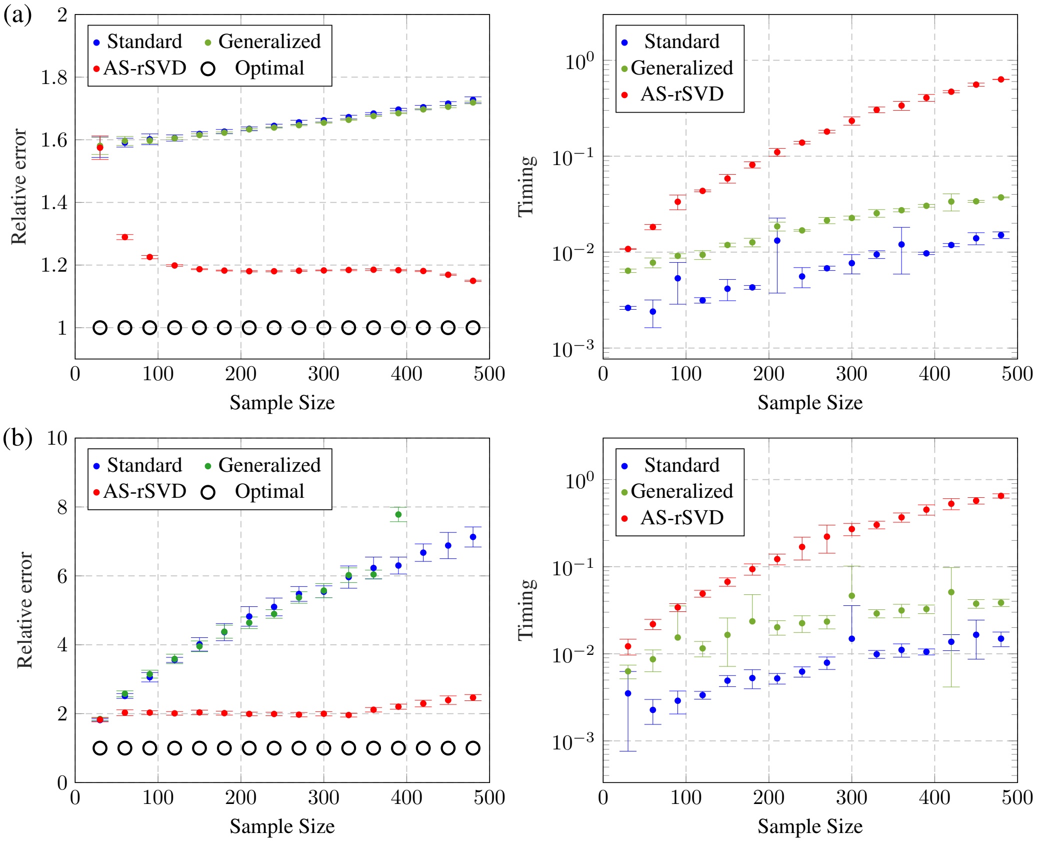

Figure 3: Left panels: Relative error (in the Frobenius norm) of the low-rank matrix approximation with respect to the number of matrix-vector products. Right panels: Run-time in seconds. (a) Polynomial singular value decay (see Equation (13) with $p = 1$). (b) Exponential singular value decay (see Equation (14) with $\delta = 0.05$).