On the Unreasonable Effectiveness of Single Vector Krylov Methods for Low-Rank Approximation

Raphael A. Meyer
New York University
ram900@nyu.edu

Cameron Musco
University of Massachusetts Amherst
cmusco@cs.umass.edu

Christopher Musco
New York University
cmusco@nyu.edu

Abstract

Krylov subspace methods are a ubiquitous tool for computing near-optimal rank k approximations of large matrices. While "large block" Krylov methods with block size *at least* k give the best known theoretical guarantees, block size one (a single vector) or a small constant is often preferred in practice. Despite their popularity, we lack theoretical bounds on the performance of such "small block" Krylov methods for low-rank approximation.

We address this gap between theory and practice by proving that small block Krylov methods essentially match all known low-rank approximation guarantees for large block methods. Via a black-box reduction we show, for example, that the standard single vector Krylov method run for t iterations obtains the same spectral norm and Frobenius norm error bounds as a Krylov method with block size $\ell \geq k$ run for $O(t/\ell)$ iterations, up to a logarithmic dependence on the smallest gap between sequential singular values. That is, for a given number of matrix-vector products, single vector methods are essentially as effective as *any choice of large block size*.

By combining our result with tail-bounds on eigenvalue gaps in random matrices, we prove that the dependence on the smallest singular value gap can be eliminated if the input matrix is perturbed by a small random matrix. Further, we show that single vector methods match the more complex algorithm of [Bakshi et al. '22], which combines the results of multiple block sizes to achieve an improved algorithm for Schatten p-norm low-rank approximation.

1 Introduction

Krylov subspace methods have been studied since the 1950s and remain our most reliable algorithms for approximating eigenvectors and singular vectors of large matrices. Krylov methods access a matrix $\mathbf{A}$ via repeated matrix multiplications (each considered an iteration of the method) either with a single vector or a block of vectors. There has been significant interest in analyzing how many iterations are required to obtain accurate eigenvector or singular vector approximations. Classic work studies both single vector [Kaniel, 1966, Paige, 1971] and block methods [Cullum and Donath, 1974, Kahan and Parlett, 1976, Saad, 1980, Saad, 2011].

More recently, there has been interest in analyzing Krylov subspace methods specifically for the downstream task of *low-rank approximation*. Since the top k singular vectors can be used to obtain an optimal rank-k approximation, the goal is to understand how many iterations are required to compute approximate singular vectors that yield a near-optimal rank-k approximation [Rokhlin et al., 2009, Halko et al., 2011b, Woodruff, 2014]. This problem differs from classical work because convergence to the actual top singular vectors is sufficient but *not necessary* for obtaining an accurate low-rank approximation [Drineas and Ipsen, 2019].

Prototypical single vector and block Krylov methods for low-rank approximation are shown in

Algorithm 1 Single Vector Krylov Method for Low-Rank Approximation

input: Matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$. Target rank k. Starting vector $\mathbf{x} \in \mathbb{R}^n$. Number of iterations t.
output: Orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{n\times k}$.

1: Compute an orthonormal basis $\mathbf{Z}$ for $\mathbf{K} = [\mathbf{x},\ (\mathbf{A}\mathbf{A}^\intercal)\mathbf{x},\ (\mathbf{A}\mathbf{A}^\intercal)^2\mathbf{x},\ \ldots,\ (\mathbf{A}\mathbf{A}^\intercal)^t\mathbf{x}]$.
2: Compute $\mathbf{U}_k$, the k top eigenvectors of $\mathbf{M} = \mathbf{Z}^\intercal\mathbf{A}\mathbf{A}^\intercal\mathbf{Z}$
3: **return** $\mathbf{Q} = \mathbf{Z}\mathbf{U}_k$.

Algorithm 2 Block Krylov Method for Low-Rank Approximation

input: Matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$. Target rank k. Starting block $\mathbf{B} \in \mathbb{R}^{n\times \ell}$. Number of iterations t.
output: Orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{n\times k}$.

1: Compute an orthonormal basis $\mathbf{Z}$ for $\mathbf{K} = [\mathbf{B},\ (\mathbf{A}\mathbf{A}^\intercal)\mathbf{B},\ (\mathbf{A}\mathbf{A}^\intercal)^2\mathbf{B}\ \ldots,\ (\mathbf{A}\mathbf{A}^\intercal)^t\mathbf{B}]$.
2: Compute $\mathbf{U}_k$, the k top eigenvectors of $\mathbf{M} = \mathbf{Z}^\intercal\mathbf{A}\mathbf{A}^\intercal\mathbf{Z}$
3: **return** $\mathbf{Q} = \mathbf{Z}\mathbf{U}_k$.

Algorithm 1 and Algorithm 2.[1] For an $n \times d$ input $\mathbf{A}$, both methods returns an $n \times k$ orthogonal $\mathbf{Q}$ so that the rank-k matrix $\mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}$ is a good approximation to $\mathbf{A}$. Ideally, it is nearly as good as $\mathbf{A}$'s optimal rank k approximation, $\mathbf{A}_k$, which is given via projection onto $\mathbf{A}$'s top k singular vectors.[2]

1.1 Large block methods and gap-free bounds

Most recent work on Krylov methods for low-rank approximation focuses on "large block" methods, where ℓ in Algorithm 2 is chosen to be $\geq k$ [Rokhlin et al., 2009, Halko et al., 2011a, Gu, 2015, Musco and Musco, 2015, Tropp, 2018, Yuan et al., 2018, Drineas et al., 2018, Tropp, 2018]. In this regime, block methods are known to quickly converge to a near-optimal low-rank approximation. For example, in just $O\left(\frac{\log(n/\varepsilon)}{\sqrt{(\sigma_k-\sigma_{\ell+1})/\sigma_k}}\right)$ iterations, Algorithm 2 initialized with an i.i.d. random Gaussian matrix $\mathbf{B} \in \mathbb{R}^{n\times\ell}$ with block size $\ell \geq k$, achieves with high probability the bound

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_\xi \leq (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_\xi \tag{1}$$

for any $\varepsilon > 0$ and $\|\cdot\|_\xi$ being either the Frobenius or spectral norm [Musco and Musco, 2015]. That is, convergence is linear with a rate depending on the square root of the relative gap from the k^{th} singular value, σ_k, to the $(\ell+1)^{\text{st}}$ singular value, $\sigma_{\ell+1}$. Even for ℓ mildly larger than k, this gap is often quite large. For example, [Halko et al., 2011b] recommends setting $\ell = k+5$ or $k+10$.

Beyond such spectrum dependent guarantees, another advantage of large block Krylov methods is that they enjoy *gap-independent* bounds, which do not involve any terms depending on $\mathbf{A}$'s spectrum. For example, a now standard result is that Algorithm 2 achieves Equation (1) in just $O(\frac{1}{\sqrt{\varepsilon}}\log(\frac{n}{\varepsilon}))$ iterations [Musco and Musco, 2015].[3] Further, this bound is essentially optimal among all methods that access $\mathbf{A}$ only through matrix-vector products [Simchowitz et al., 2018, Bakshi

[1] Algorithms 1 and 2 are examples of the simplest possible implementations of Krylov methods for low-rank approximation. In practice, various optimizations like the Lanczos recurrence are often applied, and additional care is necessary to ensure that the orthogonal basis for the Krylov subspace $\mathbf{K}$ in computed in a numerically stable way [Saad, 2011]. While an important topic, this paper is not focused on the numerical stability of Lanczos methods. All derivations assume computation in the Real RAM model of arithmetic.

[2] For simplicity, we focus on computing an approximate left singular vector subspace spanned by $\mathbf{Q} \in \mathbb{R}^{n\times k}$. If we instead care about computing right singular vectors, Algorithms 1 and 2 can be applied to $\mathbf{A}^\intercal$ instead.

[3] Randomly initialized block power method with block size k gives a similar bound, but with a suboptimal $1/\varepsilon$ rather than $1/\sqrt{\varepsilon}$ dependence on the error [Rokhlin et al., 2009, Halko et al., 2011b].

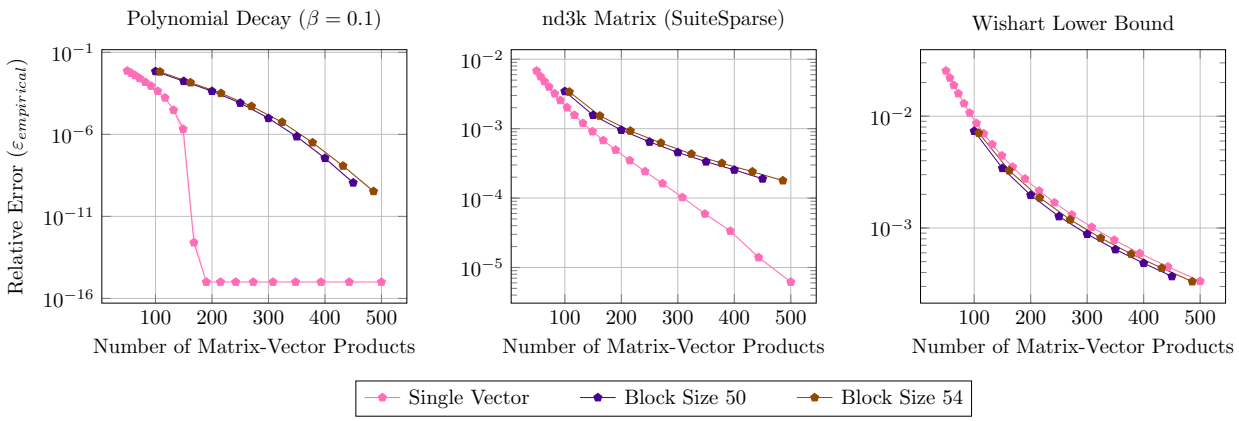

Figure 1: Comparison of the number of matrix-vector products needed for Krylov iteration to converge to an accurate rank 50 approximation under different block sizes. The left figure uses a matrix with singular values decaying polynomially; the middle figure uses a matrix from the SuiteSparse library; the right figure uses a worst-case matrix from the literature. See Section 6.5 for more details. In the first two plots, single vector Krylov outperforms large block methods. Even in the adversarially chosen hard instant on the right, it does not perform much worse.

and Narayanan, 2023]. Bounds where the iteration complexity does not depend on properties of $\mathbf{A}$, are called "universal" guarantees [Urschel, 2021]. Universal bounds are useful in applications where properties like large spectral gaps cannot be ensured, but where worst-case accuracy guarantees are still desired [Hegde et al., 2016, Li et al., 2017, Soltani and Hegde, 2018].

In contrast to large block sizes, it is *impossible* to prove gap-independent guarantees for single vector or small block Krylov iteration. To see why, consider $\mathbf{A} \in \mathbb{R}^{k \times d}$ that is all zeros, except that $\mathbf{A}_{ii} = 1$ for $i = 1, \ldots, k$. I.e., $\mathbf{A} = \begin{bmatrix} \mathbf{I}_k & \mathbf{0} \end{bmatrix}$. where $\mathbf{I}_k$ denotes the $k \times k$ identity matrix. If we run Algorithm 2 on this matrix with block size $\ell < k$, then it can be checked that the Krylov subspace $\mathbf{K}$ will have rank $\ell < k$, and thus any low-rank approximation obtained from the subspace cannot be near-optimal. In general, bounds for small block methods must depend *inversely* on the gaps between sequential singular values. In the above example, these gaps are equal to 0.

1.2 Main contribution: the virtue of small block Krylov methods

The inability of single vector and small block Krylov methods to offer gap-independent bounds has been a point of concern for the use of these methods in computing low-rank approximations [Li et al., 2017, Musco and Musco, 2015, Li and Zhang, 2015]. At the same time, in practice, low-rank approximation is frequently solved using iterative eigensolvers based on single vector or small block methods. Such methods are the standard in MATLAB, Julia, Python, and essentially all languages used for matrix computations [Lehoucq et al., 1998, Mathworks, 2023, SciPy Community, 2023]. These methods often perform very well, converging quickly to good low-rank approximations. In fact, in our experience, they typically outperform large block methods in terms of the number of matrix-vector products required to achieve a desired level of accuracy – see Figure 1.[4]

The main goal of this paper is to explain this phenomenon. We ask:

[4]The number of matrix-vector products used by an algorithm does not necessarily translate directly into the computational cost of the algorithm. For example, in many computing systems, it is faster to multiply a matrix $\mathbf{A}$ by a block of k vectors all at once, than to multiply by k vectors chosen in sequence. Nevertheless, matrix-vector products are still a valuable measure of complexity for many problems where they dominate other runtime costs.

For low-rank approximation, when and why do small block Krylov methods require the same or fewer matrix-vector multiplications than large block Krylov methods?

We answer this question in a strong way by proving that small block methods nearly match or even improve on all known theoretical guarantees on the convergence of large-block methods for low-rank approximation. In particular, up to a *logarithmic dependence* on the smallest gap between singular values, the trade-off between accuracy and number of matrix-vector products achieved by small block methods matches the trade-off achieved by large block methods. Since there are a variety of guarantees known for large block methods, this claim is broken down as a number of results throughout our paper. We state one such result as a concrete example:

Theorem 1. *For $\mathbf{A} \in \mathbb{R}^{n\times d}$, let $g_{min} = \min_{i\in\{1,\ldots,k-1\}} \frac{\sigma_i - \sigma_{i+1}}{\sigma_i}$ be the smallest relative gap among the top k singular values. For any $\varepsilon, \delta \in (0,1)$, Algorithm 1 initialized with an i.i.d. mean zero Gaussian vector $\mathbf{x}$ and run for $t = O(\frac{k}{\sqrt{\varepsilon}} \log(\frac{1}{g_{min}}) + \frac{1}{\sqrt{\varepsilon}} \log(\frac{n}{\varepsilon\delta}))$ iterations returns an orthogonal $\mathbf{Q} \in \mathbb{R}^{n\times k}$ such that, with probability at least $1-\delta$, letting $\|\cdot\|_\xi$ be the spectral or Frobenius norm,*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_\xi \leq (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_\xi.$$

As discussed, [Musco and Musco, 2015] prove that Algorithm 2 with block size k achieves an identical error bound in $O\left(\frac{\log(n/\varepsilon\delta)}{\sqrt{\varepsilon}}\right)$ iterations. This translates to $O\left(\frac{k\log(n/\varepsilon\delta)}{\sqrt{\varepsilon}}\right)$ matrix-vector products, which Theorem 1 matches, except for the dependence on $\log(1/g_{min})$. At the same time, Theorem 1 improves on the large block bound by separating the $\log(n/\varepsilon\delta)$ and k terms.

Remark. Since it is a logarithmic instead of polynomial dependence, we consider the $\log(1/g_{min})$ term to be mild for typical problems. In experiments, it appears to have little impact on the observed convergence of the single vector Krylov method (see Section 6). Indeed, except in adversarial cases, such as the identity matrix, where g_{min} truly equals 0, in finite precision, we cannot expect to resolve singular value gaps to accuracy better than machine precision. So, it is reasonable to think that in practice, this term should be at most a moderate constant. We make this intuition formal in Section 5, showing that the dependence on g_{min} can be eliminated in a smoothed analysis setting (i.e., when the input is perturbed by a small random matrix).

Our proof for Theorem 1 (with some additional results) is given in Section 3. Our approach is via a black-box reduction to the existing analysis for large block methods. In particular, we view the single-vector method of Algorithm 1 as a block Krylov method in disguise. We observe that

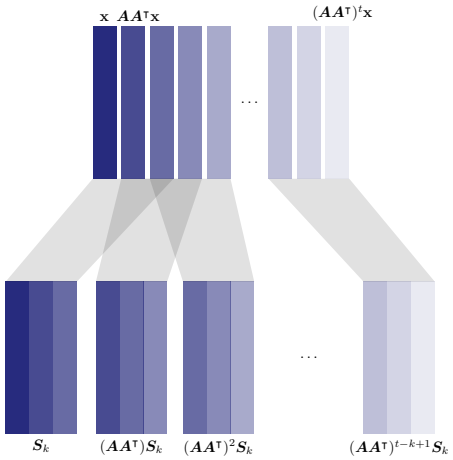

Figure 2: Our main analysis is based on the simple observation that the span of the Krylov subspace generated by a single vector Krylov method after t iterations is exactly equivalent to that generated by a block Krylov method run for $t-k+1$ iterations with starting block $\mathbf{S}_k = [\mathbf{x},\ (\mathbf{A}\mathbf{A}^\intercal)\mathbf{x},\ \ldots,\ (\mathbf{A}\mathbf{A}^\intercal)^k\mathbf{x}]$. This observation allows us to take advantage of existing results on block methods in a black-box way.

the span of the single-vector Krylov subspace $\mathbf{K}$ (Line 1 of Algorithm 1) is exactly equivalent to the span of a block Krylov subspace generated from a specific starting matrix. Concretely, suppose Algorithm 1 is run for $t \geq k$ iterations and let $\mathbf{S}_k \in \mathbb{R}^{d \times k}$ equal the first k columns of $\mathbf{K}$. I.e.,

$$\mathbf{S}_k := \begin{bmatrix} \mathbf{x} & \mathbf{A}\mathbf{A}^\intercal\mathbf{x} & (\mathbf{A}\mathbf{A}^\intercal)^2\mathbf{x} & \ldots & (\mathbf{A}\mathbf{A}^\intercal)^{k-1}\mathbf{x} \end{bmatrix}. \tag{2}$$

Then we can check that for $q = t - k + 1$:

$$\text{span}(\mathbf{K}) = \text{span}\left(\begin{bmatrix} \mathbf{S}_k & \mathbf{A}\mathbf{A}^\intercal\mathbf{S}_k & (\mathbf{A}\mathbf{A}^\intercal)^2\mathbf{S}_k & \ldots & (\mathbf{A}\mathbf{A}^\intercal)^q\mathbf{S}_k \end{bmatrix}\right). \tag{3}$$

This equivalence is visualized in Figure 2. Since both Algorithm 1 and Algorithm 2 only depend on the *span* of the Krylov subspace they generate (through $\mathbf{Z}$), the single vector method thus matches the block Krylov method run for $k-1$ fewer iterations, with the specific starting block $\mathbf{S}_k$.

With this perspective, a naive hope might be to directly appeal to prior results on block Krylov iteration to analyze the single vector method. Unfortunately, these results rely on the fact that the starting matrix $\mathbf{B}$ is chosen at random, typically with i.i.d. Gaussian or sub-Gaussian entries [Halko et al., 2011b, Musco and Musco, 2015, Bakshi et al., 2022]. In contrast, $\mathbf{S}_k$ is far from a random Gaussian matrix. Its columns are highly dependent on each other. To understand just how far $\mathbf{S}_k$ is from an ideal starting matrix, note that, generally, a block Krylov subspace with q blocks will have rank qk when $\mathbf{B}$ is a random Gaussian matrix. In contrast, the block Krylov subspace $\begin{bmatrix} \mathbf{S}_k & \mathbf{A}\mathbf{A}^\intercal\mathbf{S}_k & \ldots & (\mathbf{A}\mathbf{A}^\intercal)^q\mathbf{S}_k \end{bmatrix}$ only has rank $t = q + k - 1$.

Surprisingly, however, we are still able to show that $\mathbf{S}_k$ provides a (barely) good enough starting matrix for the block Krylov method in Algorithm 2 to succeed. To do so, we consider a natural definition of what it means to be a "good" starting matrix. At a high-level, we need $\mathbf{S}_k$ to have non-negligible inner product with all top k singular vectors of $\mathbf{A}$. While $\mathbf{S}_k$ is exponentially worse in terms of starting inner product than a random $\mathbf{B}$, this is made up for by the fact that it is far cheaper (in terms of matrix-vector products) to build a block Krylov subspace with $\mathbf{S}_k$; we can compute a degree q subspace using just $q + k$ matrix-vector products with $\mathbf{A}$. In contrast, computing a degree q block Krylov subspace with a random starting block $\mathbf{B} \in \mathbb{R}^{n \times k}$ requires qk matrix-vector products. The detailed proof is presented in Section 3.

1.3 Results and Paper Organization

Theorem 1 is our main result, and its proof is contained entirely in Section 3. In addition to this result, whose proof shows the crux of our argument that single vector methods converge quickly, we include several other bounds for single vector and small block methods. We summarize these additional results below. Section 6 contains experiments which demonstrate that our bounds are predictive of the performance of these methods in practice.

Spectrum adaptive bounds, Section 4.1. The black-box nature of Theorem 1's proof allows us to similarly adapt other results on large block Krylov methods to the single vector setting. For example, we show that Algorithm 1 matches known "spectrum dependent" bounds for large block methods. As discussed in Section 1.1, the convergence rate of these bounds depends on $g_{k \to \ell} = \frac{\sigma_k - \sigma_{\ell+1}}{\sigma_k}$, the gap between the k^{th} and $(\ell+1)^{\text{st}}$ singular values. Since this gap increases with ℓ, there is a natural tradeoff: a larger block means more matrix-vector products per iteration, but fewer iterations. Our proof shows that single vector methods match any block size $\ell \geq k$ (up to a dependence on $\log(1/g_{min})$). I.e., they automatically match the complexity of the method with *best choice of large block size*, without need for parameter tuning.

Schatten p-norm low-rank approximation, Sections 4.2 and 4.3. Recently, a combination of spectrum dependent and spectrum independent bounds have been used to give faster convergence

rates for Schatten p-norm low rank approximation. In particular, for constant p, [Bakshi et al., 2022] show how to find a low-rank approximation achieving $\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_p \leq (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_p$ using just $\tilde{O}(k/\varepsilon^{1/3})$ matrix-vector products with $\mathbf{A}$. For the Frobenius norm (i.e., $p = 2$), this is an improvement on the $\tilde{O}(k/\sqrt{\varepsilon})$ required by [Musco and Musco, 2015]. Their method requires running Algorithm 2 with multiple choices of block size, and optimizing over the best Krylov subspace. We show that, again up to a logarithmic dependence on $1/g_{min}$, the exact same guarantees can be obtained by simply running a single vector Krylov method. In concurrent work, [Bakshi and Narayanan, 2023] show a similar result without a dependence on $\log(1/g_{min})$ or $\log(n)$.

Beyond block size 1, Section 4.4. While the above results focus on the single vector Krylov method, our bounds naturally generalize to other small block sizes between 1 and k (e.g. 4, 10, or $k-1$). For general small block size b, we show that dependence on g_{min} can be replaced with a dependence on the smallest "b^{th} order" gap $g_{min,b} := \min_{i\in\{1,\ldots,k-b\}} \frac{\sigma_i - \sigma_{i+b}}{\sigma_i}$.

Removing the gap dependence, Section 5. While a dependence on singular value gaps is unavoidable for small block methods in the worst-case, the parameter seems to rarely have an impact in practice. We take a step towards explaining this observation via a *smoothed-analysis* result [Spielman and Teng, 2004, Sankar et al., 2006]. Specifically, we leverage work in random matrix theory on *eigenvalue repulsion*, which shows that small spectral gaps in a matrix are brittle: adding a tiny amount of random noise to any matrix ensures that its singular value gaps are at worst inverse polynomial in the problem parameters. Using this fact, we present bounds that replace the dependence on $\log(1/g_{min})$ in our prior results with a dependence on $\log(\frac{n\kappa_k}{\delta\varepsilon})$ for randomly perturbed matrices, where $\kappa_k = \sigma_1/\sigma_k$ measures the conditioning of the top k singular values. From an algorithm design perspective, the $\log(1/g_{min})$ can be removed even in the worst case by explicitly adding a random diagonal perturbation to $\mathbf{A}$.

Single Vector Simultaneous Iteration, Appendix G. Lastly, we describe a single vector analogue for simultaneous iteration. While it converges somewhat more slowly, this method has the advantage of using less space than the single vector Krylov method, which needs to store the span for the entire Krylov subspace. For instance, it allows us to store just k vectors while converging in $\tilde{O}(k/\varepsilon)$ iterations, in contrast to the $\tilde{O}(k/\sqrt{\varepsilon})$ iterations required by the standard single vector Krylov method. Since single vector simultaneous iteration only uses a single starting vector, its convergence still depends on $\log(1/g_{min})$.

1.4 Related Work

We briefly discuss additional prior work on low-rank approximation and Krylov methods, though since the literature is rich, so we cannot cover all relevant prior work. As discussed, early analyses of Krylov methods for approximating eigenvectors consider both single vector and block methods [Saad, 1980, Golub and Underwood, 1977, Kuczyński and Woźniakowski, 1992]. However, this work does not directly provide strong bounds for low-rank approximation, since convergence to the top singular vectors is not required to accurately solve the problem [Drineas and Ipsen, 2019].

For low-rank approximation, large block methods have been more popular. In addition to prior work already discussed, this includes work on randomized sketching methods, which can be viewed as large block Krylov methods run for one or two iterations [Martinsson et al., 2006, Cohen et al., 2015, Clarkson and Woodruff, 2013, Drineas and Mahoney, 2016, Martinsson and Tropp, 2020]. Sketching methods have become a mainstay technique in randomized numerical linear algebra.

Work on single vector or small block methods for low-rank approximation has been more sparse. [Wang et al., 2015] experimentally study small block methods, and suggest that large blocks are only worthwhile when singular value gaps are very small, when low precision suffices, or when

making many passes over a matrix is expensive. [Yuan et al., 2018] theoretically studies the related problem of singular value approximation for all block sizes, and as in our work, obtains linear convergence rates depending on $g_{k\to\ell}$. They also show superlinear rates when $\mathbf{A}$ has a sufficiently quickly decaying spectrum. While it is difficult to directly compare their results to ours on low-rank approximation, it would be interesting to consider such spectra in our setting. Finally, we note that [Allen-Zhu and Li, 2016] proves a result similar to Theorem 1 using an algorithm that in some ways is a single vector Krylov method. However, because the method iteratively restarts k times with k randomly chosen starting vectors, it ultimately returns a solution from a block Krylov subspace.

Related to our results in Section 5, we note that adding small random perturbations to avoid small singular value gaps or other conditioning issues is a technique that has been employed in several recent works focused on worst-case runtime bounds for linear algebraic problems [Boutsidis et al., 2016, Peng and Vempala, 2021, Banks et al., 2022].

2 Notation

We use capital bold letters to denote matrices, lowercase bold letters to denote vectors, and lowercase non-bold letters to denote scalars. For a matrix $\mathbf{Q}$, we let $\mathbf{q}_i$ denote the i^{th} column, $\text{span}(\mathbf{Q})$ denote its column span, and $\mathbf{Q}^\intercal$ denote its transpose. Typically, $\mathbf{A} \in \mathbb{R}^{n\times d}$ denotes our input matrix. We let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\intercal$ denote the SVD of $\mathbf{A}$, with $\mathbf{U} \in \mathbb{R}^{n\times n}, \mathbf{\Sigma} \in \mathbb{R}^{n\times n}, \mathbf{V} \in \mathbb{R}^{d\times n}$. We let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$ denote the singular values of $\mathbf{A}$ (the diagonal entries of $\mathbf{\Sigma}$). We let $\mathbf{U}_k \in \mathbb{R}^{n\times k}$ and $\mathbf{V}_k \in \mathbb{R}^{d\times k}$ denote the first k columns of $\mathbf{U}$ and $\mathbf{V}$, and let $\mathbf{\Sigma}_k \in \mathbb{R}^{k\times k}$ denote the top $k \times k$ principal submatrix of $\mathbf{\Sigma}$. Then $\mathbf{A}_k = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^\intercal$ is the best rank-k approximation to $\mathbf{A}$ in any unitarily invariant norm. When $\mathbf{A}$ is square, we let $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\intercal$ be the eigendecomposition of $\mathbf{A}$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ are the eigenvalues of $\mathbf{A}$ (the diagonal entries of $\mathbf{\Lambda}$). We often work with symmetric positive semi-definite (PSD) matrices, which have all non-negative eigenvalues. In this case, the singular values equal the eigenvalues. We also work with matrix polynomials. If $p(x) = \sum_{i=1}^q c_i x^i$ is a polynomial and if $\mathbf{A}$ is square, $p(\mathbf{A}) := \sum_{i=1}^q c_i \mathbf{A}^i$.

We let $\|\mathbf{x}\|_2$ denote the vector ℓ_2 norm, $\|\mathbf{A}\|_2$ the spectral norm, $\|\mathbf{A}\|_F$ the Frobenius norm, and $\|\mathbf{A}\|_p = (\sum_{i=1}^p \sigma_i^p)^{1/p}$ the Schatten p-norm. Wherever $\|\mathbf{A}\|_\xi$ is used, the equation holds for both the spectral and Frobenius norms. We let $[n] = \{1, \ldots, n\}$ be the set of integers between 1 and n. Finally, we let $\mathcal{N}(\mathbf{0}, \mathbf{I})$ denote the distribution over vectors whose entries are i.i.d. mean zero unit variance Gaussians. The dimension will be clear from context.

3 Proof of Theorem 1

In this section, we prove Theorem 1 by showing that $\mathbf{S}_k$ as described in Section 1.2 is a good enough starting matrix for block Krylov iteration. This proof serves as a foundation for all additional results in the paper. Throughout, we assume that $\mathbf{A} \in \mathbb{R}^{n\times n}$ is square and positive semidefinite. We shown in Appendix A that this is without loss of generality: running Algorithm 1 or Algorithm 2 on a matrix $\mathbf{C} \in \mathbb{R}^{n\times d}$ with SVD $\mathbf{C} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\intercal$ yields an identical output to running the method on the PSD matrix $\mathbf{A} = (\mathbf{C}\mathbf{C}^\intercal)^{1/2} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\intercal$. Further, all low-rank approximation and singular value approximation results guaranteed for the returned matrix $\mathbf{Q}$ directly carry over from $\mathbf{A}$ to $\mathbf{C}$.

3.1 A naive approach that actually works.

As discussed in Section 1.2, our main approach is to view the single-vector method of Algorithm 1 as a block Krylov method in disguise. Specifically, recall the matrix $\mathbf{S}_k$ and the Krylov subspace

from Equations (2) and (3). Since we now assume that $\mathbf{A}$ is PSD, we have $\mathbf{A}\mathbf{A}^\intercal = \mathbf{A}^2$, so we can write

$$\mathbf{S}_k := \begin{bmatrix}\mathbf{x} & \mathbf{A}^2\mathbf{x} & \mathbf{A}^4\mathbf{x} & \dots & \mathbf{A}^{2(k-1)}\mathbf{x}\end{bmatrix} \tag{4}$$

and

$$\text{span}(\mathbf{K}) = \text{span}\left(\begin{bmatrix}\mathbf{S}_k & \mathbf{A}^2\mathbf{S}_k & \dots & \mathbf{A}^{2q}\mathbf{S}_k\end{bmatrix}\right), \tag{5}$$

where $q := t - k + 1$. This matches the subspace spanned by $\mathbf{K}$ in Algorithm 2 with starting block $\mathbf{B} = \mathbf{S}_k$. Since the output of Algorithm 1 and Algorithm 2 depend only on the *span* of the Krylov subspace they construct, we will use this equivalence to appeal to prior results on block Krylov methods to analyze Algorithm 1. To do so, we need to show that, even though $\mathbf{S}_k$ is very much unlike the i.i.d. random starting matrices used in prior work, it still provides a good enough starting matrix for convergence to a near-optimal low-rank approximation.

Towards that end, we use a natural definition of what it means to be a "good" starting matrix that specifically will allow us to leverage results on block Krylov methods from [Musco and Musco, 2015] in a black-box way. That same definition suffices for other results as well [Woodruff, 2014, Drineas et al., 2018]. Intuitively, we require that a starting matrix $\mathbf{B}$ has nontrivial inner product with all of the top k singular vectors of $\mathbf{A}$:

Definition 1 ((k, L)-good Starting Matrix)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be a matrix with top k left singular vectors $\mathbf{U}_k \in \mathbb{R}^{n\times k}$. A matrix $\mathbf{B} \in \mathbb{R}^{n\times k}$ is a (k, L)-good starting matrix for $\mathbf{A}$ if, letting $\mathbf{Q} \in \mathbb{R}^{n\times k}$ be an orthonormal basis for* span$(\mathbf{B})$*, $\mathbf{U}_k^\intercal\mathbf{Q}$ is invertible and $\|(\mathbf{U}_k^\intercal\mathbf{Q})^{-1}\|_2^2 \le L$.*

The condition above is equivalent to requiring that all singular values of $\mathbf{U}_k^\intercal\mathbf{Q}$ are at least $\frac{1}{\sqrt{L}}$, or that all principle angles between span$(\mathbf{U}_k)$ and span$(\mathbf{B})$ have $\cos(\theta_i) \ge \frac{1}{\sqrt{L}}$ [Drineas et al., 2018].

Using Definition 1, we can immediately obtain bounds on the low-rank approximation error of a subspace $\mathbf{Q}$ returned by Algorithm 2 when run with any (k, L)-good starting matrix. We will use such bounds to analyze the single vector Krylov method of Algorithm 1, after proving that $\mathbf{S}_k$ is (k, L)-good. Consider the following bound, which depends logarithmically on L:

Imported Theorem 2 (Theorem 1 of [Musco and Musco, 2015])**.** *Let $\mathbf{B} \in \mathbb{R}^{n\times k}$ be any (k, L)-good starting matrix (Definition 1) matrix for $\mathbf{A}$. If we run Block Krylov iteration (Algorithm 2) for $q = O(\frac{1}{\sqrt{\varepsilon}}\log(\frac{nL}{\varepsilon}))$ iterations with starting block $\mathbf{B}$, then the output $\mathbf{Q} \in \mathbb{R}^{n\times k}$ satisfies*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_\xi \le (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_\xi \tag{6}$$

and, letting $\mathbf{q}_i$ be the i^{th} column of $\mathbf{Q}$,

$$\left|\mathbf{q}_i^\intercal\mathbf{A}\mathbf{A}^\intercal\mathbf{q}_i - \sigma_i(\mathbf{A})^2\right| \le \varepsilon\sigma_{k+1}(\mathbf{A})^2. \tag{7}$$

Imported Theorem 2 is implicit in [Musco and Musco, 2015], although, as stated in that work, it is specialized to when $\mathbf{B}$ is a matrix with i.i.d. Gaussian entries. In Appendix F we discuss how the more general result stated above follows from [Musco and Musco, 2015]. Imported Theorem 2 gives two different guarantees. The first bounds low-rank approximation error, in both the Frobenius and spectral norms. The second shows that the columns of $\mathbf{Q}$ can be used to estimate the top singular values of $\mathbf{A}$. This can be called a "Ritz value guarantee" or a "singular value guarantee".

A random Gaussian matrix $\mathbf{B}$ can be shown to be an $(k, O(nk))$-good starting matrix with high probability (see Lemma 4 in [Musco and Musco, 2015] or [Rudelson and Vershynin, 2010]). This is

intuitive, since $\mathbf{B}$ will span a uniformly random subspace, which has non-negligible inner product with any other fixed subspace, including the one spanned by $\mathbf{U}_k$. For a Gaussian starting block, Imported Theorem 2 therefore gives a bound of $O(\frac{1}{\sqrt{\varepsilon}}\log(n/\varepsilon))$ iterations to achieve Equations (6) and (7).

The fact that $\mathbf{S}_k$ satisfies Definition 1 is less clear. Our main technical contribution is to prove that it does, albeit with a much larger value of L than in the Gaussian case: in Section 3.2 we show that, with probability at least $1-\delta$, $\mathbf{S}_k$ is a (k, L)-good starting matrix for $L = O(\text{poly}(n/\delta g_{\min}^k))$. Here $g_{min} = \min_{i\in\{1,\ldots,k-1\}} \frac{\sigma_i - \sigma_{i+1}}{\sigma_i}$ is the minimum gap between $\mathbf{A}$'s top k singular values. Since Imported Theorem 2 depends logarithmically on L, it follows that block Krylov with starting block $\mathbf{S}_k$ needs $O\left(\frac{k}{\sqrt{\varepsilon}}\log(\frac{1}{g_{\min}}) + \frac{1}{\sqrt{\varepsilon}}\log(\frac{n}{\varepsilon\delta})\right)$ iterations to achieve Equations (6) and (7) – yielding Theorem 1.

In other words, we require k times as many iterations when starting with $\mathbf{S}_k$ instead of a fully random $\mathbf{B}$. However, notice that running Algorithm 2 for q iterations with starting block $\mathbf{S}_k$ only requires $q+k$ iterations of the single vector Krylov method from Algorithm 1, and thus $q+k$ matrix-vector products. In contrast, running the method with a random $\mathbf{B}$ requires qk matrix-vector products. Ultimately, this allows us to achieve in Theorem 1 a total complexity (in terms of matrix-vector products) that matches, and in some cases improves, the block Krylov method initialized with a Gaussian starting matrix, up to the dependence on $\log(1/g_{min})$.

3.2 Main Technical Analysis

Formally, we prove the following (k, L)-good guarantee for $\mathbf{S}_k$. In combination with Imported Theorem 2 and Equation (5), this bound yields Theorem 1. It will also serve as the basis for all of the other results discussed in Section 1.3.

Theorem 3. *Fix any PSD matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ with singular values $\sigma_1 \geq \ldots \geq \sigma_n$, let $\mathbf{g}$ be a vector with i.i.d. mean zero Gaussian entries, and let $\mathbf{S}_k = \begin{bmatrix}\mathbf{g} & \mathbf{A}^2\mathbf{g} & \mathbf{A}^4\mathbf{g} & \ldots & \mathbf{A}^{2(k-1)}\mathbf{g}\end{bmatrix}$. For any $\delta \in (0,1)$, with probability at least $1-\delta$, $\mathbf{S}_k$ is a (k, L)-good starting matrix for $\mathbf{A}$ for $L = \frac{cnk^3\log(n/\delta)}{\delta^2 g_{\min}^{4k}}$. Here $c = 2.5\pi$ is a fixed constant and $g_{min} = \min_{i\in\{1,\ldots,k-1\}} \frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+1}}$.*

To prove Theorem 3, we will need a simple bound on the minimum of k independent Gaussian random variables:

Lemma 4. *Let $g_1, \ldots, g_k \sim \mathcal{N}(0,1)$ and independent. Then, with probability at least $\geq 1-\delta$, $\min_i g_i^2 \geq \frac{2\delta^2}{\pi k^2}$.*

Proof. We have that:

$$\Pr[\min_i g_i^2 \geq t] = \left(1 - \Pr[{g_1}^2 \leq t]\right)^k = \left(1 - 2\Pr[0 \leq g_1 \leq \sqrt{t}]\right)^k \geq (1 - \sqrt{1-e^{-2t/\pi}})^k.$$

The last line uses the bound $\Pr[0 \leq g_1 \leq \sqrt{t}] \leq \frac{1}{2}\sqrt{1-e^{-2x^2/\pi}}$ from [Chu, 1955]. Setting the right hand side equal to $1-\delta$ and solving for t, we get:

$$t = \frac{\pi}{2}\ln\frac{1}{1-(1-(1-\delta)^{1/k})^2} \geq \frac{\pi}{2}(1-(1-\delta)^{1/k})^2 \geq \frac{\pi}{2}(\frac{\delta}{k})^2 = \frac{\pi\delta^2}{2k^2}.$$

In the first inequality we used that $\ln\left(\frac{1}{1-x}\right) \geq x$. In the second we used that $1-(1-x)^{1/k} \geq \frac{x}{k}$. □

Proof of Theorem 3. We first argue that $\mathbf{U}_k^\intercal \mathbf{S}_k$ is invertible. Observe that for any $\mathbf{x} \in \mathbb{R}^k$, $\mathbf{S}_k \mathbf{x} = \hat{p}(\mathbf{A}^2)\mathbf{g}$ for some degree $k-1$ polynomial $\hat{p}$ with coefficients determined by the entries in $\mathbf{x}$. We let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\intercal$ be the SVD of $\mathbf{A}$. Then

$$\mathbf{U}_k^\intercal \mathbf{S}_k \mathbf{x} = \mathbf{U}_k^\intercal \hat{p}(\mathbf{A}^2)\mathbf{g} = \mathbf{U}_k^\intercal \mathbf{U} \hat{p}(\mathbf{\Sigma}^2)\mathbf{U}^\intercal \mathbf{g} = \hat{p}(\mathbf{\Sigma}_k^2)\tilde{\mathbf{g}},$$

where $\mathbf{\Sigma}_k$ contains the top left k elements of $\mathbf{\Sigma}$ and $\tilde{\mathbf{g}} := \mathbf{U}_k^\intercal \mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ by rotational invariance of the Gaussian distribution. Note by Lemma 4 that $\min_{i \in \{1,\ldots,k\}} \tilde{g}_i^2 \geq \frac{2\delta^2}{\pi k^2}$ with probability at least $1-\delta$. So altogether, we can write

$$\|\mathbf{U}_k^\intercal \mathbf{S}_k \mathbf{x}\|_2^2 = \|\mathbf{U}_k^\intercal \hat{p}(\mathbf{A}^2)\tilde{\mathbf{g}}\|_2^2 = \sum_{i=1}^k (\hat{p}(\sigma_i^2))^2 \tilde{g}_i^2 \geq \frac{2\delta^2}{\pi k^2} \sum_{i=1}^k (\hat{p}(\sigma_i^2))^2. \tag{8}$$

Since $\hat{p}$ has degree $k-1$, if none of the top k singular values are repeated (i.e., $g_{\min} > 0$), we have that the right hand side is nonzero for any nonzero $\mathbf{x}$. Thus, $\mathbf{U}_k^\intercal \mathbf{S}_k$ is invertible.

Now, let $\mathbf{Q} \in \mathbb{R}^{n \times k}$ be any orthonormal basis for $\text{span}(\mathbf{S}_k)$, so that $\mathbf{S}_k = \mathbf{Q}\mathbf{C}$ for some invertible matrix $\mathbf{C} \in \mathbb{R}^{k \times k}$ (observe that $\mathbf{S}_k$ must be full rank since $\mathbf{U}_k^\intercal \mathbf{S}_k$ is invertible). Since $\mathbf{U}_k^\intercal \mathbf{S}_k$ is invertible, we then also know that $\mathbf{U}_k^\intercal \mathbf{Q}$ is invertible, as required by Definition 1. We also have $\mathbf{S}_k(\mathbf{U}_k^\intercal \mathbf{S}_k)^{-1} = \mathbf{Q}\mathbf{C}(\mathbf{U}_k^\intercal \mathbf{Q}\mathbf{C})^{-1} = \mathbf{Q}(\mathbf{U}_k^\intercal \mathbf{Q})^{-1}$ and therefore $\|\mathbf{S}_k(\mathbf{U}_k^\intercal \mathbf{S}_k)^{-1}\|_2^2 = \|\mathbf{Q}(\mathbf{U}_k^\intercal \mathbf{Q})^{-1}\|_2^2 = \|(\mathbf{U}_k^\intercal \mathbf{Q})^{-1}\|_2^2$. So, to prove the theorem, it suffices to bound

$$\|\mathbf{S}_k(\mathbf{U}_k^\intercal \mathbf{S}_k)^{-1}\|_2^2 = \max_{\mathbf{x}} \frac{\|\mathbf{S}_k(\mathbf{U}_k^\intercal \mathbf{S}_k)^{-1}\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} = \max_{\mathbf{x}} \frac{\|\mathbf{S}_k \mathbf{x}\|_2^2}{\|\mathbf{U}_k^\intercal \mathbf{S}_k \mathbf{x}\|_2^2} = \max_{\deg(\hat{p}) \leq k-1} \frac{\|\hat{p}(\mathbf{A}^2)\mathbf{g}\|_2^2}{\|\mathbf{U}_k^\intercal \hat{p}(\mathbf{A}^2)\mathbf{g}\|_2^2}. \tag{9}$$

We already bounded the denominator in (8). Thus, we turn to the numerator. Since $\mathbf{g}$ has i.i.d. mean zero, unit variance Gaussian entries we have for each i, $g_i^2 \leq 1 + 4\log(1/\delta)$ with probability at least $1-\delta$ by standard concentration bounds for chi-squared random variables [Laurent and Massart, 2000]. So, by a union bound, $\max_i g_i^2 \leq 5\log(n/\delta)$ for $n > 2$. We thus have:

$$\|\hat{p}(\mathbf{A}^2)\mathbf{g}\|_2^2 \leq 5\log(n/\delta) \sum_{i=1}^n (\hat{p}(\sigma_i^2))^2 \leq 5n\log(n/\delta) \cdot \max_{i \in \{1,\ldots,n\}} (\hat{p}(\sigma_i^2))^2. \tag{10}$$

Combining (10) and (8), we conclude that

$$\|\mathbf{S}_k(\mathbf{U}_k^\intercal \mathbf{S}_k)^{-1}\|_2^2 \leq \frac{5\pi n k^2 \log(\frac{n}{\delta})}{2\delta^2} \cdot \max_{\deg(\hat{p}) \leq k-1} \frac{\max_{i \in \{1,\ldots,n\}} (\hat{p}(\sigma_i^2))^2}{\sum_{i=1}^k (\hat{p}(\sigma_i^2))^2}. \tag{11}$$

We now focus on bounding the maximum in (11). Observe that if there were no gap between two of the top k singular values, then some nonzero polynomial $\hat{p}$ could make the denominator zero by equaling zero on the at most $k-1$ unique values in $\sigma_1, \ldots, \sigma_k$. So, any bound on (11) must depend on the minimum gap between singular values. Second, note that if the maximum in the numerator is achieved for $i \leq k$, then the overall ratio is trivially at most 1. So, without loss of generality, we only consider $\max_{i \in \{k+1,\ldots,n\}} (\hat{p}(\sigma_i^2))^2$ in the numerator.

To bound this ratio, we follow a similar broad approach to [Saad, 1980], who bounds a related term by expanding $\hat{p}$ as an interpolating polynomial. Formally, we write $\hat{p}$ as a Lagrange interpolating polynomial over $\sigma_1^2, \ldots, \sigma_k^2$:

$$\hat{p}(x) = \sum_{i=1}^k \phi_i \ell_i(x) \quad \text{where} \quad \phi_i := \hat{p}(\sigma_i^2), \quad \ell_i(x) := \prod_{\substack{j \in \{1,\ldots,k\} \\ j \neq i}} \frac{x - \sigma_j^2}{\sigma_i^2 - \sigma_j^2}, \quad i \in \{1, \ldots, k\}.$$

For any $0 \le x \le \sigma_k^2$ we have

$$|\ell_i(x)| \le \prod_{\substack{j\in\{1,\dots,k\}\\ j\neq i}} \frac{\sigma_j^2}{|\sigma_i^2-\sigma_j^2|} \le \prod_{\substack{j\in\{1,\dots,k\}\\ j\neq i}} \frac{\sigma_j^2}{|\sigma_i-\sigma_j|^2} \le \frac{1}{g_{min}^{2k}},$$

where the second inequality uses that $|\sigma_i^2-\sigma_j^2| \ge |\sigma_i-\sigma_j|^2$ for all $\sigma_i,\sigma_j \ge 0$. Next, we write $\boldsymbol{\phi} = [\phi_1 \ \dots \ \phi_k] = [\hat{p}(\sigma_1)^2 \ \dots \ \hat{p}(\sigma_k^2)]$ and obtain:

$$\begin{aligned}
\frac{\max_{i\in\{k+1,\dots,n\}} |\hat{p}(\sigma_i^2)|^2}{\sum_{i=1}^k (\hat{p}(\sigma_i^2))^2} \le \frac{\max_{x\in[0,\sigma_k^2]} |\hat{p}(x)|^2}{\sum_{i=1}^k (\hat{p}(\sigma_i^2))^2} &= \frac{\max_{x\in[0,\sigma_k^2]} |\sum_{i=1}^k \phi_i \ell_i(x)|^2}{\sum_{i=1}^k (\hat{p}(\sigma_i^2))^2} \\
&\le \frac{\max_{x\in[0,\sigma_k^2]} (\|\boldsymbol{\phi}\|_1 \max_i |\ell_i(x)|)^2}{\|\boldsymbol{\phi}\|_2^2} \\
&= \frac{\|\boldsymbol{\phi}\|_1^2}{\|\boldsymbol{\phi}\|_2^2} \max_{x\in[0,\sigma_k^2], i\in\{1,\dots,k\}} |\ell_i(x)|^2 \\
&\le \frac{k}{g_{min}^{4k}}.
\end{aligned}$$

Finally, we plug back into (11) to conclude that $\|(\mathbf{U}_k^\intercal \mathbf{Q})^{-1}\|_2^2 = \|\mathbf{S}_k(\mathbf{U}_k^\intercal \mathbf{S}_k)^{-1}\|_2^2 \le \frac{5\pi n k^3}{2 g_{min}^{4k} \delta^2} \log(\frac{n}{\delta})$, which completes the proof. □

3.3 Proof of Theorem 1

As mentioned, Theorem 1 follows directly by combining Imported Theorem 2 and Theorem 3. In particular, since $\mathbf{S}_k$ is an (k, L)-good starting matrix for $L \le \left(\frac{n}{\delta g_{\min}^k}\right)^c$ for a constant c, if we run the block Krylov method initialized with $\mathbf{S}_k$, then with probability at least $1-\delta$, we obtain $\mathbf{Q}$ achieving the guarantees of Theorem 1 after

$$q = O\left(\frac{1}{\sqrt{\varepsilon}} \log\left(\frac{nL}{\varepsilon}\right)\right) = O\left(\frac{k}{\sqrt{\varepsilon}} \log\left(\frac{1}{g_{\min}}\right) + \frac{1}{\sqrt{\varepsilon}} \log\left(\frac{n}{\varepsilon\delta}\right)\right) \text{ iterations.}$$

Moreover, by Equation (5), the $\mathbf{Q}$ returned by Algorithm 2 with $\mathbf{S}_k$ as the starting block, is exactly the same as the $\mathbf{Q}$ returned by the single vector Krylov method (Algorithm 1) after $q+k$ iterations. Overall, we get the following full version of Theorem 1, which includes both the low-rank approximation and singular value guarantees from Imported Theorem 2:

Theorem 1 Restated. *For $\mathbf{A} \in \mathbb{R}^{n\times d}$, let $g_{min} = \min_{i\in\{1,\dots,k-1\}} \frac{\sigma_i - \sigma_{i+1}}{\sigma_i}$. For any $\varepsilon, \delta \in (0,1)$, Algorithm 1 initialized with $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and run for $t = O(\frac{k}{\sqrt{\varepsilon}} \log(\frac{1}{g_{min}}) + \frac{1}{\sqrt{\varepsilon}} \log(\frac{n}{\varepsilon\delta}))$ iterations returns an orthogonal $\mathbf{Q} \in \mathbb{R}^{n\times k}$ such that, with probability at least $1-\delta$,*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal \mathbf{A}\|_\xi \le (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_\xi \qquad \text{and} \qquad \left|\mathbf{q}_i^\intercal \mathbf{A}\mathbf{A}^\intercal \mathbf{q}_i - \sigma_i(\mathbf{A})^2\right| \le \varepsilon \sigma_{k+1}(\mathbf{A})^2.$$

4 Additional Applications of Main Result

In this section, we leverage our analysis of the starting block $\mathbf{S}_k$ to give several results beyond Theorem 1. In Section 4.1, we start by generalizing from using $\mathbf{S}_k$ to using $\mathbf{S}_\ell$, which has $\ell \ge k$ columns, and lets us obtain faster rates of convergence when the spectrum of $\mathbf{A}$ decays between

singular values k and $\ell+1$. Next, in Section 4.2, we use the results of Theorem 1 and Section 4.1 to get a faster rate of convergence in the Frobenius norm, simplifying the algorithm of [Bakshi et al., 2022]. In Section 4.3, we generalize from the Frobenius norm to Schatten p-norms, also simplifying [Bakshi et al., 2022]. Lastly, in Section 4.4, we generalize from single vector Krylov with block size $b = 1$ to small-block Krylov with any block size $b < k$.

4.1 Faster Convergence with Spectral Decay

As discussed in Section 1, in addition to spectrum-independent bounds, we know that large block Krylov methods achieve very fast convergence when using block size $\ell \geq k$ if there is a sufficiently large gap between σ_k and $\sigma_{\ell+1}$. Formally, [Musco and Musco, 2015] show the following:

Imported Theorem 5 (Theorem 13 of [Musco and Musco, 2015])**.** *Let $\mathbf{B} \in \mathbb{R}^{n\times\ell}$ be any (ℓ, L)-good starting matrix (Definition 1) matrix for $\mathbf{A}$, for some $\ell \geq k$. If we run Block Krylov iteration (Algorithm 2) for $q = O(\frac{1}{\sqrt{g_{k\to\ell}}}\log(\frac{nL}{\varepsilon}))$ iterations with starting block $\mathbf{B}$, where $g_{k\to\ell} = \frac{\sigma_k - \sigma_{\ell+1}}{\sigma_k}$, then the output $\mathbf{Q} \in \mathbb{R}^{n\times k}$ satisfies*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_\xi \leq (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_\xi \qquad \text{and} \qquad \left|\mathbf{q}_i^\intercal\mathbf{A}\mathbf{A}^\intercal\mathbf{q}_i - \sigma_i(\mathbf{A})^2\right| \leq \varepsilon\sigma_{k+1}(\mathbf{A})^2.$$

Our second main result shows that the convergence rate of Imported Theorem 5 applies to single vector Krylov as well. In particular, we can simply apply the same idea as was used to prove Theorem 1, but with "simulated block size" $\ell \geq k$. Letting

$$\mathbf{S}_\ell := \begin{bmatrix}\mathbf{x} & \mathbf{A}^2\mathbf{x} & \mathbf{A}^4\mathbf{x} & \dots & \mathbf{A}^{2(\ell-1)}\mathbf{x}\end{bmatrix} \quad \text{and} \quad \text{span}(\mathbf{K}) = \text{span}\left(\begin{bmatrix}\mathbf{S}_\ell & \mathbf{A}^2\mathbf{S}_\ell & \dots & \mathbf{A}^{2q}\mathbf{S}_\ell\end{bmatrix}\right),$$

we observe that single vector Krylov run for t iterations exactly computes $\text{span}(\mathbf{K})$ for $q = t-\ell+1$. This lets us show that single vector Krylov run for t iterations essentially matches the convergence rate of block Krylov run for $\approx t/\ell$ iterations:

Theorem 6 (Spectral Decay Convergence)**.** *For $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\ell \geq k$, let $g_{min} = \min_{i\in\{1,\dots,\ell-1\}}\frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+1}}$ and $g_{k\to\ell} = \frac{\sigma_k - \sigma_{\ell+1}}{\sigma_k}$. For any $\varepsilon, \delta \in (0,1)$, Algorithm 1 initialized with $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and run for $t = O(\frac{\ell}{\sqrt{g_{k\to\ell}}}\log(\frac{1}{g_{\min}}) + \frac{1}{\sqrt{g_{k\to\ell}}}\log(\frac{n}{\delta\varepsilon}))$ iterations returns an orthogonal $\mathbf{Q} \in \mathbb{R}^{n\times k}$ such that, with probability at least $1-\delta$,*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_\xi \leq (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_\xi \qquad \text{and} \qquad \left|\mathbf{q}_i^\intercal\mathbf{A}\mathbf{A}^\intercal\mathbf{q}_i - \sigma_i(\mathbf{A})^2\right| \leq \varepsilon\sigma_{k+1}(\mathbf{A})^2.$$

Proof. By Theorem 3, we know that $\mathbf{S}_\ell$ is an (ℓ, L)-good starting matrix for block Krylov iteration on $\mathbf{A}$, where $L \leq (\frac{n}{\delta g_{min}^\ell})^c$ for a constant c. So, by Imported Theorem 5, we find that block Krylov iteration with starting block $\mathbf{S}_\ell$ converges in

$$q = O\left(\frac{1}{\sqrt{g_{k\to\ell}}}\log\left(\frac{nL}{\varepsilon}\right)\right) = O\left(\frac{\ell}{\sqrt{g_{k\to\ell}}}\log\left(\frac{1}{g_{min}}\right) + \frac{1}{\sqrt{g_{k\to\ell}}}\log\left(\frac{n}{\delta\varepsilon}\right)\right)$$

iterations. Moreover, by Equation (5), the $\mathbf{Q}$ returned by Algorithm 2 with $\mathbf{S}_\ell$ as the starting block is exactly the same as the $\mathbf{Q}$ returned by the single vector Krylov method (Algorithm 1) after $q+\ell$ iterations, which completes the proof. □

Comparison to Prior Bounds. In terms of the number of matrix-vector products computed, Theorem 6 can significantly improve upon the prior work for large k. [Musco and Musco, 2015] require using a block size $\ell \geq k$ for $t = O(\frac{1}{\sqrt{g_{k\to\ell}}}\log(\frac{n}{\delta\varepsilon}))$ iterations, or equivalently for $O(\frac{1}{\sqrt{g_{k\to\ell}}}\cdot$

$\ell \log(\frac{n}{\delta\varepsilon}))$ matrix-vector products. Assuming that g_{min} is a constant, our guarantee for single vector Krylov methods only requires $O(\frac{1}{\sqrt{g_{k\to\ell}}}(\ell+\log(\frac{n}{\varepsilon\delta})))$ matrix-vector products. That is, we reduce the product between ℓ and $\log(\frac{n}{\delta\varepsilon})$ into a sum, suggesting a nearly ℓ-fold speedup when we want high precision results. Further, we obtain the above guarantee for any $\ell \geq k$, meaning that single vector Krylov automatically competes with the bound for the *best possible choice* of block size, without knowing it in advance. We observe in Section 6.5 that the above theoretical advantages translate into practice – for matrices with decaying singular value spectra, we find that single vector Krylov methods substantially outperform large block methods.

Comparison to Lower Bounds. It is worth comparing Theorem 6 to the lower bound of [Simchowitz et al., 2018], who show that finding an orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{n\times k}$ such that $\sum_{i=1}^{k} \mathbf{q}_i^\intercal \mathbf{A}\mathbf{q}_i \geq (1-\varepsilon)\sum_{i=1}^{k}\sigma_i(\mathbf{A})$ requires at least $\Omega(\frac{k\log n}{\sqrt{g_{k\to k}}})$ matrix-vector products when $\varepsilon = g_{k\to k} = \frac{\sigma_k - \sigma_{k+1}}{\sigma_k}$. In comparison, applying Theorem 6 with $\ell = k$ yields an upper bound of roughly $O(\frac{k\log(1/g_{min})}{\sqrt{g_{k\to k}}} + \frac{\log n}{\sqrt{g_{k\to k}}})$, which only does not violate the lower bound of [Simchowitz et al., 2018] since $\frac{1}{g_{min}} = \text{poly}(n)$ for their input. I.e., their input suffers from very small singular value gaps.

When $k = 1$, the intuition for the $\log n$ dependence is that since a random start vector has $\frac{1}{\text{poly}(n)}$ inner product with the top singular vector, it requires $\log n$ iterations to converge. The matching upper and lower bounds of $\Theta(\frac{k\log n}{\sqrt{g_{k\to k}}})$ from [Musco and Musco, 2015] and [Simchowitz et al., 2018] suggest that the cost of rank-k approximation is simply k times the cost of rank-1 approximation, i.e., roughly $k\log n$. In fact, the situation is more nuanced. Our work suggests that, unless gaps between singular values are very small, the cost of rank-k approximation is much actually cheaper.

4.2 Improved Results for Frobenius Norm Low-Rank Approximation

Theorem 6 shows that single vector Krylov methods achieve strong spectrum-adaptive guarantees, converging at essentially the same rate as the best choice of block size, if not faster. As a concrete application of this observation, we show that single vector Krylov methods automatically match a recent result of [Bakshi et al., 2022] on Frobenius norm low-rank approximation. [Bakshi et al., 2022] propose an algorithm that combines the results of running block Krylov iteration twice, once with block size k for $\tilde{O}(\frac{1}{\varepsilon^{1/3}})$ iterations and once with block size $O(\frac{k}{\varepsilon^{1/3}})$ for $\tilde{O}(1)$ iterations. Their algorithm obtains a $(1+\varepsilon)$ optimal low-rank approximation in the Frobenius norm using $\tilde{O}(\frac{k}{\varepsilon^{1/3}})$ matrix-vector products instead of $\tilde{O}(\frac{k}{\sqrt{\varepsilon}})$, as required by Theorem 1. Since Theorem 1 and Theorem 6 show that single vector Krylov methods match the convergence rates of both block size k and block size $O(\frac{k}{\varepsilon^{1/3}})$, they therefore match the result of [Bakshi et al., 2022]. Formally we have:

Theorem 7. *For $\mathbf{A} \in \mathbb{R}^{n\times d}$, let $g_{min} = \min_{i\in\{1,\ldots,\ell-1\}} \frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+1}}$ where $\ell = \Theta(\frac{k}{\varepsilon^{1/3}})$. For any $\varepsilon, \delta \in (0,1)$, Algorithm 1 initialized with $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and run for $t = O(\frac{k}{\varepsilon^{1/3}}\log(\frac{1}{g_{min}}) + \frac{1}{\varepsilon^{1/3}}\log(\frac{n}{\delta\varepsilon}))$ iterations returns an orthogonal $\mathbf{Q} \in \mathbb{R}^{d\times k}$ such that, with probability at least $1-\delta$, we have*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_F \leq (1+\varepsilon)\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_F.$$

We formally prove Theorem 7 in Appendix B, by generalizing and formalizing an analysis stated in the introduction of [Bakshi et al., 2022]. We note that, in concurrent work, [Bakshi and Narayanan, 2023] show a similar result which has no g_{min} dependence and further removes the $\log n$ dependence. However, their bound only applies to the special case of rank-1 approximation.

4.3 Improved Rates for Schatten Norm Low-Rank Approximation

The work of [Bakshi et al., 2022] also gives bounds for low-rank approximation in general Schatten p-Norms for any $p \geq 1$. They show that by running Krylov iteration 4 times, on both $\mathbf{A}$ and $\mathbf{A}^\intercal$ and with both a relatively small block size $\ell = k$ and a relatively large block size $\ell = \tilde{O}(\frac{k}{\varepsilon^{1/3}})$, they can recover a low-rank approximation to $\mathbf{A}$ in the Schatten p-norm using $\tilde{O}(\frac{kp^{1/6}}{\varepsilon^{1/3}})$ matrix-vector multiplications. As in the case of Frobenius norm low-rank approximation, we show that we can match this entire process with a single instantiation of single vector Krylov, yielding:

Theorem 8 (Schatten-p Norm Low-Rank Approximation)**.** *For $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $p \geq 1$, let $g_{min} = \min_{i\in\{1,\ldots,\ell-1\}} \frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+1}}$ where $\ell = \Theta(\frac{k}{\varepsilon^{1/3}p^{1/3}})$. For any $\varepsilon, \delta \in (0,1)$, let $\mathbf{Q} \in \mathbb{R}^{d\times k}$ be the result of running Algorithm 1 on $\mathbf{A}^\intercal$ initialized with $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and run for*

$$t = O\left(\frac{kp^{1/6}}{\varepsilon^{1/3}}\log(\frac{1}{g_{min}}) + (\sqrt{p} + \frac{p^{1/6}}{\varepsilon^{1/3}})\log(\frac{np}{\delta\varepsilon})\right) = \tilde{O}\left(\frac{kp^{1/6}}{\varepsilon^{1/3}} + \sqrt{p}\right)$$

iterations. Let $\mathbf{Z} \in \mathbb{R}^{n\times k}$ be an orthonormal basis for $\mathbf{AQ}$. Then, with probability at least $1-\delta$,

$$\|\mathbf{A} - \mathbf{Z}\mathbf{Z}^\intercal\mathbf{A}\|_p \leq (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_p,$$

where $\|\mathbf{A}\|_p := (\sum_{i=1}^n \sigma_i(\mathbf{A})^p)^{1/p}$ is the Schatten p-norm.

We prove Theorem 8 in Appendix C. Unlike our previous results, Theorem 8 first uses single vector Krylov to compute an orthonormal basis $\mathbf{Q}$, but then outputs $\mathbf{Z}$, which is an orthonormal basis for $\mathbf{AQ}$. We suspect that this two-step process is an artifact of the analysis in [Bakshi et al., 2022], and that simply returning the output of a single vector Krylov method suffices.

Note that since the same single vector Krylov method obtains the result of Theorem 8 for any choice of p, and setting $p = O(\frac{\log n}{\varepsilon})$ closely approximates the Schatten-∞ norm (i.e., the spectral norm), we achieve the best known result for outputting a low-rank approximation *simultaneously* in all Schatten p-norms:

Corollary 9. *For $\mathbf{A} \in \mathbb{R}^{n\times d}$, let $g_{min} = \min_{i\in\{1,\ldots,\ell-1\}} \frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+1}}$ where $\ell = \Theta(\frac{k}{\log^{1/3}(n)})$. For any $\varepsilon, \delta \in (0,1)$, let $\mathbf{Q} \in \mathbb{R}^{d\times k}$ be the result of running Algorithm 1 on $\mathbf{A}^\intercal$ initialized with $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and run for*

$$t = O\left(\frac{k\log^{\frac{1}{6}}(n)}{\sqrt{\varepsilon}}\log(\frac{1}{g_{min}}) + \frac{\sqrt{\log(n)}}{\sqrt{\varepsilon}}\log(\frac{n}{\delta\varepsilon})\right) = \tilde{O}\left(\frac{k}{\sqrt{\varepsilon}}\right)$$

iterations. Let $\mathbf{Z} \in \mathbb{R}^{n\times k}$ be an orthonormal basis for $\mathbf{AQ}$. Then, for $\varepsilon_p := \varepsilon \cdot \min\{1, \frac{\sqrt{p\varepsilon}}{\log(n)}\} \leq \varepsilon$, with probability at least $1-\delta$, simultaneously for all $p \geq 1$,

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_p \leq (1+\varepsilon_p)\|\mathbf{A} - \mathbf{A}_k\|_p.$$

A similar but weaker guarantee is available from [Bakshi et al., 2022]. They show that running four Block Krylov methods with block size choices depending on p suffices to obtain a low-rank approximation in the Schatten q-norm for all $q \leq p$. If we let $p = O(\frac{\log n}{\varepsilon})$, so that $\|\mathbf{A}\|_p$ approximates the spectral norm, then, their result shows that $\tilde{O}(\frac{k}{\sqrt{\varepsilon}})$ matrix-vector products suffice to output a $(1+\varepsilon)$ relative error low-rank approximation in all Schatten norms, including the Frobenius norm. In contrast, Theorem 7 shows that running single vector Krylov for $\tilde{O}(\frac{k}{\sqrt{\varepsilon}})$ iterations actually gives a better relative error of $(1+\varepsilon^{3/2})$ in the Frobenius norm.

4.4 General Bounds for Small-Block Methods

Our approach to analyzing single vector Krylov also extends to 'small-block' Krylov methods, which use block size $1 < b < k$. Such methods are common in practice as they help avoid slow convergence due to nearby singular values: intuitively, we expect a block size of b to be effective even when the input $\mathbf{A}$ has clusters of at most b very close singular values. Additionally, parallelism often lets us compute multiple matrix-vector products with $\mathbf{A}$ just as quickly as computing a single matrix-vector product, with incentives the use of a block size > 1.

We outline results for small block methods in this section, but defer proofs to Appendix E. To start, we generalize the notion of a gap between singular values:

Definition 2. *Fix block size $b \in [k]$. For each $i \in [k]$, we let $\mathcal{N}_i \subset [k] \setminus \{i\}$ be the indices of the $b-1$ singular values other than i that minimize $|\frac{\sigma_i - \sigma_j}{\sigma_j}|$. Then let $g_{min,b}$ be the b^{th}-order gap of $\mathbf{A}$:*

$$g_{min,b} = \min_{i \in [k]} \min_{j \in [k] \setminus \mathcal{N}_i, j \neq i} \left| \frac{\sigma_i - \sigma_j}{\sigma_j} \right|.$$

If $b = k$, then the sets $\mathcal{N}_i$ have no terms, and we define $g_{min,b} := 1$.

If $b = 1$, then we see that the sets $\mathcal{N}_i$ are empty, and we recover $g_{min,b} = g_{min}$ as defined in Theorem 1. If $b = k$, then $g_{min,b}$ is just 1, which matches the fact that block size k does not require a gap dependence [Musco and Musco, 2015]. If e.g., $b = 2$ and $\mathbf{A}$ has two identical singular values, then we may still have $g_{min,b} > 0$, and so an algorithm can depend on $\frac{1}{g_{min,b}}$ without risking total failure. With this characterization of small-block gaps, we prove the following generalization of Theorem 3 in Appendix E:

Theorem 10. *Fix any PSD $\mathbf{A} \in \mathbb{R}^{n \times n}$ and block size $b \in \{1, \ldots, k\}$. Let $\mathbf{G} \in \mathbb{R}^{n \times b}$ be a matrix with i.i.d. $\mathcal{N}(0,1)$ entries, and let $\mathbf{S}_r = \begin{bmatrix} \mathbf{G} & \mathbf{A}^2\mathbf{G} & \mathbf{A}^4\mathbf{G} & \ldots & \mathbf{A}^{2(r-1)}\mathbf{G} \end{bmatrix}$, where $r = k - b + 1$. For any $\delta \in (0,1)$, with probability at least $(1-\delta)$, there exists a matrix $\mathbf{Q} \in \mathbb{R}^{n \times k}$ that lies in the span of $\mathbf{S}_r$ and is a (k, L)-good starting matrix for $\mathbf{A}$ for $L = \frac{cb^2k^2n\log(1/\delta)}{\delta^2 g_{min,b}^{4(k-b)}}$. Here, c is a fixed constant.*

Above, the starting block $\mathbf{S}_r$ has $rb \approx b(k-b)$ columns, which can be larger than k. Theorem 10 shows that there exists a matrix $\mathbf{Q} \in \mathbb{R}^{n \times k}$ in the span of $\mathbf{S}_r$ which satisfies the (k, L)-good property of Definition 1. Since the Krylov subspace generated by $\mathbf{Q}$ lies within the Krylov subspace generated by $\mathbf{S}_r$, we can thus use Imported Theorem 2 and Imported Theorem 5 to obtain guarantees for the subspace generated by $\mathbf{S}_r$. For a formal argument, see Appendix E. Overall, by plugging in the value of $L = \text{poly}(\frac{n}{\delta g_{min,b}^{k-b}})$ into these theorems, we achieve similar guarantees as for single vector Krylov, but with a dependence on $g_{min,b}$. For instance, we generalize Theorem 1, obtaining:

Theorem 11. *For $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $b \leq k$, let $g_{min,b}$ as in Definition 2. For any $\varepsilon, \delta \in (0,1)$, Algorithm 2 initialized with i.i.d. $\mathcal{N}(0,1)$ matrix $\mathbf{G} \in \mathbb{R}^{n \times b}$ and run for $t = O(\frac{k-b}{\sqrt{\varepsilon}} \log(\frac{1}{g_{min,b}}) + \frac{1}{\sqrt{\varepsilon}} \log(\frac{n}{\delta\varepsilon}))$ iterations returns an orthogonal $\mathbf{Q} \in \mathbb{R}^{n \times k}$ such that, with probability at least $1 - \delta$,*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_\xi \leq (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_\xi \qquad \text{and} \qquad \left|\mathbf{q}_i^\intercal \mathbf{A}\mathbf{A}^\intercal \mathbf{q}_i - \sigma_i(\mathbf{A})^2\right| \leq \varepsilon\sigma_{k+1}(\mathbf{A})^2.$$

In particular, this requires $O(\frac{b(k-b)}{\sqrt{\varepsilon}} \log(\frac{1}{g_{min,b}}) + \frac{b}{\sqrt{\varepsilon}} \log(\frac{n}{\delta\varepsilon}))$ matrix-vector products with $\mathbf{A}$.

For constant $b = O(1)$, the above theorem recovers the same asymptotic matrix-vector complexity as Theorem 1 but with an improved dependence on singular value gaps. For $k - b = O(1)$, it nearly matches the matrix-vector complexity of block Krylov with block size k, but with a very mild gap dependence. When $b = \frac{k}{2}$, we get the worst of both worlds, needing $m = \tilde{O}(\frac{k^2}{\sqrt{\varepsilon}})$ matrix-vector products. This may be a limitation of our proof techniques: it could be possible to show that the matrix-vector complexity scales linearly in k for any block size.

Further, observe that by using Theorem 10, the spectrum-adaptive results of Section 4.1, the fast Frobenius results of Section 4.2, and the fast Schatten norm results of Section 4.3 also apply in the general block size $b \leq k$ case. Wherever those theorems have g_{min}, we replace this with $g_{min,b}$, where $g_{min,b}$ is defined now across the top $\ell - b$ singular values instead of the top $k - b$.

5 Random Perturbation Analysis

Theorems 1, 6, 8 and 11 all show that single vector or small-block Krylov methods match or improve the existing bounds for large block methods, up to a factor of $\log(\frac{1}{g_{min}})$. Single vector methods cannot avoid this dependence in general: if one of $\mathbf{A}$'s top singular values is repeated, so that $g_{min} = 0$, then the Krylov subspace can be rank deficient and fail to converge to the top subspace of $\mathbf{A}$. To address this possible point of failure, we take a smoothed-analysis approach, showing that if our input instance is subjected to a small random perturbation, the dependence on g_{min} can be removed. The key tool we leverage is a line of work from random matrix theory called *eigenvalue repulsion*, which shows that small eigenvalue gaps are brittle: by adding a tiny amount of random noise to any matrix, we can ensure that its eigenvalues are well separated [Minami, 1996, Nguyen et al., 2017, Beenakker, 1997].

5.1 Gap-Independent Bounds for PSD Matrices

Our first result shows that we can remove the dependence on g_{min} when our input matrix $\mathbf{A}$ is PSD. In Section 5.2, we show that this result gives some bounds for non-PSD matrices as well. Our result leverages the following eigenvalue repulsion bound, which we derive in Appendix D.1 using a result of [Minami, 1996].

Lemma 12. *Fix symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\delta \in (0, 1)$, and $\Delta \leq \|\mathbf{A}\|_2$. Let $\mathbf{D} \in \mathbb{R}^{n \times n}$ be a diagonal matrix whose entries are uniformly distributed in $[-\Delta, +\Delta]$. Then, letting $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{D}$ and letting C denote some universal constant, with probability at least $1 - \delta$,*

$$\min_{i \in [n-1]} \frac{|\lambda_i(\tilde{\mathbf{A}}) - \lambda_{i+1}(\tilde{\mathbf{A}})|}{|\lambda_{i+1}(\tilde{\mathbf{A}})|} \geq \frac{\delta}{Cn^2} \cdot \frac{\Delta^2}{\|\mathbf{A}\|_2^2}.$$

That is, by adding a small amount of noise to the diagonal of $\mathbf{A}$, we can ensure its eigenvalue gaps are polynomially large in the problem parameters. While Lemma 12 holds for general symmetric matrices, we will apply it specifically to PSD $\mathbf{A}$ in our analysis, since for indefinite matrices, having large eigenvalue gaps does not necessarily imply having *singular value gaps*, which are required for our convergence results for single vector Krylov methods. It would be interesting to prove an analogous result to Lemma 12 that applies directly to singular values and use it to generalize our results beyond PSD matrices.

Naturally, the larger the perturbation parameter Δ in Lemma 12, the larger the resulting gaps. Thus, Δ gives a tradeoff between runtime and accuracy: higher Δ leads to larger gaps and thus faster convergence. However, it will also make the result of our algorithms less accurate. In

Appendix D.2 we show that picking Δ such that $\|\mathbf{D}\|_2 \leq \frac{\varepsilon\sigma_{k+1}(\mathbf{A})}{n}$ suffices to guarantee that a near-optimal low-rank approximation of $\tilde{\mathbf{A}}$ is also near optimal for $\mathbf{A}$:

Lemma 13. *Let $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{D}$ where $\|\mathbf{D}\|_2 \leq \frac{\varepsilon}{3n}\sigma_{k+1}(\mathbf{A})$ and $\varepsilon \in (0,1)$. Fix any $\mathbf{Q} \in \mathbb{R}^{n\times k}$ with orthonormal columns $\mathbf{q}_1, \ldots, \mathbf{q}_k$. Then, with probability at least $1-\delta$,*

1. *If $|\mathbf{q}_i^\intercal \tilde{\mathbf{A}}\tilde{\mathbf{A}}^\intercal \mathbf{q}_i - \sigma_i(\tilde{\mathbf{A}})^2| \leq \varepsilon\sigma_{k+1}(\tilde{\mathbf{A}})^2$, then $|\mathbf{q}_i^\intercal \mathbf{A}\mathbf{A}^\intercal \mathbf{q}_i - \sigma_i(\mathbf{A})^2| \leq 8\varepsilon\sigma_i(\mathbf{A})^2$.*
2. *If $\|\tilde{\mathbf{A}} - \mathbf{Q}\mathbf{Q}^\intercal\tilde{\mathbf{A}}\|_2 \leq (1+\varepsilon)\|\tilde{\mathbf{A}} - \tilde{\mathbf{A}}_k\|_2$, then $\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_2 \leq (1+2\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_2$.*
3. *If $\|\tilde{\mathbf{A}} - \mathbf{Q}\mathbf{Q}^\intercal\tilde{\mathbf{A}}\|_F \leq (1+\varepsilon)\|\tilde{\mathbf{A}} - \tilde{\mathbf{A}}_k\|_F$, then $\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_F \leq (1+4\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_F$.*

In particular, since we pick $\mathbf{D}$ to be diagonal, we have $\|\mathbf{D}\|_2 \leq \Delta \leq \frac{\varepsilon\sigma_{k+1}(\mathbf{A})}{3n}$. Lemmas 12 and 13 then together imply the following gap-independent variant of Theorem 1 for PSD matrices.

Corollary 14 (Gap-Independent Convergence)**.** *For PSD $\mathbf{A} \in \mathbb{R}^{n\times n}$, let $\kappa_k = \frac{\sigma_1}{\sigma_k}$ and $\Delta = \frac{\varepsilon\sigma_{k+1}}{3n}$. For any $\varepsilon, \delta \in (0, \frac{1}{2})$, let $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{D}$ where $\mathbf{D} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with entries drawn uniformly and i.i.d. from $[-\Delta, \Delta]$. Then, Algorithm 1 run on $\tilde{\mathbf{A}}$ initialized with $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and run for $t = O(\frac{k}{\sqrt{\varepsilon}}\log(\frac{n\kappa_k}{\delta\varepsilon}))$ iterations returns orthogonal $\mathbf{Q} \in \mathbb{R}^{n\times k}$ such that, with probability at least $1-\delta$,*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_\xi \leq (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_\xi \qquad \text{and} \qquad \left|\mathbf{q}_i^\intercal \mathbf{A}\mathbf{A}^\intercal \mathbf{q}_i - \sigma_i(\mathbf{A})^2\right| \leq \varepsilon\sigma_{k+1}(\mathbf{A})^2.$$

Proof. Proving this result simply requires showing that Lemma 12 implies gaps between $\mathbf{A}$'s singular values. This is not immediate since, even if we assume $\mathbf{A}$ is PSD, $\tilde{\mathbf{A}}$ might not be, so could have negative eigenvalues. An eigenvalue at η and another eigenvalue at $-\eta$, which would give a singular value gap of 0, which we need to rule out. To do so, note that $\|\mathbf{D}\|_2 = \max_i |\mathbf{D}_{i,i}| \leq \Delta < \varepsilon\sigma_{k+1}(\mathbf{A})$. So by assuming that $\mathbf{A}$ is PSD, we know that for any negative eigenvalue $\lambda_i(\tilde{\mathbf{A}})$ of $\tilde{\mathbf{A}}$,

$$|\lambda_i(\tilde{\mathbf{A}})| \leq |\lambda_i(\tilde{\mathbf{A}}) - \lambda_i(\mathbf{A})| \leq \|\mathbf{D}\|_2 < \varepsilon\sigma_{k+1}(\mathbf{A}) \leq \sigma_{k+1}(\tilde{\mathbf{A}}).$$

The last inequality assumes $\varepsilon \leq \frac{1}{2}$ to say that $\sigma_{k+1}(\tilde{\mathbf{A}}) \geq \sigma_{k+1}(\mathbf{A}) - \|\mathbf{D}\|_2 \geq (1-\varepsilon)\sigma_{k+1}(\mathbf{A}) \geq \varepsilon\sigma_{k+1}(\mathbf{A})$. So, we know that if $\lambda_i(\tilde{\mathbf{A}}) < 0$ then any singular value associated with $\lambda_i(\tilde{\mathbf{A}})$ is not one of the top k singular values of $\tilde{\mathbf{A}}$. So, the top k singular values of $\tilde{\mathbf{A}}$ must all be associated with nonnegative eigenvalues. That is $\sigma_i(\tilde{\mathbf{A}}) = \lambda_i(\tilde{\mathbf{A}})$ for $i \in [k]$. And so, we find that

$$g_{min} = \min_{i\in\{1,\ldots,k-1\}} \frac{|\sigma_i(\tilde{\mathbf{A}}) - \sigma_{i+1}(\tilde{\mathbf{A}})|}{\sigma_{i+1}(\tilde{\mathbf{A}})} = \min_{i\in\{1,\ldots,k-1\}} \frac{|\lambda_i(\tilde{\mathbf{A}}) - \lambda_{i+1}(\tilde{\mathbf{A}})|}{|\lambda_{i+1}(\tilde{\mathbf{A}})|} \geq \frac{\delta}{Cn^2} \cdot \frac{\Delta^2}{\|\mathbf{A}\|_2^2}.$$

To complete the theorem, we then plug in $\Delta = \frac{\varepsilon\sigma_{k+1}}{3n}$, getting

$$g_{min} \geq \frac{\delta}{Cn^2} \cdot \frac{\frac{\varepsilon^2\sigma_{k+1}^3}{9n^2}}{\sigma_1^2} = \frac{\delta\varepsilon^2}{9Cn^4\kappa_k^2}.$$

We then appeal to Theorem 1 to get the final iteration complexity of $t = O(\frac{k}{\sqrt{\varepsilon}}\log(\frac{1}{g_{min}}) + \frac{1}{\sqrt{\varepsilon}}\log(\frac{n}{\varepsilon\delta})) = O(\frac{k}{\sqrt{\varepsilon}}\log(\frac{n\kappa_k}{\delta\varepsilon}))$. □

Up to a logarithmic dependence on κ_k, Corollary 14 exactly matches the gap-independent low-rank approximation for the block Krylov method [Musco and Musco, 2015]. The same approach can be generalized to give analogs to Theorem 6, Theorem 7, Theorem 8, and Corollary 9 with a dependence on κ_k instead of g_{min}.

Corollary 14 can be interpreted in several ways. In practice, it is unlikely that adding random noise is in fact needed to break small singular value gaps. Noise inherent in the input matrix or due to roundoff error will generally suffice to rule out the existence of tiny singular value gaps. Thus, the corollary can be thought of as a smoothed-analysis result [Spielman and Teng, 2004, Sankar et al., 2006], showing that single vector Krylov methods display gap-independent convergence even on input instances which are tiny random perturbations of potentially worst-case instances. Alternatively, when rigorous worst-case guarantees are required, we could actually run Algorithm 1 on the true input matrix $\mathbf{A}$ with a random diagonal perturbation added. Since $\mathbf{D}$ is diagonal, the runtime of matrix-vector products with $\tilde{\mathbf{A}}$ will generally be dominated by the runtime of matrix-vector products with $\mathbf{A}$, so this is very efficient. We experimentally explore the convergence of this perturbed iteration on a PSD matrix in Section 6.4.

5.2 Gap-Independent Bounds for Rectangular and Indefinite Inputs

We next observe that Corollary 14 gives results for non-PSD matrices as well, at least for low-rank approximation with respect to the spectral norm. In this case, we can run a single vector Krylov method on a perturbation of PSD matrix $\mathbf{A}\mathbf{A}^\intercal$. This suffices because the spectral norm guarantee has the property that a near-optimal basis for approximating $\mathbf{A}\mathbf{A}^\intercal$ is also a near-optimal for $\mathbf{A}$:

Lemma 15. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$. Let $\mathbf{Q} \in \mathbb{R}^{n\times k}$ be a matrix with orthonormal columns such that $\|\mathbf{A}\mathbf{A}^\intercal - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\mathbf{A}^\intercal\|_2 \leq (1+\varepsilon)\|\mathbf{A}\mathbf{A}^\intercal - (\mathbf{A}\mathbf{A}^\intercal)_k\|_2$. Then, $\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_2 \leq (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_2$.*

Proof. Let $\mathbf{P} := \mathbf{I} - \mathbf{Q}\mathbf{Q}^\intercal$. Observe that $\|\mathbf{A}\mathbf{A}^\intercal - (\mathbf{A}\mathbf{A}^\intercal)_k\|_2 = \sigma_{k+1}(\mathbf{A})^2$, so we are given that $\|\mathbf{P}\mathbf{A}\mathbf{A}^\intercal\|_2 \leq (1+\varepsilon)\ \sigma_{k+1}(\mathbf{A})^2$. Next note that $\mathbf{P}$ is a projection matrix, which can only decrease spectral norms, so we have

$$\|\mathbf{P}\mathbf{A}\|_2^2 = \|(\mathbf{P}\mathbf{A})(\mathbf{P}\mathbf{A})^\intercal\|_2 = \|\mathbf{P}\mathbf{A}\mathbf{A}^\intercal\mathbf{P}\|_2 \leq \|\mathbf{P}\mathbf{A}\mathbf{A}^\intercal\|_2 \leq (1+\varepsilon)\ \sigma_{k+1}(\mathbf{A})^2.$$

Taking the square root of both sides, and noting that $\sqrt{1+\varepsilon} \leq 1+\varepsilon$, we conclude that $\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_2 \leq (1+\varepsilon)\sigma_{k+1}(\mathbf{A}) = (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_2$. □

Combining Lemma 15 with Lemma 15 we obtain the following result:

Corollary 16. *For $\mathbf{A} \in \mathbb{R}^{n\times d}$, let $\kappa_k = \frac{\sigma_1}{\sigma_k}$ and $\Delta = \frac{\varepsilon\sigma_{k+1}^2}{3n}$. For any $\varepsilon, \delta \in (0, \frac{1}{2})$, let $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{A}^\intercal + \mathbf{D}$ where $\mathbf{D} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with entries drawn uniformly and i.i.d. from $[-\Delta, \Delta]$. Then, Algorithm 1 run on $\tilde{\mathbf{A}}$ initialized with $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and run for $t = O(\frac{k}{\sqrt{\varepsilon}}\log(\frac{n\kappa_k}{\delta\varepsilon}))$ iterations returns an orthogonal $\mathbf{Q} \in \mathbb{R}^{n\times k}$ such that, with probability at least $1-\delta$,*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_2 \leq (1+\varepsilon)\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_2.$$

Note that it is not clear that Lemma 15 extends beyond the spectral norm, e.g., to the Frobenius norm. Nevertheless, we suspect that a result comparable to Corollary 16 should hold for all other error metrics considered in this paper.

6 Numerical Experiments

In this section we validate the core findings of our theoretical results with numerical experiments. We focus on four key findings. In Section 6.2, we verify that the dependence of single vector Krylov on the sequential gap size g_{min} is in fact logarithmic, matching the theoretical bounds of Sections 3 and 4. In Section 6.3, we show that using a small block size $b > 1$ can ameliorate this

gap dependence, by replacing $\log(1/g_{min})$ with $\log(1/g_{min,b})$ as shown in Theorem 11. Relatedly, in Section 6.4 we show that a small random perturbation of the input matrix can break up overlapping singular values and lead to much faster convergence of the single vector Krylov method, matching our theoretical findings from Section 5. In Section 6.5 we compare single vector and large block Krylov methods. We find that for a wide range of matrices, single vector Krylov methods significantly outperform larger block methods. However, for some very specific worst case instances, large block methods can perform better.

Finally, while our theoretical bounds ignore issues of numerical stability, in Section 6.6, we observe empirically that small block methods tend to have significantly more issues with stability than large block methods. Exploring this issue further in future work would be very interesting.

6.1 Experimental Set Up

To control against stability issues in our primary experiments, we implement algorithms Algorithm 1 and Algorithm 2 using a full reothogonalization strategy to keep the Krylov subspace close to orthogonal. At every iteration of the (block) Krylov iteration, we orthogonalize the most recently generated column (resp. block) of the Krylov subspace against all previous columns (resp. blocks) using the modified Gram-Schmidt process. At the next iteration, this column (resp. block) is multiplied by $\mathbf{A}$ to produce a new column (resp. block) of the Kyrlov subspace. See Section 6.6 for more details or our code, which is available on GitHub[5].

We also note that, like most standard implementations, including those based on the Lanczos recurrence, our code implements Algorithm 1 run for t iterations using $t+1$ matrix-vector products with $\mathbf{A}\mathbf{A}^\intercal$. To see why this is possible, note that the matrix $\mathbf{A}\mathbf{A}^\intercal\mathbf{Z}$ computed on Line 2 of the algorithm can be formed "on-the-fly" as we generate the Krylov subspace. Let $\mathbf{z}_i$ be the i^{th} column in $\mathbf{Z}$. At each iteration we already compute $\mathbf{A}\mathbf{A}^\intercal\mathbf{z}_i$ for column $\mathbf{z}_i$ to form column $\mathbf{z}_{i+1}$, so can just store this result to form $\mathbf{A}\mathbf{A}^\intercal\mathbf{Z}$. Similarly, implementing Algorithm 2 requires $(t+1)\ell$ matrix-vector products with $\mathbf{A}\mathbf{A}^\intercal$.

Throughout our experiments, we only report low-rank approximation error in terms of the Frobenius norm, as we found convergence in the spectral and Frobenius norms typically matched quite closely. In particular, letting $\mathbf{Q}$ be the output of Algorithm 1 or Algorithm 2, we report

$$\varepsilon_{empirical} := \frac{\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_F - \|\mathbf{A} - \mathbf{A}_k\|_F}{\|\mathbf{A} - \mathbf{A}_k\|_F}.$$

Our theory describes how $\varepsilon_{empiricial}$ should change as a function of the number of iterations, the block size, and the singular value gaps of $\mathbf{A}$.

Lastly, we note that, since Krylov methods initialized with random Gaussian vectors are invariant to rotation, without loss of generality we test on diagonal input matrices for all synthetic data experiments. Each matrix's diagonal entries correspond to its singular values.

6.2 Verifying Gap Dependence

We first empirically show that single vector Krylov has a logarithmic dependence on the minimum sequential gap size $g_{min} = \min_{i\in[k-1]} \frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+1}}$, as predicted by the bounds of Sections 3 and 4. We consider an exponentially decaying spectrum with parameter $\alpha = 1.1$ whose singular values are all nearly repeated, with gap sizes varying between $g_{min} \in [10^{-10}, 1]$. That is, letting our vector of

[5] https://github.com/RaphaelArkadyMeyerNYU/SingleVectorKrylov

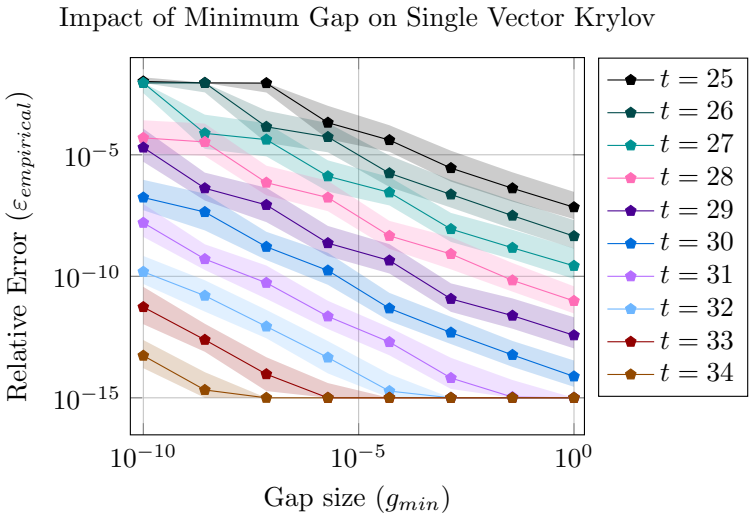

Figure 3: Low-rank approximation error vs. minimum gap size for diagonal $\mathbf{A} \in \mathbb{R}^{1000\times 1000}$ with singular values $\boldsymbol{\sigma}$ as described in Equation (12). For 8 different gap sizes logarithmically spaced in $[10^{-10}, 1]$, we run single vector Krylov for $t = 25, 26, \ldots, 34$ iterations with target rank $k = 10$. The median of $\varepsilon_{empirical}$ over 500 independent trials is plotted, with the 25^{th} and 75^{th} quartiles shaded in. When $\varepsilon_{empirical} < 10^{-15}$, we plot it at 10^{-15} so that the log/log plot does not degenerate. As expected given the theoretical bound of Theorem 6, we see a linear relationship between the log relative error and log gap size.

singular values be denoted $\boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_n]$ so that $\mathbf{A} = \mathrm{diag}(\boldsymbol{\sigma})$, and fixing $n = 1000$, we let

$$\boldsymbol{\sigma} = \left[1, \quad \tfrac{1}{1+g_{min}}, \quad \alpha^{-1}, \quad \tfrac{\alpha^{-1}}{1+g_{min}}, \quad \alpha^{-2}, \quad \tfrac{\alpha^{-2}}{1+g_{min}}, \quad \ldots \quad \alpha^{-499}, \quad \tfrac{\alpha^{-499}}{1+g_{min}}\right]. \tag{12}$$

Since this matrix has a fast decay of singular values, we expect the performance of single vector Krylov to follow the spectral decay rate of Theorem 6. That is, fixing the dimension d and failure probability δ, we should expect the number of iterations to scale as:

$$t \propto \tfrac{\ell}{\sqrt{g_{k\to\ell}}} \log\left(\tfrac{1}{g_{min}}\right) + \tfrac{1}{\sqrt{g_{k\to\ell}}} \log\left(\tfrac{1}{\varepsilon}\right).$$

Rearranging this expression, we can equivalently expect

$$\begin{aligned} \log(\varepsilon) &= -t\sqrt{g_{k\to\ell}} - \ell \log(g_{min}) \\ &= \quad -C_t \quad - \ell \log(g_{min}) \end{aligned}$$

where $C_t := t\sqrt{g_{k\to\ell}}$ is independent of g_{min}. So, if we plot $\varepsilon_{empirical}$ versus g_{min} on a log/log plot for a fixed value of t, we should see a line with negative slope. Further, since the vertical offset of these lines are $C_t \propto t$, we should expect that increasing t should shift these lines downwards and proportionally to t. We see this behavior exactly in Figure 3.

6.3 Verifying the Effect of Block Size on Gap Dependence

We next show that when $\mathbf{A}$ has very small singular value gaps (or even exactly overlapping singular values), the dependence on $\log(1/g_{\min})$ can be avoided by using a small constant block size b. This lets us instead depend on $\log(1/g_{min,b})$, as in the analysis of Theorem 11.

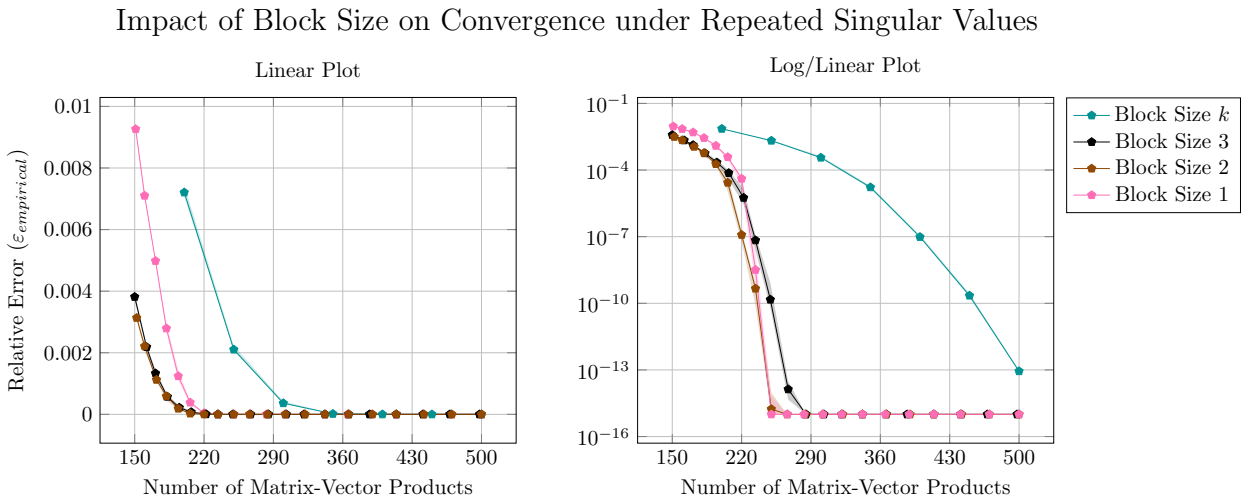

Figure 4: Low-rank approximation error vs. number of matrix-vector products for diagonal $\mathbf{A} \in \mathbb{R}^{1000\times 1000}$ with repeated top singular values as in Equation (13). We run Krylov iteration with target rank $k = 50$ for block sizes $b = 1, 2, 3, 50$. The median of $\varepsilon_{empirical}$ over 10 independent trials is plotted, with the 25^{th} and 75^{th} quartiles shaded in. Note that these quartiles are very close together in this example. When $\varepsilon_{empirical} < 10^{-15}$, we plot it at 10^{-15} so that the plot does not degenerate. The plot on the left has a linear y-axis and highlights the performance for early iterations, while the plot on the right uses a logarithmic y-axis and highlights the performance for later iterations. Overall, we see that block size $b = 2$, which is just large enough to avoid the repeated pairs of singular values in $\mathbf{A}$'s spectrum, performs best. Single vector Krylov is handicapped by the repeated singular values, especially at early iterations. Large block Krylov converges at a significantly slower rate than the small block variants.

We focus on when $\mathbf{A}$ has pairs, but not triplets, of exactly overlapping singular values. In this case, block Krylov with block size $b = 2$ should perform well, since it should not suffer due to the overlapping singular values. Further, it should match or outperform larger block methods. To show this, we construct an exponentially decaying spectrum with parameter $\alpha = 1.005$ and whose top $k = 50$ singular values are each repeated, with sequential gap size $g_{min} = 0$. Formally, we choose 1000 singular values as follows:

$$\begin{aligned} \boldsymbol{\sigma}_A &= \begin{bmatrix} 1 & 1 & \alpha^{-1} & \alpha^{-1} & \alpha^{-2} & \alpha^{-2} & \dots & \alpha^{-25} & \alpha^{-25} \end{bmatrix} \\ \boldsymbol{\sigma}_B &= \begin{bmatrix} \alpha^{-26} & \alpha^{-27} & \alpha^{-28} & \dots & \alpha^{-975} \end{bmatrix} \\ \boldsymbol{\sigma} &= \begin{bmatrix} \boldsymbol{\sigma}_A & \boldsymbol{\sigma}_B \end{bmatrix}. \end{aligned} \tag{13}$$

In theory, single vector Krylov should completely fail in this case, only capturing a $k/2$-dimensional subspace of the span of the top k singular vectors. Due to finite precision roundoff, the method nevertheless converges. However, it is still significantly handicapped by the repeated singular values.

In Figure 4 we plot the low-rank approximation error vs. number of matrix-vector products of single vector Krylov and block Krylov with block sizes 2, 3, and 50, for target rank $k = 50$. We show both y-linear and y-logarithmic plots to highlight the performance at early and later iterations. We see that block size 2 performs the best across the board, and that block size 3 is only mildly worse. Due to the repeated singular values, single vector Krylov performs worse, especially for the early iterations. It becomes competitive with block size 3 eventually. In contrast, the full block size k method converges much more slowly.

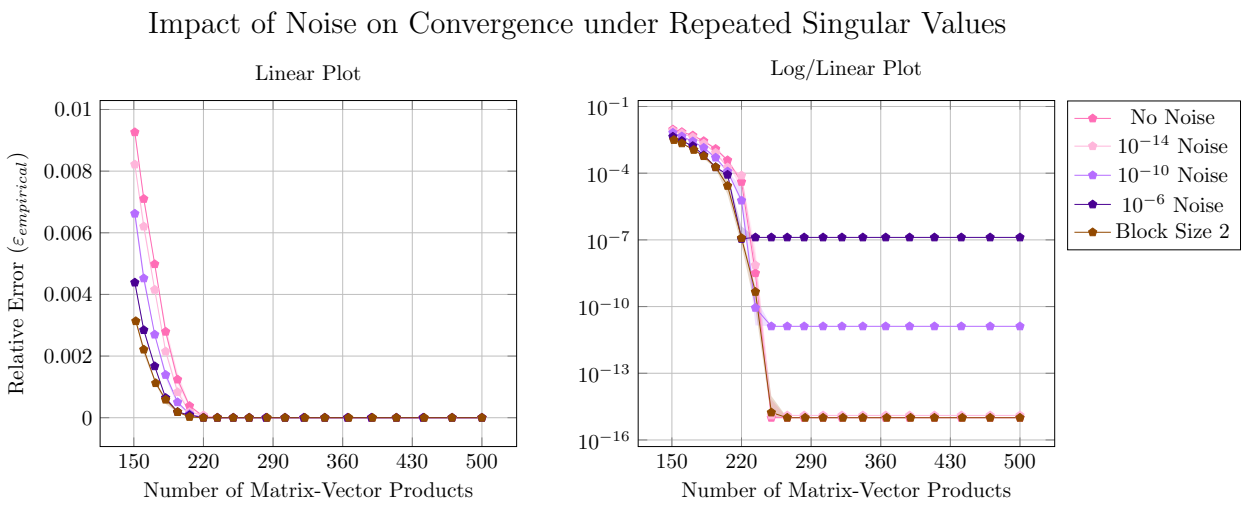

Figure 5: Low-rank approximation error vs. number of matrix-vector products for diagonal $\mathbf{A} \in \mathbb{R}^{1000\times 1000}$ with repeated top singular values as described in Equation (13). We run the single vector Krylov method with target rank $k = 50$, with varying levels of noise added to $\mathbf{A}$ to separate its repeated singular values. In particular, following Corollary 14, we run the method on $\mathbf{A} + \mathbf{D}$ where $\mathbf{D} \in \mathbb{R}^{1000\times 1000}$ is a random diagonal matrix with entries drawn uniformly on $[-\Delta, \Delta]$, for $\Delta \in \{10^{-6}, 10^{-10}, 10^{-14}, 0\}$. Low-rank approximation error is measured with respect to the original input $\mathbf{A}$. For comparison, we also show performance of block Krylov with block size $b = 2$ with no noise added. This was the optimal block size for this spectrum as seen in Figure 4. The median of $\varepsilon_{empirical}$ over 10 independent trials is plotted, with the 25^{th} and 75^{th} quartiles shaded in. Note that these quartiles are very close together in this example. When $\varepsilon_{empirical} < 10^{-15}$, we plot it at 10^{-15} so that the plot does not degenerate. The plot on the left has a linear y-axis and highlights the performance for early iterations, while the plot on the right uses a logarithmic y-axis and highlights the performance for later iterations. We can see that by adding a small random perturbation, we can improve the convergence of single vector Krylov to nearly match that of block Krylov with $b = 2$. Larger noise leads to faster convergence but also larger final error.

6.4 Verifying the Effect of Random Perturbations on Gap Dependence

Next, we show that adding a small amount of random noise to break up small singular value gaps can also make single vector Krylov converge more quickly, verifying the results of Section 5. We use the same matrix as in Section 6.3, with spectrum given in Equation (13). In Figure 5, we show that adding noise to $\mathbf{A}$ the order of 10^{-6} leads to single vector Krylov converging nearly as quickly as the optimal $b = 2$ block Krylov method as seen in Section 6.3. This noise does limit our eventual accuracy at convergence, which can be seen clearly in our logarithmic error plot. Changing the magnitude of the noise lets us interpolate between fast convergence and high accuracy.

6.5 Effect of Block Size in Convergence

We next present a wider comparison of how the choice of block size for Krylov iteration effects convergence to a near-optimal low-rank approximation. We fix target rank $k = 50$ and compare block sizes 1, 2, 3, 50, and 54. Block size 1 corresponds to the single vector Krylov method. Block sizes 2 and 3 should be more resilient to a pairs or triplets of very close singular values, respectively. Block sizes k and $k + 4$ are recommended by prior theoretical work on Krylov Iteration for low-

rank approximation [Musco and Musco, 2015, Tropp, 2018]. We consider eight input matrices. All synthetic inputs are 1000×1000 and diagonal.

1. **Exponential Decay**: $\sigma_i = \alpha^{-i}$ for $\alpha \in \{1.001,\ 1.01,\ 1.1\}$
2. **Polynomial Decay**: $\sigma_i = i^{-\beta}$ for $\beta \in \{0.1,\ 0.5,\ 1.5\}$
3. **Repeated Singular Values**: A matrix with each of its top k singular values repeated, as defined in Equation (13).
4. **Wishart Lower Bound**: $\sigma_i = \sqrt{1 - (\frac{i}{1000})^2}$. This is an approximation of the spectrum of $\mathbf{I} - \frac{1}{5n}\mathbf{G}^\intercal\mathbf{G}$ where $\mathbf{G}^{1000\times 1000}$ has i.i.d $\mathcal{N}(0,1)$ entries. This matrix is used as a lower bound instance for rank-1 low-rank approximation in [Bakshi et al., 2022].
5. **nd3k, appu, human_gene_2, exdata_1**: Various real-world matrices arbitrarily chosen from SuiteSparse [Davis and Hu, 2011].

We can see the results of these experiments in Figure 6. We see that for all except the repeated singular value and Wishart lower bound matrices, single vector Krylov dominates. For the repeated singular value matrix, as in Section 6.3, we see that block size 2 dominates again. For the Wishart lower bound matrix from [Bakshi et al., 2022], we see that large block methods marginally (though consistently) outperform small block methods. This lower bound instance is designed to force Krylov methods to converge at a rate of $\frac{1}{\varepsilon^{1/3}}$, instead of at the spectral decay rate $\frac{1}{\sqrt{g_{k\to\ell}}}$. This seems to makes the rate of convergence of single vector Krylov slower than block Krylov, since we pay a $\log(1/g_{min})$ dependence, while only benefiting a small amount from separating the k and $\log(n/\varepsilon)$ dependence (see Theorem 1). In contrast, the other figures show matrices where the spectral decay rate controls convergence, where single vector Krylov still pays $\log(1/g_{min})$ but seems to see performance gains from being able to simulate general block sizes and from separating the $\log(n/\varepsilon)$ dependence from the (simulated) block size dependence (see Theorem 6).

6.6 Block Size and Numerical Stability

It is well known that Krylov methods can suffer from numerical stability issues [Golub and Van Loan, 2013, Meurant and Strakoš, 2006]. In particular, the iterates $(\mathbf{A}\mathbf{A}^\intercal)^t\mathbf{g}$ approach the same vector (the top singular vector of $\mathbf{A}$) as t grows large. So, $\mathbf{K}$ becomes ill-conditioned. So far, we have focused on convergence guarantees and ignored numerical stability. As discussed, our implementations use orthogonalization to keep $\mathbf{K}$ well-conditioned at all iterations. That is, for single vector Krylov, at every iteration, we compute $(\mathbf{A}\mathbf{A}^\intercal)\mathbf{k}_{t-1}$ where $\mathbf{k}_{t-1}$ is the last column in the Krylov matrix $\mathbf{K}$. Then we project $(\mathbf{A}\mathbf{A}^\intercal)\mathbf{k}_{t-1}$ away from all of the previous columns $\mathbf{k}_1, \ldots, \mathbf{k}_{t-1}$ via modified Gram-Schmidt, and store the resulting vector as $\mathbf{k}_t$. We do the same for block Krylov, where we compute $(\mathbf{A}\mathbf{A}^\intercal)\begin{bmatrix}\mathbf{k}_{t-b-1} & \ldots & \mathbf{k}_{t-1}\end{bmatrix}$ and add the resulting b columns to $\mathbf{K}$ iteratively via modified Gram-Schmidt.

In practice, Krylov implementations typically spend less effort orthogonalizing at each step. For example, they are commonly implemented via the Lanczos method, where $\mathbf{k}_t$ is only projected away from $\mathbf{k}_{t-1}$ and $\mathbf{k}_{t-2}$. In infinite precision, this is equivalent to projecting away from all previous columns [Golub and Van Loan, 2013]. Similar ideas can be applied to block Krylov methods [Rokhlin et al., 2009, Saad, 1980]. While such methods are highly efficient, when using them, $\mathbf{K}$ can lose orthogonality. This can lead to slower convergence or necessitate modifications such as restarts or reorthogonalization [Calvetti et al., 1994, Paige, 1972, Parlett, 1998].

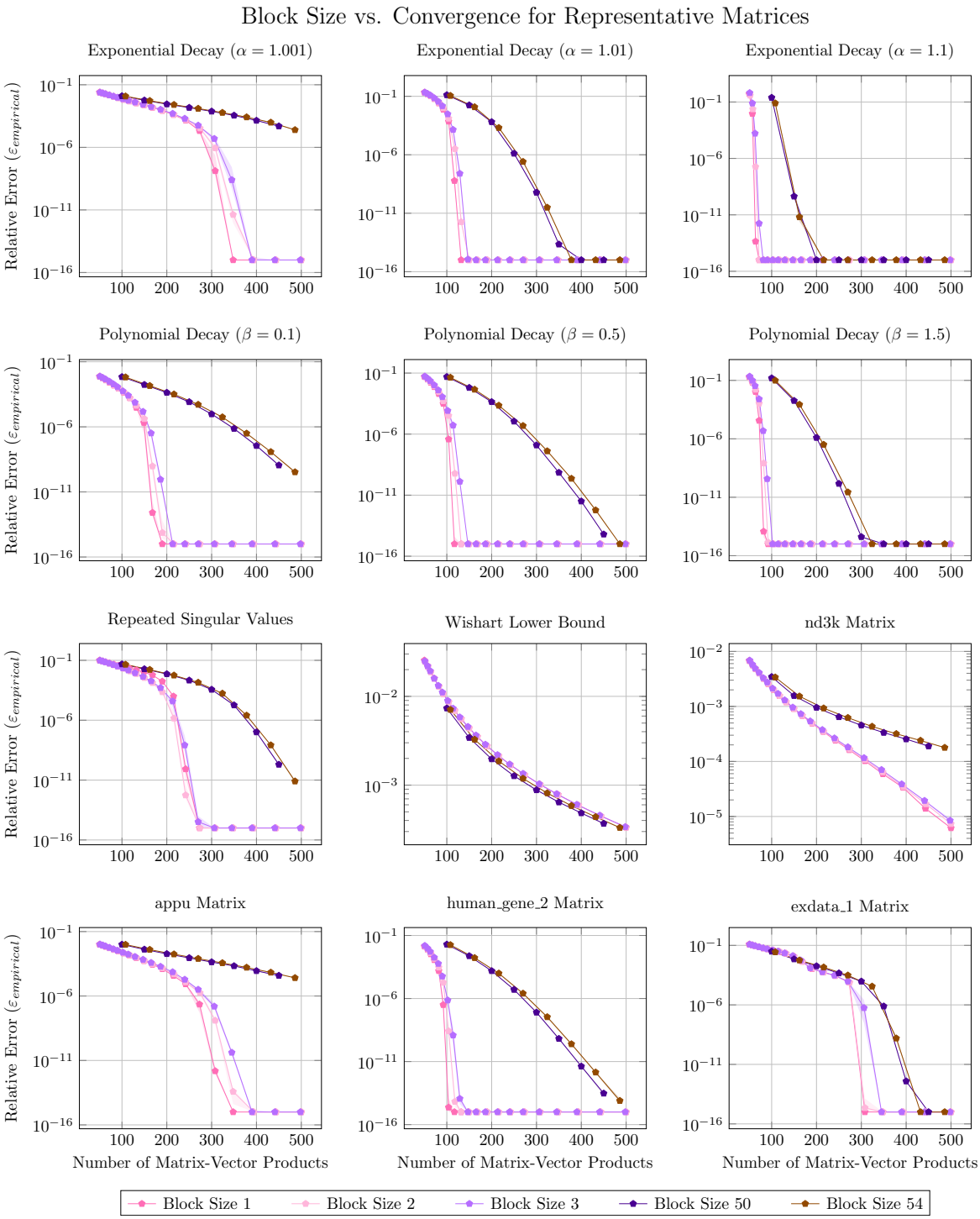

Figure 6: Low-rank approximation error vs. number of matrix-vector products for Krylov iteration with various block sizes on synthetic and real-world input matrices as described in Section 6.5. In all cases, the target rank is set to $k = 50$. The median of $\varepsilon_{empirical}$ is plotted over 10 independent trials, with the 25^{th} and 75^{th} quartiles shaded in. Note that these quartiles are very close together in most plots. When $\varepsilon_{empirical} < 10^{-15}$, we plot it at 10^{-15} so that the plot does not degenerate.

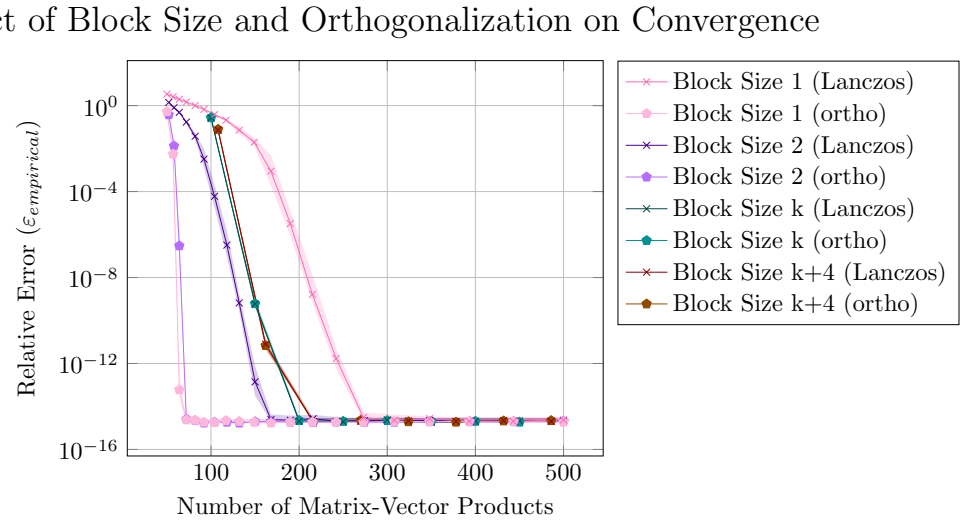

Figure 7: Low-rank approximation error vs. number of matrix-vector products for diagonal $\mathbf{A} \in \mathbb{R}^{1000\times 1000}$ with singular values $\sigma_i = 1.1^{-i}$ and target rank $k = 50$. We run for block sizes $b \in \{1, 2, k, k+4\}$ both with Lanczos and with full orthogonalization. The median of $\varepsilon_{empirical}$ over 100 independent trials is plotted, with the 25^{th} and 75^{th} quartiles shaded in. Notice that for this experiment, single vector Krylov converges the fastest with full orthogonalization and slowest with Lanczos, and that large block methods have no real gap between full orthogonalization and Lanczos. I.e., large block methods are much more stable even with partial orthogonalization.

Intuitively, comparing single vector or small block Krylov to large block Krylov with a fixed size Krylov subspace $\mathbf{K}$, we expect that single vector and small block Krylov will be more susceptible to conditioning issues, since they require more iterations to reach the same sized subspace. Thus, we should expect partial orthogonalization methods like Lanczos to lead to slower convergence for these methods as compared to large block methods. With full orthogonalization, we should instead expect to see small block methods dominate. We see this trend exactly in Figure 7. An interesting extension to our work would be to more closely study the stability of Krylov methods for low-rank approximation, and to develop a more clear theoretical understanding of the advantages of large block methods in this regard.

Acknowledgements

This work was supported by NSF Awards 2046235 and 2045590. Cameron Musco thanks Alex Breuer for helpful conversations in 2016 that helped inspire the initial idea behind this work. Raphael Meyer was partially supported by a GAANN fellowship from the US Department of Education. He thanks Tyler Chen, Apoorv Singh, and Axel Elaldi for help designing the figures. He also thanks Kathryn Lund, Kirk Soodhalter, and David Persson for helpful conversations contextualizing this work. He further thanks David Persson for help formalizing the results on single vector simultaneous iteration.

References

[Aizenman et al., 2017] Aizenman, M., Peled, R., Schenker, J., Shamis, M., and Sodin, S. (2017). Matrix regularizing effects of Gaussian perturbations. *Communications in Contemporary Mathematics*, 19(03):1750028.

[Allen-Zhu and Li, 2016] Allen-Zhu, Z. and Li, Y. (2016). LazySVD: Even faster SVD decomposition yet without agonizing pain. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*.

[Bakshi et al., 2022] Bakshi, A., Clarkson, K. L., and Woodruff, D. P. (2022). Low-rank approximation with $1/\varepsilon^{1/3}$ matrix-vector products. In *Proceedings of the 54th Annual ACM Symposium on Theory of Computing (STOC)*.

[Bakshi and Narayanan, 2023] Bakshi, A. and Narayanan, S. (2023). Krylov methods are (nearly) optimal for low-rank approximation. *arXiv:2304.03191*.

[Banks et al., 2022] Banks, J., Garza-Vargas, J., Kulkarni, A., and Srivastava, N. (2022). Pseudospectral shattering, the sign function, and diagonalization in nearly matrix multiplication time. *Foundations of Computational Mathematics*, pages 1–89.

[Beenakker, 1997] Beenakker, C. W. (1997). Random-matrix theory of quantum transport. *Reviews of modern physics*, 69(3):731.

[Boutsidis et al., 2016] Boutsidis, C., Woodruff, D. P., and Zhong, P. (2016). Optimal principal component analysis in distributed and streaming models. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing (STOC)*.

[Calvetti et al., 1994] Calvetti, D., Reichel, L., and Sorensen, D. C. (1994). An implicitly restarted Lanczos method for large symmetric eigenvalue problems. *Electronic Transactions on Numerical Analysis*, 2(1):21.

[Chu, 1955] Chu, J. T. (1955). On bounds for the normal integral. *Biometrika*, 42(1/2):263–265.

[Clarkson and Woodruff, 2013] Clarkson, K. L. and Woodruff, D. P. (2013). Low rank approximation and regression in input sparsity time. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing (STOC)*.

[Cohen et al., 2015] Cohen, M., Elder, S., Musco, C., Musco, C., and Persu, M. (2015). Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC)*.

[Cullum and Donath, 1974] Cullum, J. and Donath, W. (1974). A block Lanczos algorithm for computing the q algebraically largest eigenvalues and a corresponding eigenspace of large, sparse, real symmetric matrices. In *IEEE Conference on Decision and Control including the 13th Symposium on Adaptive Processes*, pages 505–509.

[Davis and Hu, 2011] Davis, T. A. and Hu, Y. (2011). The University of Florida sparse matrix collection. *ACM Transactions on Mathematical Software*, 38(1).

[Drineas et al., 2018] Drineas, P., Ipsen, I. C., Kontopoulou, E.-M., and Magdon-Ismail, M. (2018). Structural convergence results for approximation of dominant subspaces from block Krylov spaces. *SIAM Journal on Matrix Analysis and Applications*, 39(2):567–586.

[Drineas and Ipsen, 2019] Drineas, P. and Ipsen, I. C. F. (2019). Low-rank matrix approximations do not need a singular value gap. *SIAM Journal on Matrix Analysis and Applications*, 40(1):299–319.

[Drineas and Mahoney, 2016] Drineas, P. and Mahoney, M. W. (2016). RandNLA: Randomized numerical linear algebra. *Communications of the ACM*, 59(6).

[Golub and Underwood, 1977] Golub, G. and Underwood, R. (1977). The block Lanczos method for computing eigenvalues. *Mathematical Software III*, (3):361–377.

[Golub and Van Loan, 2013] Golub, G. and Van Loan, C. (2013). *Matrix Computations*. Johns Hopkins University Press, 4th edition.

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

[Halko et al., 2011a] Halko, N., Martinsson, P.-G., Shkolnisky, Y., and Tygert, M. (2011a). An algorithm for the principal component analysis of large data sets. *SIAM Journal on Scientific Computing*, 33(5):2580–2594.

[Halko et al., 2011b] Halko, N., Martinsson, P.-G., and Tropp, J. (2011b). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288.

[Hegde et al., 2016] Hegde, C., Indyk, P., and Schmidt, L. (2016). Fast recovery from a union of subspaces. *Advances in Neural Information Processing Systems 29 (NeurIPS)*, 29.

[Huang and Tikhomirov, 2020] Huang, H. and Tikhomirov, K. (2020). A remark on the smallest singular value of powers of Gaussian matrices. *Electronic Communications in Probability*, 25:1–8.

[Kahan and Parlett, 1976] Kahan, W. and Parlett, B. (1976). How far should you go with the Lanczos process? In *Sparse Matrix Computations*, pages 131–144. Academic Press.

[Kaniel, 1966] Kaniel, S. (1966). Estimates for some computational techniques in linear algebra. *Mathematics of Computation*, 20(95):369–378.

[Kuczyński and Woźniakowski, 1992] Kuczyński, J. and Woźniakowski, H. (1992). 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.

[Laurent and Massart, 2000] Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338.

[Lehoucq et al., 1998] Lehoucq, R. B., Sorensen, D. C., and Yang, C. (1998). *ARPACK users' guide: solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*. SIAM.

[Li et al., 2017] Li, H., Linderman, G. C., Szlam, A., Stanton, K. P., Kluger, Y., and Tygert, M. (2017). Algorithm 971: An implementation of a randomized algorithm for principal component analysis. *ACM Transations on Mathemathical Software*, 43(3).

[Li and Zhang, 2015] Li, R.-C. and Zhang, L.-H. (2015). Convergence of the block Lanczos method for eigenvalue clusters. *Numerische Mathematik*, 131(1):83–113.

[Martinsson et al., 2006] Martinsson, P.-G., Rokhlin, V., and Tygert, M. (2006). A randomized algorithm for the approximation of matrices. Technical Report 1361, Yale University.

[Martinsson and Tropp, 2020] Martinsson, P.-G. and Tropp, J. A. (2020). Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica*, 29:403–572.

[Mathworks, 2023] Mathworks (2023). Matlab svds documentation. https://www.mathworks.com/help/matlab/ref/svds.html.

[Meurant and Strakoš, 2006] Meurant, G. and Strakoš, Z. (2006). The Lanczos and conjugate gradient algorithms in finite precision arithmetic. *Acta Numerica*, 15:471–542.

[Minami, 1996] Minami, N. (1996). Local fluctuation of the spectrum of a multidimensional Anderson tight binding model. *Communications in mathematical physics*, 177:709–725.

[Musco and Musco, 2015] Musco, C. and Musco, C. (2015). Randomized block Krylov methods for stronger and faster approximate singular value decomposition. In *Advances in Neural Information Processing Systems 28 (NeurIPS)*. Full version at arXiv:1504.05477.

[Nguyen et al., 2017] Nguyen, H., Tao, T., and Vu, V. (2017). Random matrices: tail bounds for gaps between eigenvalues. *Probability Theory and Related Fields*, 167(3):777–816.

[Paige, 1971] Paige, C. C. (1971). *The computation of eigenvalues and eigenvectors of very large sparse matrices*. PhD thesis, University of London.

[Paige, 1972] Paige, C. C. (1972). Computational variants of the Lanczos method for the eigenproblem. *IMA Journal of Applied Mathematics*, 10(3):373–381.

[Parlett, 1998] Parlett, B. N. (1998). *The symmetric eigenvalue problem*. SIAM.

[Peng and Vempala, 2021] Peng, R. and Vempala, S. (2021). Solving sparse linear systems faster than matrix multiplication. In *Proceedings of the 32nd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*.

[Rokhlin et al., 2009] Rokhlin, V., Szlam, A., and Tygert, M. (2009). A randomized algorithm for principal component analysis. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1100–1124.

[Rudelson and Vershynin, 2010] Rudelson, M. and Vershynin, R. (2010). Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM)*, volume 3, pages 1576–1602.

[Saad, 1980] Saad, Y. (1980). On the rates of convergence of the Lanczos and the block-Lanczos methods. *SIAM Journal on Numerical Analysis*, 17(5):687–706.

[Saad, 2011] Saad, Y. (2011). *Numerical Methods for Large Eigenvalue Problems: Revised Edition*, volume 66.

[Sankar et al., 2006] Sankar, A., Spielman, D. A., and Teng, S.-H. (2006). Smoothed analysis of the condition numbers and growth factors of matrices. *SIAM Journal on Matrix Analysis and Applications*, 28(2):446–476.

[SciPy Community, 2023] SciPy Community (2023). SciPy v1.10.1 Manual: scipy.sparse.linalg.svds. https://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.linalg.svds.html.

[Simchowitz et al., 2018] Simchowitz, M., El Alaoui, A., and Recht, B. (2018). Tight query complexity lower bounds for PCA via finite sample deformed Wigner law. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1249–1259.

[Soltani and Hegde, 2018] Soltani, M. and Hegde, C. (2018). Fast low-rank matrix estimation for ill-conditioned matrices. In *2018 IEEE International Symposium on Information Theory (ISIT)*.

[Spielman and Teng, 2004] Spielman, D. A. and Teng, S.-H. (2004). Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM)*, 51(3):385–463.

[Tropp, 2018] Tropp, J. A. (2018). Analysis of randomized block Krylov methods.

[Urschel, 2021] Urschel, J. C. (2021). Uniform error estimates for the Lanczos method. *SIAM Journal on Matrix Analysis and Applications*, 42(3):1423–1450.

[Wang et al., 2015] Wang, S., Zhang, Z., and Zhang, T. (2015). Improved analyses of the randomized power method and block Lanczos method. *arXiv:1508.06429*.

[Wegner, 1981] Wegner, F. (1981). Bounds on the density of states in disordered systems. *Zeitschrift für Physik B Condensed Matter*, 44(1):9–15.

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

[Yuan et al., 2018] Yuan, Q., Gu, M., and Li, B. (2018). Superlinear convergence of randomized block Lanczos algorithm. In *IEEE International Conference on Data Mining (ICDM)*, pages 1404–1409.

A Reduction to the Positive Semidefinite Case

In our analysis, we can assume without loss of generality that the input $\mathbf{A}$ is a square PSD matrix. To see why, for any $\mathbf{C} \in \mathbb{R}^{n\times d}$, let $\mathbf{A} = (\mathbf{C}\mathbf{C}^\intercal)^{1/2}$. Observe that $\mathbf{A} \in \mathbb{R}^{n\times n}$ is PSD. Further, observe that since $\mathbf{A}^2 = \mathbf{A}\mathbf{A}^\intercal = \mathbf{C}\mathbf{C}^\intercal$, Algorithm 1 and Algorithm 2 yield identical outputs for $\mathbf{A}$ and $\mathbf{C}$.

Expanding the SVD $\mathbf{C} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\intercal$, we can have $\mathbf{A} = (\mathbf{C}\mathbf{C}^\intercal)^{1/2} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\intercal$. Thus, $\mathbf{A}$ and $\mathbf{C}$ have identical singular values and $\|\mathbf{A} - \mathbf{A}_k\|_\xi = \|\mathbf{C} - \mathbf{C}_k\|_\xi$ for any unitarily invariant norm $\|\cdot\|_\xi$ (including the spectral and Frobenius norms) and any k. Additionally, for any $\mathbf{Q} \in \mathbb{R}^{n\times k}$,

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_\xi = \|\mathbf{U}\mathbf{\Sigma} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{U}\mathbf{\Sigma}\|_\xi = \|\mathbf{C} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{C}\|_\xi.$$

Thus, any bound on $\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_\xi$ in terms of $\|\mathbf{A} - \mathbf{A}_k\|_\xi$ holds identically for $\mathbf{C}$. Finally, for any $\mathbf{q} \in \mathbb{R}^n$, $\mathbf{q}^\intercal\mathbf{C}\mathbf{C}^\intercal\mathbf{q} = \mathbf{q}^\intercal\mathbf{A}^2\mathbf{q}$. Thus, any bound on $\mathbf{q}^\intercal\mathbf{A}^2\mathbf{q}$ in terms of $\sigma_i(\mathbf{A})$ holds identically for $\mathbf{C}$.

B Frobenius Low-Rank Approximation with $\varepsilon^{-1/3}$ Dependence

In this section, we prove Theorem 7. This analysis closely follows the intuition given in the introduction of [Bakshi et al., 2022]. We have the following:

Theorem 7 Restated. *For $\mathbf{A} \in \mathbb{R}^{n\times d}$, let $g_{min} = \min_{i\in\{1,\ldots,\ell-1\}} \frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+1}}$ where $\ell = \Theta(\frac{k}{\varepsilon^{1/3}})$. For any $\varepsilon, \delta \in (0,1)$, Algorithm 1 initialized with $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and run for $t = O(\frac{k}{\varepsilon^{1/3}}\log(\frac{1}{g_{min}}) + \frac{1}{\varepsilon^{1/3}}\log(\frac{n}{\delta\varepsilon}))$ iterations returns an orthogonal $\mathbf{Q} \in \mathbb{R}^{n\times k}$ such that, with probability at least $1-\delta$,*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_F \leq (1+\varepsilon)\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_F.$$

Proof. We first recall that we have two different guarantees for Algorithm 1. Fix some $\gamma > 0$. By Theorem 1, if we run Algorithm 1 for $t = O(\frac{k}{\sqrt{\gamma}}\log(\frac{1}{g_{min}}) + \frac{1}{\sqrt{\gamma}}\log(\frac{n}{\delta\gamma}))$ iterations, then with probability at least $1-\delta$,

$$|\mathbf{q}_i^\intercal \mathbf{A}\mathbf{A}^\intercal \mathbf{q}_i - \sigma_i^2| \leq \gamma\sigma_{k+1}^2.$$

Fix some $\ell \geq k$, and let $g_{k\to\ell} = \frac{\sigma_k - \sigma_{\ell+1}}{\sigma_k}$. Then, by Theorem 6, if we run Algorithm 1 for $t = O(\frac{\ell}{\sqrt{g_{k\to\ell}}}\log(\frac{1}{g_{min}}) + \frac{1}{\sqrt{g a p_{k\to\ell}}}\log(\frac{n}{\delta\eta}))$ iterations, then with probability at least $1-\delta$ we again have

$$|\mathbf{q}_i^\intercal \mathbf{A}\mathbf{A}^\intercal \mathbf{q}_i - \sigma_i^2| \leq \eta\sigma_{k+1}^2.$$

We will show that we can get error ε in Frobenius norm by taking $\gamma = \varepsilon^{2/3}$ and $\eta = \frac{\varepsilon}{k}$. In particular, we run a case-analysis between either large-tailed or small-tailed spectra of $\mathbf{A}$.

Small Tailed Case: Suppose $\|\mathbf{A} - \mathbf{A}_k\|_F^2 \leq \frac{k}{\varepsilon^{1/3}}\sigma_{k+1}^2$. Then $\mathbf{A}$ must have a fast spectral decay. In particular, let $\ell = (1 + \frac{4}{\varepsilon^{1/3}})k = O(\frac{k}{\varepsilon^{1/3}})$. Then σ_ℓ is substantially smaller than σ_k:

$$\frac{k}{\varepsilon^{1/3}}\sigma_{k+1}^2 \geq \|\mathbf{A} - \mathbf{A}_k\|_F^2 = \sum_{i=k+1}^{n} \sigma_i^2 \geq (\ell - k)\sigma_\ell^2 = \frac{4k}{\varepsilon^{1/3}}\sigma_\ell^2.$$

That is, $\sigma_\ell \leq \frac{\sigma_{k+1}}{2}$, so that $\sqrt{g_{k\to\ell}} = \sqrt{\frac{\sigma_k - \sigma_{\ell+1}}{\sigma_k}} \geq \frac{1}{\sqrt{2}}$, so the spectral-decay analysis of Theorem 6 says that $t = O(\frac{k}{\varepsilon^{1/3}}\log(\frac{1}{g_{min}}) + \log(\frac{n}{\delta\eta}))$ iterations suffice to get the singular value guarantee $\|\mathbf{q}_i^\intercal\mathbf{A}\|_2^2 \in \sigma_i^2 \pm \eta\sigma_{k+1}^2$. Since $\mathbf{Q}\mathbf{Q}^\intercal = \sum_{i=1}^k \mathbf{q}_i\mathbf{q}_i^\intercal$ is a sum of orthogonal projection matrices,

$$\begin{aligned}
\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_F^2 &= \|\mathbf{A}\|_F^2 - \sum_{i=1}^k \|\mathbf{q}_i\mathbf{q}_i^\intercal\mathbf{A}\|_F^2 && \text{(Matrix Pythagoras)}\\
&= \|\mathbf{A}\|_F^2 - \sum_{i=1}^k \|\mathbf{q}_i^\intercal\mathbf{A}\|_2^2 \\
&\leq \|\mathbf{A}\|_F^2 - \sum_{i=1}^k \sigma_i^2 + \eta k\sigma_{k+1}^2 && \text{(Singular Value Guarantee)}\\
&= \|\mathbf{A} - \mathbf{A}_k\|_F^2 + \eta k\sigma_{k+1}^2 \\
&\leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + \eta k\|\mathbf{A} - \mathbf{A}_k\|_F^2 \\
&= (1+\eta k)\|\mathbf{A} - \mathbf{A}_k\|_F^2.
\end{aligned}$$

So, taking $\eta = \frac{\varepsilon}{k}$ for a total iteration count of $t = O(\frac{k}{\varepsilon^{1/3}}\log(\frac{1}{g_{min}}) + \log(\frac{n}{\delta\varepsilon}))$ suffices in this case.

Large Tailed Case: Suppose $\|\mathbf{A} - \mathbf{A}_k\|_F^2 > \frac{k}{\varepsilon^{1/3}}\sigma_{k+1}^2$. Since the tail is large, even a low-accuracy singular value guarantee still ensures a good Frobenius norm guarantee. In particular, we take the

gap-independent analysis of Theorem 1 with $\gamma = \varepsilon^{2/3}$, so that $\|\mathbf{q}_i^\intercal \mathbf{A}\|_2^2 \in \sigma_i^2 \pm \varepsilon^{2/3}\sigma_{k+1}^2$, and we get

$$\begin{aligned}
\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_F^2 &= \|\mathbf{A}\|_F^2 - \sum_{i=1}^{k} \|\mathbf{q}_i\mathbf{q}_i^\intercal\mathbf{A}\|_F^2 && \text{(Matrix Pythagoras)}\\
&= \|\mathbf{A}\|_F^2 - \sum_{i=1}^{k} \|\mathbf{q}_i^\intercal\mathbf{A}\|_2^2\\
&\leq \|\mathbf{A}\|_F^2 - \sum_{i=1}^{k} \sigma_i^2 + \varepsilon^{2/3}k\sigma_{k+1}^2 && \text{(Singular Value Guarantee)}\\
&= \|\mathbf{A} - \mathbf{A}_k\|_F^2 + \varepsilon^{2/3}k\sigma_{k+1}^2\\
&\leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + \varepsilon\|\mathbf{A} - \mathbf{A}_k\|_F^2 && (\sigma_{k+1}^2 < \tfrac{\varepsilon^{1/3}}{k}\|\mathbf{A} - \mathbf{A}_k\|_F^2)\\
&= (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2.
\end{aligned}$$

So, since $\gamma = \varepsilon^{2/3}$ here, we achieve error ε under Frobenius norm with $t = O(\frac{k}{\varepsilon^{1/3}}\log(\frac{1}{g_{min}}) + \frac{1}{\varepsilon^{1/3}}\log(\frac{n}{\delta\varepsilon}))$.

Putting it Together. So, in either case, running $t = \tilde{O}(\frac{k}{\varepsilon^{1/3}})$ iterations suffices to obtain a $(1+\varepsilon)$ optimal low-rank approximation in the Frobenius norm. Further, the algorithm used in the two cases is identical, so (unlike [Bakshi et al., 2022]) we do not have to detect which case we are in and alter the algorithm accordingly. We simply run single vector Krylov. □

C Schatten Norm Low-Rank Approx. with $\varepsilon^{-1/3}$ Dependence

In this section, we prove Theorem 8 by arguing that running Algorithm 1 once effectively simulates running Algorithm 5.4 from [Bakshi et al., 2022]. We have the following:

Theorem 8 Restated. *For $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $p \geq 1$, let $g_{min} = \min_{i\in\{1,\ldots,\ell-1\}} \frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+1}}$ where $\ell = \Theta(\frac{k}{\varepsilon^{1/3}p^{1/3}})$. For any $\varepsilon, \delta \in (0,1)$, let $\mathbf{Q} \in \mathbb{R}^{d\times k}$ be the result of running Algorithm 1 on $\mathbf{A}^\intercal$ initialized with $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and run for*

$$t = O\left(\frac{kp^{1/6}}{\varepsilon^{1/3}}\log(\frac{1}{g_{min}}) + (\sqrt{p} + \frac{p^{1/6}}{\varepsilon^{1/3}})\log(\frac{np}{\delta\varepsilon})\right) = \tilde{O}\left(\frac{kp^{1/6}}{\varepsilon^{1/3}} + \sqrt{p}\right)$$

iterations. Let $\mathbf{Z} \in \mathbb{R}^{n\times k}$ be an orthonormal basis for $\mathbf{AQ}$. Then, with probability at least $1-\delta$,

$$\|\mathbf{A} - \mathbf{Z}\mathbf{Z}^\intercal\mathbf{A}\|_p \leq (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_p,$$

where $\|\mathbf{A}\|_p := (\sum_{i=1}^n \sigma_i(\mathbf{A})^p)^{1/p}$ is the Schatten p-norm.

Proof. Note that [Bakshi et al., 2022] outputs orthonormal $\mathbf{Q} \in \mathbb{R}^{d\times k}$ with $\|\mathbf{A} - \mathbf{A}\mathbf{Q}\mathbf{Q}^\intercal\|$ bounded, rather than $\mathbf{Q} \in \mathbb{R}^{n\times k}$ with $\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|$ bounded. However, we can translate their analysis to the later case simply by running their algorithms on $\mathbf{A}^\intercal$. Thus we consider this case going forward. The first two lines of Algorithm 5.4 in [Bakshi et al., 2022] run Block Krylov Iteration twice on $\mathbf{A}^\intercal$. First, they let $\mathbf{W}_1 \in \mathbb{R}^{d\times k}$ be the result of using block size $\ell_1 = k$ and running until the gap-independent rate gives a singular value guarantee (i.e. Equation (7)) with relative error at most $\gamma_1 = \frac{\varepsilon^{2/3}}{p^{1/3}}$. For single vector Krylov, by Theorem 1, this takes

$$O\left(\frac{k}{\sqrt{\gamma_1}}\log(\frac{1}{g_{min}}) + \frac{1}{\sqrt{\gamma_1}}\log(\frac{n}{\gamma_1\delta})\right) = O\left(\frac{kp^{1/6}}{\varepsilon^{1/3}}\log(\frac{1}{g_{min}}) + \frac{p^{1/6}}{\varepsilon^{1/3}}\log(\frac{np}{\varepsilon\delta})\right)$$

iterations. Second, they let $\mathbf{W}_2 \in \mathbb{R}^{d\times k}$ be the result of running with block size $\ell_2 = O(\frac{k}{\varepsilon^{1/3}p^{1/3}})$ for enough iterations so that if $g_{k\to\ell_2} \geq \frac{1}{p}$, then block Krylov would achieve error $\gamma_2 = \text{poly}(\frac{\varepsilon}{n})$[6]. For single vector Krylov, by Theorem 6, this takes

$$O\left(\frac{\ell_2}{\sqrt{g_{k\to\ell_2}}}\log(\frac{1}{g_{min}}) + \frac{1}{\sqrt{g_{k\to\ell_2}}}\log(\frac{n}{\gamma_2\delta})\right) = O\left(\frac{kp^{1/6}}{\varepsilon^{1/3}}\log(\frac{1}{g_{min}}) + \sqrt{p}\log(\frac{n}{\varepsilon\delta})\right)$$

iterations. Note that Algorithm 1 outputs a single matrix $\mathbf{Q}$ that achieves the guarantees needed by both $\mathbf{W}_1$ and $\mathbf{W}_2$.

Next, we consider the third and fourth lines of Algorithm 5.4 in [Bakshi et al., 2022]. The third line runs block Krylov on $\mathbf{A}$ directly to estimate several of its singular values. The fourth line uses those estimated singular values to determine if we should return an orthogonal basis for $\mathbf{W}_1^\intercal\mathbf{A}^\intercal$ or $\mathbf{W}_2^\intercal\mathbf{A}^\intercal$.[7] Since we have $\mathbf{W}_1 = \mathbf{W}_2 = \mathbf{Q}$, we can ignore the tests in the third and fourth lines, and just always return a basis for $\mathbf{Q}^\intercal\mathbf{A}^\intercal$. So, overall, we compute a matrix $\mathbf{Z}$ with the exact same guarantees as the [Bakshi et al., 2022] by only using a one instance of single vector Krylov. □

D Eigenvalue Repulsion Corollaries

This appendix covers the proofs needed for Corollary 14. First, we take a result of [Minami, 1996] and use it to prove a gap on the eigenvalues of symmetric matrices. Second, we show that a optimal projection matrix that achieves near-optimal low-rank approximation on the perturbation of $\mathbf{A}$ must also achieve near-optimal low-rank approximation on $\mathbf{A}$ itself.

D.1 Proof of Lemma 12

We first import a result of [Minami, 1996], originally studied in relation to the Wegner Estimate [Wegner, 1981]. This result is also given as Equation (1.11) in [Aizenman et al., 2017]:

Imported Theorem 17. *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be symmetric, and let $\mathbf{D} \in \mathbb{R}^{n\times n}$ be diagonal, with entries drawn i.i.d. from a distribution with pdf $p(\cdot)$. Then, for any interval $\mathcal{I} \subset \mathbb{R}$,*

$$\Pr[\mathbf{A}+\mathbf{D} \text{ has at least 2 eigenvalues in } \mathcal{I}] \leq C(\|p\|_\infty|\mathcal{I}|n)^2,$$

for some universal constant C, where $|\mathcal{I}|$ is the length of $\mathcal{I}$, and where $\|p\|_\infty = \max_{t\in\mathbb{R}} p(t)$.

Lemma 12 Restated. *Fix symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\delta \in (0,1)$, and $\Delta \leq \|\mathbf{A}\|_2$. Let $\mathbf{D} \in \mathbb{R}^{n\times n}$ be a diagonal matrix whose entries are uniformly distributed in $[-\Delta,+\Delta]$. Then, letting $\tilde{\mathbf{A}} = \mathbf{A}+\mathbf{D}$ and letting C denote some universal constant, with probability at least $1-\delta$,*

$$\min_{i\in[n-1]} \frac{|\lambda_i(\tilde{\mathbf{A}}) - \lambda_{i+1}(\tilde{\mathbf{A}})|}{|\lambda_{i+1}(\tilde{\mathbf{A}})|} \geq \frac{\delta}{Cn^2}\cdot\frac{\Delta^2}{\|\mathbf{A}\|_2^2}.$$

Proof. Let $R := 2\|\mathbf{A}\|_2$. Since $\Delta < \|\mathbf{A}\|_2$, we know that $\|\tilde{\mathbf{A}}\|_2 \leq \|\mathbf{A}\|_2 + \|\mathbf{D}\|_2 \leq 2\|\mathbf{A}\|_2 = R$. Let $\gamma > 0$ be a number to be fixed later. Then define

$$\mathcal{I}_i := (-R+i\gamma) \pm \gamma = [(-R+\gamma i)-\gamma, (-R+\gamma i)+\gamma]$$

[6] They write $\gamma_2 = \varepsilon$ in the algorithm but in Equation (5.21) we can see they actually want this smaller error. Since gap-dependent rate depends on $\log(\frac{n}{\gamma_2})$, shrinking γ_2 from ε to $\text{poly}(\frac{\varepsilon}{n})$ does not change the asymptotic complexity.

[7] In a personal communication with the authors of [Bakshi et al., 2022], we confirmed there is a typo in the current arXiv version of the paper, where the algorithm says to return a matrix $\mathbf{Z}_1$. It should return an orthonormal basis for $\mathbf{W}_1^\intercal\mathbf{A}^\intercal$ instead.

for $i = 1, \dots, m$, where $m = \frac{4\|\mathbf{A}\|_2}{\gamma} - 1$. These are intervals of width 2γ that overlap and cover the range $[-R, R]$. For instance, we have $\mathcal{I}_1 = [-R, -R + 2\gamma]$, $\mathcal{I}_2 = [-R + \gamma, -R + 2\gamma]$, and $\mathcal{I}_3 = [-R + 2\gamma, R + 4\gamma]$, so that $\mathcal{I}_2$ overlaps with $\mathcal{I}_1$ and $\mathcal{I}_3$. In particular, if $\tilde{\mathbf{A}}$ has two eigenvalues that are additively γ close, so that $|\lambda_i(\tilde{\mathbf{A}}) - \lambda_{i+1}(\tilde{\mathbf{A}})| \le \gamma$, then we know that $\lambda_i(\tilde{\mathbf{A}})$ and $\lambda_{i+1}(\tilde{\mathbf{A}})$ both lie in some $\mathcal{I}_j$. Therefore, we can write

$$\begin{aligned}
\Pr[\exists i : |\lambda_i(\tilde{\mathbf{A}}) - \lambda_{i+1}(\tilde{\mathbf{A}})| < \gamma] &\le \Pr[\text{at least two eigenvalues of } \tilde{\mathbf{A}} \text{ lie in some } \mathcal{I}_j] \\
&\le \sum_{j=1}^{m} \Pr[\text{at least two eigenvalues of } \tilde{\mathbf{A}} \text{ lie in } \mathcal{I}_j] \\
&\le Cm \left(\frac{1}{2\Delta} \cdot 2\gamma \cdot n \right)^2 && \text{(Imported Theorem 17)} \\
&\le \frac{4C\gamma n^2 \|\mathbf{A}\|_2}{\Delta^2} && \text{(using that } m \le \tfrac{4\|\mathbf{A}\|_2}{\gamma}\text{)} \\
&= \delta,
\end{aligned}$$

where the last line holds if we fix $\gamma = \frac{\Delta^2 \delta}{4Cn^2\|\mathbf{A}\|_2}$. That is, with probability at least $1 - \delta$, we know that $|\lambda_i(\tilde{\mathbf{A}}) - \lambda_{i+1}(\tilde{\mathbf{A}})| \ge \frac{\Delta^2\delta}{4Cn^2\|\mathbf{A}\|_2}$ for all $i = 1, \dots, n-1$. Lastly, we take

$$\frac{|\lambda_i(\tilde{\mathbf{A}}) - \lambda_{i+1}(\tilde{\mathbf{A}})|}{|\lambda_{i+1}(\tilde{\mathbf{A}})|} \ge \frac{|\lambda_i(\tilde{\mathbf{A}}) - \lambda_{i+1}(\tilde{\mathbf{A}})|}{\|\tilde{\mathbf{A}}\|_2} \ge \frac{\frac{\Delta^2\delta}{4Cn^2\|\mathbf{A}\|_2}}{2\|\mathbf{A}\|_2} = \frac{\Delta^2\delta}{8Cn^2\|\mathbf{A}\|_2^2},$$

which completes the proof. □

D.2 Proof of Lemma 13

We next show that a small enough perturbation of $\mathbf{A}$ suffices to give approximate SVD results for $\mathbf{A}$ itself.

Lemma 13 Restated. *Let $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{D}$ where $\|\mathbf{D}\|_2 \le \frac{\varepsilon}{3n}\sigma_{k+1}(\mathbf{A})$ and $\varepsilon \in (0, 1)$. Fix any $\mathbf{Q} \in \mathbb{R}^{n \times k}$ with orthonormal columns $\mathbf{q}_1, \dots, \mathbf{q}_k$. Then, with probability at least $1 - \delta$,*

1. *If $|\mathbf{q}_i^\intercal \tilde{\mathbf{A}}\tilde{\mathbf{A}}^\intercal \mathbf{q}_i - \sigma_i(\tilde{\mathbf{A}})^2| \le \varepsilon\sigma_{k+1}(\tilde{\mathbf{A}})^2$, then $|\mathbf{q}_i^\intercal \mathbf{A}\mathbf{A}^\intercal \mathbf{q}_i - \sigma_i(\mathbf{A})^2| \le 8\varepsilon\sigma_i(\mathbf{A})^2$.*
2. *If $\|\tilde{\mathbf{A}} - \mathbf{Q}\mathbf{Q}^\intercal\tilde{\mathbf{A}}\|_2 \le (1+\varepsilon)\|\tilde{\mathbf{A}} - \tilde{\mathbf{A}}_k\|_2$, then $\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_2 \le (1+2\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_2$.*
3. *If $\|\tilde{\mathbf{A}} - \mathbf{Q}\mathbf{Q}^\intercal\tilde{\mathbf{A}}\|_F \le (1+\varepsilon)\|\tilde{\mathbf{A}} - \tilde{\mathbf{A}}_k\|_F$, then $\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_F \le (1+4\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_F$.*

Proof.

Singular Value Guarantee. First note that for any real a, b, c such that $|a - b| \le c$ and $c < b$, we have $|a^2 - b^2| \le 3bc$. This follows from expanding $(b-c)^2 < a^2 < (b+c)^2$ and applying the AMGM inequality. Then note that $|\sigma_i(\tilde{\mathbf{A}}) - \sigma_i(\mathbf{A})| \le \|\tilde{\mathbf{A}} - \mathbf{A}\|_2 = \|\mathbf{D}\|_2$. We then find that for $i \le k+1$,

$$|\sigma_i^2(\tilde{\mathbf{A}}) - \sigma_i^2(\mathbf{A})| \le 3\sigma_i(\mathbf{A})\|\mathbf{D}\|_2 \le \varepsilon\sigma_i(\mathbf{A})\sigma_{k+1}(\mathbf{A}) \le \varepsilon\sigma_i(\mathbf{A})^2.$$

Similarly note that $|\|\tilde{\mathbf{A}}\mathbf{q}_i\|_2 - \|\mathbf{A}\mathbf{q}_i\|_2| \le \|\mathbf{D}\|_2$ and $\|\tilde{\mathbf{A}}\mathbf{q}_i\|_2 \le 2\sigma_i(\tilde{\mathbf{A}}) \le 4\sigma_i(\mathbf{A})$, so we have

$$|\|\tilde{\mathbf{A}}\mathbf{q}_i\|_2^2 - \|\mathbf{A}\mathbf{q}_i\|_2^2| \le 3\|\tilde{\mathbf{A}}\mathbf{q}_i\|_2\|\mathbf{D}\|_2 \le 4\varepsilon\sigma_i(\mathbf{A})^2.$$

Which completes this part by triangle inequality:

$$\begin{aligned}|\mathbf{q}_i^\intercal \mathbf{A}\mathbf{A}^\intercal \mathbf{q}_i - \sigma_i(\mathbf{A})^2| &\le |\mathbf{q}_i^\intercal \tilde{\mathbf{A}}\tilde{\mathbf{A}}^\intercal \mathbf{q}_i - \sigma_i(\mathbf{A})^2| + 4\varepsilon\sigma_i(\mathbf{A})^2 \\ &\le |\mathbf{q}_i^\intercal \tilde{\mathbf{A}}\tilde{\mathbf{A}}^\intercal \mathbf{q}_i - \sigma_i(\tilde{\mathbf{A}})^2| + 7\varepsilon\sigma_i(\mathbf{A})^2 \\ &\le 8\varepsilon\sigma_i(\mathbf{A})^2.\end{aligned}$$

Spectral Norm Guarantee. Here, first note that $\|\tilde{\mathbf{A}} - \tilde{\mathbf{A}}_k\|_2 = \sigma_{k+1}(\tilde{\mathbf{A}}) \le (1+\varepsilon)\sigma_{k+1}(\mathbf{A})$ by the prior analysis on the singular value guarantee. Next, we use the fact that $\mathbf{I} - \mathbf{Q}\mathbf{Q}^\intercal$ is a projection to simplify

$$\|\tilde{\mathbf{A}} - \mathbf{Q}\mathbf{Q}^\intercal\tilde{\mathbf{A}}\|_2 = \|(\mathbf{I} - \mathbf{Q}\mathbf{Q}^\intercal)(\mathbf{A}+\mathbf{D})\|_2 \ge \|(\mathbf{I} - \mathbf{Q}\mathbf{Q}^\intercal)\mathbf{A}\|_2 - \|\mathbf{D}\|_2,$$

and since $\|\mathbf{D}\|_2 \le \varepsilon\sigma_{k+1}(\mathbf{A})$, we have $\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_2 \le \|\tilde{\mathbf{A}} - \mathbf{Q}\mathbf{Q}^\intercal\tilde{\mathbf{A}}\|_2 + \|\mathbf{D}\|_2 \le (1+2\varepsilon)\sigma_{k+1}(\mathbf{A})$, which completes this part of the lemma.

Frobenius Norm Guarantee. Here, first note that $\|\tilde{\mathbf{A}} - \tilde{\mathbf{A}}_k\|_F^2 \le (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2$, since

$$\|\tilde{\mathbf{A}} - \tilde{\mathbf{A}}_k\|_F^2 = \sum_{i=k+1}^{n} \sigma_i^2(\tilde{\mathbf{A}}) \le \sum_{i=k+1}^{n} (\sigma_i(\mathbf{A}) + \|\mathbf{D}\|_2)^2 \le \|\mathbf{A} - \mathbf{A}_k\|_F^2 + \sum_{i=k+1}^{n} (\|\mathbf{D}\|_2^2 + 2\sigma_i(\mathbf{A})\|\mathbf{D}\|_2),$$

where we can further upper bound

$$\sum_{k+1}^{n}(\|\mathbf{D}\|_2^2 + 2\sigma_i(\mathbf{A})\|\mathbf{D}\|_2) \le n(\tfrac{\varepsilon\sigma_{k+1}(\mathbf{A})}{3n})^2 + 2n\sigma_{k+1}(\mathbf{A})\tfrac{\varepsilon\sigma_{k+1}(\mathbf{A})}{3n} \le \varepsilon\sigma_{k+1}^2(\mathbf{A}) \le \varepsilon\|\mathbf{A} - \mathbf{A}_k\|_F^2.$$

Next, using the fact that $\mathbf{P} = \mathbf{I} - \mathbf{Q}\mathbf{Q}^\intercal$ is a projection matrix, we simplify

$$\|\tilde{\mathbf{A}} - \mathbf{Q}\mathbf{Q}^\intercal\tilde{\mathbf{A}}\|_F = \|\mathbf{P}(\mathbf{A}+\mathbf{D})\|_F \ge \|\mathbf{P}\mathbf{A}\|_F - \|\mathbf{P}\mathbf{D}\|_F,$$

and since $\|\mathbf{P}\mathbf{D}\|_F \le \|\mathbf{P}\|_F\|\mathbf{D}\|_2 \le \sqrt{d}\tfrac{\varepsilon}{3n}\sigma_{k+1}(\mathbf{A}) \le \varepsilon\|\mathbf{A} - \mathbf{A}_k\|_F$, we have

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_F \le \|\tilde{\mathbf{A}} - \mathbf{Q}\mathbf{Q}^\intercal\tilde{\mathbf{A}}\|_F + \|\mathbf{P}\mathbf{D}\|_F \le ((1+\varepsilon)^2 + \varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_F,$$

which completes the proof since $((1+\varepsilon)^2 + \varepsilon) \le 1 + 4\varepsilon$ for $\varepsilon \in (0,1)$. □

E Krylov Analysis with Small Blocks

This section proves Theorem 11. We first define the starting matrix that we will simulate block Krylov iteration on, analogously to what we use in Section 3:

$$\mathbf{S}_r := \begin{bmatrix}\mathbf{G} & \mathbf{A}^2\mathbf{G} & \mathbf{A}^4\mathbf{G} & \dots & \mathbf{A}^{2(r-1)}\mathbf{G}\end{bmatrix} \qquad \mathbf{K} \overset{\text{span}}{=} \begin{bmatrix}\mathbf{S}_r & \mathbf{A}^2\mathbf{S}_r & \mathbf{A}^4\mathbf{S}_r & \dots & \mathbf{A}^{2q}\mathbf{S}_r\end{bmatrix}$$

where $\ell = rb$ is the simulated block size (so we assume the integer r has $rb > k$) and where $q = t - r + 1$ denotes the number of simulated block-Krylov iterations run. Our proof will set $r = k - b + 1$. Notably, this means that $\mathbf{S}_r$ can have more than k columns, which is not allowed by the definition of (k, L)-good in Definition 1. So, we first present a generalization of Definition 1 that also suffices for convergence under Imported Theorem 2 and Imported Theorem 5. In particular, it suffices for $\mathbf{S}_r$ to contain a $n \times k$ size (k, L)-good matrix within its span. The Krylov subspace generated by this matrix will be contained in the subspace generated by $\mathbf{S}_r$, and thus any guarantees that hold for it apply to $\mathbf{S}_r$ as well. See Appendix F for a formal argument.

Definition 3 ((k, L)-good Starting Matrix (Generalized))**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be a matrix with top k left singular vectors $\mathbf{U}_k \in \mathbb{R}^{n\times k}$. A matrix $\mathbf{B} \in \mathbb{R}^{n\times \ell}$ is a (k, L)-good starting matrix for $\mathbf{A}$ if $\|(\mathbf{U}_k^\intercal \mathbf{Q})^{-1}\|_2^2 \leq L$ for some orthonormal $\mathbf{Q} \in \mathbb{R}^{n\times k}$ whose columns lie in* span$(\mathbf{B})$.

Observe that when the starting block has exactly $\ell = k$ columns, Definition 3 matches Definition 1 exactly. To prove the (k, L)-good guarantee for small block methods, we first note an equivalent formulation of the generalized (k, L)-good guarantee:

Lemma 18. *$\mathbf{B} \in \mathbb{R}^{n\times \ell}$ is (k, L)-good for $\mathbf{A} \in \mathbb{R}^{n\times n}$ if and only if there exists a matrix $\mathbf{M} \in \mathbb{R}^{\ell\times k}$ with $\mathbf{U}_k^\intercal \mathbf{B}\mathbf{M} = \mathbf{I}$ and $\|\mathbf{B}\mathbf{M}\|_2^2 \leq L$.*

Proof. Suppose we are given such an $\mathbf{M}$. Since span$(\mathbf{B}\mathbf{M})$ is a k-dimensional subspace of span$(\mathbf{B})$, we can let $\mathbf{Q}$ be an orthonormal basis for $\mathbf{B}\mathbf{M}$. Then we have $\mathbf{B}\mathbf{M} = \mathbf{Q}\mathbf{X}$ for some invertible $\mathbf{X} \in \mathbb{R}^{k\times k}$. Since $\mathbf{U}_k^\intercal \mathbf{Q}\mathbf{X} = \mathbf{U}_k^\intercal \mathbf{B}\mathbf{M} = \mathbf{I}$, we have $\mathbf{X}^{-1} = \mathbf{U}_k^\intercal \mathbf{Q}$. And so, we have $\|(\mathbf{U}_k^\intercal \mathbf{Q})^{-1}\|_2^2 = \|\mathbf{X}\|_2^2 = \|\mathbf{Q}\mathbf{X}\|_2^2 = \|\mathbf{B}\mathbf{M}\|_2^2 \leq L$.

In the other direction, we are given $\mathbf{Q}$ which spans a subspace of span$(\mathbf{B})$. So, for any invertible matrix $\mathbf{X} \in \mathbb{R}^{k\times k}$, we can write $\mathbf{B}\mathbf{M} = \mathbf{Q}\mathbf{X}$ for some $\mathbf{M} \in \mathbb{R}^{\ell\times k}$. If we take $\mathbf{X} = (\mathbf{U}_k^\intercal \mathbf{Q})^{-1}$, then we have $\mathbf{U}_k^\intercal \mathbf{B}\mathbf{M} = \mathbf{U}_k^\intercal \mathbf{Q}\mathbf{X} = \mathbf{I}$ and $\|\mathbf{B}\mathbf{M}\|_2^2 = \|\mathbf{Q}\mathbf{X}\|_2^2 = \|\mathbf{X}\|_2^2 = \|(\mathbf{U}_k^\intercal \mathbf{Q})^{-1}\|_2^2 \leq L$. □

With this foundation in place, we prove the following guarantee on $\mathbf{S}_r$:

Theorem 10 Restated. *Fix any PSD matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ with singular values $\sigma_1 \geq \ldots \geq \sigma_n$, $k \in [n]$, and $b \in [k]$. Let $r = k - b + 1$. Let $\mathbf{G} \in \mathbb{R}^{n\times b}$ be a matrix i.i.d. $\mathcal{N}(0, 1)$ entries, and let $\mathbf{S}_r = \begin{bmatrix}\mathbf{G} & \mathbf{A}^2\mathbf{G} & \mathbf{A}^4\mathbf{G} & \ldots & \mathbf{A}^{2(r-1)}\mathbf{G}\end{bmatrix}$. For any $\delta \in (0, 1)$, with probability at least $(1 - \delta)$, $\mathbf{S}_r$ is a (k, L)-good starting matrix for $\mathbf{A}$ for $L = \frac{cb^2k^2n\log(1/\delta)}{\delta^2 g_{min,b}^{4(k-b)}}$. Here c is a fixed constant.*

Proof. We prove the generalized (k, L)-good bound by constructing a matrix $\mathbf{M} \in \mathbb{R}^{\ell\times k}$ with $\mathbf{U}_k^\intercal \mathbf{S}_r\mathbf{M} = \mathbf{I}$ and then applying Lemma 18. Consider first the matrix $\mathbf{U}^\intercal \mathbf{S}_r\mathbf{M}$. Writing the SVD $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\intercal$, we can define $\hat{\mathbf{G}} := \mathbf{U}^\intercal \mathbf{G} \in \mathbb{R}^{n\times b}$, which is distributed as a iid $\mathcal{N}(0, 1)$ matrix because of the rotational invariance of the Gaussian. We can then write $\mathbf{U}^\intercal \mathbf{A}^{2i}\mathbf{G} = \mathbf{U}^\intercal \mathbf{U}\mathbf{\Sigma}^{2i}\mathbf{U}^\intercal \mathbf{G} = \mathbf{\Sigma}^{2i}\hat{\mathbf{G}}$. Thus,

$$\mathbf{U}^\intercal \mathbf{S}_r = \begin{bmatrix}\hat{\mathbf{G}} & \mathbf{\Sigma}^2\hat{\mathbf{G}} & \mathbf{\Sigma}^4\hat{\mathbf{G}} & \ldots & \mathbf{\Sigma}^{2(r-1)}\hat{\mathbf{G}}\end{bmatrix}.$$

If we let $\hat{\mathbf{g}}_i$ be the i^{th} column of $\hat{\mathbf{G}}$ and permute the columns of $\mathbf{U}^\intercal \mathbf{G}$, we can write $\mathbf{U}^\intercal \mathbf{S}_r$ as

$$\mathbf{U}^\intercal \mathbf{S}_r\mathbf{P} = \left[\hat{\mathbf{g}}_1\ \mathbf{\Sigma}^2\hat{\mathbf{g}}_1\ \ldots\ \mathbf{\Sigma}^{2(r-1)}\hat{\mathbf{g}}_1 \mid \hat{\mathbf{g}}_2\ \mathbf{\Sigma}^2\hat{\mathbf{g}}_2\ \ldots\ \mathbf{\Sigma}^{2(r-1)}\hat{\mathbf{g}}_2 \mid \ldots \mid \hat{\mathbf{g}}_b\ \mathbf{\Sigma}^2\hat{\mathbf{g}}_b\ \ldots\ \mathbf{\Sigma}^{2(r-1)}\hat{\mathbf{g}}_b\right].$$

where $\mathbf{P}$ is the permutation matrix that reorders the columns as such. So, for some vector $\mathbf{m}_i$, we can decompose $\mathbf{P}^\intercal \mathbf{m}_i = \begin{bmatrix}\mathbf{c}_1 \\ \ldots \\ \mathbf{c}_b\end{bmatrix}$ for $\mathbf{c}_j \in \mathbb{R}^r$, and let $p_{i,j}(t)$ be the degree $r - 1$ polynomial with coefficients $\mathbf{c}_j$. Then, we get

$$\mathbf{U}^\intercal \mathbf{S}_r\mathbf{m}_i = \mathbf{U}^\intercal \mathbf{S}_r\mathbf{P}\mathbf{P}^\intercal \mathbf{m}_i = p_{i,1}(\mathbf{\Sigma}^2)\hat{\mathbf{g}}_1 + p_{i,2}(\mathbf{\Sigma}^2)\hat{\mathbf{g}}_2 + \ldots + p_{i,b}(\mathbf{\Sigma}^2)\hat{\mathbf{g}}_b. \tag{14}$$

And so, the matrix $\mathbf{U}^\intercal \mathbf{S}_r\mathbf{M}$ is the concatenation of k such vectors $\mathbf{U}^\intercal \mathbf{S}_r\mathbf{m}_1, \ldots, \mathbf{U}^\intercal \mathbf{S}_r\mathbf{m}_k$.

Observe that $\mathbf{U}_k^\intercal \mathbf{S}_r\mathbf{M}$ is just the top k rows of $\mathbf{U}^\intercal \mathbf{S}_r\mathbf{M}$. So to apply Lemma 18, we need to find polynomials $p_{1,1}, \ldots, p_{k,b}$ which make the top k rows of $\mathbf{U}^\intercal \mathbf{S}_r\mathbf{M}$ into an identity matrix. I.e., we need to show $\mathbf{U}_k^\intercal \mathbf{S}_r\mathbf{m}_i = \mathbf{e}_i$ for $i = 1, \ldots, k$. We do this by first designing the degree $k - b$ filter polynomials $f_1, \ldots, f_k$ by

$$f_i(\sigma_j^2) = \begin{cases} 1 & j = i \\ 0 & j \notin \mathcal{N}_i \cup \{i\} \end{cases}$$

That is, f_i is 1 at σ_i^2 and is 0 on the squares of the $k-b+1$ singular values furthest from σ_i. The squares are there because the Krylov subspace uses polynomial of $\mathbf{\Sigma}^2$. We then take

$$p_{i,j}(t) := f_i(t) z_{i,j}$$

for some values of $z_{i,j}$ which will be specified later. Note these polynomials are all degree $r-1 = k-b$, which is why we constrain $r = k-b+1$. If we plug this into Equation (14), we get

$$\begin{aligned}\mathbf{U}^\intercal \mathbf{S}_r \mathbf{m}_i &= p_{i,1}(\mathbf{\Sigma}^2)\,\hat{\mathbf{g}}_1 + p_{i,2}(\mathbf{\Sigma}^2)\,\hat{\mathbf{g}}_2 + \ldots + p_{i,b}(\mathbf{\Sigma}^2)\,\hat{\mathbf{g}}_b \\ &= f_i(\mathbf{\Sigma}^2)\ (\hat{\mathbf{g}}_1 z_{i,1} + \hat{\mathbf{g}}_2 z_{i,2} + \ldots + \hat{\mathbf{g}}_b z_{i,b}) \\ &= f_i(\mathbf{\Sigma}^2)\,\hat{\mathbf{G}}\mathbf{z}_i,\end{aligned}$$

where $\mathbf{z}_i = \begin{bmatrix} z_{i,1} & \ldots & z_{i,b}\end{bmatrix}^\intercal$. So, we need to find a choice of $\mathbf{z}_i \in \mathbb{R}^b$ such that $\mathbf{U}_k^\intercal \mathbf{S}_r \mathbf{m}_i = \mathbf{e}_i$. Letting $\hat{\mathbf{G}}_k \in \mathbb{R}^{k\times b}$ be the top k rows of $\hat{\mathbf{G}}$, we write

$$\mathbf{U}_k^\intercal \mathbf{S}_r \mathbf{m}_i = f_i(\mathbf{\Sigma}_k^2)\,\hat{\mathbf{G}}_k \mathbf{z}_i.$$

Recalling that f_i is a filter polynomial, we know that $[\mathbf{U}_k^\intercal \mathbf{S}_r \mathbf{m}_i]_j = 0$ for $j \notin \mathcal{N}_i \cup \{i\}$. So, we just need to pick $\mathbf{z}_i \in \mathbb{R}^b$ such that $[\mathbf{U}_k^\intercal \mathbf{S}_r \mathbf{m}_i]_j = 0$ for $j \in \mathcal{N}_i$ and $[\mathbf{U}_k^\intercal \mathbf{S}_r \mathbf{m}_i]_i = 1$. That is, we need to pick $\mathbf{z}_i$ such that the product $\hat{\mathbf{G}}_k \mathbf{z}_i$ is zero for the $b-1$ rows in $\mathcal{N}_i$ and is 1 for row i. This is just a linear system, so we pick the unique $\mathbf{z}_i = [\hat{\mathbf{G}}_k]^{-1}_{\mathcal{N}_i\cup\{i\}}\mathbf{e}_i$, where $[\hat{\mathbf{G}}_k]_{\mathcal{N}_i\cup\{i\}}$ just selects the b rows indexed by $\mathcal{N}_i \cup \{i\}$.

Once we have these $\mathbf{z}_i$ vectors, we get $\mathbf{U}_k^\intercal \mathbf{S}_r \mathbf{m}_i = f_i(\mathbf{\Sigma}_k^2)\,\hat{\mathbf{G}}_k \mathbf{z}_i = \mathbf{e}_i$, and so $\mathbf{U}_k^\intercal \mathbf{S}_r \mathbf{M} = \mathbf{I}$, as required by Lemma 18. Now we just have to bound $\|\mathbf{S}_r\mathbf{M}\|_2^2$. We do this by bounding $\|\mathbf{S}_r\mathbf{M}\|_2^2 = \|\mathbf{U}^\intercal\mathbf{S}_r\mathbf{M}\|_2^2 \le \|\mathbf{U}^\intercal\mathbf{S}_r\mathbf{M}\|_F^2 \le \sum_{i=1}^k \|\mathbf{U}^\intercal\mathbf{S}_r\mathbf{m}_i\|_2^2$.

Let $\mathbf{\Sigma}_{-k} \in \mathbb{R}^{(n-k)\times(n-k)}$ be the bottom $n-k$ singular values of $\mathbf{A}$, and let $\hat{\mathbf{G}}_{-k} \in \mathbb{R}^{(n-k)\times b}$ bottom $n-k$ rows of $\hat{\mathbf{G}}$. Then, we can decompose

$$\mathbf{U}^\intercal \mathbf{S}_r \mathbf{m}_i = f_i(\mathbf{\Sigma}^2)\,\hat{\mathbf{G}}\mathbf{z}_i = \begin{bmatrix} f_i(\mathbf{\Sigma}_k^2)\hat{\mathbf{G}}_k\mathbf{z}_i \\ f_i(\mathbf{\Sigma}_{-k}^2)\hat{\mathbf{G}}_{-k}\mathbf{z}_i\end{bmatrix} = \begin{bmatrix} \mathbf{e}_i \\ f_i(\mathbf{\Sigma}_{-k}^2)\,\hat{\mathbf{G}}_{-k}[\hat{\mathbf{G}}_k]^{-1}_{\mathcal{N}_i\cup\{i\}}\mathbf{e}_i\end{bmatrix}.$$

We now bound the lower elements. First, we look at $\hat{\mathbf{G}}_{-k}[\hat{\mathbf{G}}_k]^{-1}_{\mathcal{N}_i\cup\{i\}}\mathbf{e}_i$, which is the product of a $(n-k)\times b$ Gaussian matrix and inverse of an independent $b\times b$ Gaussian matrix. Since these matrices are independent, we can directly bound

$$\|\hat{\mathbf{G}}_{-k}[\hat{\mathbf{G}}_k]^{-1}_{\mathcal{N}_i\cup\{i\}}\mathbf{e}_i\|_2 \le \|\hat{\mathbf{G}}_{-k}\|_2 \cdot \frac{1}{\sigma_{min}([\hat{\mathbf{G}}_k]_{\mathcal{N}_i\cup\{i\}})} \cdot \|\mathbf{e}\|_2 \le O\left(\frac{b\sqrt{kn}}{\delta}\sqrt{\log\left(\frac{1}{\delta}\right)}\right), \tag{15}$$

where we use the bound $\|\mathbf{G}_{-k}\|_2 \le O(\sqrt{(n-k)b\ln(\frac{1}{\delta})})$ from Equation (2.3) of [Rudelson and Vershynin, 2010] and the bound $\sigma_{min}([\hat{\mathbf{G}}_k]_{\mathcal{N}_i\cup\{i\}}) \ge \Omega(\frac{\delta}{k\sqrt{b}})$. This latter bound comes from a union-bound argument using the fact that any square Gaussian $\tilde{\mathbf{G}} \in \mathbb{R}^{b\times b}$ has $\Pr[\sigma_{min}(\tilde{\mathbf{G}}) \ge \Omega(\frac{\delta}{\sqrt{b}})] \ge 1-\delta$ [Huang and Tikhomirov, 2020], and that $[\hat{\mathbf{G}}_k]_{\mathcal{N}_1\cup\{1\}}, \ldots, [\hat{\mathbf{G}}_k]_{\mathcal{N}_k\cup\{k\}}$ are all square Gaussian matrices. Next, we bound $f_i(\mathbf{\Sigma}_{-k}^2)$ by writing f_i as a Lagrange interpolating polynomial:

$$f_i(t) = \prod_{j\in[k]\setminus\mathcal{N}_i, j\neq i} \frac{t-\sigma_j^2}{\sigma_i^2-\sigma_j^2}.$$

We can then use our gap assumption to bound f_i on all singular values below σ_k. That is, for $0 \le t \le \sigma_k$ we have

$$|f_i(t)| \le \prod_{j\in[k]\setminus\mathcal{N}_i, j\neq i} \left|\frac{\sigma_j^2}{\sigma_i^2-\sigma_j^2}\right| \le \prod_{j\in[k]\setminus\mathcal{N}_i, j\neq i} \left|\frac{\sigma_j}{\sigma_i-\sigma_j}\right|^2 \le \frac{1}{g_{min,b}^{2(k-b)}}.$$

Putting this altogether, we then find that

$$\begin{aligned}\|f_i(\mathbf{\Sigma}_{-k}^2)\,\hat{\mathbf{G}}_{-k}[\hat{\mathbf{G}}_k]^{-1}_{\mathcal{N}_i\cup\{i\}}\mathbf{e}_i\|_2^2 &\le \|f_i(\mathbf{\Sigma}_{-k}^2)\|_2^2 \;\; \|\hat{\mathbf{G}}_{-k}[\hat{\mathbf{G}}_k]^{-1}_{\mathcal{N}_i\cup\{i\}}\mathbf{e}_i\|_2^2 \\ &\le O\left(\frac{b^2kn}{g_{min,b}^{4(k-b)}\delta^2}\log\left(\frac{1}{\delta}\right)\right).\end{aligned}$$

And therefore

$$\|\mathbf{S}_r\mathbf{M}\|_2^2 \le \|\mathbf{U}^\intercal\mathbf{S}_r\mathbf{M}\|_F^2 \le O\left(1+\frac{b^2k^2n}{g_{min,b}^{4(k-b)}\delta^2}\log\left(\frac{1}{\delta}\right)\right),$$

which completes the proof by Lemma 18. □

F Convergence Results from [Musco and Musco, 2015]

In this section, we show how Imported Theorem 2 and Imported Theorem 5 follow from [Musco and Musco, 2015]. We prove the result using a generalization of a (k, L)-good starting block (Definition 1), which is useful in the small block analysis of Appendix E:

Definition 3 Restated ((k, L)-good Starting Matrix (Generalized))**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be a matrix with top k left singular vectors $\mathbf{U}_k \in \mathbb{R}^{n\times k}$. A matrix $\mathbf{B} \in \mathbb{R}^{n\times\ell}$ is a (k, L)-good starting matrix for $\mathbf{A}$ if $\|(\mathbf{U}_k^\intercal\mathbf{Q})^{-1}\|_2^2 \le L$ for some orthonormal $\mathbf{Q} \in \mathbb{R}^{n\times k}$ that lies in* span($\mathbf{B}$).

Note that if $\mathbf{B}$ has exactly k columns then this generalized definition exactly matches Definition 1. This is the case for the analysis of single vector Krylov in Section 3, where we take $\mathbf{B} = \mathbf{S}_k$ with exactly k columns and show that $\mathbf{S}_k$ is (k, L)-good. In Appendix E, we show an equivalent formulation which is easier to use in our analysis:

Lemma 18 Restated. *$\mathbf{B} \in \mathbb{R}^{n\times\ell}$ is (k, L)-good for $\mathbf{A}$ if and only if there exists a matrix $\mathbf{M} \in \mathbb{R}^{\ell\times k}$ with $\mathbf{U}_k^\intercal\mathbf{B}\mathbf{M} = \mathbf{I}$ and $\|\mathbf{B}\mathbf{M}\|_2^2 \le L$.*

Given this definition, we will show that Imported Theorem 2 and Imported Theorem 5 follow from [Musco and Musco, 2015]. Before diving in, we first state a guarantee on the polynomials used in [Musco and Musco, 2015]. They use polynomials of the form $p(x) := \frac{(1+\gamma)\alpha}{T_q(1+\gamma)}T_q(\frac{x}{\alpha})$ for some $\alpha > 0$, $\gamma \in (0, 1)$, and $q \in \mathbb{N}$, and where $T_q(x)$ is the degree q Chebyshev polynomial of the first kind. We first show that such polynomials are monotonic for large enough x:

Lemma 19. *Let $p(x) = \frac{(1+\gamma)\alpha}{T_q(1+\gamma)}T_q(\frac{x}{\alpha})$ where $\alpha > 0$ and $\gamma \in (0, 1)$. Then, $\max_{x\in[0,\alpha]} p(x) = p(\alpha)$ and $p(x)$ is monotonically increasing on (α, ∞).*

Proof. The result follows from two well-known properties of T_q: that T_q is monotonically increasing on $(1, \infty)$ and that $\max_{t\in[0,1]} T_q(t) = T_q(1) = 1$. Since $\gamma > 0$, we know that $T_q(1 + \gamma) > T_q(1) = 1$, so we have $\frac{(1+\gamma)\alpha}{T_q(1+\gamma)} > 0$. Therefore, we get that $p(x)$ is monotonically increasing on (α, ∞), and that $\max_{x\in[0,\alpha]} p(x) = p(\alpha)$. □

In order to relate the generalized (k, L)-good definition to the convergence analysis of [Musco and Musco, 2015], we rely on a slight generalization of a theorem used in the appendix of [Musco and Musco, 2015]:

Imported Lemma 20 (Lemma 48 of [Woodruff, 2014]). *Let $\mathbf{P} = \mathbf{P}\mathbf{U}\mathbf{U}^\intercal + \mathbf{E} \in \mathbb{R}^{n\times n}$ be a low-rank factorization of $\mathbf{A} \in \mathbb{R}^{n\times n}$, with $\mathbf{U} \in \mathbb{R}^{n\times k}$ and $\mathbf{U}^\intercal\mathbf{U} = \mathbf{I}_k$. Let $\mathbf{B} \in \mathbb{R}^{n\times \ell}$ $(\ell \geq k)$ be any matrix with $\text{rank}(\mathbf{U}^\intercal\mathbf{S}) = \text{rank}(\mathbf{U}) = k$. Let $\mathbf{M} \in \mathbb{R}^{\ell\times k}$ with $\mathbf{U}^\intercal\mathbf{B}\mathbf{M} = \mathbf{I}_k$. Let $\mathbf{C} = \mathbf{A}\mathbf{S} \in \mathbb{R}^{n\times \ell}$. Then,*

$$\|\mathbf{P} - \Pi_{\mathbf{C},k}(\mathbf{P})\|_F^2 \leq \|\mathbf{E}\|_F^2 + \|\mathbf{E}\mathbf{B}\mathbf{M}\|_F^2.$$

Here, $\Pi_{\mathbf{C},k}(\mathbf{P}) = \mathbf{Y}_{opt}\mathbf{Y}_{opt}^\intercal\mathbf{A} \in \mathbb{R}^{n\times n}$ is the best rank k approximation to $\mathbf{A}$ in the column space of $\mathbf{C}$ So, $\mathbf{Y}_{opt} \in \mathbb{R}^{n\times k}$ is an orthogonal matrix that lies in the column span of $\mathbf{C}$.

Lemma 48 from [Woodruff, 2014] is stated with $\mathbf{M} = (\mathbf{U}^\intercal\mathbf{B})^+$, but the proof only requires that $\mathbf{U}^\intercal\mathbf{B}\mathbf{M} = \mathbf{I}$. We now prove the imported theorems:

Imported Theorem 2 Restated (Theorem 1 of [Musco and Musco, 2015]). *Let $\mathbf{B} \in \mathbb{R}^{n\times \ell}$ be any (k, L)-good starting matrix (Definition 3) matrix for $\mathbf{A}$. If we run Block Krylov iteration (Algorithm 2) for $q = O(\frac{1}{\sqrt{\varepsilon}}\log(\frac{nL}{\varepsilon}))$ iterations with starting block $\mathbf{B}$, then the output $\mathbf{Q} \in \mathbb{R}^{n\times k}$ satisfies*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_\xi \leq (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_\xi \qquad \text{and} \qquad \left|\mathbf{q}_i^\intercal\mathbf{A}\mathbf{A}^\intercal\mathbf{q}_i - \sigma_i(\mathbf{A})^2\right| \leq \varepsilon\sigma_{k+1}(\mathbf{A})^2.$$

Proof. To recover this guarantee from [Musco and Musco, 2015], we recover Properties 1-3 of their Lemma 9. We start by recovering Property 1. We let $p_1(x) = \frac{(1+\gamma)\alpha}{T_q(1+\gamma)}T_q(\frac{x}{\alpha})$ where $\alpha = \sigma_{k+1}(\mathbf{A})$, $\gamma = \frac{\varepsilon}{2}$, and $q = O(\frac{1}{\sqrt{\varepsilon}}\log(\frac{nL}{\varepsilon}))$. Let $\mathbf{M} \in \mathbb{R}^{\ell\times k}$ be the matrix guaranteed to exist by Lemma 18. We then instantiate Imported Lemma 20 with $\mathbf{P} = p_1(\mathbf{A})$, $\mathbf{B} = \mathbf{B}$, $\mathbf{M} = \mathbf{M}$, and $\mathbf{U} = \mathbf{U}_k$ being the top k singular vectors of $\mathbf{A}$. By Lemma 19, since $\alpha = \sigma_{k+1}(\mathbf{A})$, we know that the top k singular vectors of $\mathbf{A}$ are also the top k singular vectors of $p_1(\mathbf{A})$, and therefore that $\mathbf{P}\mathbf{U}\mathbf{U}^\intercal = p_1(\mathbf{A})_k$. We then get that some orthonormal $\mathbf{Y}_1 \in \mathbb{R}^{n\times k}$ in the span of $p(\mathbf{A})\mathbf{B}$ has

$$\begin{aligned}
\|p_1(\mathbf{A}) - \mathbf{Y}_1\mathbf{Y}_1^\intercal p_1(\mathbf{A})\|_F^2 &\leq \|p_1(\mathbf{A}) - p_1(\mathbf{A})_k\|_F^2 + \|(p_1(\mathbf{A}) - p_1(\mathbf{A})_k)\mathbf{B}\mathbf{M}\|_F^2 \\
&\leq \|p_1(\mathbf{A}) - p_1(\mathbf{A})_k\|_F^2 + \|p_1(\mathbf{A}) - p_1(\mathbf{A})_k\|_F^2\|\mathbf{B}\mathbf{M}\|_2^2 \\
&\leq (L+1)\|p_1(\mathbf{A}) - p_1(\mathbf{A})_k\|_F^2 \\
&= (L+1)\sum_{i=k+1}^{n} p(\sigma_i(\mathbf{A}))^2 \\
&\leq (L+1)n \cdot \frac{4\sigma_{k+1}(\mathbf{A})^2}{2^{2q\sqrt{\gamma}}} \qquad \text{(Lemma 5 of [Musco and Musco, 2015])} \\
&\leq \frac{\varepsilon}{2}\sigma_{k+1}(\mathbf{A})^2, \qquad (16)
\end{aligned}$$

where the last line uses $q \geq \frac{1}{2\sqrt{\gamma}}\log_2(\frac{8(L+1)n}{\varepsilon}) = O(\frac{1}{\sqrt{\varepsilon}}\log(\frac{Ln}{\varepsilon}))$. The inequality $\|p_1(\mathbf{A}) - \mathbf{Y}_1\mathbf{Y}_1^\intercal p_1(\mathbf{A})\| \leq \frac{\varepsilon}{2}\sigma_{k+1}(\mathbf{A})^2$ is exactly Equation (5) on page 11 of [Musco and Musco, 2015]. Once Equation (5) is achieved, the rest of the proof of Property 1 then follows without any alteration.

We next move onto proving properties 2 and 3. The proofs of properties 2 and 3 in [Musco and Musco, 2015] both involve defining the matrix $\mathbf{A}_{outer}$. If $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\intercal$ is the SVD of $\mathbf{A}$, then $\mathbf{A}_{outer} := \mathbf{U}\mathbf{\Sigma}_{outer}\mathbf{U}^\intercal$ where $\mathbf{\Sigma}_{outer}$ contains all the singular values of $\mathbf{A}$ with either $\sigma_i(\mathbf{A}) \geq \sigma_k(\mathbf{A})$ or $\sigma_i(\mathbf{A}) < \frac{1}{1+\varepsilon/2}\sigma_k(\mathbf{A})$. All the other singular values are set to equal zero. Crucially, the top k singular vectors of $\mathbf{A}$ are still the top k singular vectors of $\mathbf{A}_{outer}$.

We let $p_2(x) = \frac{(1+\gamma)\alpha}{T_q(1+\gamma)} T_q(\frac{x}{\alpha})$ where $\alpha = \frac{1}{1+\varepsilon/2}\sigma_{k+1}(\mathbf{A})$, $\gamma = \frac{\varepsilon}{2}$, and $q = O(\frac{1}{\sqrt{\varepsilon}}\log(\frac{nL}{\varepsilon}))$. We next instantiate Imported Lemma 20 with $\mathbf{P} = p_2(\mathbf{A}_{outer})$, $\mathbf{B} = \mathbf{B}$, $\mathbf{M} = \mathbf{M}$, and $\mathbf{U} = \mathbf{U}_k$. By Lemma 19, since $\alpha \leq \sigma_k(\mathbf{A}_{outer})$, we know that the top k singular vectors of $\mathbf{A}_{outer}$ are also the top k singular vectors of $p_2(\mathbf{A}_{outer})$, and therefore that $\mathbf{PUU}^\intercal = p_2(\mathbf{A}_{outer})_k$. We then let $\mathbf{Y}_{outer} \in \mathbb{R}^{n\times k}$ be the orthogonal basis that constructs $\Pi_{\mathbf{C},k}(\mathbf{A})$, so that

$$\begin{aligned}\|p_2(\mathbf{A}_{outer}) - \mathbf{Y}_{outer}\mathbf{Y}_{outer}^\intercal p_2(\mathbf{A}_{outer})\|_F^2 &\leq \|p_2(\mathbf{A}_{outer}) - p_2(\mathbf{A}_{outer})_k\|_F^2 \\ &\quad + \|(p_2(\mathbf{A}_{outer}) - p_2(\mathbf{A}_{outer})_k)\mathbf{BM}\|_F^2 \\ &\leq (L+1)\|p_2(\mathbf{A}_{outer}) - p_2(\mathbf{A}_{outer})\|_F^2 \\ &\leq \frac{\varepsilon}{2}\sigma_{k+1}(\mathbf{A}_{outer})^2.\end{aligned}$$

Where the inequalities follow from the same logic as earlier, for Equation (16). We again find ourselves at Equation (5) of [Musco and Musco, 2015], and follow the rest of the proof on page 11 to get the following guarantee:

$$\|(\mathbf{A}_{outer})_k\|_F^2 - \|\mathbf{Y}_{outer}\mathbf{Y}_{outer}^\intercal(\mathbf{A}_{outer})_k\|_F^2 \leq \frac{\varepsilon}{2}\sigma_{k+1}(\mathbf{A}_{outer})^2.$$

Since $(\mathbf{A}_{outer})_k = \mathbf{A}_k$, the above inequality then recovers Equation (10) of [Musco and Musco, 2015]. Given this proof of Equation (10), the rest of the proof of Property 2 holds as written. Equation (10) is also used on page 14 of [Musco and Musco, 2015] to prove Property 3. The rest of the proof of Property 3 also holds without alteration given Equation (10).

Overall, we have recovered the proofs of Properties 1-3 of Lemma 9 of [Musco and Musco, 2015]. That is, we have shown that Lemma 9 holds for block Krylov iteration starting from (k, L)-good starting block $\mathbf{B}$ with $q = O(\frac{1}{\sqrt{\varepsilon}}\log(\frac{nL}{\varepsilon}))$ iterations. The proofs in Section 6.2 of [Musco and Musco, 2015] then show that Properties 1-3 suffice to achieve the spectral, Frobenius, and singular value guarantees. □

We now move onto Imported Theorem 5 from [Musco and Musco, 2015], which depends on spectral decay. Note there are two different variables ℓ and ℓ_0 in this context. In order to perform rank-k approximation, we want to recover a convergence bound in terms of $g_{k\to\ell}$ for some $\ell \geq k$. Our starting block $\mathbf{B} \in \mathbb{R}^{n\times\ell_0}$ has $\ell_0 \geq \ell$ columns and is an (ℓ, L)-good starting block. We need to consider this case because, when running block size $b \leq k$ Krylov iteration, the analysis in Appendix E considers a simulated starting block $\mathbf{S}_r$ which uses $\ell_0 \approx b(\ell - b) \geq \ell$ columns to simulate block size ℓ Krylov iteration.

Imported Theorem 5 Restated (Theorem 13 of [Musco and Musco, 2015]). *Let $\mathbf{B} \in \mathbb{R}^{n\times\ell_0}$ be any (ℓ, L)-good starting matrix (Definition 3) matrix for $\mathbf{A}$, for some $\ell \geq k$. If we run Block Krylov iteration (Algorithm 2) for $q = O(\frac{1}{\sqrt{g_{k\to\ell}}}\log(\frac{nL}{\varepsilon}))$ iterations with starting block $\mathbf{B}$, where $g_{k\to\ell} = \frac{\sigma_k - \sigma_{\ell+1}}{\sigma_k}$, then the output $\mathbf{Q} \in \mathbb{R}^{n\times k}$ satisfies*

$$\|\mathbf{A} - \mathbf{QQ}^\intercal\mathbf{A}\|_\xi \leq (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_\xi \qquad \text{and} \qquad \left|\mathbf{q}_i^\intercal\mathbf{AA}^\intercal\mathbf{q}_i - \sigma_i(\mathbf{A})^2\right| \leq \varepsilon\sigma_{k+1}(\mathbf{A})^2.$$

Proof. We again show that Properties 1-3 of Lemma 9 of [Musco and Musco, 2015] hold, but now with $q = \frac{1}{\sqrt{g_{k\to\ell}}}\log(\frac{nL}{\varepsilon})$. Following Section 7 of [Musco and Musco, 2015], we will prove Property 1 for all $l \in [k]$, which in turn implies that Properties 2 and 3 hold. To begin, let $p_3(x) = \frac{(1+\gamma)\alpha}{T_q(1+\gamma)} T_q(\frac{x}{\alpha})$ where $\alpha = \sigma_{\ell+1}(\mathbf{A})$, $\gamma = g_{k\to\ell}$, and $q = O(\frac{1}{\sqrt{g_{k\to\ell}}}\log(\frac{nL}{\varepsilon}))$. We let $\mathbf{M} \in \mathbb{R}^{\ell_0\times\ell}$ be the matrix guaranteed to exist by Lemma 18. Then, noticing that the top ℓ singular vectors of $\mathbf{A}$ are also

the top ℓ eigenvectors of $p_3(\mathbf{A})$, we again appeal to Imported Lemma 20 with $\mathbf{P} = p_3(\mathbf{A})$, $\mathbf{B} = \mathbf{B}$, $\mathbf{M} = \mathbf{M}$, and $\mathbf{U} = \mathbf{U}_k$. We get that some orthogonal $\mathbf{Y}_3 \in \mathbb{R}^{n\times\ell}$ in the span of $p_3(\mathbf{A})\mathbf{B}$ has:

$$\begin{aligned}
\|p_3(\mathbf{A}) - \mathbf{Y}_3\mathbf{Y}_3^\intercal p_3(\mathbf{A})\|_F^2 &\leq \|p_3(\mathbf{A}) - p_3(\mathbf{A})_\ell\|_F^2 + \|(p_3(\mathbf{A}) - p_3(\mathbf{A})_\ell)\mathbf{B}\mathbf{M}\|_F^2 \\
&\leq \|p_3(\mathbf{A}) - p_3(\mathbf{A})_\ell\|_F^2 + \|p_3(\mathbf{A}) - p_3(\mathbf{A})_\ell\|_F^2\|\mathbf{B}\mathbf{M}\|_2^2 \\
&\leq (L+1)\|p_3(\mathbf{A}) - p_3(\mathbf{A})_\ell\|_F^2 \\
&= (L+1)\sum_{i=\ell+1}^{n} p(\sigma_i(\mathbf{A}))^2 \\
&\leq (L+1)n \cdot \frac{4\sigma_{\ell+1}(\mathbf{A})^2}{2^{2q\sqrt{\gamma}}} \qquad \text{(Lemma 5 of [Musco and Musco, 2015])} \\
&\leq \frac{\varepsilon}{2}\sigma_{\ell+1}(\mathbf{A})^2,
\end{aligned}$$

This recovers Equation (5) on page 10 of [Musco and Musco, 2015], and from there the rest of the proof of Property 1 holds. As discussed on page 16 of [Musco and Musco, 2015], Property 1 holds here for all $l \in [k]$, and therefore Properties 2 and 3 also hold. Then, the analysis in Section 6.2 shows how Properties 1-3 imply the spectral, Frobenius, and singular value guarantees. □

G Single Vector Simultaneous Iteration

We briefly present a single vector algorithm for low-rank approximation based on the standard simultaneous iteration algorithm. Simultaneous iteration, or block power method, extracts a low-rank approximation from the span of $\mathbf{K} = \mathbf{A}^{2q}\mathbf{B}$ where $\mathbf{B}$ is a starting block. We present a prototypical pseudocode for simultaneous iteration in Algorithm 3. [Musco and Musco, 2015] show that Algorithm 3 converges from any (k, L)-good starting block $\mathbf{B}$, giving the two following theorems, which follow from the same arguments as in Appendix F and [Musco and Musco, 2015]:

Algorithm 3 Simultaneous Iteration for Low-Rank Approximation

input: Matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$. Target rank k. Starting block $\mathbf{B} \in \mathbb{R}^{n\times\ell}$. Number of iterations t.
output: Orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{n\times k}$.

1: Compute an orthonormal basis $\mathbf{Z}$ for $\mathbf{K} = (\mathbf{A}\mathbf{A}^\intercal)^t\mathbf{B}$.
2: Compute $\mathbf{U}_k$, the k top eigenvectors of $\mathbf{M} = \mathbf{Z}^\intercal\mathbf{A}\mathbf{A}^\intercal\mathbf{Z}$
3: **return** $\mathbf{Q} = \mathbf{Z}\mathbf{U}_k$.

Algorithm 4 Single Vector Simultaneous Iteration for Low-Rank Approximation

input: Matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$. Target rank k. Starting vector $\mathbf{x} \in \mathbb{R}^n$. Number of iterations t. Memory budget $\ell \geq k$.
output: Orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{n\times k}$.

1: Compute an orthonormal basis $\mathbf{Z}$ for $\mathbf{K} = [\,(\mathbf{A}\mathbf{A}^\intercal)^{t-\ell+1}\mathbf{x},\ (\mathbf{A}\mathbf{A}^\intercal)^{t-\ell+2}\mathbf{x},\ \ldots,\ (\mathbf{A}\mathbf{A}^\intercal)^t\mathbf{x}\,]$.
2: Compute $\mathbf{U}_k$, the k top eigenvectors of $\mathbf{M} = \mathbf{Z}^\intercal\mathbf{A}\mathbf{A}^\intercal\mathbf{Z}$
3: **return** $\mathbf{Q} = \mathbf{Z}\mathbf{U}_k$.

Imported Theorem 21 (Theorem 1 of [Musco and Musco, 2015]). *Let $\mathbf{B} \in \mathbb{R}^{n\times\ell}$ be any (k, L)-good starting matrix (Definition 3) matrix for $\mathbf{A}$. If we run Simultaneous Iteration (Algorithm 3)*

for $q = O(\frac{1}{\varepsilon}\log(\frac{nL}{\varepsilon}))$ *iterations with starting block* $\mathbf{B}$*, then the output* $\mathbf{Q} \in \mathbb{R}^{n\times k}$ *satisfies*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_\xi \leq (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_\xi \qquad \textit{and} \qquad \left|\mathbf{q}_i^\intercal\mathbf{A}\mathbf{A}^\intercal\mathbf{q}_i - \sigma_i(\mathbf{A})^2\right| \leq \varepsilon\sigma_{k+1}(\mathbf{A})^2.$$

Imported Theorem 22 (Theorem 13 of [Musco and Musco, 2015])**.** *Let* $\mathbf{B} \in \mathbb{R}^{n\times\ell_0}$ *be any* (ℓ, L)*-good starting matrix (Definition 3) matrix for* $\mathbf{A}$*, for some* $\ell \geq k$*. If we run Simultaneous Iteration (Algorithm 3) for* $q = O(\frac{1}{g_{k\to\ell}}\log(\frac{nL}{\varepsilon}))$ *iterations with starting block* $\mathbf{B}$*, where* $g_{k\to\ell} = \frac{\sigma_k - \sigma_{\ell+1}}{\sigma_k}$*, then the output* $\mathbf{Q} \in \mathbb{R}^{n\times k}$ *satisfies*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_\xi \leq (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_\xi \qquad \textit{and} \qquad \left|\mathbf{q}_i^\intercal\mathbf{A}\mathbf{A}^\intercal\mathbf{q}_i - \sigma_i(\mathbf{A})^2\right| \leq \varepsilon\sigma_{k+1}(\mathbf{A})^2.$$

When compared to block Krylov Iteration, subspace iteration uses less memory, but converges slower. Specifically, in the theorems above we obtain a dependence on $1/\varepsilon$ and $1/g_{k\to\ell}$ in comparison to $1/\sqrt{\varepsilon}$ and $1/\sqrt{g_{k\to\ell}}$ in the comparable Imported Theorem 2 and Imported Theorem 5 for block Krylov iteration.

In Algorithm 4, we present a "single vector" variant of simultaneous iteration that similarly saves memory over the single vector Krylov method from Algorithm 1. When run for t iterations, instead of storing the entire length t Krylov subspace as in Algorithm 1, the method only stores the last $\ell \geq k$ columns for a specified memory budget ℓ.

Taking $\ell = k$, we can analyze Algorithm 4 by letting $\mathbf{S}_k = \begin{bmatrix}\mathbf{x} & \mathbf{A}^2\mathbf{x} & \mathbf{A}^4\mathbf{x} & \ldots & \mathbf{A}^{2(k-1)}\mathbf{x}\end{bmatrix}$ and noticing that Algorithm 3 run with starting matrix $\mathbf{S}_k$ for q iterations produces the matrix $\mathbf{K} = \mathbf{A}^{2q}\mathbf{S}_k = \begin{bmatrix}\mathbf{A}^{2q}\mathbf{x} & \mathbf{A}^{2(q+1)}\mathbf{x} & \ldots & \mathbf{A}^{2(q+k)}\mathbf{x}\end{bmatrix}$. Since Theorem 3 already tells us that $\mathbf{S}_k$ is (k, L)-good for $\mathbf{A}$, applying Imported Theorem 21 immediately gives the following convergence guarantees for Algorithm 4:

Theorem 23. *For* $\mathbf{A} \in \mathbb{R}^{n\times d}$*, let* $g_{min} = \min_{i\in\{1,\ldots,k-1\}}\frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+1}}$*. For any* $\varepsilon, \delta \in (0,1)$*, Algorithm 4 initialized with* $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ *and run for* $t = O(\frac{k}{\varepsilon}\log(\frac{1}{g_{min}}) + \frac{1}{\varepsilon}\log(\frac{n}{\varepsilon\delta}))$ *iterations with memory budget* k *returns an orthogonal* $\mathbf{Q} \in \mathbb{R}^{n\times k}$ *such that, with probability at least* $1-\delta$*,*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_\xi \leq (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_\xi \qquad \textit{and} \qquad \left|\mathbf{q}_i^\intercal\mathbf{A}\mathbf{A}^\intercal\mathbf{q}_i - \sigma_i(\mathbf{A})^2\right| \leq \varepsilon\sigma_{k+1}(\mathbf{A})^2.$$

Additionally, if Algorithm 4 is run with memory budget $\ell \geq k$, we can obtain a spectrum dependent convergence bound comparable to the result proven in Section 4.1 for single vector Krylov iteration. Specifically, since Theorem 3 already tells us that $\mathbf{S}_\ell$ is (ℓ, L)-good for $\mathbf{A}$, applying Imported Theorem 22 immediately gives the following convergence guarantees for Algorithm 4:

Theorem 24. *For* $\mathbf{A} \in \mathbb{R}^{n\times d}$ *and* $\ell \geq k$*, let* $g_{min} = \min_{i\in\{1,\ldots,\ell-1\}}\frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+1}}$ *and* $g_{k\to\ell} = \frac{\sigma_k - \sigma_{\ell+1}}{\sigma_k}$*. For any* $\varepsilon, \delta \in (0,1)$*, Algorithm 4 initialized with* $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ *and run for* $t = O(\frac{\ell}{g_{k\to\ell}}\log(\frac{1}{g_{\min}}) + \frac{1}{g_{k\to\ell}}\log(\frac{n}{\delta\varepsilon}))$ *iterations with memory budget* ℓ *returns an orthogonal* $\mathbf{Q} \in \mathbb{R}^{n\times k}$ *such that, with probability at least* $1-\delta$*,*

$$\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\intercal\mathbf{A}\|_\xi \leq (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_\xi \qquad \textit{and} \qquad \left|\mathbf{q}_i^\intercal\mathbf{A}\mathbf{A}^\intercal\mathbf{q}_i - \sigma_i(\mathbf{A})^2\right| \leq \varepsilon\sigma_{k+1}(\mathbf{A})^2.$$