ACCELERATING POWER METHOD WITH FAST SKETCHING FOR STRONGER LOW-RANK APPROXIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The power method is one of the most fundamental tools for extracting top principal components from data through low-rank matrix approximation. Yet, when the target rank is large, the cost of matrix multiplication associated with this procedure becomes a major bottleneck. We develop an algorithmic and theoretical framework for accelerating the power method using fast sketching, which is a popular paradigm in randomized linear algebra. Our framework leads to simple and provably efficient methods for singular value decomposition, low-rank factorization, and Nyström approximation, which attain strong numerical performance on benchmark problems. The key novelty in our analysis is the use of regularized spectral approximation, a property of fast sketching methods which proves more flexible in generalizing power method guarantees than traditional arguments.

1 INTRODUCTION

Computing top singular values and singular vectors of a large matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is one of the most fundamental tasks in data science. A simple yet powerful approach for doing this is the *power method*. In its most basic form, the power method starts from a random (say, Gaussian) vector $\boldsymbol{\omega} \in \mathbb{R}^n$. Then, it repeatedly applies matrix $\mathbf{A}$ and its transpose, obtaining a new vector $\mathbf{y} = (\mathbf{A}\mathbf{A}^\top)^q\mathbf{A}\boldsymbol{\omega}$. As q grows large, this vector becomes more and more closely aligned with the top singular vector of $\mathbf{A}$. A natural extension of this strategy when seeking top k singular vectors is to start from a random Gaussian matrix $\boldsymbol{\Omega} \in \mathbb{R}^{n\times O(k)}$, and then compute an orthonormal basis for the columns of a powered-up matrix $\mathbf{Y} = (\mathbf{A}\mathbf{A}^\top)^q\mathbf{A}\boldsymbol{\Omega}$. The resulting matrix $\mathbf{Q} = \mathrm{orth}(\mathbf{Y})$, known as the *range finder*, defines a subspace that is closely aligned with the span of the top k singular vectors of $\mathbf{A}$. This simple procedure offers strong theoretical guarantees for its approximation quality and lies at the core of widely used algorithms such as Randomized SVD (Halko et al., 2011) as well as many other low-rank factorization techniques (Rokhlin et al., 2010; Gu, 2015; Tropp & Webber, 2023). However, as k gets large, the matrix multiplication cost associated with repeatedly applying $\mathbf{A}$ becomes a major bottleneck of this approach.

Another powerful tool for low-rank approximation is *fast sketching*, which is widely used to avoid expensive matrix multiplication in dimensionality reduction algorithms (Sarlos, 2006; Clarkson & Woodruff, 2017; Woodruff, 2014). For this approach, instead of a dense Gaussian matrix $\boldsymbol{\Omega}$, we use a carefully constructed random matrix $\mathbf{S} \in \mathbb{R}^{n\times O(k)}$ to compute a so-called sketch $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{S}$ of the matrix $\mathbf{A}$, where the associated matrix product can be computed almost as quickly as it takes to read matrix $\mathbf{A}$, thanks to the structure imposed on $\mathbf{S}$ (e.g., because $\mathbf{S}$ is an extremely sparse random matrix (Clarkson & Woodruff, 2017)). Since $\mathbf{A}\mathbf{S}$ can be viewed as a rough version of the power method initialized from $\mathbf{S}$ (by setting $q = 0$), fast sketching has been used to construct range finders and low-rank factorizations under highly constrained computational budgets. A seemingly natural refinement of this would be to combine fast sketching with the power method by computing $(\mathbf{A}\mathbf{A}^\top)^q\mathbf{A}\mathbf{S}$ for some $q \geq 1$, however subsequent applications of $\mathbf{A}$ no longer take advantage of fast sketching, thereby largely erasing the benefit of this approach (Tropp & Webber, 2023).

Another natural strategy for accelerating the power method with fast sketching is to replace each application of $\mathbf{A}$ with an application of a sketch $\tilde{\mathbf{A}}$ (while still initializing with a Gaussian matrix $\boldsymbol{\Omega}$). This leads to the *sketched power method* for constructing a range finder (Chang & Yang, 2026):

$$\mathbf{Q} := \mathrm{orth}\big((\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top)^q\tilde{\mathbf{A}}\boldsymbol{\Omega}\big), \qquad \text{where} \quad \tilde{\mathbf{A}} = \mathbf{A}\mathbf{S} \in \mathbb{R}^{m\times O(l)}, \quad \boldsymbol{\Omega} \in \mathbb{R}^{n\times O(k)}, \quad l \gg k. \tag{1}$$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This strategy has the potential to be effective as long as the sketch $\tilde{\mathbf{A}}$ is sufficiently larger than $\boldsymbol{\Omega}$ (i.e., $l \gg k$), so that it carries enough information about the spectrum of $\mathbf{A}$, but also sufficiently smaller than $\mathbf{A}$ (i.e., $l \ll n$), to reduce the computational cost of repeated matrix multiplications. Crucially, the cost of fast sketching methods is not affected by choosing a very large sketch size, so we can attain this regime essentially for free. The intuition is that, while the sketched power method no longer converges to a perfect alignment with the top-k subspace of $\mathbf{A}$ as q grows large, it may reach a sufficiently good approximation much faster than the classical power method due to smaller per-iteration cost. Surprisingly, despite the simplicity of this procedure and significant prior work on low-rank approximation via both fast sketching (e.g., see Woodruff, 2014) and the power method (e.g., see Tropp & Webber, 2023), little is known about the effectiveness of the sketched power method, except in the special case where $\mathbf{S}$ is Gaussian (Chang & Yang, 2026) which largely erases any computational benefits of the approach. See Appendix A for a detailed discussion of related work.

1.1 Our contributions

In this work, we develop a theoretical framework for analyzing the effectiveness of the sketched power method under a variety of fast sketching algorithms, and show how it can be used to improve the efficiency of standard tools for top-k SVD and low-rank factorization.

1. **Sketch-Powered Range Finder**. We show that (1) yields an effective algorithm for constructing a range finder $\mathbf{Q}$ whose approximation error is not much larger than if it was computed directly as $\mathbf{Q} = \mathrm{orth}(\mathbf{AS})$. Thus, as long as the sketch size is sufficiently larger than the target rank (i.e., $l \gg k$), fast sketching is guaranteed to make the first several steps of the power method more efficient at attaining a strong approximation.
2. **Sketch-Powered Low-Rank Factorization**. The simplest strategy for using the range finder to obtain a low-rank factorization is the Randomized SVD algorithm (Halko et al., 2011). However, this approach still requires one additional expensive matrix product. We show how to avoid this bottleneck by incorporating (1) into another commonly used low-rank factorization technique (generalized Nyström, (Nakatsukasa, 2020)), and analyze the resulting approximation guarantee.
3. **Sketch-Powered Nyström Approximation**. We further optimize our approach for the task of factorizing positive semidefinite matrices. By incorporating sketched power method into the classical Nyström approximation, we obtain additional computational gains for this task.

In particular, our results yield the following simple and remarkable conclusion (Theorem 2.3):

Given $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $k \le O(\sqrt{n})$, sketch-powered low-rank factorization takes $\tilde{O}(n^2)$ time to compute a rank $O(k)$ factorization $\widehat{\mathbf{A}} = \mathbf{YX}$ such that $\|\mathbf{A} - \widehat{\mathbf{A}}\| \le \sqrt{k+1} \cdot \sigma_{k+1}(\mathbf{A})$.

Without sketched power method, the commonly used low-rank factorization schemes either only attain a weaker $\|\mathbf{A} - \widehat{\mathbf{A}}\| = O(\sqrt{n/k}) \cdot \sigma_{k+1}(\mathbf{A})$ approximation guarantee in the same time budget, or they invoke the classical power method which leads to a higher $\tilde{O}(n^2 k)$ time complexity. See Table 1 for a detailed comparison. We confirm these computational gains empirically by comparing Randomized SVD with and without sketched power method on large matrices from benchmark datasets, as well as similarly evaluating our sketch-powered low-rank factorization algorithm.

1.2 Overview of the techniques

Classical analysis of the power method and range finder construction typically follows one of two common strategies: Either it relies on the special rotational invariance properties of the Gaussian distribution (Halko et al., 2011; Chang & Yang, 2026), or it builds off of the *subspace embedding* property (Sarlos, 2006; Cohen et al., 2016), which is one of the corner-stones of fast sketching analysis (Woodruff, 2014). However, neither of these approaches appears directly suitable to the sketched power method, perhaps owing to the fact that it combines two different sources of randomness: a structured sketching matrix $\mathbf{S}$ and a dense Gaussian initialization $\boldsymbol{\Omega}$.

We depart from the classical analysis, and instead build on a different property of fast sketching called *regularized spectral approximation*, which holds when a sketch $\tilde{\mathbf{A}} = \mathbf{AS}$ satisfies $\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top + \lambda\mathbf{I} \approx_\varepsilon$

$\mathbf{A}\mathbf{A}^\top + \lambda\mathbf{I}$ for a suitably chosen regularization parameter $\lambda \geq 0$ and accuracy $\varepsilon \in [0, 1]$. This property has been used in prior works primarily in the context of sketching for kernel ridge regression or regularized least squares (Alaoui & Mahoney, 2015; Musco & Musco, 2017; Rudi et al., 2018; Garg & Dereziński, 2025), however it seems less obviously suitable for the analysis of the power method, since there is no explicit regularization occurring in the algorithm.

We first show that a simple argument based on the regularized spectral approximation property can be used to recover classical approximation guarantees for the range finder and the power method, when initialized not only with a Gaussian starting matrix, but also with many fast sketching matrices (Lemma 3.3). We then extend this argument to the sketched power method, by showing that a subspace approximation guarantee with respect to $\tilde{\mathbf{A}}$ can be translated to a guarantee with respect to $\mathbf{A}$, up to an additive error that depends only on the regularized spectral approximation quality of $\tilde{\mathbf{A}}$ (Theorem 3.5). Finally, in order to turn this approach into a proper low-rank factorization, we further extend the analysis to the generalized Nyström approximation by introducing another fast sketching matrix and building on sketching guarantees for generalized regression (Cohen et al., 2016).

Notation. For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ we use $\|\mathbf{A}\|$ and $\|\mathbf{A}\|_F$ to denote the spectral and Frobenius norm, respectively, with $\sigma_i(\mathbf{A})$ denoting its ith largest singular value and $\mathbf{A}^\dagger$ denoting the Moore-Penrose pseudoinverse. For a positive semidefinite (psd) matrix $\mathbf{M}$, we use $\lambda_i(\mathbf{M})$ to denote its ith largest eigenvalue, and we use $\mathbf{M} \preceq \mathbf{N}$ to denote the Loewner ordering between symmetric matrices. We let $\mathrm{orth}(\mathbf{A})$ denote the matrix with orthonormal columns spanning the range of $\mathbf{A}$.

2 Main Results

In this section, we incorporate sketched power method into three low-rank approximation algorithms, and provide theoretical guarantees for the running time and approximation error of these methods.

Sketch-Powered Range Finder. We start with the most basic version of our approach, which is constructing the range finder $\mathbf{Q}$ as described in Section 1. This is formally given in Algorithm 1.

Algorithm 1 Sketch-Powered Range Finder

inputs: $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{S} \in \mathbb{R}^{n\times r_1}$, $\mathbf{\Omega} \in \mathbb{R}^{r_1\times r_2}$, $q \in \mathbb{N}$
output: $\mathbf{Q} \in \mathbb{R}^{m\times n}$

1: Compute $\mathbf{Y} := (\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top)^q\tilde{\mathbf{A}}\mathbf{\Omega}$, where $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{S}$.
2: Orthonormalize the columns of $\mathbf{Y}$ to obtain $\mathbf{Q} \in \mathbb{R}^{m\times r_2}$.
3: **return** $\mathbf{Q}$

Comparison with Classical Range Finder. If we let $\mathbf{S}$ be the identity matrix, then this recovers the classical range finder construction. Here, for simplicity of presentation, we omit the intermediate re-orthonormalization step which is often performed in practice after each iteration of the power method for the sake of numerical stability (this does not affect the output in exact arithmetic). The guarantee for the classical range finder construction with block size $r_2 = O(k)$ is (Halko et al., 2011):

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\top\mathbf{A}\|^2 \leq \Big(2\sigma_{k+1}^2(\mathbf{A}) + \frac{1}{k}\sum_{i>k}\sigma_i^2(\mathbf{A})\Big)^{\frac{1}{2q+1}} \leq \Big(\frac{\min\{m,n\}}{k}\Big)^{\frac{1}{2q+1}}\sigma_{k+1}^2(\mathbf{A}),$$

where $\|\mathbf{A}-\mathbf{Q}\mathbf{Q}^\top\mathbf{A}\|^2$ measures how much of the spectrum of $\mathbf{A}$ is captured in the column span of $\mathbf{Q}$. By choosing $q = \tilde{O}(1/\varepsilon)$, we can bound this error by $(1+\varepsilon)\sigma_{k+1}^2(\mathbf{A})$. However, when k is large this requires expensive matrix multiplication which takes $\tilde{O}(\mathrm{nnz}(\mathbf{A})k/\varepsilon)$ time for $\mathbf{A}$ with $\mathrm{nnz}(\mathbf{A})$ non-zero entries, or $\tilde{O}(mnk/\varepsilon)$ for a dense $\mathbf{A}$. To avoid this cost, one can choose $q = 0$ and replace $\mathbf{\Omega}$ with a fast sketching matrix, but then the squared error can be as large as $\frac{\min\{m,n\}}{k}\sigma_{k+1}^2(\mathbf{A})$.

Our result. By introducing fast sketching directly into the power method, we can substantially optimize this trade-off. In particular, relying on sparse random sketching matrices $\mathbf{S}$ (Clarkson & Woodruff, 2017; Nelson & Nguyên, 2013; Chenakkod et al., 2025), we obtain the following guarantee for Sketch-Powered Range Finder (see Section 3 for details).

Table 1: Runtime and guarantees for computing an $n \times n$ rank k factorization $\hat{\mathbf{A}} = \mathbf{YX}$ such that $\|\mathbf{A} - \hat{\mathbf{A}}\| \leq \alpha \cdot \sigma_{k+1}(\mathbf{A})$, where $n \geq l \geq k$ and l is the parameter of sketched power method. We note that Chang & Yang (2026) consider a similar algorithm to ours, however their analysis crucially relies on the rotational invariance of Gaussian sketching for the matrix $\mathbf{S}$, which makes their algorithm even more expensive than vanilla power method (see Appendix A for details).

Method	Runtime $\tilde{O}(\cdot)$	Factor α	Source
Power method	$\text{nnz}(\mathbf{A})k/\varepsilon + nk^2/\varepsilon$	$1+\varepsilon$	(Halko et al., 2011)
Krylov iteration	$\text{nnz}(\mathbf{A})k/\sqrt{\varepsilon} + nk^2/\varepsilon$	$1+\varepsilon$	(Musco & Musco, 2015)
Fast sketching	$\text{nnz}(\mathbf{A}) + nk^2$	$O(\sqrt{n/k})$	(Cohen et al., 2016)
Gaussian sketched power method	$\text{nnz}(\mathbf{A})l + nlk$	$O(\sqrt{n/l})$	(Chang & Yang, 2026)
Fast sketched power method	$\text{nnz}(\mathbf{A}) + nlk$	$O(\sqrt{n/l})$	This work (Theorem 2.3)

Theorem 2.1. *Given matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, $1 \leq k \leq l \leq \min\{m,n\}$ and $\varepsilon \in (0, 1/2]$, Algorithm 1 implemented using a sparse matrix $\mathbf{S}$ of size $r_1 = \tilde{O}(l/\varepsilon^2)$, a Gaussian matrix $\mathbf{\Omega}$ of size $r_2 = O(k)$, and $q = \tilde{O}(1/\varepsilon)$, in time $\tilde{O}(\text{nnz}(\mathbf{A})/\varepsilon + mlk/\varepsilon^3)$ computes $\mathbf{Q} \in \mathbb{R}^{m\times O(k)}$ such that, with probability $1 - k^{-O(1)}$,*

$$\|\mathbf{A} - \mathbf{QQ}^\top\mathbf{A}\|^2 \leq (1+\varepsilon)\sigma_{k+1}^2(\mathbf{A}) + \frac{\varepsilon}{l}\sum_{i>l}\sigma_i^2(\mathbf{A}).$$

Remark 2.2. Due to the presence of the additive error term, it is often better to optimize this bound by choosing a large l, rather than a small ε. For example, given $k \leq O(\sqrt{n})$, we can obtain $\|\mathbf{A} - \mathbf{QQ}^\top\mathbf{A}\| \leq \sqrt{k+1}\cdot\sigma_{k+1}(\mathbf{A})$ in time $\tilde{O}(mn)$ with $l = O(n/k)$ and $\varepsilon = \Theta(1)$. This is better than the $\sqrt{\min\{m,n\}/k}\cdot\sigma_{k+1}(\mathbf{A})$ bound attained by the classical range finder in the same time.

Sketch-Powered Low-Rank Factorization. We can obtain a low-rank factorization from the range finder $\mathbf{Q}$ by computing $\mathbf{B} = \mathbf{Q}^\top\mathbf{A}$ and returning factors $\mathbf{Q}, \mathbf{B}$. For instance, this strategy is employed by the Randomized SVD algorithm (Halko et al., 2011) to compute $\mathbf{U\Sigma V}^\top \approx \mathbf{A}$ by performing the SVD of $\mathbf{B} = \tilde{\mathbf{U}}\mathbf{\Sigma V}^\top$ and computing $\mathbf{U} = \mathbf{Q}\tilde{\mathbf{U}}$, where $\mathbf{\Sigma}$ is diagonal and $\tilde{\mathbf{U}}, \mathbf{V}$ are orthonormal. Thus, the sketched power method can effectively accelerate Randomized SVD, as shown numerically in Section 4. Yet, computing $\mathbf{B}$ still requires one expensive matrix product which can be a bottleneck.

We get around this issue by incorporating sketched power method into a slightly modified low-rank factorization scheme (Nakatsukasa, 2020), which uses a second fast sketching matrix to effectively approximate the projection of $\mathbf{A}$ onto the subspace defined by $\mathbf{Q}$. See Algorithm 2.

Algorithm 2 Sketch-Powered Low-Rank Factorization

inputs: $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{S}_1 \in \mathbb{R}^{n\times r_1}$, $\mathbf{S}_2 \in \mathbb{R}^{m\times r_1}$, $\mathbf{\Omega} \in \mathbb{R}^{r_1\times r_2}$ and $q \in \mathbb{N}$

output: factors $\mathbf{Y} \in \mathbb{R}^{m\times r_2}$ and $\mathbf{X} \in \mathbb{R}^{r_2\times n}$ defining $\hat{\mathbf{A}} = \mathbf{YX}$

1: Compute $\mathbf{Y} := (\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top)^q\tilde{\mathbf{A}}\mathbf{\Omega}$, where $\tilde{\mathbf{A}} = \mathbf{AS}_1$.
2: Compute $\mathbf{X} := (\mathbf{S}_2^\top\mathbf{Y})^\dagger(\mathbf{S}_2^\top\mathbf{A})$.
3: **return** $\mathbf{Y}, \mathbf{X}$

Theorem 2.3. *Given matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, $1 \leq k \leq l \leq \min\{m,n\}$ and $\varepsilon \in (0, 1/2]$, Algorithm 2 implemented using sparse matrices $\mathbf{S}_1, \mathbf{S}_2$ of size $r_1 = \tilde{O}(l/\epsilon^2)$, a Gaussian matrix $\mathbf{\Omega}$ of size $r_2 = O(k)$, and $q = \tilde{O}(1/\varepsilon)$, in time $\tilde{O}(\text{nnz}(\mathbf{A})/\varepsilon + mlk/\varepsilon^3 + nlk/\varepsilon^2)$ computes $\mathbf{Y} \in \mathbb{R}^{m\times O(k)}$ and $\mathbf{X} \in \mathbb{R}^{O(k)\times n}$ such that, with probability $1 - k^{-O(1)}$,*

$$\|\mathbf{A} - \mathbf{YX}\|^2 \leq (1+\varepsilon)\sigma_{k+1}^2(\mathbf{A}) + \frac{\varepsilon}{l}\sum_{i>l}\sigma_i^2(\mathbf{A}).$$

Remark 2.4. In Table 1, we provide a detailed comparison with prior works. For example, given a dense $n \times n$ matrix $\mathbf{A}$ and $k \leq O(\sqrt{n})$, in time $\tilde{O}(n^2)$ Algorithm 2 finds $\mathbf{Y}, \mathbf{X}^\top \in \mathbb{R}^{n\times O(k)}$ such that $\|\mathbf{A} - \mathbf{YX}\| \leq \sqrt{k+1}\cdot\sigma_{k+1}(\mathbf{A})$. This guarantee is not attained by existing approaches.

Sketch-Powered Nyström Approximation. When the matrix $\mathbf{A}$ is symmetric positive semidefinite (psd), then it is natural to seek a low-rank factorization that is also psd. A standard approach for doing this is the so-called Nyström approximation (Williams & Seeger, 2000; Gittens & Mahoney, 2013; Alaoui & Mahoney, 2015; Musco & Musco, 2017; Rudi et al., 2018), where we use a sketching matrix $\mathbf{S}$ to construct factors $\mathbf{C} = \mathbf{AS}$ and $\mathbf{W} = \mathbf{S}^\top\mathbf{AS}$ such that $\widehat{\mathbf{A}} = \mathbf{CW}^\dagger\mathbf{C}^\top$ forms a low-rank approximation of $\mathbf{A}$. In Algorithm 3, we show how to incorporate sketched power method into this procedure by relying on a large intermediate Nyström approximation $\tilde{\mathbf{C}}, \tilde{\mathbf{W}}$ to construct a smaller (but still accurate) one. An added computational advantage of this construction is that the power method is performed on the matrix $\tilde{\mathbf{W}}$ which has both dimensions reduced rather than just one.

Algorithm 3 Sketch-Powered Nyström Approximation

inputs: psd $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{S} \in \mathbb{R}^{n\times r_1}$, $\mathbf{\Omega} \in \mathbb{R}^{r_1\times r_2}$, $q \in \mathbb{N}$
output: factors $\mathbf{C} \in \mathbb{R}^{n\times r_2}$ and $\mathbf{W} \in \mathbb{R}^{r_2\times r_2}$ defining $\widehat{\mathbf{A}} = \mathbf{CW}^\dagger\mathbf{C}^\top$

1: Compute $\tilde{\mathbf{C}} := \mathbf{AS}$ and $\tilde{\mathbf{W}} := \mathbf{S}^\top\mathbf{AS}$.
2: Compute $\mathbf{Y} := \tilde{\mathbf{W}}^q\mathbf{\Omega}$.
3: Compute $\mathbf{C} := \tilde{\mathbf{C}}\mathbf{Y}$ and $\mathbf{W} := \mathbf{Y}^\top\tilde{\mathbf{W}}\mathbf{Y}$.
4: **return** $\mathbf{C}, \mathbf{W}$

Theorem 2.5. *Given a psd matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, $1 \le k \le l \le n$ and $\varepsilon \in (0, 1/2]$, Algorithm 3 implemented using a sparse matrix $\mathbf{S}$ of size $r_1 = \tilde{O}(l/\epsilon^2)$, a Gaussian matrix $\mathbf{\Omega}$ of size $r_2 = O(k)$, and $q = \tilde{O}(1/\varepsilon)$, in time $\tilde{O}(\mathrm{nnz}(\mathbf{A})/\varepsilon + nlk/\varepsilon^2 + l^2k/\varepsilon^5)$ computes $\mathbf{C} \in \mathbb{R}^{n\times O(k)}$ and a psd matrix $\mathbf{W} \in \mathbb{R}^{O(k)\times O(k)}$ such that, with probability $1 - k^{-O(1)}$,*

$$\|\mathbf{A} - \mathbf{CW}^\dagger\mathbf{C}^\top\| \le (1+\varepsilon)\lambda_{k+1}(\mathbf{A}) + \frac{\varepsilon}{l}\sum_{i>l}\lambda_i(\mathbf{A}).$$

3 Analysis via Regularized Spectral Approximation

Our analysis relies on the following approximation property for a sketch, which has been previously used for low-rank approximation in certain specialized settings such as kernel ridge regression (Alaoui & Mahoney, 2015; Musco & Musco, 2017; Rudi et al., 2018), but is not typically associated with guarantees for range finder construction.

Definition 3.1. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\lambda \ge 0$, and $\varepsilon \in [0, 1/2]$. For a matrix $\mathbf{S} \in \mathbb{R}^{n\times r}$, we say that $\mathbf{AS}$ is a λ-regularized ε-spectral approximation of $\mathbf{A}$ if

$$(1-\varepsilon)\big(\mathbf{AA}^\top + \lambda\mathbf{I}\big) \preceq \mathbf{ASS}^\top\mathbf{A}^\top + \lambda\mathbf{I} \preceq (1+\varepsilon)\big(\mathbf{AA}^\top + \lambda\mathbf{I}\big).$$

This property can be satisfied by most standard sketching matrices with an appropriate choice of dimension r and regularizer λ, as shown in the following lemma which is an immediate consequence of Dereziński & Sidford (2026, Lemma 12) combined with the discussion in Cohen et al. (2016, Sections 2.1.1-2.1.3).

Lemma 3.2. *Given $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\varepsilon, \delta \in (0, 1)$, and $1 \le k \le \mathrm{rank}(\mathbf{A})$, consider:*

$$\lambda = \frac{1}{k}\sum_{i>k}\sigma_i^2(\mathbf{A}). \tag{2}$$

The following distributions over a sketching matrix $\mathbf{S} \in \mathbb{R}^{n\times r}$ have the property that $\mathbf{AS}$ is a λ-regularized ε-spectral approximation of $\mathbf{A}$ with probability $1 - \delta$:

1. *$\mathbf{S}$ has independent sub-Gaussian entries scaled by $1/\sqrt{r}$, for example $\mathbf{S}_{ij} \sim \mathcal{N}(0, 1/r)$ or $\mathbf{S}_{ij} \sim \pm 1/\sqrt{r}$, and the sketch size satisfies $r \ge O((k + \log(1/\delta))/\varepsilon^2)$.*
2. *$\mathbf{S}$ is a CountSketch-type matrix with s random non-zero entries per row of the form $\pm 1/\sqrt{s}$, and one of the following holds:*
   (a) $r \ge O((k + \log(1/\varepsilon\delta))/\varepsilon^2)$ and $s \ge O(\log^2(k/\delta)/\varepsilon + \log^3(k/\delta))$;
   (b) $r \ge O(k\log(k/\delta)/\varepsilon^2)$ and $s \ge O(\log(k/\delta)/\varepsilon)$.

3. $\mathbf{S}$ *is a Subsampled Randomized Hadamard Transform, i.e.,* $\mathbf{S} = \frac{1}{\sqrt{r}}\mathbf{D}\mathbf{H}\mathbf{I}_{:,S}$*, where* $\mathbf{D}$ *is diagonal with random* ± 1 *entries,* $\mathbf{H}$ *is a Hadamard matrix, and* S *is a uniformly random subset of* $\{1, ..., n\}$ *with size* $r \geq O((k + \log(n/\delta)) \log(k/\delta)/\varepsilon^2)$.

We start by showing how the classical guarantees for range finder construction can be easily recovered up to constant factors from the regularized spectral approximation property. We consider this result to be folklore, as similar arguments have appeared in prior work, although in different contexts (Dereziński & Sidford, 2026).

Lemma 3.3. *Given matrix* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *and* $\mathbf{\Omega} \in \mathbb{R}^{n\times r}$*, let* $\mathbf{Q} = \mathrm{orth}(\mathbf{A}\mathbf{\Omega})$*. If* $\mathbf{A}\mathbf{\Omega}$ *is a* λ*-regularized* $1/2$*-spectral approximation of* $\mathbf{A}$ *for some* $\lambda \geq 0$*, then*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\top\mathbf{A}\|^2 \leq 2\lambda \quad \textit{and} \quad \|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\top\mathbf{A}\|_F^2 \leq 2 \cdot \min_l \Big\{ l\lambda + \sum_{i>l} \sigma_i^2(\mathbf{A}) \Big\}.$$

In particular, if λ *is as in* (2) *for some* $1 \leq k \leq n$ *(e.g.,* $\mathbf{\Omega} \in \mathbb{R}^{n\times O(k)}$ *is a Gaussian matrix), then:*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\top\mathbf{A}\|^2 \leq \frac{2}{k}\sum_{i>k} \sigma_i^2(\mathbf{A}) \quad \textit{and} \quad \|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\top\mathbf{A}\|_F^2 \leq 4\sum_{i>k} \sigma_i^2(\mathbf{A}).$$

Remark 3.4. While the above statement corresponds only to the power method with $q = 0$, one can easily extend it to $q > 0$ by replacing $\mathbf{A}$ with $(\mathbf{A}\mathbf{A}^\top)^q\mathbf{A}$ (Halko et al., 2011).

The key idea is to show the following property of $\mathbf{P}_{\mathbf{A}\mathbf{\Omega}} := \mathbf{Q}\mathbf{Q}^\top$ (projection onto the span of $\mathbf{Q}$): $\mathbf{I} - \mathbf{P}_{\mathbf{A}\mathbf{\Omega}} \preceq \lambda(\mathbf{A}\mathbf{\Omega}\mathbf{\Omega}^\top\mathbf{A}^\top + \lambda\mathbf{I})^{-1} \preceq 2\lambda(\mathbf{A}\mathbf{A}^\top + \lambda\mathbf{I})^{-1}$. This is obtained by combining the regularized spectral approximation property with the following two observations:

1. $\mathbf{I} = (\mathbf{A}\mathbf{\Omega}\mathbf{\Omega}^\top\mathbf{A}^\top + \lambda\mathbf{I})(\mathbf{A}\mathbf{\Omega}\mathbf{\Omega}^\top\mathbf{A}^\top + \lambda\mathbf{I})^{-1}$, and
2. $\mathbf{A}\mathbf{\Omega}\mathbf{\Omega}^\top\mathbf{A}^\top(\mathbf{A}\mathbf{\Omega}\mathbf{\Omega}^\top\mathbf{A}^\top + \lambda\mathbf{I})^{-1} \preceq \mathbf{A}\mathbf{\Omega}\mathbf{\Omega}^\top\mathbf{A}^\top(\mathbf{A}\mathbf{\Omega}\mathbf{\Omega}^\top\mathbf{A}^\top)^\dagger = \mathbf{P}_{\mathbf{A}\mathbf{\Omega}}$.

The full proof of Lemma 3.3 can be found in Appendix F.1. Our main technical result is essentially an extension of Lemma 3.3 to the sketched power method. Note that we provide bounds for both the spectral and Frobenius norm error, as both are needed later for Theorem 2.3. We provide the proof sketch for each bound in Sections 3.1 and 3.2, respectively.

Theorem 3.5. *Given* $\mathbf{A} \in \mathbb{R}^{m\times n}$*,* $\mathbf{S} \in \mathbb{R}^{n\times r_1}$*,* $\mathbf{\Omega} \in \mathbb{R}^{r_1\times r_2}$*, and* $q, \lambda_1, \lambda_2 \geq 0$*, suppose that:*

1. $\mathbf{A}\mathbf{S}$ *is a* λ_1*-regularized* ε*-spectral approximation of* $\mathbf{A}$ *for some* $\varepsilon \in [0, 1/2]$*;*
2. $\mathbf{B}\mathbf{\Omega}$ *is a* λ_2*-regularized* $1/2$*-spectral approximation of* $\mathbf{B} := (\mathbf{A}\mathbf{S}(\mathbf{A}\mathbf{S})^\top)^q\mathbf{A}\mathbf{S}$.

Then, $\mathbf{Q} = \mathrm{orth}(\mathbf{B}\mathbf{\Omega})$ *satisfies:*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\top\mathbf{A}\|^2 \leq (1 + 2\varepsilon)\Big((2\lambda_2)^{\frac{1}{2q+1}} + \varepsilon\lambda_1\Big), \tag{3}$$

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\top\mathbf{A}\|_F^2 \leq \min_r \Big\{ 8r\big(\lambda_2^{\frac{1}{2q+1}} + \lambda_1\big) + \sum_{i>r} \sigma_i^2(\mathbf{A}) \Big\}. \tag{4}$$

Intuitively, by choosing matrix $\mathbf{S}$ to be sufficiently larger than $\mathbf{\Omega}$, we can ensure that $\lambda_1 \ll \lambda_2$, so that performing a few steps of power iteration reduces the overall error until $\lambda_2^{1/(2q+1)} \approx \lambda_1$. Following this intuition and selecting the sketching matrices according to Lemma 3.2 yields Theorem 2.1.

Proof of Theorem 2.1. Let $\lambda_1 = \frac{1}{l}\sum_{i>l} \sigma_i^2(\mathbf{A})$. By Lemma 3.2, we can generate an $n \times r_1$ sparse random matrix $\mathbf{S}$ with $s = O(\log(l/\delta)/\varepsilon)$ non-zero entries per row such that with probability 1 - δ, $\tilde{\mathbf{A}} := \mathbf{A}\mathbf{S}$ is a λ_1-regularized ε-spectral approximation of $\mathbf{A}$ for $r_1 = O(l \log(l/\delta)/\varepsilon^2)$, so

$$(1 - \varepsilon)\big(\sigma_i^2(\mathbf{A}) + \lambda_1\big) \leq \sigma_i^2(\tilde{\mathbf{A}}) + \lambda_1 \leq (1 + \varepsilon)\big(\sigma_i^2(\mathbf{A}) + \lambda_1\big). \tag{5}$$

Let $\mathbf{B} = (\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top)^q\tilde{\mathbf{A}}$ and $\lambda_2 = \frac{1}{k}\sum_{i>k} \sigma_i^2(\mathbf{B})$. By Lemma 3.2, if $\mathbf{\Omega}$ is an $r_1 \times O(k + \log 1/\delta)$ Gaussian matrix, then with probability $1 - \delta$, $\mathbf{B}\mathbf{\Omega}$ is a λ_2-regularized $1/2$-spectral approximation of $\mathbf{B}$. Choosing $\delta = k^{-O(1)}$, we absorb it into the constant factors throughout the analysis.

To get the stated bound after applying Theorem 3.5, we express $(2\lambda_2)^{\frac{1}{(2q+1)}}$ in terms of $\sigma_i^2(\mathbf{A})$. Note that, $\sigma_i^2(\mathbf{B}) = \sigma_i(\tilde{\mathbf{A}})^{2(2q+1)} \leq \big((1+\varepsilon)\sigma_i(\mathbf{A})^2 + \lambda_1\varepsilon\big)^{2q+1}$ by (5). Thus, $(2\lambda_2)^{\frac{1}{(2q+1)}}$ is bounded by the ℓ_{2q+1} norm of a vector with coordinates $(1+\varepsilon)\sigma_i(\mathbf{A})^2 + \lambda_1\varepsilon$ for $i > k$. In Lemma F.1 we show that for $q = \Omega(\log(m)/\varepsilon)$, this ℓ_{2q+1} norm is essentially the same as the largest coordinate:

$$(2\lambda_2)^{\frac{1}{(2q+1)}} \leq (1+4\varepsilon)\sigma_{k+1}(\mathbf{A})^2 + 2\lambda_1\varepsilon.$$

Then, by Theorem 3.5, we have,

$$\begin{aligned}\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\top\mathbf{A}\|^2 &\leq (1+2\varepsilon)\Big((2\lambda_2)^{\frac{1}{(2q+1)}} + \varepsilon\lambda_1\Big)\\ &\leq (1+2\varepsilon)\big((1+4\varepsilon)\sigma_{k+1}(\mathbf{A})^2 + 3\lambda_1\varepsilon\big)\\ &\leq (1+10\varepsilon)\sigma_{k+1}(\mathbf{A})^2 + 6\lambda_1\varepsilon.\end{aligned}$$

Replacing ε by $\varepsilon/10$ recovers the bound. We can compute $\mathbf{AS}$ in time $O(\mathrm{nnz}(\mathbf{A})\log(l)/\varepsilon)$, while $(\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top)^q\tilde{\mathbf{A}}\mathbf{\Omega}$ takes $O(qmr_1r_2) = \tilde{O}(mlk/\varepsilon^3)$ time. Constructing $\mathbf{Q}$ takes $O(mk^2)$ time. □

Our psd low-rank factorization guarantee (Theorem 2.5) follows via a reduction from Theorem 2.1 by observing that $\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^\top = \mathbf{A}^{1/2}\mathbf{Q}\mathbf{Q}^\top\mathbf{A}^{1/2}$ when $\mathbf{Q}$ is the orthonormal basis returned by Algorithm 1 when applied to $\mathbf{A}^{1/2}$ with sketches $\mathbf{S}, \mathbf{\Omega}$. The full proof is in Appendix B.

Obtaining our general low-rank factorization guarantee (Theorem 2.3) requires additional effort, as it does not follow from Theorem 2.1. Here, we rely on both the spectral and Frobenius norm bounds in Theorem 3.5, as well as a sketching guarantee for generalized regression (see Appendix F.3).

Lemma 3.6 (Generalized regression (Cohen et al., 2016))**.** *Given $\mathbf{A} \in \mathbb{R}^{n\times d}, \mathbf{B} \in \mathbb{R}^{n\times p}$, consider the generalized regression problem of finding $\mathbf{X}$ to minimize $\|\mathbf{AX} - \mathbf{B}\|$. Let $\mathbf{S} \in \mathbb{R}^{n\times r}$ be one of the matrices from Lemma 3.2 for some $k \geq d$. Then, $\widetilde{\mathbf{X}} := (\mathbf{S}^\top\mathbf{A})^\dagger\mathbf{S}^\top\mathbf{B}$ with probability $1-\delta$ satisfies,*

$$\|\mathbf{A}\widetilde{\mathbf{X}} - \mathbf{B}\|^2 \leq (1+\varepsilon^2)\|\mathbf{P}_\mathbf{A}\mathbf{B} - \mathbf{B}\|^2 + \frac{\varepsilon^2}{k}\|\mathbf{P}_\mathbf{A}\mathbf{B} - \mathbf{B}\|_F^2.$$

Proof of Theorem 2.3. Let $\mathbf{B} := (\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top)^q\tilde{\mathbf{A}}$, where $\tilde{\mathbf{A}} = \mathbf{AS}_1$, and recall that $\mathbf{Y} = \mathbf{B\Omega}$. Let $\mathbf{S}_1$ have $s = O(\log(l/\delta)/\varepsilon)$ non-zero entries per row with $r_1 = O(l\log(l/\delta)/\varepsilon^2)$, and let $\mathbf{\Omega}$ be Gaussian with $r_2 = O(k)$. Then, by Theorem 3.5 (via Lemma 3.2), we get, with probability $1 - 2\delta = 1 - k^{-O(1)}$, where δ is chosen to be $k^{-O(1)}$ as in the proof of Theorem 2.1,

$$\|\mathbf{A} - \mathbf{P}_\mathbf{Y}\mathbf{A}\|^2 \leq (1+2\varepsilon)\Big((2\lambda_2)^{\frac{1}{2q+1}} + \varepsilon\lambda_1\Big), \tag{6}$$

$$\|\mathbf{A} - \mathbf{P}_\mathbf{Y}\mathbf{A}\|_F^2 \leq \min_r\Big\{8r(\lambda_2^{\frac{1}{2q+1}} + \lambda_1) + \sum_{i>r}\sigma_i^2(\mathbf{A})\Big\}. \tag{7}$$

where $\lambda_1 = \frac{1}{l}\sum_{i>l}\sigma_i^2(\mathbf{A})$ and $\lambda_2 = \frac{1}{k}\sum_{i>k}\sigma_i^2(\mathbf{B})$ as in the proof of Theorem 2.1. Let $\mathbf{S}_2 \in \mathbb{R}^{m\times r_1}$ have the same sparsity and sketching dimension as $\mathbf{S}_1$. Then, by Lemma 3.6, applied to the generalized regression problem $\min_\mathbf{X}\|\mathbf{YX} - \mathbf{A}\|$ with $\mathbf{X} := \big(\mathbf{S}_2^\top\mathbf{Y}\big)^\dagger\mathbf{S}_2^\top\mathbf{A}$,

$$\|\mathbf{YX} - \mathbf{A}\|^2 \leq (1+\varepsilon^2)\|\mathbf{P}_\mathbf{Y}\mathbf{A} - \mathbf{A}\|^2 + \frac{\varepsilon^2}{l}\|\mathbf{P}_\mathbf{Y}\mathbf{A} - \mathbf{A}\|_F^2.$$

Using (6), (7), and the bound $(2\lambda_2)^{\frac{1}{2q+1}} \leq (1+4\varepsilon)\sigma_{k+1}(\mathbf{A})^2 + 2\lambda_1\varepsilon$ from Lemma F.1,

$$\begin{aligned}\|\mathbf{YX} - \mathbf{A}\|^2 &\leq (1+\varepsilon^2)\big((1+10\varepsilon)\sigma_{k+1}(\mathbf{A})^2 + 6\lambda_1\varepsilon\big) + \frac{\varepsilon^2}{l}\left(8l(\lambda_2^{1/2q+1} + \lambda_1) + \sum_{i>l}\sigma_i^2(\mathbf{A})\right)\\ &\leq (1+40\varepsilon)\sigma_{k+1}(\mathbf{A})^2 + \frac{29\varepsilon}{l}\sum_{i>l}\sigma_i^2(\mathbf{A}).\end{aligned}$$

Replacing ε by $\varepsilon/40$, we are done. The matrix $\mathbf{Y}$ can be computed in time $\tilde{O}\big(\mathrm{nnz}(\mathbf{A}) + mlk/\varepsilon^3\big)$, whereas forming $\mathbf{X} = \big(\mathbf{S}_2^\top\mathbf{Y}\big)^\dagger\mathbf{S}_2^\top\mathbf{A}$ takes $\tilde{O}\big(\mathrm{nnz}(\mathbf{A}) + mk + lk^2/\varepsilon^2 + nlk/\varepsilon^2\big)$ time. □

3.1 Proof Sketch of Theorem 3.5: Spectral Norm Bound

We now return to the proof of our main technical result. We start with the spectral norm $\|\mathbf{A}-\mathbf{Q}\mathbf{Q}^\top\mathbf{A}\|$ where $\mathbf{Q} = \mathrm{orth}(\mathbf{B}\boldsymbol{\Omega})$ and $\mathbf{B} = (\mathbf{A}\mathbf{S}(\mathbf{A}\mathbf{S})^\top)^q\mathbf{A}\mathbf{S}$. Intuitively, the key difference between this and the classical power iteration is that we get a subspace which aligns more closely with the subspace spanned by the top k singular vectors of $\tilde{\mathbf{A}} := \mathbf{A}\mathbf{S}$ instead of $\mathbf{A}$. This means that if we project $\tilde{\mathbf{A}}$ onto the column space of $\mathbf{Q}$, we can get a good low-rank approximation of $\tilde{\mathbf{A}}$.

However, we need a low-rank approximation of $\mathbf{A}$, and not $\tilde{\mathbf{A}}$. So we use the regularized spectral approximation guarantee to show that the approximation error of the projection $\mathbf{P}_{\mathbf{B}\boldsymbol{\Omega}} := \mathbf{Q}\mathbf{Q}^\top$ on $\mathbf{A}$ is bounded by the approximation error of the projection $\mathbf{P}_{\mathbf{B}\boldsymbol{\Omega}}$ on $\tilde{\mathbf{A}}$. Formally, since $\tilde{\mathbf{A}}$ is a (ε, λ_1)-regularized spectral approximation of $\mathbf{A}$, we have, $\mathbf{A}\mathbf{A}^\top \preceq \frac{1}{1-\varepsilon}(\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top + \varepsilon\lambda_1\mathbf{I})$, which gives

$$(\mathbf{I}-\mathbf{P}_{\mathbf{B}\boldsymbol{\Omega}})\mathbf{A}\mathbf{A}^\top(\mathbf{I}-\mathbf{P}_{\mathbf{B}\boldsymbol{\Omega}}) \preceq (1+2\varepsilon)(\mathbf{I}-\mathbf{P}_{\mathbf{B}\boldsymbol{\Omega}})\big(\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top + \varepsilon\lambda_1\mathbf{I}\big)(\mathbf{I}-\mathbf{P}_{\mathbf{B}\boldsymbol{\Omega}})$$

where we use that when $\varepsilon \le 1/2$, $\frac{1}{1-\varepsilon} \le 1+2\varepsilon$. A careful application of Jensen's inequality gives

$$\|(\mathbf{I}-\mathbf{P}_{\mathbf{B}\boldsymbol{\Omega}})\mathbf{A}\|^2 \le (1+2\varepsilon)\left(\left\|(\mathbf{I}-\mathbf{P}_{\mathbf{B}\boldsymbol{\Omega}})\big(\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top\big)^{2q+1}(\mathbf{I}-\mathbf{P}_{\mathbf{B}\boldsymbol{\Omega}})\right\|^{\frac{1}{2q+1}} + \varepsilon\lambda_1\right).$$

The norm under the $\frac{1}{2q+1}$ exponent simplifies to $\|(\mathbf{I}-\mathbf{P}_{\mathbf{B}\boldsymbol{\Omega}})\mathbf{B}\|^2$, which can be bounded using Lemma 3.3, obtaining $\|(\mathbf{I}-\mathbf{P}_{\mathbf{B}\boldsymbol{\Omega}})\mathbf{B}\|^2 \le 2\lambda_2$. See the complete proof in Appendix C.

3.2 Proof Sketch of Theorem 3.5: Frobenius Norm Bound

We now turn to bounding the Frobenius norm $\|\mathbf{A}-\mathbf{Q}\mathbf{Q}^\top\mathbf{A}\|_F^2$, which is essential for our low-rank factorization guarantee (Theorem 2.3).

Let $\mathbf{P}_\perp := \mathbf{I}-\mathbf{P}_{\mathbf{B}\boldsymbol{\Omega}} = \mathbf{I}-\mathbf{Q}\mathbf{Q}^\top$. We bound the effect of $\mathbf{P}_\perp$ on $\mathbf{A}$ in two ways. By the Pythagorean theorem, we have $\|\mathbf{P}_\perp\mathbf{A}\|_F^2 = \|\mathbf{P}_\perp\mathbf{A}_r\|_F^2 + \|\mathbf{P}_\perp\mathbf{A}_{>r}\|_F^2$, where $\mathbf{A}_r$ represents the matrix formed by the top r principal vectors of $\mathbf{A}$ and $\mathbf{A}_{>r} := \mathbf{A}-\mathbf{A}_r$. The second term, $\|\mathbf{P}_\perp\mathbf{A}_{>r}\|_F^2$ can be at most $\|\mathbf{A}_{>r}\|_F^2 = \|\mathbf{A}-\mathbf{A}_r\|_F^2 = \sum_{i>r}\sigma_i^2(\mathbf{A})$, so the main challenge is bounding the first term.

Using the SVD of $\mathbf{A}_r$, we get, $\|(\mathbf{I}-\mathbf{P}_{\mathbf{B}\boldsymbol{\Omega}})\mathbf{A}_r\|_F^2 = \sum_{i=1}^r \sigma_i^2(\mathbf{A})\,\mathbf{u}_i^\top(\mathbf{I}-\mathbf{P}_{\mathbf{B}\boldsymbol{\Omega}})\mathbf{u}_i$, where $\mathbf{u}_i$ are the left singular vectors of $\mathbf{A}$. Following the proof of Lemma 3.3, since $\mathbf{B}\boldsymbol{\Omega}$ is a λ_2-regularized $1/2$-spectral approximation of $\mathbf{B}$, we have, $\mathbf{I}-\mathbf{P}_{\mathbf{B}\boldsymbol{\Omega}} \preceq 2\lambda_2(\mathbf{B}\mathbf{B}^\top + \lambda_2\mathbf{I})^{-1}$. It follows that

$$\|(\mathbf{I}-\mathbf{P}_{\mathbf{B}\boldsymbol{\Omega}})\mathbf{A}_r\|_F^2 = \sum_{i=1}^r \sigma_i^2(\mathbf{A})\,\mathbf{u}_i^\top(\mathbf{I}-\mathbf{P}_{\mathbf{B}\boldsymbol{\Omega}})\mathbf{u}_i \le \sum_{i=1}^r \sigma_i^2(\mathbf{A})\,2\lambda_2\mathbf{u}_i^\top\big(\mathbf{B}\mathbf{B}^\top+\lambda_2\mathbf{I}\big)^{-1}\mathbf{u}_i. \quad (8)$$

However, we are faced with the mismatch that $\mathbf{u}_i$ are the left singular vectors of $\mathbf{A}$, but we have them acting on $\mathbf{B}$, a power of $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{S}$, instead. Ideally we want, $\mathbf{u}_i^\top\big(\mathbf{B}\mathbf{B}^\top+\lambda_2\mathbf{I}\big)^{-1}\mathbf{u}_i \propto 1/\sigma_i^2(\mathbf{A})$ to kill the contribution of the large singular values, but a direct expansion of $\mathbf{u}_i^\top\big(\mathbf{B}\mathbf{B}^\top+\lambda_2\mathbf{I}\big)^{-1}\mathbf{u}_i$ gives us many $\sigma_j^2(\mathbf{B}) = \sigma_j\big((\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top)^{2q+1}\big)$ terms in the denominator.

To overcome this, we work with $\widetilde{\mathbf{B}} := (\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top)^q\tilde{\mathbf{A}}/\theta^{2q+1}$, a scaled version of $\mathbf{B}$. This does not change the algorithm since $\mathbf{P}_{\mathbf{B}\boldsymbol{\Omega}} = \mathbf{P}_{\widetilde{\mathbf{B}}\boldsymbol{\Omega}}$. Then, for the indices with $\sigma_j^2(\tilde{\mathbf{A}}) \ge \theta^2$ we have,

$$\sigma_j^2(\widetilde{\mathbf{B}}) = \frac{\sigma_j\big((\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top)^{(2q+1)}\big)}{\theta^{2(2q+1)}} \ge \left(\frac{\sigma_j^2(\tilde{\mathbf{A}})}{\theta^2}\right)^{2q+1} \ge \frac{\sigma_j^2(\tilde{\mathbf{A}})}{\theta^2}.$$

Thus, when working with $\widetilde{\mathbf{B}}$ instead of $\mathbf{B}$, we are able to replace the $\sigma_j^2(\widetilde{\mathbf{B}})$ terms that arise in (8) with $\sigma_j^2(\tilde{\mathbf{A}})$ when $\sigma_j^2(\tilde{\mathbf{A}})$ is large. This allows us to use the regularized spectral approximation guarantee for $\tilde{\mathbf{A}}$ with $\lambda_1' = \lambda_1 + \lambda_2^{1/2q+1}$ (where λ_1' is chosen to simplify the resulting calculations) obtaining,

$$\sigma_i^2(\mathbf{A})\,\mathbf{u}_i^\top(\mathbf{I}-\mathbf{P}_{\widetilde{\mathbf{B}}\boldsymbol{\Omega}})\mathbf{u}_i \le \sigma_i^2(\mathbf{A})\cdot\frac{C\lambda_1'}{\sigma_i^2(\mathbf{A})+\lambda_1'} = O(\lambda_1'),$$

since a given λ_1-regularized ε-spectral approximation $\tilde{\mathbf{A}}$ is also a λ_1'-regularized ε-spectral approximation for $\lambda_1' \ge \lambda_1$. Further details can be found in the full proof in Appendix D.

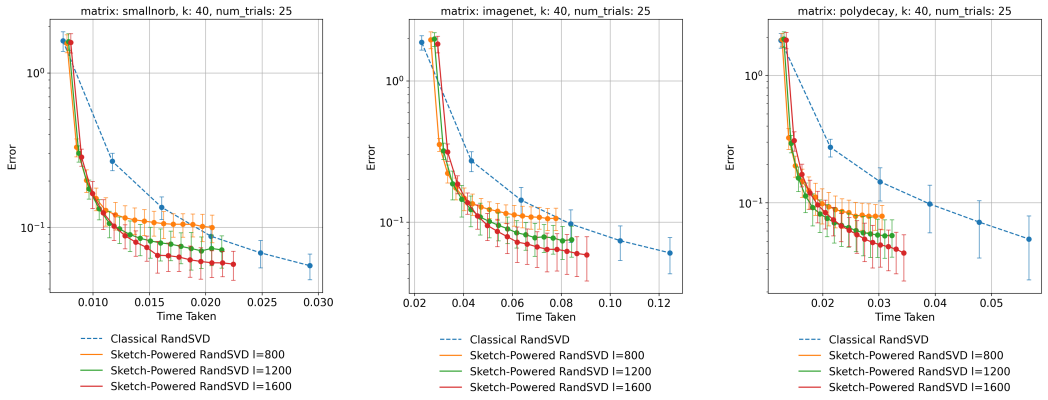

Figure 1: Approximation error vs time for Classical Randomized SVD (RandSVD) and Sketch-Powered RandSVD for (from left to right) `smallnorb`, `ImageNet`, and `polydecay` matrices.

4 Experiments

To validate our theory, we compare the numerical performance of sketched power method to its classical counterpart on three large data matrices. In this section, we evaluate a sketch-powered version of Randomized SVD, as it is arguably the most common application of the power method (Halko et al., 2011) (see Appendix E for more experiments). Concretely, we use Algorithm 1 to construct the range finder $\mathbf{Q}$, and then use that to compute an approximate SVD, $\mathbf{A} \approx \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{Q}\mathbf{Q}^\top\mathbf{A}$, as described in Section 2. Focusing on the regime where $\varepsilon = \Theta(1)$, we let $r_1 = l$ and $r_2 = k$.

The matrices we consider are: (1) `smallnorb` (LeCun et al., 2004) (size 24300×9216), (2) `ImageNet` (Deng et al., 2009) (size 20800×50176 matrix containing 50176 red channel pixel values of 20800 images resized to pixel size 224×224) and (3) `polydecay`, a synthetic matrix of size 30000×15000 generated by multiplying on both sides a diagonal matrix with polynomially decaying values $30000/i$ by random orthogonal matrices of the appropriate dimension. We let the sketching matrix $\mathbf{S} \in \mathbb{R}^{n\times l}$ in Algorithm 1 be the CountSketch matrix (Charikar et al., 2002; Clarkson & Woodruff, 2017) with one non-zero per row, and we use a Gaussian matrix $\mathbf{\Omega} \in \mathbb{R}^{l\times k}$ for the power method. The implementation was done using an A100 GPU in Google Colab.

The results of the comparison for target rank $k = 40$ are in Figure 1. For each iteration $q = 0, 1, \ldots$ we plot the average relative error $\|\mathbf{A} - \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\|/\sigma_{k+1}(\mathbf{A}) - 1$ for that iterate averaged over 25 trials versus the average time taken for that iterate. We vary the sketch size among $l \in \{800, 1200, 1600\}$.

Since the sketched power iterations are faster, we plot errors for 15 iterations against only 5 iterations for the classical power method. While the approximation error of the Sketch-Powered RandSVD does eventually start to plateau, the method reaches a moderately small relative error (say, 0.1) much sooner than the Classical RandSVD does. Such moderately small error is sufficient for many machine learning applications (Gittens & Mahoney, 2013), and the faster iteration time of the sketched power method more than makes up for the added overhead cost of forming the sketch $\mathbf{AS}$. As expected, we observe that there is some trade-off between the iteration speed and error plateau when varying l.

In Appendix E, we also provide a numerical evaluation of the performance of Sketch-Powered Low-Rank Factorization (Algorithm 2). Here, we observe similar computational improvements from using sketched power method as in the case of RandSVD.

5 Conclusions

We develop an algorithmic and theoretical framework for accelerating the power method using fast sketching. We use this framework to design and analyze efficient algorithms for singular value decomposition, low-rank factorization, and Nyström approximation. We demonstrate the effectiveness of this approach on several benchmark data matrices.

AI USE STATEMENT

In this work, we used generative AI tools only for LaTeX typesetting, translating mathematical algorithms presented in the paper into Python code and pre-processing datasets. We did not use generative AI tools to formulate mathematical claims or assist in the writing of proofs. We take full responsibility for the contents of the paper.

REFERENCES

Ahmed Alaoui and Michael W Mahoney. Fast randomized kernel ridge regression with statistical guarantees. *Advances in neural information processing systems*, 28, 2015.

Ainesh Bakshi, Kenneth L Clarkson, and David P Woodruff. Low-rank approximation with $1/\epsilon^{1/3}$ matrix-vector products. *arXiv preprint arXiv:2202.05120*, 2022.

Chao Chang and Yuning Yang. Improving sketching algorithms for low-rank matrix approximation via sketch-power iterations. *SIAM Journal on Matrix Analysis and Applications*, 47(3):1122–1165, 2026.

Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *Proceedings of the 29th International Colloquium on Automata, Languages and Programming*, ICALP '02, pp. 693–703, Berlin, Heidelberg, 2002. Springer-Verlag. ISBN 3540438645.

Tyler Chen, Ethan N Epperly, Raphael A Meyer, Christopher Musco, and Akash Rao. Does block size matter in randomized block krylov low-rank approximation? In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1026–1046. SIAM, 2026.

Shabarish Chenakkod, Michał Dereziński, and Xiaoyu Dong. Optimal oblivious subspace embeddings with near-optimal sparsity. In *52nd International Colloquium on Automata, Languages, and Programming (ICALP 2025)*, 2025.

Kenneth L Clarkson and David P Woodruff. Low-rank approximation and regression in input sparsity time. *Journal of the ACM (JACM)*, 63(6):1–45, 2017.

Michael B. Cohen, Jelani Nelson, and David P. Woodruff. Optimal Approximate Matrix Product in Terms of Stable Rank. In *43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016)*, volume 55, pp. 11:1–11:14, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

Michał Dereziński and Michael W Mahoney. Recent and upcoming developments in randomized numerical linear algebra for machine learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6470–6479, 2024.

Michal Dereziński and Aaron Sidford. Approaching optimality for solving dense linear systems with low-rank structure. In *ACM-SIAM Symposium on Discrete Algorithms*, 2026.

Petros Drineas and Michael W Mahoney. Randnla: randomized numerical linear algebra. *Communications of the ACM*, 59(6):80–90, 2016.

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.

Sachin Garg and Michał Dereziński. Faster low-rank approximation and kernel ridge regression via the block-nyström method. In *The Thirty Eighth Annual Conference on Learning Theory*, pp. 2291–2325. PMLR, 2025.

Alex Gittens and Michael Mahoney. Revisiting the nystrom method for improved large-scale machine learning. In *International Conference on Machine Learning*, pp. 567–575. PMLR, 2013.

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

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

Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. *Advances in neural information processing systems*, 27, 2014.

Moritz Hardt and Aaron Roth. Beyond worst-case analysis in private singular vector computation. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 331–340, 2013.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 665–674, 2013.

Jacek Kuczyński and Henryk Woźniakowski. Estimating the largest eigenvalue by the power and lanczos algorithms with a random start. *SIAM Journal on Matrix Analysis and Applications*, 13(4): 1094–1122, 1992.

Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pp. II–104. IEEE, 2004.

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

Raphael Meyer, Cameron Musco, and Christopher Musco. On the unreasonable effectiveness of single vector krylov methods for low-rank approximation. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 811–845. SIAM, 2024.

Ioannis Mitliagkas, Constantine Caramanis, and Prateek Jain. Memory limited, streaming pca. *Advances in neural information processing systems*, 26, 2013.

Riley Murray, James Demmel, Michael W Mahoney, N Benjamin Erichson, Maksim Melnichenko, Osman Asif Malik, Laura Grigori, Piotr Luszczek, Michał Dereziński, Miles E Lopes, et al. Randomized numerical linear algebra: A perspective on the field with an eye to software. *arXiv preprint arXiv:2302.11474*, 2023.

Cameron Musco and Christopher Musco. Randomized block krylov methods for stronger and faster approximate singular value decomposition. *Advances in neural information processing systems*, 28, 2015.

Cameron Musco and Christopher Musco. Recursive sampling for the nystrom method. *Advances in neural information processing systems*, 30, 2017.

Cameron Musco and David P Woodruff. Sublinear time low-rank approximation of positive semidefinite matrices. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 672–683. IEEE, 2017.

Yuji Nakatsukasa. Fast and stable randomized low-rank matrix approximation. *arXiv preprint arXiv:2009.11392*, 2020.

Jelani Nelson and Huy L Nguyên. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 ieee 54th annual symposium on foundations of computer science*, pp. 117–126. IEEE, 2013.

Vladimir Rokhlin, Arthur Szlam, and Mark Tygert. A randomized algorithm for principal component analysis. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1100–1124, 2010.

Alessandro Rudi, Daniele Calandriello, Luigi Carratino, and Lorenzo Rosasco. On fast leverage score sampling and optimal learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *2006 47th annual IEEE symposium on foundations of computer science (FOCS'06)*, pp. 143–152. IEEE, 2006.

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

Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. *Advances in neural information processing systems*, 13, 2000.

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

A Related Work

Early work in numerical linear algebra focused on using the power method for estimating the largest singular vector of a matrix (Kuczyński & Woźniakowski, 1992). The modern analysis for using the power method to construct the range finder and Randomized SVD was developed by Rokhlin et al. (2010), and then extended and popularized by Halko et al. (2011), among others (Gu, 2015; Tropp & Webber, 2023). In this context, the power method is also often referred to as subspace iteration or power iteration. When seeking high-accuracy low-rank approximations, i.e., with error $(1+\varepsilon)\sigma_{k+1}(\mathbf{A})$ for $\varepsilon \ll 1$, the power method can be accelerated using Krylov subspaces (Musco & Musco, 2015; Bakshi et al., 2022; Chen et al., 2026; Meyer et al., 2024), which requires fewer iterations but has a higher per-iteration cost. In this work, we focus on computationally constrained environments where even the per-iteration cost of the classical power method is a bottleneck, and so we are primarily interested in the $\varepsilon = \Omega(1)$ regime.

Several prior works have considered approximate versions of the power method. In particular, Hardt & Price (2014) study a noisy version of the power method, where each application of $\mathbf{A}$ is accompanied by some additive noise, a model that arises in applications to matrix completion (Jain et al., 2013), streaming algorithms (Mitliagkas et al., 2013), and differential privacy (Hardt & Roth, 2013). While Hardt & Price (2014) introduce constraints on the additive noise that scale with the spectral gap of $\mathbf{A}$, in our work, we are dealing with a very specific type of noise that comes from sketching, and are able to avoid dependence on any spectral gaps.

In a nearly concurrent work, Chang & Yang (2026) consider a sketched variant of the power method which is very similar to (1). However, even though they recommend the use of fast sketching matrices in practice, their theoretical analysis crucially requires the sketching matrices to be Gaussian, since they rely on the rotational invariance property of the Gaussian distribution. This limits any computational benefits of the approach (other than in single-pass streaming models). See Table 1 for a detailed runtime comparison. Our theoretical framework can be viewed as avoiding this key limitation through the use of the regularized spectral approximation argument, thus addressing a natural question arising from Chang & Yang (2026).

Fast sketching has been extensively used more broadly in the area of Randomized Numerical Linear Algebra (RandNLA, (Woodruff, 2014; Drineas & Mahoney, 2016; Martinsson & Tropp, 2020; Murray et al., 2023; Dereziński & Mahoney, 2024)), as well as specifically in the context of low-rank approximation, e.g., Sarlos (2006); Drineas et al. (2006); Clarkson & Woodruff (2017); Halko et al. (2011); Nakatsukasa (2020). Particularly relevant to our analysis is the line of works that has relied on the regularized spectral approximation property of sketching. This notion was first used for analyzing sub-sampling based methods in the context of Nyström approximation (Alaoui & Mahoney, 2015; Musco & Musco, 2017; Rudi et al., 2018; Garg & Dereziński, 2025) and psd low-rank approximation (Musco & Woodruff, 2017). In contrast, for sparse and structured sketching matrices such as those considered in this work, a more common approach has been to rely on the oblivious subspace embedding (OSE) property, e.g., Sarlos (2006); Woodruff (2014); Cohen et al. (2016). In fact, one can relatively straightforwardly pass from a slightly more general *OSE moments* property (Cohen et al., 2016) to regularized spectral approximation, e.g., as described by Dereziński & Sidford (2026).

B Proof of Theorem 2.5

Proof of Theorem 2.5. Let $\mathbf{Q}$ denote the orthonormal basis returned by Algorithm 1 when applied to $\mathbf{A}^{1/2}$ with sketches $\mathbf{S}, \mathbf{\Omega}$. Then $\mathbf{Q}$ is the orthonormalization of the columns of $\mathbf{K} := \big(\mathbf{A}^{1/2}\mathbf{S}(\mathbf{A}^{1/2}\mathbf{S})^\top\big)^q\mathbf{A}^{1/2}\mathbf{S}\mathbf{\Omega} = \mathbf{A}^{1/2}\mathbf{S}(\mathbf{S}^\top\mathbf{A}\mathbf{S})^q\mathbf{\Omega} = \mathbf{A}^{1/2}\mathbf{S}\mathbf{Y}$. Thus $\mathbf{K}^\top\mathbf{K} = \mathbf{Y}^\top\tilde{\mathbf{W}}\mathbf{Y}$ and therefore,

$$\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^\top = \mathbf{A}\mathbf{S}\mathbf{Y}\big(\mathbf{Y}^\top\tilde{\mathbf{W}}\mathbf{Y}\big)^\dagger\mathbf{Y}^\top\mathbf{S}^\top\mathbf{A} = \mathbf{A}^{1/2}\mathbf{K}\big(\mathbf{K}^\top\mathbf{K}\big)^\dagger\mathbf{K}^\top\mathbf{A}^{1/2} = \mathbf{A}^{1/2}\mathbf{Q}\mathbf{Q}^\top\mathbf{A}^{1/2}.$$

Applying Theorem 2.1 with $\mathbf{A}^{1/2}$ in place of $\mathbf{A}$ gives

$$\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^\top\| = \|\big(\mathbf{I} - \mathbf{Q}\mathbf{Q}^\top\big)\mathbf{A}^{1/2}\|^2 \leq (1+\varepsilon)\sigma_{k+1}^2(\mathbf{A}^{1/2}) + \frac{\varepsilon}{l}\sum_{i>l}\sigma_i^2(\mathbf{A}^{1/2}).$$

Since $\mathbf{A}$ is psd, $\sigma_i^2(\mathbf{A}^{1/2}) = \sigma_i(\mathbf{A}) = \lambda_i(\mathbf{A})$ for every i, which yields the bound. Since $\mathbf{S}$ is sparse, the matrices $\tilde{\mathbf{C}} = \mathbf{A}\mathbf{S}$ and $\tilde{\mathbf{W}} = \mathbf{S}^\top\tilde{\mathbf{C}}$ can be computed in $\tilde{O}(\mathrm{nnz}(\mathbf{A})/\varepsilon)$ and $\tilde{O}\big(ln/\varepsilon^3\big)$ time, respectively. Computing $\mathbf{Y} = \tilde{\mathbf{W}}^q\mathbf{\Omega}$ takes $\tilde{O}\big(kl^2/\varepsilon^5\big)$ time, and forming $\mathbf{C} = \tilde{\mathbf{C}}\mathbf{Y}$ takes $O\big(kln/\varepsilon^2\big)$ time. Computing $\mathbf{W} = \mathbf{Y}^\top\tilde{\mathbf{W}}\mathbf{Y}$ takes $O\big(kl^2/\varepsilon^4 + k^2l/\varepsilon^2\big)$ time. □

C Proof of Theorem 3.5: Spectral Norm Bound

Since $\mathbf{A}\mathbf{S}$ is an λ_1-regularized ε-spectral approximation of $\mathbf{A}$, we have,

$$(1-\varepsilon)\big(\mathbf{A}\mathbf{A}^\top + \lambda_1\mathbf{I}\big) \preceq \mathbf{A}\mathbf{S}(\mathbf{A}\mathbf{S})^\top + \lambda_1\mathbf{I} \preceq (1+\varepsilon)\big(\mathbf{A}\mathbf{A}^\top + \lambda_1\mathbf{I}\big) \tag{9}$$

which gives,

$$\mathbf{A}\mathbf{A}^\top \preceq \frac{1}{1-\varepsilon}\big(\mathbf{A}\mathbf{S}(\mathbf{A}\mathbf{S})^\top + \varepsilon\lambda_1\mathbf{I}\big)$$

Let $\mathbf{P}$ be any orthogonal projection. Then,

$$\mathbf{P}\mathbf{A}\mathbf{A}^\top\mathbf{P} \preceq \frac{1}{1-\varepsilon}\mathbf{P}\big(\mathbf{A}\mathbf{S}(\mathbf{A}\mathbf{S})^\top + \varepsilon\lambda_1\mathbf{I}\big)\mathbf{P}$$

Let $\mathbf{x}$ be a unit vector at which $\mathbf{P}\mathbf{A}\mathbf{A}^\top\mathbf{P}$ attains its norm. Then,

$$\|\mathbf{P}\mathbf{A}\|^2 \leq \frac{1}{1-\varepsilon}\mathbf{x}^\top\mathbf{P}\big(\mathbf{A}\mathbf{S}(\mathbf{A}\mathbf{S})^\top + \varepsilon\lambda_1\mathbf{I}\big)\mathbf{P}\mathbf{x}$$

Let $\mathbf{A}\mathbf{S}(\mathbf{A}\mathbf{S})^\top + \varepsilon\lambda_1\mathbf{I} = \mathbf{U}(\mathbf{D} + \varepsilon\lambda_1\mathbf{I})\mathbf{U}^\top$ be the eigendecomposition. Then,

$$\begin{aligned}\mathbf{x}^\top\mathbf{P}\big(\mathbf{A}\mathbf{S}(\mathbf{A}\mathbf{S})^\top + \varepsilon\lambda_1\mathbf{I}\big)\mathbf{P}\mathbf{x} &= \mathbf{x}^\top\mathbf{P}\mathbf{U}(\mathbf{D} + \varepsilon\lambda_1\mathbf{I})\mathbf{U}^\top\mathbf{P}\mathbf{x} \\ &= \sum_{i=1}^{m}\big(\sigma_i(\mathbf{A}\mathbf{S})^2 + \varepsilon\lambda_1\big)(\mathbf{U}^\top\mathbf{P}\mathbf{x})_i^2\end{aligned}$$

and,

$$\begin{aligned}
\|\mathbf{PA}\|^{2(2q+1)} &\le \frac{1}{(1-\varepsilon)^{2q+1}}\left(\sum_{i=1}^{m}(\sigma_i(\mathbf{AS})^2+\varepsilon\lambda_1)(\mathbf{U}^\top\mathbf{Px})_i^2\right)^{2q+1}\\
&= \left(\frac{\|\mathbf{U}^\top\mathbf{Px}\|_2^2}{1-\varepsilon}\right)^{2q+1}\left(\sum_{i=1}^{m}(\sigma_i(\mathbf{AS})^2+\varepsilon\lambda_1)\frac{(\mathbf{U}^\top\mathbf{Px})_i^2}{\|\mathbf{U}^\top\mathbf{Px}\|_2^2}\right)^{2q+1}\\
&\le \left(\frac{\|\mathbf{U}^\top\mathbf{Px}\|_2^2}{1-\varepsilon}\right)^{2q+1}\left(\sum_{i=1}^{m}(\sigma_i(\mathbf{AS})^2+\varepsilon\lambda_1)^{2q+1}\frac{(\mathbf{U}^\top\mathbf{Px})_i^2}{\|\mathbf{U}^\top\mathbf{Px}\|_2^2}\right) \quad \text{(Jensen)}\\
&\le \left(\frac{1}{1-\varepsilon}\right)^{2q+1}\|\mathbf{U}^\top\mathbf{Px}\|_2^{2q}\left(\sum_{i=1}^{m}(\sigma_i(\mathbf{AS})^2+\varepsilon\lambda_1)^{2q+1}(\mathbf{U}^\top\mathbf{Px})_i^2\right)\\
\implies \|\mathbf{PA}\|^2 &\le \left(\frac{1}{1-\varepsilon}\right)\left(\sum_{i=1}^{m}(\sigma_i(\mathbf{AS})^2+\varepsilon\lambda_1)^{2q+1}(\mathbf{U}^\top\mathbf{Px})_i^2\right)^{\frac{1}{2q+1}}\\
&\le \left(\frac{1}{1-\varepsilon}\right)\left(\left(\sum_{i=1}^{m}\sigma_i(\mathbf{AS})^{2(2q+1)}(\mathbf{U}^\top\mathbf{Px})_i^2\right)^{\frac{1}{2q+1}}+\left(\sum_{i=1}^{m}(\varepsilon\lambda_1)^{2q+1}(\mathbf{U}^\top\mathbf{Px})_i^2\right)^{\frac{1}{2q+1}}\right)\\
&\le \left(\frac{1}{1-\varepsilon}\right)\left(\left\|\mathbf{P}\big(\mathbf{AS}(\mathbf{AS})^\top\big)^{2q+1}\mathbf{P}\right\|^{\frac{1}{2q+1}}+\varepsilon\lambda_1\right)
\end{aligned}$$

where the second inequality after taking the $(2q+1)^{\text{th}}$ root follows from Minkowski's inequality (triangle inequality for L_q norms).

Now, $\left\|\mathbf{P}\big(\mathbf{AS}(\mathbf{AS})^\top\big)^{2q+1}\mathbf{P}\right\| = \left\|\mathbf{P}\big(\mathbf{AS}(\mathbf{AS})^\top\big)^{q}\mathbf{AS}\right\|^2$. Thus,

$$\|\mathbf{PA}\|^2 \le \left(\frac{1}{1-\varepsilon}\right)\left(\left\|\mathbf{P}\big(\mathbf{AS}(\mathbf{AS})^\top\big)^{q}\mathbf{AS}\right\|^{\frac{2}{(2q+1)}}+\varepsilon\lambda_1\right)$$

Let $\mathbf{B} = \big(\mathbf{AS}(\mathbf{AS})^\top\big)^q\mathbf{AS}$. Consider the secondary sketch $\mathbf{B\Omega}$ and let $\mathbf{P} = \mathbf{I} - \mathbf{P}_{\mathbf{B\Omega}}$, where $\mathbf{P}_{\mathbf{B\Omega}}$ is the orthogonal projection onto the column span of $\mathbf{B\Omega}$. Then,

$$\|(\mathbf{I}-\mathbf{P}_{\mathbf{B\Omega}})\mathbf{A}\|^2 \le \left(\frac{1}{1-\varepsilon}\right)\Big(\|(\mathbf{I}-\mathbf{P}_{\mathbf{B\Omega}})\mathbf{B}\|^{\frac{2}{(2q+1)}}+\varepsilon\lambda_1\Big)$$

If $\mathbf{B\Omega}$ is a λ_2-regularized $1/2$-spectral approximation of $\mathbf{B}$, by Lemma 3.3, $\|(\mathbf{I}-\mathbf{P}_{\mathbf{B\Omega}})\mathbf{B}\|^2 \le 2\lambda_2$. So we get, when $\varepsilon \le 1/2$,

$$\|(\mathbf{I}-\mathbf{P}_{\mathbf{B\Omega}})\mathbf{A}\|^2 \le (1+2\varepsilon)\Big((2\lambda_2)^{\frac{1}{(2q+1)}}+\varepsilon\lambda_1\Big)$$

D Proof of Theorem 3.5: Frobenius Norm Bound

Let $\mathbf{P} = \mathbf{QQ}^\top$. We want to estimate $\|(\mathbf{I}-\mathbf{P})\mathbf{A}\|_F^2$. Let $\mathbf{A} = \mathbf{A}_r + \mathbf{A}_{>r}$ be the split given by the SVD of $\mathbf{A}$, where $\mathbf{A}_r = \sum_{i=1}^{r}\sigma_i(\mathbf{A})\mathbf{u}_i\mathbf{v}_i^\top$ retains the top r principal directions and $\mathbf{A}_{>r} = \sum_{i=r+1}^{p}\sigma_i(\mathbf{A})\mathbf{u}_i\mathbf{v}_i^\top$ contains the remaining directions where p is the rank of $\mathbf{A}$. Since $\{\mathbf{v}_i\}_{i=1}^r$ and $\{\mathbf{v}_i\}_{i=r+1}^p$ are orthogonal sets of right singular vectors, the cross terms vanish and we have

$$\begin{aligned}
\|(\mathbf{I}-\mathbf{P})\mathbf{A}\|_F^2 &= \|(\mathbf{I}-\mathbf{P})(\mathbf{A}_r+\mathbf{A}_{>r})\|_F^2\\
&= \|(\mathbf{I}-\mathbf{P})\mathbf{A}_r\|_F^2 + 2\langle(\mathbf{I}-\mathbf{P})\mathbf{A}_r, (\mathbf{I}-\mathbf{P})\mathbf{A}_{>r}\rangle_F + \|(\mathbf{I}-\mathbf{P})\mathbf{A}_{>r}\|_F^2\\
&= \|(\mathbf{I}-\mathbf{P})\mathbf{A}_r\|_F^2 + \|(\mathbf{I}-\mathbf{P})\mathbf{A}_{>r}\|_F^2
\end{aligned}$$

where the last equality uses $\langle (\mathbf{I}-\mathbf{P})\mathbf{A}_r, (\mathbf{I}-\mathbf{P})\mathbf{A}_{>r}\rangle_F = \operatorname{tr}\big((\mathbf{I}-\mathbf{P})\mathbf{A}_{>r}\mathbf{A}_r^\top(\mathbf{I}-\mathbf{P})\big) = 0$, since $\mathbf{A}_{>r}\mathbf{A}_r^\top = \sum_{i=1}^r\sum_{j=r+1}^p \sigma_i(\mathbf{A})\sigma_j(\mathbf{A})\mathbf{u}_j\mathbf{v}_j^\top\mathbf{v}_i\mathbf{u}_i^\top = 0$ by orthogonality of right singular vectors.

For the first term, we expand using $\mathbf{A}_r = \sum_{i=1}^r \sigma_i(\mathbf{A})\mathbf{u}_i\mathbf{v}_i^\top$:

$$\begin{aligned}
\|(\mathbf{I}-\mathbf{P})\mathbf{A}_r\|_F^2 &= \operatorname{tr}\big((\mathbf{I}-\mathbf{P})\mathbf{A}_r\mathbf{A}_r^\top(\mathbf{I}-\mathbf{P})\big) \\
&= \operatorname{tr}\left((\mathbf{I}-\mathbf{P})\sum_{i=1}^r\sum_{j=1}^r \sigma_i(\mathbf{A})\sigma_j(\mathbf{A})\mathbf{u}_i\mathbf{v}_i^\top\mathbf{v}_j\mathbf{u}_j^\top(\mathbf{I}-\mathbf{P})\right) \\
&= \sum_{i=1}^r \sigma_i(\mathbf{A})^2\ \operatorname{tr}\big((\mathbf{I}-\mathbf{P})\mathbf{u}_i\mathbf{u}_i^\top(\mathbf{I}-\mathbf{P})\big) \\
&= \sum_{i=1}^r \sigma_i(\mathbf{A})^2\,\mathbf{u}_i^\top(\mathbf{I}-\mathbf{P})^2\mathbf{u}_i = \sum_{i=1}^r \sigma_i(\mathbf{A})^2\,\mathbf{u}_i^\top(\mathbf{I}-\mathbf{P})\mathbf{u}_i
\end{aligned}$$

where the third equality uses orthonormality of the right singular vectors $\mathbf{v}_i^\top\mathbf{v}_j = \delta_{ij}$, and the last uses $(\mathbf{I}-\mathbf{P})^2 = \mathbf{I}-\mathbf{P}$.

For the second term, since $\mathbf{I}-\mathbf{P}$ is an orthogonal projection we have $(\mathbf{I}-\mathbf{P})^2 = \mathbf{I}-\mathbf{P} \preceq \mathbf{I}$, so

$$\begin{aligned}
\|(\mathbf{I}-\mathbf{P})\mathbf{A}_{>r}\|_F^2 &= \operatorname{tr}\big((\mathbf{I}-\mathbf{P})\mathbf{A}_{>r}\mathbf{A}_{>r}^\top(\mathbf{I}-\mathbf{P})\big) \\
&= \operatorname{tr}\big(\mathbf{A}_{>r}^\top(\mathbf{I}-\mathbf{P})^2\mathbf{A}_{>r}\big) \\
&\le \operatorname{tr}\big(\mathbf{A}_{>r}^\top\mathbf{A}_{>r}\big) = \|\mathbf{A}_{>r}\|_F^2 = \sum_{i=r+1}^p \sigma_i(\mathbf{A})^2.
\end{aligned}$$

Since we are given that, $\mathbf{AS}$ is a λ_1-regularized ε-spectral approximation of $\mathbf{A}$, we have,

$$(1-\varepsilon)\big(\mathbf{A}\mathbf{A}^\top + \lambda_1\mathbf{I}\big) \preceq \mathbf{AS}(\mathbf{AS})^\top + \lambda_1\mathbf{I} \preceq (1+\varepsilon)\big(\mathbf{A}\mathbf{A}^\top + \lambda_1\mathbf{I}\big)$$

However, observe that the above approximation still holds if λ_1 is replaced by $\lambda_1' \ge \lambda_1$. So letting $\lambda_1' = \lambda_1 + \lambda_2^{1/2q+1}$, we may assume

$$(1-\varepsilon)\big(\mathbf{A}\mathbf{A}^\top + \lambda_1'\mathbf{I}\big) \preceq \mathbf{AS}(\mathbf{AS})^\top + \lambda_1'\mathbf{I} \preceq (1+\varepsilon)\big(\mathbf{A}\mathbf{A}^\top + \lambda_1'\mathbf{I}\big) \tag{10}$$

Recall that $\mathbf{P} = \mathbf{P}_{\mathbf{B\Omega}}$, where $\mathbf{P}_{\mathbf{B\Omega}}$ is orthogonal projection onto the column span of $\mathbf{B\Omega}$ and $\mathbf{B} = \big(\mathbf{AS}(\mathbf{AS})^\top\big)^q\mathbf{AS}$. Instead of working with $\mathbf{B}$, we will instead work with a scaled version of $\mathbf{B}$ defined as a power of a scaled version of $\mathbf{AS}$. Let $\widetilde{\mathbf{B}} := \frac{1}{\theta^{2q+1}}\big(\mathbf{AS}(\mathbf{AS})^\top\big)^q\mathbf{AS}$ for $\theta = \sqrt{\lambda_1'}$. This is not a problem since $\mathbf{P}_{\mathbf{B\Omega}} = \mathbf{P}_{\widetilde{\mathbf{B}}\mathbf{\Omega}}$.

If $\mathbf{B\Omega}$ is a λ_2-regularized $1/2$-spectral approximation of $\mathbf{B}$, then $\widetilde{\mathbf{B}}\mathbf{\Omega}$ is a $\lambda_2/\theta^{2(2q+1)}$-regularized $1/2$-spectral approximation of $\widetilde{\mathbf{B}}$. Since $\theta^{2(2q+1)} = \lambda_1'^{2q+1} \ge \lambda_2$, we may assume that $\widetilde{\mathbf{B}}\mathbf{\Omega}$ is a 1-regularized $1/2$-spectral approximation of $\widetilde{\mathbf{B}}$.

Thus, by the proof of Lemma 3.3, $\mathbf{I}-\mathbf{P} \preceq 2(\widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^\top + \mathbf{I})^{-1}$.

For any fixed i, this gives $\mathbf{u}_i^\top(\mathbf{I}-\mathbf{P})\mathbf{u}_i \le 2\,\mathbf{u}_i^\top(\widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^\top+\mathbf{I})^{-1}\mathbf{u}_i$. We now bound $\mathbf{u}_i^\top(\widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^\top+\mathbf{I})^{-1}\mathbf{u}_i$,

$$\begin{aligned}
\mathbf{u}_i^\top(\widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^\top+\mathbf{I})^{-1}\mathbf{u}_i &= \mathbf{u}_i^\top\mathbf{W}\left(\frac{1}{\theta^{2(2q+1)}}\mathbf{D}^{2q+1}+\mathbf{I}\right)^{-1}\mathbf{W}^\top\mathbf{u}_i \\
&= \sum_{j:\sigma_j(\mathbf{AS})\ge\theta}\frac{(\mathbf{W}^\top\mathbf{u}_i)_j^2}{(\sigma_j(\mathbf{AS})/\theta)^{2(2q+1)}+1} \\
&\quad+\sum_{j:\sigma_j(\mathbf{AS})<\theta}\frac{(\mathbf{W}^\top\mathbf{u}_i)_j^2}{(\sigma_j(\mathbf{AS})/\theta)^{2(2q+1)}+1}, \\
&\le \sum_{j:\sigma_j(\mathbf{AS})\ge\theta}\frac{(\mathbf{W}^\top\mathbf{u}_i)_j^2}{(\sigma_j(\mathbf{AS})/\theta)^{2}}+\sum_{j:\sigma_j(\mathbf{AS})<\theta}\frac{(\mathbf{W}^\top\mathbf{u}_i)_j^2}{1} \\
&= \theta^2\left(\sum_{j:\sigma_j(\mathbf{AS})\ge\theta}\frac{(\mathbf{W}^\top\mathbf{u}_i)_j^2}{\sigma_j(\mathbf{AS})^2}+\sum_{j:\sigma_j(\mathbf{AS})<\theta}\frac{(\mathbf{W}^\top\mathbf{u}_i)_j^2}{\theta^2}\right) \\
&\le \theta^2\left(\sum_{j:\sigma_j(\mathbf{AS})\ge\theta}\frac{(\mathbf{W}^\top\mathbf{u}_i)_j^2}{\sigma_j(\mathbf{AS})^2}+\sum_{j:\sigma_j(\mathbf{AS})<\theta}\frac{(\mathbf{W}^\top\mathbf{u}_i)_j^2}{\theta^2}\right)
\end{aligned}$$

where $\mathbf{AS}(\mathbf{AS})^\top = \mathbf{W}\mathbf{D}\mathbf{W}^\top$ is the eigendecomposition with $\mathbf{D} = \mathrm{diag}(\sigma_1(\mathbf{AS})^2,\dots,\sigma_{\widehat{m}}(\mathbf{AS})^2)$ (for $\widehat{m} := \min\{m, r_1\}$), so $\widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^\top = \theta^{-2(2q+1)}\mathbf{W}\mathbf{D}^{2q+1}\mathbf{W}^\top$. Also, by (10),

$$\left(\mathbf{AS}(\mathbf{AS})^\top+\lambda_1'\mathbf{I}\right)^{-1} \preceq \frac{1}{1-\varepsilon}\left(\mathbf{A}\mathbf{A}^\top+\lambda_1'\mathbf{I}\right)^{-1}.$$

Hence, for each left singular vector $\mathbf{u}_i$ of $\mathbf{A}$,

$$\mathbf{u}_i^\top\left(\mathbf{AS}(\mathbf{AS})^\top+\lambda_1'\mathbf{I}\right)^{-1}\mathbf{u}_i \le \frac{1}{1-\varepsilon}\,\mathbf{u}_i^\top\left(\mathbf{A}\mathbf{A}^\top+\lambda_1'\mathbf{I}\right)^{-1}\mathbf{u}_i = \frac{1}{(1-\varepsilon)(\sigma_i(\mathbf{A})^2+\lambda_1')}$$

Equivalently, using the eigendecomposition of $\mathbf{AS}(\mathbf{AS})^\top$,

$$\begin{aligned}
\mathbf{u}_i^\top\left(\mathbf{AS}(\mathbf{AS})^\top+\lambda_1'\mathbf{I}\right)^{-1}\mathbf{u}_i &= \mathbf{u}_i^\top\mathbf{W}(\mathbf{D}+\lambda_1'\mathbf{I})^{-1}\mathbf{W}^\top\mathbf{u}_i \\
&= \sum_{j:\sigma_j(\mathbf{AS})^2\ge\lambda_1'}\frac{(\mathbf{W}^\top\mathbf{u}_i)_j^2}{\sigma_j(\mathbf{AS})^2+\lambda_1'}+\sum_{j:\sigma_j(\mathbf{AS})^2<\lambda_1'}\frac{(\mathbf{W}^\top\mathbf{u}_i)_j^2}{\sigma_j(\mathbf{AS})^2+\lambda_1'} \\
&\ge \sum_{j:\sigma_j(\mathbf{AS})^2\ge\lambda_1'}\frac{(\mathbf{W}^\top\mathbf{u}_i)_j^2}{2\sigma_j(\mathbf{AS})^2}+\sum_{j:\sigma_j(\mathbf{AS})^2<\lambda_1'}\frac{(\mathbf{W}^\top\mathbf{u}_i)_j^2}{2\lambda_1'} \\
2\cdot\mathbf{u}_i^\top\left(\mathbf{AS}(\mathbf{AS})^\top+\lambda_1'\mathbf{I}\right)^{-1}\mathbf{u}_i &\ge \sum_{j:\sigma_j(\mathbf{AS})^2\ge\lambda_1'}\frac{(\mathbf{W}^\top\mathbf{u}_i)_j^2}{\sigma_j(\mathbf{AS})^2}+\sum_{j:\sigma_j(\mathbf{AS})^2<\lambda_1'}\frac{(\mathbf{W}^\top\mathbf{u}_i)_j^2}{\lambda_1'}
\end{aligned}$$

Thus, we have,

$$\sum_{j:\sigma_j(\mathbf{AS})^2\ge\lambda_1'}\frac{(\mathbf{W}^\top\mathbf{u}_i)_j^2}{\sigma_j(\mathbf{AS})^2}+\sum_{j:\sigma_j(\mathbf{AS})^2<\lambda_1'}\frac{(\mathbf{W}^\top\mathbf{u}_i)_j^2}{\lambda_1'} \le \frac{2}{(1-\varepsilon)(\sigma_i(\mathbf{A})^2+\lambda_1')}$$

Since $\theta^2 = \lambda_1'$, we get,

$$\mathbf{u}_i^\top(\widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^\top+\mathbf{I})^{-1}\mathbf{u}_i \le \frac{2\theta^2}{(1-\varepsilon)(\sigma_i(\mathbf{A})^2+\lambda_1')}$$

Thus,

$$\begin{aligned}\|(\mathbf{I}-\mathbf{P})\mathbf{A}_r\|_F^2 &= \sum_{i=1}^{r}\sigma_i(\mathbf{A})^2\,\mathbf{u}_i^\top(\mathbf{I}-\mathbf{P})\mathbf{u}_i \\ &\le \sum_{i=1}^{r}\sigma_i(\mathbf{A})^2\,2\,\mathbf{u}_i^\top(\widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^\top+\mathbf{I})^{-1}\mathbf{u}_i \\ &\le \sum_{i=1}^{r}\sigma_i(\mathbf{A})^2\,2\cdot\frac{2\theta^2}{(1-\varepsilon)(\sigma_i(\mathbf{A})^2+\lambda_1')} \\ &\le \frac{4r\theta^2}{(1-\varepsilon)}\end{aligned}$$

Combining with the bound for $\|(\mathbf{I}-\mathbf{P})\mathbf{A}_{>r}\|_F^2$, we get,

$$\begin{aligned}\|(\mathbf{I}-\mathbf{P})\mathbf{A}\|_F^2 &= \|(\mathbf{I}-\mathbf{P})\mathbf{A}_r\|_F^2+\|(\mathbf{I}-\mathbf{P})\mathbf{A}_{>r}\|_F^2 \\ &\le \sum_{i>r}\sigma_i^2(\mathbf{A})+8r(\lambda_1+\lambda_2^{1/2q+1})\end{aligned}$$

when $\varepsilon < 1/2$.

E Additional Experiments

In this section, we compare the numerical performance of Sketch-Powered Low-Rank Factorization (Algorithm 2) against the version of Algorithm 2 that does iterations without sketching, i.e. when $\mathbf{S}_1 = \mathbf{I}$. This baseline (denoted simply as "Low-Rank Factorization" in our plots) precisely matches the so-called Generalized Nyström algorithm as described, e.g., by Nakatsukasa (2020), except this algorithm is stated there without power iteration (i.e., $q = 0$). Generalized Nyström is designed to avoid the expensive matrix multiplication step $\mathbf{Q}^\top\mathbf{A}$ that occurs in Randomized SVD. Because of this, prior work would have little reason to incorporate full power method into this procedure, as that would introduce even more expensive matrix multiplications, erasing most of this gain. Nevertheless, for completeness, in our experiments we show the performance of this baseline for a range of $q \ge 0$.

We consider the same target rank $k = 40$ and the same matrices `smallnorb`, `ImageNet` and `polydecay` as described in Section 4 and, as earlier, plot the average relative error with respect to the average time taken for a particular iterate in Figure 2 for $l \in \{800, 1200, 1600\}$. Note that in this experiment the baseline algorithm also depends on the parameter $r_1 = l$, which corresponds to the size of the sketching matrix $\mathbf{S}_2$.

Similar to what we observed in Figure 1 for Sketch-Powered Randomized SVD, Algorithm 2 reaches a relative error ≈ 0.1 faster than the baseline algorithm, and then at a certain point the error starts to plateau. Increasing the intermediate sketch size l decreases the level at which the relative error plateaus, as predicted by Theorem 2.3.

F Miscellaneous Lemmas

F.1 Upper Bound on Projection

Proof of Lemma 3.3. Let $\mathbf{P}_{\mathbf{A}\boldsymbol{\Omega}} := \mathbf{Q}\mathbf{Q}^\top$. Using the fact that $\mathbf{P}_{\mathbf{A}\boldsymbol{\Omega}} = \mathbf{A}\boldsymbol{\Omega}\boldsymbol{\Omega}^\top\mathbf{A}^\top(\mathbf{A}\boldsymbol{\Omega}\boldsymbol{\Omega}^\top\mathbf{A}^\top)^\dagger$, we have

$$\begin{aligned}\mathbf{I}-\mathbf{P}_{\mathbf{A}\boldsymbol{\Omega}} &= (\mathbf{A}\boldsymbol{\Omega}\boldsymbol{\Omega}^\top\mathbf{A}^\top+\lambda\mathbf{I})(\mathbf{A}\boldsymbol{\Omega}\boldsymbol{\Omega}^\top\mathbf{A}^\top+\lambda\mathbf{I})^{-1}-\mathbf{P}_{\mathbf{A}\boldsymbol{\Omega}} \\ &= \mathbf{A}\boldsymbol{\Omega}\boldsymbol{\Omega}^\top\mathbf{A}^\top(\mathbf{A}\boldsymbol{\Omega}\boldsymbol{\Omega}^\top\mathbf{A}^\top+\lambda\mathbf{I})^{-1}-\mathbf{P}_{\mathbf{A}\boldsymbol{\Omega}}+\lambda(\mathbf{A}\boldsymbol{\Omega}\boldsymbol{\Omega}^\top\mathbf{A}^\top+\lambda\mathbf{I})^{-1} \\ &\preceq \lambda(\mathbf{A}\boldsymbol{\Omega}\boldsymbol{\Omega}^\top\mathbf{A}^\top+\lambda\mathbf{I})^{-1} \preceq 2\lambda(\mathbf{A}\mathbf{A}^\top+\lambda\mathbf{I})^{-1}.\end{aligned}$$

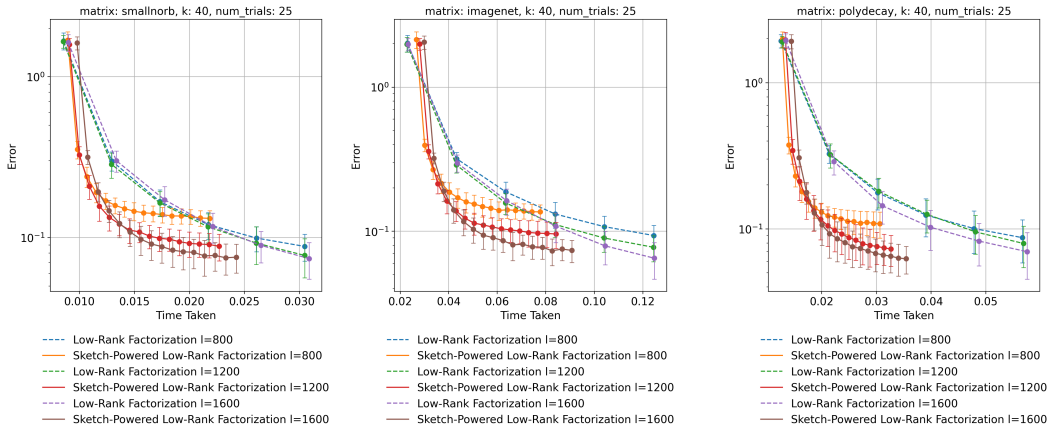

Figure 2: Approximation error vs time for Sketch-Powered Low-Rank Factorization (Algorithm 2) compared with the same algorithm used without sketching the power method, for (from left to right) `smallnorb`, `ImageNet`, and `polydecay` matrices.

where the second to last inequality follows from the fact that $\mathbf{A}\boldsymbol{\Omega}\boldsymbol{\Omega}^\top\mathbf{A}^\top(\mathbf{A}\boldsymbol{\Omega}\boldsymbol{\Omega}^\top\mathbf{A}^\top + \lambda\mathbf{I})^{-1} \preceq \mathbf{P}_{\mathbf{A}\boldsymbol{\Omega}}$ and the last inequality follows from the regularized spectral approximation guarantee. So,

$$\|\mathbf{A} - \mathbf{P}_{\mathbf{A}\boldsymbol{\Omega}}\mathbf{A}\|^2 = \|\mathbf{A}^\top(\mathbf{I} - \mathbf{P}_{\mathbf{A}\boldsymbol{\Omega}})\mathbf{A}\| \leq 2\lambda\|\mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top + \lambda\mathbf{I})^{-1}\mathbf{A}\| \leq 2\lambda.$$

Similarly, we get $\|\mathbf{A} - \mathbf{P}_{\mathbf{A}\boldsymbol{\Omega}}\mathbf{A}\|_F^2 \leq 2\lambda\,\mathrm{tr}(\mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top + \lambda\mathbf{I})^{-1}\mathbf{A})$. Finally, for any $l \geq 0$, we have

$$\mathrm{tr}(\mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top + \lambda\mathbf{I})^{-1}\mathbf{A}) = \sum_i \frac{\sigma_i^2(\mathbf{A})}{\sigma_i^2(\mathbf{A}) + \lambda} \leq l + \frac{1}{\lambda}\sum_{i>l}\sigma_i^2(\mathbf{A}),$$

which concludes the proof. □

F.2 Error Term Estimate

Lemma F.1. *For $\mathbf{A} \in \mathbb{R}^{m\times n}$, let $\mathbf{AS}$ be a λ_1-regularized ε-spectral approximation of $\mathbf{A}$ for some $\varepsilon \in [0, 1/2]$ and let $\widehat{m} := \mathrm{rank}(\mathbf{AS})$. Then, for $k > 0$ and $q \geq \log(2\widehat{m})/2\varepsilon$,*

$$\left(\frac{2}{k}\sum_{i>k}\sigma_i(\mathbf{AS})^{2(2q+1)}\right)^{\frac{1}{2q+1}} \leq (1+4\varepsilon)\sigma_{k+1}(\mathbf{A})^2 + 2\lambda_1\varepsilon$$

Proof. By the spectral approximation guarantee,

$$(1-\varepsilon)\big(\mathbf{A}\mathbf{A}^\top + \lambda_1\mathbf{I}\big) \preceq \mathbf{AS}(\mathbf{AS})^\top + \lambda_1\mathbf{I} \preceq (1+\varepsilon)\big(\mathbf{A}\mathbf{A}^\top + \lambda_1\mathbf{I}\big)$$
$$(1-\varepsilon)\big(\sigma_i(\mathbf{A})^2 + \lambda_1\big) \leq \sigma_i(\mathbf{AS})^2 + \lambda_1 \leq (1+\varepsilon)\big(\sigma_i(\mathbf{A})^2 + \lambda_1\big)$$

So, for $\widehat{m} = \mathrm{rank}(\mathbf{AS})$,

$$\left(\sum_{i>k}\sigma_i(\mathbf{AS})^{2(2q+1)}\right)^{\frac{1}{2q+1}} \leq \left(\sum_{i=k+1}^{\widehat{m}}\big((1+\varepsilon)\sigma_i(\mathbf{A})^2 + \lambda_1\varepsilon\big)^{(2q+1)}\right)^{\frac{1}{2q+1}}$$
$$\leq (1+\varepsilon)\left(\sum_{i=k+1}^{\widehat{m}}\sigma_i(\mathbf{A})^{2(2q+1)}\right)^{\frac{1}{2q+1}} + \lambda_1\varepsilon\widehat{m}^{1/2q+1}$$
$$\left(\frac{2}{k}\sum_{i>k}\sigma_i(\mathbf{AS})^{2(2q+1)}\right)^{\frac{1}{2q+1}} \leq (1+\varepsilon)\left(\frac{2}{k}\sum_{i=k+1}^{\widehat{m}}\sigma_i(\mathbf{A})^{2(2q+1)}\right)^{\frac{1}{2q+1}} + \lambda_1\varepsilon\left(\frac{2\widehat{m}}{k}\right)^{1/2q+1}$$

Then, for $q = \log(2\widehat{m})/2\varepsilon$, $\left(2\widehat{m}/k\right)^{1/2q+1} \leq (2\widehat{m})^{(\varepsilon/\log(2\widehat{m}))} \leq e^{\varepsilon} \leq 1 + 2\varepsilon \leq 2$ (since $\varepsilon \leq 1/2$). Also,

$$\begin{aligned}\left(\frac{2}{k}\sum_{i=k+1}^{\widehat{m}}\sigma_i(\mathbf{A})^{2(2q+1)}\right)^{\frac{1}{2q+1}} &\leq \sigma_{k+1}(\mathbf{A})^2\left(\frac{2}{k}\sum_{i=k+1}^{\widehat{m}}\frac{\sigma_i(\mathbf{A})^{2(2q+1)}}{\sigma_{k+1}(\mathbf{A})^{2(2q+1)}}\right)^{\frac{1}{2q+1}} \\ &\leq \sigma_{k+1}(\mathbf{A})^2(2\widehat{m})^{\frac{1}{2q}} \\ &\leq (1+2\varepsilon)\sigma_{k+1}(\mathbf{A})^2\end{aligned}$$

where we used that $\sigma_i(\mathbf{A}) \leq \sigma_{k+1}(\mathbf{A})$ for all $i > k$. Thus,

$$\begin{aligned}(2\lambda_2)^{\frac{1}{(2q+1)}} &= \left(\frac{2}{k}\sum_{i>k}\sigma_i(\mathbf{AS})^{2(2q+1)}\right)^{\frac{1}{2q+1}} \\ &\leq (1+\varepsilon)(1+2\varepsilon)\sigma_{k+1}(\mathbf{A})^2 + 2\lambda_1\varepsilon \\ &\leq (1+4\varepsilon)\sigma_{k+1}(\mathbf{A})^2 + 2\lambda_1\varepsilon\end{aligned}$$

for $\varepsilon < 1/2$. □

F.3 Proof of Lemma 3.6

Lemma 3.6 relies on the following OSE moments property defined in Cohen et al. (2016).

Definition F.2 (OSE moments restated for right sketches, (Cohen et al., 2016), Definition 4)**.** A distribution $\mathcal{D}$ over sketching matrices $\mathbf{S} \in \mathbb{R}^{n\times r}$ has $(\varepsilon, \delta, d, p)$-OSE moments if, for every matrix $\mathbf{U} \in \mathbb{R}^{n\times d}$ with orthonormal columns,

$$\mathbb{E}_{\mathbf{S}\sim\mathcal{D}}\left\|\mathbf{U}^\top\mathbf{S}\mathbf{S}^\top\mathbf{U} - \mathbf{I}_d\right\|^p < \varepsilon^p\delta.$$

By the discussion in Sections 2.1.1-2.1.3 of Cohen et al. (2016), we note that the sketching matrices listed in Lemma 3.2 satisfy the $(\varepsilon/2\sqrt{2}, \delta', 2k, p)$-OSE moment property for some $p \geq 2$ (The $2\sqrt{2}$ factor attached to ε only influences the sketch size lower bounds in Lemma 3.2 by constants, which get absorbed in the $O(\cdot)$ notation).

By Cohen et al. (2016, Theorem 6), if $\mathbf{S} \sim \mathcal{D}$ has the $(\varepsilon/2\sqrt{2}, \delta', 2k, p)$-OSE moments property, then with probability at least $1 - \delta'$, $\mathbf{S}^\top$ satisfies the $(k, \varepsilon/2\sqrt{2})$ approximate spectral norm matrix multiplication property (Cohen et al., 2016, Defintion 2), for $\mathbf{U_A}$ and $\mathbf{P}_{\bar{\mathbf{A}}}\mathbf{B} = (\mathbf{I} - \mathbf{P_A})\mathbf{B}$,

$$\left\|(\mathbf{S}^\top\mathbf{U_A})^\top\mathbf{S}^\top\mathbf{P}_{\bar{\mathbf{A}}}\mathbf{B} - \mathbf{A}^\top\mathbf{P}_{\bar{\mathbf{A}}}\mathbf{B}\right\| \leq \frac{\varepsilon}{2\sqrt{2}}\sqrt{\left(\|\mathbf{U_A}\|^2 + \frac{\|\mathbf{U_A}\|_F^2}{k}\right)\left(\|\mathbf{P}_{\bar{\mathbf{A}}}\mathbf{B}\|^2 + \frac{\|\mathbf{P}_{\bar{\mathbf{A}}}\mathbf{B}\|_F^2}{k}\right)}. \tag{11}$$

where $\mathbf{U_A}$ is an orthogonal matrix spanning the column space of $\mathbf{A}$ and $\mathbf{P_A}$ is the orthogonal projection onto the column span of $\mathbf{A}$.

The same OSE moment property, applied to $\mathbf{U_A}$ (which has rank at most k, so $\|\mathbf{U_A}\|_F^2/k \leq 1$), gives

$$\left\|(\mathbf{S}^\top\mathbf{U_A})^\top(\mathbf{S}^\top\mathbf{U_A}) - \mathbf{I}\right\| \leq \frac{\varepsilon}{2\sqrt{2}}\left(\|\mathbf{U_A}\|^2 + \|\mathbf{U_A}\|_F^2/k\right) \leq \frac{1}{2} \tag{12}$$

with probability at least $1 - \delta'$ for $\varepsilon \leq 1/2$. Equivalently, $\mathbf{S}^\top$ is a $1/2$-subspace embedding for the column space of $\mathbf{A}$.

Thus we can apply Cohen et al. (2016, Theorem 7), with Π in the statement of their theorem replaced by $\mathbf{S}^\top$ and $\sqrt{\varepsilon/8}$ in their theorem replaced by $\varepsilon/2\sqrt{2}$ to get the claimed bound after letting $\delta' = \delta/2$ and taking a union bound for (11) and (12) to hold simultaneously and absorbing constant factors into the sketch size lower bounds in Lemma 3.2.