Low-Rank Approximation with $1/\epsilon^{1/3}$ Matrix-Vector Products

Ainesh Bakshi
Carnegie Mellon University
Pittsburgh, USA
abakshi@cs.cmu.edu

Kenneth L. Clarkson
IBM
Almaden, USA
klclarks@us.ibm.com

David P. Woodruff
Carnegie Mellon University
Pittsburgh, USA
dwoodruf@cs.cmu.edu

ABSTRACT

We study iterative methods based on Krylov subspaces for low-rank approximation under any Schatten-p norm. Here, given access to a matrix $\mathbf{A}$ through matrix-vector products, an accuracy parameter ϵ, and a target rank k, the goal is to find a rank-k matrix $\mathbf{Z}$ with orthonormal columns such that $\left\|\mathbf{A}\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p} \leq (1+\epsilon)\min_{\mathbf{U}^\top\mathbf{U}=\mathbf{I}_k}\left\|\mathbf{A}\left(\mathbf{I}-\mathbf{U}\mathbf{U}^\top\right)\right\|_{\mathcal{S}_p}$, where $\|\mathbf{M}\|_{\mathcal{S}_p}$ denotes the ℓ_p norm of the the singular values of $\mathbf{M}$. For the special cases of $p=2$ (Frobenius norm) and $p=\infty$ (Spectral norm), Musco and Musco (NeurIPS 2015) obtained an algorithm based on Krylov methods that uses $\tilde{O}(k/\sqrt{\epsilon})$ matrix-vector products, improving on the naïve $\tilde{O}(k/\epsilon)$ dependence obtainable by the power method, where $\tilde{O}(\cdot)$ suppresses $\mathrm{poly}(\log(dk/\epsilon))$ factors.

Our main result is an algorithm that uses only $\tilde{O}(kp^{1/6}/\epsilon^{1/3})$ matrix-vector products, and works for *all*, not necessarily constant, $p \geq 1$. For $p=2$ our bound improves the previous $\tilde{O}(k/\epsilon^{1/2})$ bound to $\tilde{O}(k/\epsilon^{1/3})$. Since the Schatten-$p$ and Schatten-∞ norms of any matrix are the same up to a $1+\epsilon$ factor when $p \geq (\log d)/\epsilon$, our bound recovers the result of Musco and Musco for $p=\infty$. Further, we prove a matrix-vector query lower bound of $\Omega(1/\epsilon^{1/3})$ for *any* fixed constant $p \geq 1$, showing that surprisingly $\tilde{\Theta}(1/\epsilon^{1/3})$ is the optimal complexity for constant k.

To obtain our results, we introduce several new techniques, including optimizing over *multiple Krylov subspaces* simultaneously, and *pinching inequalities* for partitioned operators. Our lower bound for $p \in [1,2]$ uses the *Araki-Lieb-Thirring* trace inequality, whereas for $p>2$, we appeal to a *norm-compression* inequality for *aligned partitioned operators*. As our algorithms only require matrix-vector product access, they can be applied in settings where alternative techniques such as sketching cannot, e.g., to covariance matrices, Hessians defined implicitly by a neural network, and arbitrary polynomials of a matrix.

CCS CONCEPTS

• **Theory of computation** → **Mathematical optimization**; **Machine learning theory**.

KEYWORDS

Low-rank Approximation, Krylov Methods, Matrix-Vector Product Model, Schatten norms

STOC '22, June 20–24, 2022, Rome, Italy

ACM ISBN 978-1-4503-9264-8/22/06.
https://doi.org/10.1145/3519935.3519988

ACM Reference Format:
Ainesh Bakshi, Kenneth L. Clarkson, and David P. Woodruff. 2022. Low-Rank Approximation with $1/\epsilon^{1/3}$ Matrix-Vector Products. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing (STOC '22), June 20–24, 2022, Rome, Italy.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3519935.3519988

1 INTRODUCTION

Iterative methods, and in particular Krylov subspace methods, are ubiquitous in scientific computing. Algorithms such as power iteration, Golub-Kahan Bidiagonalization, Arnoldi iteration, and the Lanczos iteration, are used in basic subroutines for matrix inversion, solving linear systems, linear programming, low-rank approximation, and numerous other fundamental linear algebra primitives [24, 42]. A common technique in the analysis of Krylov methods is the use of Chebyshev polynomials, which can be applied to the singular values of a matrix to implement an approximate interval or step function [27, 40]. Further, Chebyshev polynomials reduce the degree required to accurately approximate such functions, leading to significantly fewer iterations and faster running time. In this paper we investigate the power of Krylov methods for low-rank approximation in the matrix-vector product model.

The Matrix-Vector Product Model. In this model, there is an underlying matrix $\mathbf{A}$, which is often implicit, and for which the only access to $\mathbf{A}$ is via matrix-vector products. Namely, the algorithm chooses a query vector v^1, obtains the product $\mathbf{A}\cdot v^1$, chooses the next query vector v^2, which is any randomized function of v^1 and $\mathbf{A}\cdot v^1$, then receives $\mathbf{A}\cdot v^2$, and so on. If $\mathbf{A}$ is a non-symmetric matrix, we assume access to products of the form $\mathbf{A}^\top v$ as well. We refer to the minimal number q of queries needed by the algorithm to solve a problem with constant probability as the *query complexity*. We note that upper bounds on the query complexity immediately translate to running time bounds for the RAM model, when $\mathbf{A}$ is explicit, since a matrix-vector product can be implemented in $\mathsf{nnz}(\mathbf{A})$ time, i.e., the number of non-zero entries in the matrix. Since this model captures a large family of iterative methods, it is natural to ask whether Krylov subspace based methods yield optimal algorithms, where the complexity measure of interest is the number of matrix-vector products.

This model and related vector-matrix-vector query models were formalized for a number of problems in [38, 45], though the model is standard for measuring efficiency in scientific computing and numerical linear algebra, see, e.g., [6]; in that literature, methods that use only matrix-vector products are called *matrix-free*. Subsequently, for the problem of estimating the top eigenvector, nearly tight bounds were obtained in [9, 44]. Also, for the problem of estimating the trace of a positive semidefinite matrix, tight bounds were obtained in [29] (see, also [51], where tight bounds were shown in the related vector-matrix-vector query model). For recovering a planted clique from a random graph, upper and lower bounds were obtained in [37]. In the non-adaptive setting, where $v^1, \ldots, v^q$, are chosen before making any queries to $\mathbf{A}$, this is equivalent to the *sketching model*, which is thoroughly studied on its own (see, e.g., [34, 52]), and in the context of data streams [22, 32].

Why is the matrix A implicit? A small query complexity q leads to an algorithm running in time $O(T(\mathbf{A}) \cdot q + P(n, d, q))$, where $T(\mathbf{A})$ is the time to multiply the $n \times d$ matrix $\mathbf{A}$ by an arbitrary vector, and $P(n, d, q)$ is the time needed to form the queries and process the query responses, which is typically small. When the matrix $\mathbf{A}$ is given as a list of $\mathsf{nnz}(\mathbf{A})$ non-zero entries, then $T(\mathbf{A}) \leq \mathsf{nnz}(\mathbf{A})$. However, in many problems $\mathbf{A}$ is not given explicitly, and it is too expensive to write $\mathbf{A}$ down. Indeed, one may be given $\mathbf{A}$ but want to compute a low-rank approximation to the "covariance" (Gram) matrix $\mathbf{A}^\top\mathbf{A}$, and computing $\mathbf{A}^\top\mathbf{A}$ is too slow [31]. More generally, one may be given $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ and a function $f : \mathbb{R} \to \mathbb{R}$, and want to compute matrix-vector products with the generalized matrix function $f(\mathbf{A}) = \mathbf{U}f(\mathbf{\Sigma})\mathbf{V}^\top$, where $\mathbf{U}$ has orthonormal columns, $\mathbf{V}^\top$ has orthonormal rows, $\mathbf{\Sigma}$ is a diagonal matrix, and f is applied entry-wise to each entry on the diagonal.

The covariance matrix corresponds to $f(x) = x^2$, and other common functions f include the matrix exponential $f(x) = e^x$ and low-degree polynomials. For instance, when $\mathbf{A}$ is the adjacency matrix of an undirected graph, $f(x) = x^3/6$ is used to count the number of triangles [4, 49]. Yet another example is when $\mathbf{A}$ is the Hessian $\mathbf{H}$ of a neural network with a huge number of parameters, for which it is often impossible to compute or store the entire Hessian [16]. Typically $\mathbf{H} \cdot v$, for any chosen vector v, is computed using Pearlmutter's trick [35]. However, even with Pearlmutter's trick and distributed computation on modern GPUs, it takes 20 hours to compute the eigendensity of a single Hessian $\mathbf{H}$ with respect to the cross-entropy loss on the CIFAR-10 dataset from a set of fixed weights for ResNet-18 [21], which has approximately 11 million parameters [16, 19]. This time is directly proportional to the number of matrix-vector products, and therefore minimizing this quantity is crucial.

Algorithms and Lower Bounds for Low-Rank Approximation. The low-rank approximation problem is well studied in numerical linear algebra, with countless applications to clustering, data mining, principal component analysis, recommendation systems, and many more. (For surveys on low-rank approximation, see the monographs [20, 26, 52] and references therein.) In this problem, given an implicit $n \times d$ matrix $\mathbf{A}$, the goal is to output a matrix $\mathbf{Z} \in \mathbb{R}^{d \times k}$ with orthonormal columns such that

$$\left\|\mathbf{A}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_X \leq (1+\epsilon) \min_{\mathbf{U}:\mathbf{U}^\top\mathbf{U}=\mathbf{I}_k} \left\|\mathbf{A}\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\right\|_X, \quad (1)$$

where $\|\cdot\|_X$ denotes some norm. Note that given $\mathbf{Z}$, one can compute $\mathbf{A}\mathbf{Z}$ with an additional k queries, which will be negligible, and then $(\mathbf{A}\mathbf{Z}) \cdot \mathbf{Z}^\top$ is a rank-k matrix written in factored form, i.e., as the product of an $n \times k$ matrix and a $k \times d$ matrix. Among other things, low-rank approximation provides (1) a compression of $\mathbf{A}$ from nd parameters to $(n+d)k$ parameters, (2) faster matrix-vector products, since $\mathbf{A}\mathbf{Z} \cdot \mathbf{Z}^\top \cdot y$ can be computed in $O((n+d)k)$ time for an arbitrary vector y, as opposed to the $O(nd)$ time needed to compute $\mathbf{A} \cdot y$, and (3) de-noising, as often matrices $\mathbf{A}$ are close to low-rank (e.g., they are the product of latent factors) but only high rank due to noise.

Despite its tremendous importance, the optimal matrix-vector product complexity of low-rank approximation is unknown for any commonly used norm. The best known upper bound is due to Musco and Musco [30], who achieve $\tilde{O}(k/\epsilon^{1/2})$ queries[1] for both the case when $\|\cdot\|_X$ is the commonly studied Frobenius norm $\|\mathbf{B}\|_F = \left(\sum_{i,j} \mathbf{B}_{i,j}^2\right)^{1/2}$ as well as when $\|\cdot\|_X$ is the Spectral (operator) norm $\|\mathbf{B}\|_2 = \sup_{\|y\|_2=1} \|\mathbf{B}y\|_2$.

On the lower bound front, there is a trivial lower bound of k, since $\mathbf{A}$ may be full rank and achieving (1) requires k matrix-vector products since one must reconstruct the column span of $\mathbf{A}$ exactly. However, *no lower bounds in terms of the approximation factor ϵ were known.* We note that Simchowitz, Alaoui and Recht [44] prove lower bounds for approximating the top r eigenvalues of a symmetric matrix; however these guarantees are incomparable to those that follow from a low-rank approximation, even when the norm $\|\cdot\|_X$ is the operator norm.

Relationship to the Sketching Literature. Low-rank approximation has been extensively studied in the sketching literature which, when $\mathbf{A}$ is given explicitly, can achieve $O(\mathsf{nnz}(\mathbf{A}))$ time both for the Frobenius norm [14, 28, 33], as well as for Schatten-p norms [23]. However, these works require reading all of the entries in $\mathbf{A}$, and thus do not apply to any of the settings mentioned above. Further, the matrix-vector query model is especially important for problems such as trace estimation, where a low-rank approximation is used to first reduce the variance [29]. As trace estimation is often applied to implicit matrices, e.g., in computing Stochastic Lanczos Quadrature (SLQ) for Hessian eigendensity estimation [16], in studying the effects of batch normalization and residual connections in neural networks [54], and in computing a disentanglement regularizer for deep generative models [36], sketching algorithms for low-rank approximation often do not apply.

Another important application is low-rank approximation of covariance matrices [31], for which the covariance matrix is not given explicitly. Here, we have a data matrix $\mathbf{A}$ and we want a low-rank approximation for $\mathbf{A}\mathbf{A}^\top$. Even when $\mathbf{S}$ is a sparse sketching matrix, the matrix $\mathbf{S}\mathbf{A}$ is no longer sparse, and one needs to multiply $\mathbf{S}\mathbf{A}$ by $\mathbf{A}^\top$ to obtain a sketch of $\mathbf{S}\mathbf{A}\mathbf{A}^\top$, which is a dense matrix-matrix multiplication. Moreover, when viewed in the matrix-vector product model, sketching algorithms obtain provably worse query complexity than existing iterative algorithms (see Table 1 for a comparison). Further, as modern GPUs often do not exploit sparsity, *even when the matrix* $\mathbf{A}$ *is given, a GPU may not be able to take advantage of sparse queries*, which means the total time taken is proportional to the number of matrix-vector products.

Motivating Schatten-p Norms. The Schatten norms for $1 \leq p < 2$ are more robust than the Frobenius norm, as they dampen the effect of large singular values. In particular, the Schatten-1 norm, also known as the nuclear norm, has been widely used for robust PCA [10, 53, 55] as well as a convex relaxation of matrix

[1] We let $\tilde{O}(f) = f \cdot \mathrm{poly}(\log(dk/\epsilon))$.

rank in matrix completion [11, 12], low-dimensional Euclidean embeddings [39, 41, 48], image denoising [17, 18] and tensor completion [56]. In contrast, for $p > 2$, Schatten norms are more sensitive to large singular values and provide an approximation to the operator norm. In particular, for a rank r matrix, it is easy to see that setting $p = \log(r)/\eta$ yields a $(1+\eta)$-approximation to the operator norm (i.e., $p = \infty$). While the Block Krylov algorithm of Musco and Musco [30] implies a matrix-vector query upper bound of $\tilde{O}\left(k/\epsilon^{1/2}\right)$ for Schatten-∞ low-rank approximation, the exact complexity of this problem remains an outstanding open problem. When $p > 2$, we can interpolate between Frobenius and operator norm, and setting p to be a large fixed constant can be a proxy for Schatten-∞ low-rank approximation, with significantly fewer matrix-vector products (see Theorem 4.2).

Our Central Question. The main question of our work is:

What is the matrix-vector product complexity of low-rank approximation for the Frobenius norm, and more generally, for other matrix norms?

1.1 Our Results

We begin by stating our results for Frobenius and more generally, Schatten-p norm low-rank approximation for any $p \geq 1$; see Table 1 for a summary.

Theorem 1.1 (Query Upper Bound, informal Theorem 4.2). *Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, a target rank $k \in [d]$, an accuracy parameter $\epsilon \in (0,1)$ and any (not necessarily constant) $p \in [1, O(\log(d)/\epsilon)]$, there exists an algorithm that uses $\tilde{O}\left(kp^{1/6}/\epsilon^{1/3}\right)$ matrix-vector products and outputs a $d \times k$ matrix $\mathbf{Z}$ with orthonormal columns such that with probability at least 99/100,*

$$\left\|\mathbf{A}\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p} \leq (1+\epsilon)\min_{\mathbf{U}:\ \mathbf{U}^\top\mathbf{U}=\mathbf{I}_k}\left\|\mathbf{A}\left(\mathbf{I}-\mathbf{U}\mathbf{U}^\top\right)\right\|_{\mathcal{S}_p}.$$

When $p \geq \log(d)/\epsilon$, we get $\tilde{O}\left(k/\sqrt{\epsilon}\right)$ matrix-vector products.

We note that for Frobenius norm low-rank approximation (Schatten p for $p = 2$), we improve the prior matrix-vector product bound of $\tilde{O}(k/\epsilon^{1/2})$ by Musco and Musco [30] to $\tilde{O}(k/\epsilon^{1/3})$. For Schatten-$p$ low-rank approximation for $p \in [1,2)$, we improve work of Li and Woodruff [23] who require query complexity at least $\Omega(k^{2/p}/\epsilon^{4/p+1})$, which is a polynomial factor worse in both k and $1/\epsilon$ than our $\tilde{O}(k/\epsilon^{1/3})$ bound.

For $p > 2$, [23] obtain a query complexity of $\Omega(\min(n,d)^{1-2/p})$. We drastically improve this to $\tilde{O}(k/\epsilon^{1/3})$, which does not depend on d or n at all. Setting $p = \log(d)/\epsilon$ suffices to obtain a $(1+\epsilon)$-approximation to the spectral norm ($p = \infty$), and we obtain an $\tilde{O}\left(k/\sqrt{\epsilon}\right)$ query algorithm, matching the best known bounds for spectral low-rank approximation [30]. When $p > \log(d)/\epsilon$, we can simply run Block Krylov for $p = \infty$.

Remark 1.2 (Comments on the RAM Model). Although our focus is on minimizing the number of matrix-vector products, which is the key resource in the applications described above, our bounds also improve the running time of low-rank approximation algorithms when the matrix $\mathbf{A}$ has a small number of non-zero entries and is explicitly given. For simplicity, we state our bounds and those of previous work without using algorithms for fast matrix multiplication; similar improvements hold when using such algorithms. When $\mathsf{nnz}(\mathbf{A}) = O(n)$, for Frobenius norm low-rank approximation, work in the sketching literature, and in particular [5] (building off of [14, 15, 33]), achieves $O(nk^2/\epsilon)$ time. In contrast, in this setting our runtime is $\tilde{O}(nk^2/\epsilon^{2/3})$. Similarly, for Schatten-p low-rank approximation for $p \in [1,2)$, the previous best [23] requires $\tilde{\Omega}(nk^{4/p}/\epsilon^{(8/p-2)})$ time, while for $p > 2$ [23] requires $\tilde{\Omega}(nd^{2(1-2/p)}(k/\epsilon)^{4/p})$ time. In both cases our runtime is only $\tilde{O}(nk^2p^{1/3}/\epsilon^{2/3})$. We obtain analogous improvements when the sparsity $\mathsf{nnz}(\mathbf{A})$ is allowed to be $n(k/\epsilon)^C$ for a small constant $C > 0$.

Next, we state our lower bounds on the matrix-vector query complexity of Schatten-p low-rank approximation.

Theorem 1.3 (Query Lower Bound for constant p, informal Theorem 5.1 and Theorem 5.4). *Given $\varepsilon > 0$, and a fixed constant $p \geq 1$, there exists a distribution $\mathcal{D}$ over $n \times n$ matrices such that for $\mathbf{A} \sim \mathcal{D}$, any algorithm that with at least constant probability outputs a unit vector v such that $\left\|\mathbf{A}\left(\mathbf{I}-vv^\top\right)\right\|_{\mathcal{S}_p}^p \leq (1+\varepsilon)\min_{\|u\|_2=1}\left\|\mathbf{A}\left(\mathbf{I}-uu^\top\right)\right\|_{\mathcal{S}_p}^p$ must perform $\Omega(1/\varepsilon^{1/3})$ matrix-vector queries to $\mathbf{A}$.*

Remark 1.4. We note that this is the first lower bound as a function of ϵ for this problem, even for the well-studied case of $p = 2$, achieving an $\Omega(1/\epsilon^{1/3})$ bound, which is tight for any constant k, simultaneously for all constant $p \geq 1$.

Remark 1.5. Braverman, Hazan, Simchowitz and Woodworth [9] and Simchowitz, Alaoui and Recht [44] establish eigenvalue estimation lower bounds that we use in our arguments, but their results do not directly imply low-rank approximation lower bounds for any matrix norm that we are aware of, including spectral low-rank approximation, i.e., $p = \infty$.

Matrix Polynomials and Streaming Algorithms. Since our algorithms are based on iterative methods, they generalize naturally to low-rank approximations of matrices of the form $\mathbf{A}\left(\mathbf{A}^\top\mathbf{A}\right)^\ell$ and $\left(\mathbf{A}^\top\mathbf{A}\right)^\ell$ for any integer ℓ, given $\mathbf{A}$ as input. We defer the details to the full version.

Since we work in the matrix-vector model, our algorithms naturally extend to the multi-pass turnstile streaming setting. Notably, for $p > 2$, with $O\left(\log(d/\epsilon)p^{1/6}/\epsilon^{1/3}\right)$ passes we are able to improve the $\tilde{O}\left(n\left(\frac{kn^{1-2/p}}{\epsilon^2}+\frac{k^{2/p}+n^{1-2/p}}{\epsilon^{2+2/p}}\right)\right)$ memory bound of [23] to $\tilde{O}\left(nk/\epsilon^{1/3}\right)$.

1.2 Open Questions

We note that our lower bounds are tight only when the target rank k and Schatten norm p are fixed constants. In particular, it is open to obtain matrix-vector lower bounds that grow as a function of k, p and $1/\epsilon$. For the important special case of Spectral low-rank approximation ($p = \infty$), it is open to obtain any lower bound that grows as a function of $1/\epsilon$, even when the target rank $k = 1$. We also note that improving our upper bound to even $p^{1/6-o(1)}$ would imply a faster algorithm for Spectral low-rank approximation, addressing the main open question in [52].

Problem	Frobenius	Schatten-p, $p \in [1, 2)$	Schatten-p, $p > 2$
Sketching [13, 23]	$\Theta(k/\epsilon)$	$\Omega(k^{2/p}/\epsilon^{4/p+1})$	$\Omega(\min(n, d)^{1-2/p})$
Block Krylov [30]	$\tilde{O}(k/\epsilon^{1/2})$	N/A	N/A
Our Upper Bound	$\tilde{O}(k/\epsilon^{1/3})$	$\tilde{O}(k/\epsilon^{1/3})$	$\tilde{O}(kp^{1/6}/\epsilon^{1/3})$
Our Lower Bound	$\Omega(1/\epsilon^{1/3})$	$\Omega(1/\epsilon^{1/3})$	$\Omega(1/\epsilon^{1/3})$

Figure 1: Prior Upper and Lower Bounds on the Matrix Vector Product Complexity for Frobenius and Schatten-p low-rank Approximation. The poly(k/ϵ) factors in prior sketching work for Schatten-p are not explicit, but we have computed lower bounds on them to illustrate our improvements. Our bounds are optimal, up to logarithmic factors, for constant k. For $p > \log(d)/\epsilon$, spectral low-rank approximation [30] implies an $\tilde{O}\left(k/\sqrt{\epsilon}\right)$ upper bound.

2 TECHNICAL OVERVIEW

For our technical overview, we drop polylogarithmic factors appearing in the analysis and assume the input $\mathbf{A}$ is a symmetric $n \times n$ matrix (we handle arbitrary $n \times d$ matrices in Section 4).

2.1 Algorithms for Low-Rank Approximation

We first describe our algorithm for the special case of rank-1 approximation in the Frobenius norm, i.e., $p = 2$. Our algorithm is inspired by the Block Krylov algorithm of Musco and Musco [30]. Briefly, their algorithm begins with a random starting vector g (block size is 1) and computes the Krylov subspace $\mathbf{K} = [\mathbf{A}g; \mathbf{A}^2 g; \dots; \mathbf{A}^q g]$, for $q = O\left(1/\epsilon^{1/2}\right)$. Next, their algorithm computes an orthonormal basis for the column span of $\mathbf{K}$, denoted by a matrix $\mathbf{Q}$, and outputs the top singular vector of $\mathbf{Q}^\top \mathbf{A}^2 \mathbf{Q}$, denoted by z (see Algorithm 4.5 for a formal description). It follows from Theorem 1, guarantee (1) in [30] that

$$\left\|\mathbf{A}\left(\mathbf{I} - zz^\top\right)\right\|_F^2 \leq (1+\epsilon) \min_{\|u\|_2=1} \left\|\mathbf{A}\left(\mathbf{I} - uu^\top\right)\right\|_F^2, \tag{2}$$

and it is easy to see that this algorithm requires $\Theta\left(1/\epsilon^{1/2}\right)$ matrix-vector products. A naïve analysis requires an $O(1/\epsilon)$-degree polynomial in the matrix $\mathbf{A}$ to obtain (2), while [30] use Chebyshev polynomials to approximate the threshold function between first and second singular value, and save a quadratic factor in the degree. The guarantee in (2) then follows from observing that the best vector in the Krylov subspace is at least as good as the one that exists using Chebyshev polynomial approximation.

> **Algorithm 2.1** (Algorithm Sketch for Frobenius LRA).
> **Input:** An $n \times n$ symmetric matrix $\mathbf{A}$, accuracy parameter $0 < \varepsilon < 1$.
> (1) Run Block Krylov for $O\left(1/\epsilon^{1/3}\right)$ iterations with a random starting vector g. Let z_1 be the resulting output.
> (2) Run Block Krylov for $O(\log(n/\epsilon))$ iterations, but initialize with an $n \times b$ random matrix $\mathbf{G}$, where $b = O\left(1/\epsilon^{1/3}\right)$. Let z_2 be the resulting output.
> **Output:** $z = \arg\max_{z_1, z_2}\left(\|\mathbf{A}z_1\|_2^2, \|\mathbf{A}z_2\|_2^2\right)$.

Our starting point is the observation that while we require degree $\Theta\left(1/\epsilon^{1/2}\right)$ to separate the first and second singular values, if any subsequent singular value is sufficiently separated from σ_1, a significantly smaller degree polynomial suffices. In the context of Krylov methods, this translates to the intuition that starting with a matrix $\mathbf{G}$ with b columns (block size is b) should result in fewer iterations to find some vector in the top b subspace of $\mathbf{A}$. On the other hand, if no such singular value exists, the norm of the tail must be large and we can get away with a less accurate solution. We show that we can indeed exploit this trade-off by running Block Krylov on two different scales in parallel and then combine the solution. In particular, we use Algorithm 2.1.

Algorithm 2.1 captures the extreme points of the trade-off between the size of the starting matrix and the number of iterations, such that the total number of matrix-vector products is at most $\tilde{O}(1/\epsilon^{1/3})$. Further, we can compute the squared Euclidean norms of $\mathbf{A}z_1$ and $\mathbf{A}z_2$ with an additional matrix-vector product, and it remains to analyze the Frobenius cost of projecting $\mathbf{A}$ on the subspace $\mathbf{I} - zz^\top$, where z is the unit vector output by Algorithm 2.1.

Using gap-independent guarantees for Block Krylov (see Lemma 4.3 for a formal statement), it follows that with $O\left(1/\epsilon^{1/3}\right)$ iterations, we have

$$\|\mathbf{A}z_1\|_2^2 \geq \sigma_1^2(\mathbf{A}) - \epsilon^{2/3}\sigma_2^2(\mathbf{A}). \tag{3}$$

In contrast, using gap-dependent guarantees (see Lemma 4.4) for Block Krylov initialized with block size b, it follows that for any $\gamma > 0$, running $q = \log(1/\gamma) \cdot \sqrt{\sigma_1(\mathbf{A})/(\sigma_1(\mathbf{A}) - \sigma_b(\mathbf{A}))}$ iterations results in z_2 such that

$$\|\mathbf{A}z_2\|_2^2 \geq \sigma_1^2(\mathbf{A}) - \gamma\sigma_2^2(\mathbf{A}). \tag{4}$$

If $\sigma_b(\mathbf{A}) \leq \sigma_1(\mathbf{A})/2$, we can set $\gamma = \epsilon/n$ in Equation (4) to obtain a highly accurate solution. Further, regardless of the input instance, Step 3 in Algorithm 2.1 ensures that we get the best of both guarantees, (3) and (4). Then, observing that $\mathbf{I} - zz^\top$ is an orthogonal projection matrix (see Definition 3.1) and using the Pythagorean Theorem for Euclidean space we have:

$$\left\|\mathbf{A}\left(\mathbf{I} - zz^\top\right)\right\|_F^2 = \|\mathbf{A}\|_F^2 - \left\|\mathbf{A}zz^\top\right\|_F^2 = \|\mathbf{A}\|_F^2 - \|\mathbf{A}z\|_2^2, \tag{5}$$

where the second inequality follows from unitary invariance (see Fact 3.8) of the Frobenius norm and that the squared Frobenius norm of a rank-1 matrix $\mathbf{A}z$ (vector) is equal to its squared Euclidean norm. If it happens that $\sigma_2(\mathbf{A}) \leq \sigma_1(\mathbf{A})/2$, i.e., a constant gap exists between the first two singular values, then since guarantee (4) implies that $\|\mathbf{A}z\|_2^2 \geq \sigma_1^2(\mathbf{A}) - (\epsilon/n)\sigma_2^2(\mathbf{A})$, we can plug this into (5) to yield a $(1 + \epsilon/n)$-approximate solution. Hence, we focus on instances where $\sigma_2(\mathbf{A}) > \sigma_1(\mathbf{A})/2$.

Consider the case where the Frobenius norm of the tail is large, i.e., $\|\mathbf{A}-\mathbf{A}_1\|_F^2 \geq \sigma_2^2(\mathbf{A})/\epsilon^{1/3}$, where $\mathbf{A}_1$ is the best rank-1 approximation to $\mathbf{A}$. Then we only require an $\epsilon^{2/3}$-approximate solution (plugging guarantee (3) into (5)) since

$$\left\|\mathbf{A}\left(\mathbf{I}-z_1 z_1^\top\right)\right\|_F^2 \leq \|\mathbf{A}\|_F^2 - \sigma_1^2(\mathbf{A}) + \epsilon^{2/3}\sigma_2^2(\mathbf{A}) \leq \|\mathbf{A}-\mathbf{A}_1\|_F^2 + \epsilon\,\|\mathbf{A}-\mathbf{A}_1\|_F^2 . \tag{6}$$

Otherwise, $\sum_{i=2}^n \sigma_i^2(\mathbf{A}) < \sigma_2^2(\mathbf{A})/\epsilon^{1/3}$, which implies that there is a constant gap between the second and b-th singular values, where $b = O\left(1/\epsilon^{1/3}\right)$. To see this, observe if $\sigma_b(\mathbf{A}) > \sigma_2(\mathbf{A})/4$, then $\sum_{i=2}^n \sigma_i^2(\mathbf{A}) \geq \sum_{i=2}^b \sigma_i^2(\mathbf{A}) \geq b\sigma_2^2(\mathbf{A})/4$, which is a contradiction when $b > 10/\epsilon^{1/3}$, and thus $\sigma_b(\mathbf{A}) \leq \sigma_2(\mathbf{A})/4 < \sigma_1/2$. Now we can apply guarantee (4) with $q = O(\log(n/\epsilon))$ and conclude $\|\mathbf{A}z\|_2^2 \geq \sigma_1^2(\mathbf{A}) - (\epsilon/n)\sigma_2^2(\mathbf{A})$, yielding a highly accurate solution yet again. Overall, this suffices to obtain a $(1+\epsilon)$-approximate solution with $\tilde{O}(1/\epsilon^{1/3})$ matrix-vector queries.

Challenges in generalizing to Schatten $p \neq 2$ and rank $k > 1$. The outline above crucially relies on the norm of interest being Frobenius. In particular, we use the Pythagorean Theorem to analyze the cost of the candidate solution in Equation (5); however, the Pythagorean Theorem does not hold for non-Euclidean spaces. Therefore, a priori, it is unclear how to analyze the Schatten-p norm of a candidate rank-1 approximation. A proxy for the Pythagorean Theorem that holds for Schatten-p norms is Mahler's operator inequality (see Fact 3.11), which is in the right direction but holds only for $p \geq 2$, whereas we would like to handle all $p \geq 1$. Separately, for $p > 2$, the case where the tail is small corresponds to $\|\mathbf{A}-\mathbf{A}_1\|_{\mathcal{S}_p}^p \leq \sigma_2^p(\mathbf{A})/\epsilon^{1/3}$. Therefore, naïvely extending the above argument requires picking a block size that scales proportional to $O\left(2^p/\epsilon^{1/3}\right)$ to induce a constant gap between σ_1 and σ_b, and the number of matrix-vector products scales exponentially in p.

Finally, in the above outline, we also crucially use that $\left\|\mathbf{A}zz^\top\right\|_F^2 = \|\mathbf{A}z\|_2^2$. Observe that this no longer holds if we replace z with a matrix $\mathbf{Z}$ that has k orthonormal columns. Therefore, it remains unclear how to relate $\|\mathbf{AZ}\|_{\mathcal{S}_p}^p$ to $\left\|\mathbf{AZ}_{*,i}\right\|_2^2$, yet the vector-by-vector error guarantee obtained by Block Krylov (see Lemmas 4.3 and 4.4) only bounds the latter.

Handling all Schatten-p Norms and $k > 1$. We modify our algorithm to run Block Krylov on $\mathbf{A}^\top$ and obtain a orthonormal matrix $\mathbf{W}$ such that for all $i \in [k]$,

$$\left\|\mathbf{A}^\top \mathbf{W}_{*,i}\right\|^2 \geq \sigma_i^2(\mathbf{A}) - \gamma\sigma_{k+1}^2(\mathbf{A}), \tag{7}$$

for some $\gamma > 0$. We then analyze the cost $\left\|\mathbf{A}\left(\mathbf{I}-\mathbf{ZZ}^\top\right)\right\|_{\mathcal{S}_p}^p$, where $\mathbf{Z}$ is a basis for $\mathbf{A}^\top\mathbf{W}$. Our key insight is to interpret the input matrix $\mathbf{A}$ as a partitioned operator (block matrix) and invoke *pinching inequalities* for such operators. Pinching inequalities were originally introduced to understand unitarily invariant norms over direct sums of Hilbert spaces [43, 50]. In our setting, given a block matrix $\mathbf{M} = \begin{pmatrix}\mathbf{M}^{(1)} & \mathbf{M}^{(2)}\\ \mathbf{M}^{(3)} & \mathbf{M}^{(4)}\end{pmatrix}$, the *pinching inequality* (see Fact 3.13) implies that for all $p \geq 1$,

$$\|\mathbf{M}\|_{\mathcal{S}_p}^p \geq \left\|\mathbf{M}^{(1)}\right\|_{\mathcal{S}_p}^p + \left\|\mathbf{M}^{(4)}\right\|_{\mathcal{S}_p}^p . \tag{8}$$

A priori, it is unclear how to use Equation (8) to bound $\|\mathbf{A}(\mathbf{I}-\mathbf{ZZ}^\top)\|_{\mathcal{S}_p}^p$. First, we establish a general inequality for the Schatten norm of a matrix times an orthogonal projection. Let $\mathbf{P}$ and $\mathbf{Q}$ be any $n\times n$ orthogonal projection matrices with rank k (see Definition 3.1). Then, we prove (see Lemma 4.6 for details) that for any matrix $\mathbf{A}$,

$$\|\mathbf{A}\|_{\mathcal{S}_p}^p \geq \|\mathbf{PAQ}\|_{\mathcal{S}_p}^p + \|(\mathbf{I}-\mathbf{P})\,\mathbf{A}\,(\mathbf{I}-\mathbf{Q})\|_{\mathcal{S}_p}^p . \tag{9}$$

To obtain this inequality, we use a rotation argument along with the fact that the Schatten-p norms are unitarily invariant to show that $\|\mathbf{A}\|_{\mathcal{S}_p}^p = \left\|\begin{pmatrix}\mathbf{A}^{(1)} & \mathbf{A}^{(2)}\\ \mathbf{A}^{(3)} & \mathbf{A}^{(4)}\end{pmatrix}\right\|_{\mathcal{S}_p}^p$, where $\left\|\mathbf{A}^{(1)}\right\|_{\mathcal{S}_p} = \|\mathbf{PAQ}\|_{\mathcal{S}_p}$ and $\left\|\mathbf{A}^{(4)}\right\|_{\mathcal{S}_p} = \|(\mathbf{I}-\mathbf{P})\,\mathbf{A}\,(\mathbf{I}-\mathbf{Q})\|_{\mathcal{S}_p}$, and then we can apply Equation (8) to the block matrix above.

Once we have established Equation (9), we can set $\mathbf{P} = \mathbf{WW}^\top$ and set $\mathbf{Q} = \mathbf{ZZ}^\top$ to be the projection matrix corresponding to the column span of $\mathbf{A}^\top\mathbf{WW}^\top$. Then, we have that $\mathbf{PAQ} = \mathbf{WW}^\top\mathbf{A}$ and $(\mathbf{I}-\mathbf{P})\,\mathbf{A}\,(\mathbf{I}-\mathbf{Q}) = \mathbf{A}\left(\mathbf{I}-\mathbf{ZZ}^\top\right)$, and combined with (9) this yields

$$\left\|\mathbf{A}\left(\mathbf{I}-\mathbf{ZZ}^\top\right)\right\|_{\mathcal{S}_p}^p \leq \|\mathbf{A}\|_{\mathcal{S}_p}^p - \left\|\mathbf{WW}^\top\mathbf{A}\right\|_{\mathcal{S}_p}^p . \tag{10}$$

To obtain a bound on $\left\|\mathbf{WW}^\top\mathbf{A}\right\|_{\mathcal{S}_p}^p$, we appeal to the per-vector guarantees in Equation (7). However, translating from ℓ_2^2 error to $\sigma_p^p\left(\mathbf{W}^\top\mathbf{A}\right)$ incurs a mixed guarantee (see Lemma 4.7 for details):

$$\left\|\mathbf{WW}^\top\mathbf{A}\right\|_{\mathcal{S}_p}^p \geq \|\mathbf{A}_k\|_{\mathcal{S}_p}^p - O(\gamma p)\sum_{i\in[k]}\sigma_{k+1}^2(\mathbf{A})\,\sigma_i^{p-2}(\mathbf{A}) .$$

To use this bound, we require $\sigma_1(\mathbf{A})$ to be comparable to $\sigma_{k+1}(\mathbf{A})$ and thus we require an involved case analysis, which appears in the proof of Theorem 4.2.

Avoiding an exponential dependence on p. Our main insight here is that we do not require a block size that induces a constant gap between singular values. Instead, we first observe that if the block size b is large enough such that $\sigma_b \leq \sigma_2/(1+1/p)$, then $O(\log(n/\epsilon)\sqrt{p})$ iterations suffice to obtain a vector z such that $\|\mathbf{A}z\|_2^2 \geq \sigma_1^2(\mathbf{A}) - (\epsilon/n)\,\sigma_2^2(\mathbf{A})$. Therefore, we can trade-off the threshold for the Schatten norm of the tail with the number of iterations as follows: if $\|\mathbf{A}-\mathbf{A}_1\|_{\mathcal{S}_p}^p \leq \frac{1}{p^{1/3}\epsilon^{1/3}}\sigma_2^p(\mathbf{A})$, then setting $b = (1+1/p)^p/(\epsilon p)^{1/3} = \Theta(1/(\epsilon p)^{1/3})$ suffices to induce a gap of $1+1/p$ with block size b. The total number of matrix-vector products is $O(b\cdot\log(n/\epsilon)\sqrt{p}) = \tilde{O}(p^{1/6}/\epsilon^{1/3})$, since p can be assumed to be at most $(\log n)/\epsilon$. Otherwise, $\|\mathbf{A}-\mathbf{A}_1\|_{\mathcal{S}_p}^p > \frac{1}{p^{1/3}\epsilon^{1/3}}\sigma_2^p(\mathbf{A})$, and we only require a $(1+\epsilon^{2/3}/p^{1/3})$-approximate solution instead (compare with Equation (6)). Using gap-independent bounds (see Lemma 4.3), it suffices to start with block size 1 and run $O\left(\log(n/\epsilon)p^{1/6}/\epsilon^{1/3}\right)$ iterations to obtain a $(1+\epsilon^{2/3}/p^{1/3})$-approx. solution.

Avoiding a Gap-Dependent Bound. We note that even when there is a constant gap between the first and second singular values, and the per vector guarantee is highly accurate, i.e., for all $i \in [k]$, $\left\|\mathbf{A}\mathbf{Z}_{*,i}\right\|^2 \geq \sigma_i^2(\mathbf{A}) - \text{poly}\left(\frac{\epsilon}{d}\right)\sigma_{k+1}^2(\mathbf{A})$, it is not clear how to lower bound $\|\mathbf{A}\mathbf{Z}\|_{\mathcal{S}_p}^p$ in Equation 10. In general, the best bound we can obtain using the above equation is

$$\|\mathbf{A}\mathbf{Z}\|_{\mathcal{S}_p}^p \geq \|\mathbf{A}_k\|_{\mathcal{S}_p}^p - O\left(\frac{\epsilon}{\text{poly}(d)}\right)\sigma_{k+1}^2 \cdot \sum_{i \in [k]} \sigma_i^{p-2}, \tag{11}$$

which may be vacuous when the top k singular values are significantly larger than σ_{k+1} and $p > 2$. One could revert to a gap-dependent bound, where the error is in terms of the gap between σ_1 and σ_{k+1}, which one could account for by running an extra factor of $O(\log(\sigma_1/\sigma_{k+1}))$ iterations.

To avoid this gap-dependent bound, we split $\mathbf{A}$ into a head part $\mathbf{A}_H$ and a tail part $\mathbf{A}_T$, such that $\mathbf{A}_H$ has all singular values that are at least $(1+1/d)\,\sigma_{k+1}$ and $\mathbf{A}_T$ has the remaining singular values. We then bound $\left\|\mathbf{A}_H\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p}$ and $\left\|\mathbf{A}_T\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p}$ separately. Repeating the above analysis, we can obtain Equation (11) for $\mathbf{A}_T$ instead, and since all singular values larger than σ_{k+1} in $\mathbf{A}_T$ are bounded, we can obtain $\left\|\mathbf{A}_T\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p}^p \leq O(\epsilon k/\text{poly}(d))\,\sigma_{k+1}^p$. To adapt the analysis for $\mathbf{A}_T$ and obtain this bound, we use Cauchy's interlacing theorem to relate the j-th singular value of $\mathbf{A}_T\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)$ to the (i^*+j)-th singular value of $\mathbf{A}\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)$, where i^* is the rank of $\mathbf{A}_H$. We lower bound the (i^*+j)-th singular value of $\mathbf{A}\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)$ using the per vector guarantee of [30].

To bound $\left\|\mathbf{A}_H\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p}$, we observe it has rank at most k and thus

$$\begin{aligned}\left\|\mathbf{A}_H\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p} &\leq \sqrt{k}\cdot\left\|\mathbf{A}_H\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)\right\|_F \\ &= \sqrt{k}\cdot\sqrt{\|\mathbf{A}_H\|_F^2 - \|\mathbf{A}_H\mathbf{Z}\|_F^2},\end{aligned}$$

and we show how to bound this term in Section 4. Intuitively, while the k-dimensional subspace that we find can "swap out" singular vectors corresponding to singular values σ_i for which σ_i is very close to σ_{k+1}, since they serve equally well for a Schatten-p low-rank approximation, for singular values σ_i that are a bit larger than σ_{k+1}, the k-dimensional subspace we find cannot do this. More precisely, if y is a singular vector of $\mathbf{A}_H$ with singular value σ_i, then the projection of y onto the k-dimensional subspace that our algorithm finds (namely, $\mathbf{Z}$) must be at least $1-\sigma_{k+1}^2/((\sigma_i^2-\sigma_{k+1}^2)\text{poly}(d))$, which suffices to bound the above since the additive error is inversely proportional to σ_i^2 when $\sigma_i^2 \gg \sigma_{k+1}^2$, and so the very tiny additive error negates the effect of very large singular values.

2.2 Query Lower Bounds

Our lower bounds rely on the hardness of estimating the smallest eigenvalue of a Wishart ensemble (see Definition 3.15), as established in recent work of Braverman, Hazan, Simchowitz and Woodworth [9]. In particular, [9] show that for a $d \times d$ instance $\mathbf{W}$ of a Wishart ensemble, estimating $\lambda_d(\mathbf{W})$ (minimum eigenvalue) to additive error $1/d^2$ requires $\Omega(d)$ adaptive matrix-vector product queries (see Theorem 3.1 in [9]). To obtain hardness for Schatten-p low-rank approximation, we show that when $d = \Theta\left(1/\epsilon^{1/3}\right)$, any candidate unit vector z that satisfies $\left\|(\mathbf{I}-\mathbf{W}/5)\left(\mathbf{I}-zz^\top\right)\right\|_{\mathcal{S}_p}^p \leq (1+\epsilon)\min_{\|u\|_2=1}\left\|(\mathbf{I}-\mathbf{W}/5)\left(\mathbf{I}-uu^\top\right)\right\|_{\mathcal{S}_p}^p$, can be used to obtain an estimate $\hat{\lambda}_d = \frac{5}{p}\left(1-\|(\mathbf{I}-\mathbf{W}/5)\,z\|_2^p\right)$ such that $\hat{\lambda}_d = (1 \pm 1/d^2)\lambda_d\,(\mathbf{I}-\mathbf{W}/5)$. Let $\mathbf{A} = (\mathbf{I}-\mathbf{W}/5)$. To show our query lower bound, in contrast to the analysis of our algorithm, the challenge is now to lower bound $\left\|\mathbf{A}\left(\mathbf{I}-zz^\top\right)\right\|_{\mathcal{S}_p}^p$ in terms of $\|\mathbf{A}\|_{\mathcal{S}_p}^p$ and $\|\mathbf{A}z\|_2^p$ (contrast with Equation (10)).

Projection Cost via Araki-Lieb-Thirring. First, we note that the case of $p = 2$ is easy given the Pythagorean theorem. For $p \in [1,2)$, we can establish an inequality fairly straightforwardly: using the trace inner product definition of Schatten-p (see Definition 3.7) norms, we have,

$$\left\|\mathbf{A}\left(\mathbf{I}-zz^\top\right)\right\|_{\mathcal{S}_p}^p = \text{Tr}\left(\left(\left(\mathbf{I}-zz^\top\right)^2\mathbf{A}^2\left(\mathbf{I}-zz^\top\right)^2\right)^{p/2}\right), \tag{12}$$

Since $p/2 \in [1/2, 1)$, we can use the reverse *Araki-Lieb-Thirring* inequality (see Fact 3.10) to show that

$$\begin{aligned}&\text{Tr}\left(\left(\left(\mathbf{I}-zz^\top\right)^2\mathbf{A}^2\left(\mathbf{I}-zz^\top\right)^2\right)^{p/2}\right)\\ &\geq \text{Tr}\left(\left(\mathbf{I}-zz^\top\right)\mathbf{A}^p\left(\mathbf{I}-zz^\top\right)\right)\\ &\geq \|\mathbf{A}\|_{\mathcal{S}_p}^p - \left\|\mathbf{A}zz^\top\right\|_{\mathcal{S}_p}^p\end{aligned} \tag{13}$$

where we use the cyclicity of the trace and again use reverse *Araki-Lieb-Thirring* (Fact 3.10) to show that

$$\text{Tr}\left(\left(zz^\top\right)^{\frac{p}{2}}\left(\mathbf{A}^2\right)^{\frac{p}{2}}\left(zz^\top\right)^{\frac{p}{2}}\right) \leq \text{Tr}\left(\left(zz^\top\mathbf{A}^2zz^\top\right)^{p/2}\right) = \left\|\mathbf{A}zz^\top\right\|_{\mathcal{S}_p}^p.$$

Since we have $\left\|\mathbf{A}zz^\top\right\|_{\mathcal{S}_p}^p = \|\mathbf{A}z\|_2^p$, we conclude $\left\|\mathbf{A}\left(\mathbf{I}-zz^\top\right)\right\|_{\mathcal{S}_p}^p \geq \|\mathbf{A}\|_{\mathcal{S}_p}^p - \left\|\mathbf{A}zz^\top\right\|_2^p$. This approach only works for $p \in [1,2)$; for $p > 2$ the application of *Araki-Lieb-Thirring* is reversed in Equation 13 (since $p/2 > 1$, see Fact 3.10) and we no longer get a lower bound on the cost in Equation 12. We therefore require a new approach.

Projection Cost via Norm Compression. Recall, z is the unit vector output by our candidate low-rank approximation and let $y = \mathbf{A}z/\|\mathbf{A}z\|_2$. We yet again interpret the input matrix $\mathbf{A}$ as a partitioned operator by considering the projection of $\mathbf{A}$ onto $zz^\top$, $yy^\top$ and the projection away from these rank-1 subspaces. In particular, let $\mathbf{I}-yy^\top = \mathbf{Y}\mathbf{Y}^\top$, and $\mathbf{I}-zz^\top = \mathbf{Z}\mathbf{Z}^\top$, where $\mathbf{Y}$ and $\mathbf{Z}$ have orthonormal columns. Then, using a rotation argument, we show that

$$\|\mathbf{A}\|_{\mathcal{S}_p} = \left\|\begin{pmatrix} y^\top\mathbf{A}z & y^\top\mathbf{A}\mathbf{Z} \\ \mathbf{Y}^\top\mathbf{A}z & \mathbf{Y}^\top\mathbf{A}\mathbf{Z}\end{pmatrix}\right\|_{\mathcal{S}_p}.$$

We define the p-compression of $\mathbf{A}$, $\mathbf{C}_{\mathbf{A},p}$:

$$\mathbf{C}_{\mathbf{A},p} = \begin{pmatrix} \left\|y^\top\mathbf{A}z\right\|_{\mathcal{S}_p} & \left\|y^\top\mathbf{A}\mathbf{Z}\right\|_{\mathcal{S}_p} \\ \left\|\mathbf{Y}^\top\mathbf{A}z\right\|_{\mathcal{S}_p} & \left\|\mathbf{Y}^\top\mathbf{A}\mathbf{Z}\right\|_{\mathcal{S}_p}\end{pmatrix}.$$

To relate the norms of $\mathbf{A}$ and $\mathbf{C}_{\mathbf{A},p}$, we consider Audenaert's Norm Compression Conjecture [2], a question in functional analysis concerning operator inequalities (see also [3]):

Conjecture 2.2 (Schatten-p Norm Compression). *Let* $\mathbf{M}$ *be a partitioned operator (block matrix) such that* $\mathbf{M} = \begin{pmatrix} \mathbf{M}_1 & \mathbf{M}_2 \\ \mathbf{M}_3 & \mathbf{M}_4 \end{pmatrix}$. *Let* $\mathbf{C}_{\mathbf{M},p} = \begin{pmatrix} \|\mathbf{M}_1\|_{\mathcal{S}_p} & \|\mathbf{M}_2\|_{\mathcal{S}_p} \\ \|\mathbf{M}_3\|_{\mathcal{S}_p} & \|\mathbf{M}_4\|_{\mathcal{S}_p} \end{pmatrix}$ *be a* 2×2 *matrix that denotes the Schatten-p compression of* $\mathbf{M}$ *for any* $p \geq 1$. *Then,* $\|\mathbf{M}\|_{\mathcal{S}_p} \geq \|\mathbf{C}_{\mathbf{M},p}\|_{\mathcal{S}_p}$ *if* $1 \leq p \leq 2$, *and* $\|\mathbf{M}\|_{\mathcal{S}_p} \leq \|\mathbf{C}_{\mathbf{M},p}\|_{\mathcal{S}_p}$ *if* $2 \leq p < \infty$.

We could simply appeal to this conjecture to obtain that for all $p > 2$,

$$\|\mathbf{A}\|_{\mathcal{S}_p} \leq \|\mathbf{C}_{\mathbf{A},p}\|_{\mathcal{S}_p} = \left\| \begin{pmatrix} \|yy^\top \mathbf{A} zz^\top\|_{\mathcal{S}_p} & \|yy^\top \mathbf{A}\left(I - zz^\top\right)\|_{\mathcal{S}_p} \\ \|\left(\mathbf{I} - yy^\top\right)\mathbf{A} zz^\top\|_{\mathcal{S}_p} & \|\left(\mathbf{I} - yy^\top\right)\mathbf{A}\left(\mathbf{I} - zz^\top\right)\|_{\mathcal{S}_p} \end{pmatrix} \right\|_{\mathcal{S}_p}. \tag{14}$$

However, for our choice of y, $\|yy^\top \mathbf{A}\left(I - zz^\top\right)\|_{\mathcal{S}_p} = 0$. With padding and rotation arguments, we can then reduce our problem to a block matrix where the blocks in each row are aligned, i.e., each row is a scalar multiple of a fixed matrix (see Lemma 5.6). Then, we can use one of the few special cases of Conjecture 2.2 for aligned operators which has actually been proved, and appears in Fact 3.14. We can thus unconditionally obtain the inequality in Equation (14).

Now that we have reduced to the case where we have a 2×2 matrix with 3 non-zero entries, we would like to bound its Schatten-p norm. We explicitly compute the singular values of $\mathbf{C}_{\mathbf{A},p}$ (see Fact 5.7), and then use the structure of the instance to directly lower bound $\|\mathbf{A}z\|_2^p$ as follows:

$$\|\mathbf{A}z\|_2^p + \left(1 + O\left(\epsilon^{2p/3}\right)\right) \|\mathbf{A} - \mathbf{A}_1\|_{\mathcal{S}_p}^p \geq \|\mathbf{C}_{\mathbf{A},p}\|_{\mathcal{S}_p}^p \geq \|\mathbf{A}\|_{\mathcal{S}_p}^p, \tag{15}$$

where the last inequality follows from Equation (14). Since we understand the spectrum of the matrix $\mathbf{A}$, we can explicitly compute all the terms in (15) above and show that we can obtain an accurate estimate of the minimum singular value of $\mathbf{A}$ from $\|\mathbf{A}z\|_2^p$. See details in Section 5.2.

3 PRELIMINARIES

Given an $n \times d$ matrix $\mathbf{A}$ with rank r, and $n \geq d$, we can compute its singular value decomposition, denoted by $SVD(\mathbf{A}) = \mathbf{U}\Sigma\mathbf{V}^\top$, such that $\mathbf{U}$ is an $n \times r$ matrix with orthonormal columns, $\mathbf{V}^\top$ is an $r \times d$ matrix with orthonormal rows and Σ is an $r \times r$ diagonal matrix. The entries along the diagonal are the singular values of $\mathbf{A}$, denoted by $\sigma_1, \sigma_2 \ldots \sigma_r$. Given an integer $k \leq r$, we define the truncated singular value decomposition of $\mathbf{A}$ that zeros out all but the top k singular values of $\mathbf{A}$, i.e., $\mathbf{A}_k = \mathbf{U}\Sigma_k\mathbf{V}^\top$, where Σ_k has only k non-zero entries along the diagonal. It is well-known that the truncated SVD computes the best rank-k approximation to $\mathbf{A}$ under any unitarily invariant norm, but in particular for any Schatten-p norm (defined below), we have $\mathbf{A}_k = \min_{\text{rank}(\mathbf{X})=k} \|\mathbf{A} - \mathbf{X}\|_{\mathcal{S}_p}$. More generally, for any matrix $\mathbf{M}$, we use the notation $\mathbf{M}_k$ and $\mathbf{M}_{\setminus k}$ to denote the first k components and all but the first k components respectively. We use $\mathbf{M}_{i,*}$ and $\mathbf{M}_{*,j}$ to refer to the i^{th} row and j^{th} column of $\mathbf{M}$ respectively.

We use the notation $\mathbf{I}_k$ to denote a *truncated identity matrix*, that is, a square matrix with its top k diagonal entries equal to one, and all other entries zero. The dimension of $\mathbf{I}_k$ will be determined by context.

Definition 3.1 (Orthogonal Projection Matrices)**.** Given a $d \times d$ symmetric matrix $\mathbf{P}$ and $k \in [d]$, $\mathbf{P}$ is a rank-k orthogonal projection matrix if $\text{rank}(\mathbf{P}) = k$ and $\mathbf{P}^2 = \mathbf{P}$.

It follows from the above definition that $\mathbf{P}$ has eigenvalues that are either 0 or 1 and admits a singular value decomposition of the form $\mathbf{U}\mathbf{U}^\top$ where $\mathbf{U}$ has k orthonormal columns.

Definition 3.2 (Unitary Matrices)**.** Given a symmetric matrix $\mathbf{U} \in \mathbb{R}^{d \times d}$ we say $\mathbf{U}$ is a unitary matrix if $\mathbf{U}^\top\mathbf{U} = \mathbf{U}\mathbf{U}^\top = \mathbf{I}$.

Definition 3.3 (Rotation Matrices)**.** Given a symmetric matrix $\mathbf{R} \in \mathbb{R}^{d \times d}$ we say $\mathbf{R}$ is a rotation matrix if $\mathbf{R}$ is unitary and $\det(\mathbf{R}) = 1$.

Fact 3.4 (Courant-Fischer for Singular Values). *Given an* $n \times d$ *matrix* $\mathbf{A}$ *with singular values* $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d$, *the following holds: for all* $i \in [d]$,

$$\sigma_i = \max_{S:\ dim(S)=i} \ \min_{x \in S:\ \|x\|_2=1} \|x^\top \mathbf{A}\|_2.$$

Fact 3.5 (Weyl's Inequality for Singular Values (see Exercise 22 [47])). *Given* $n \times d$ *matrices* $\mathbf{X}, \mathbf{Y}$, *for any* $i, (j-1) \in [d]$ *such that* $i + j \leq d$,

$$\sigma_{i+j}\left(\mathbf{X} + \mathbf{Y}\right) \leq \sigma_i(\mathbf{X}) + \sigma_{j+1}(\mathbf{Y}).$$

Fact 3.6 (Bernoulli's Inequality). *For any* $x, p \in \mathbb{R}$ *such that* $x \geq -1$ *and* $p \geq 1$, $(1+x)^p \geq 1 + px$.

Schatten Norms and Trace Inequalities. We recall some basic facts for Schatten-p norms. We also require the following trace and operator inequalities.

Definition 3.7 (Schatten-p Norm)**.** Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d$ be the singular values of $\mathbf{A}$. Then, for any $p \in [0, \infty)$, the Schatten-p norm of $\mathbf{A}$ is defined as

$$\|\mathbf{A}\|_{\mathcal{S}_p} = \text{Tr}\left(\left(\mathbf{A}^\top\mathbf{A}\right)^{p/2}\right)^{1/p} = \left(\sum_{i \in [d]} \sigma_i^p(\mathbf{A})\right)^{1/p}.$$

Fact 3.8 (Schatten-p norms are Unitarily Invariant). *Given an* $n \times d$ *matrix* $\mathbf{M}$, *for any* $m \times n$ *matrix* $\mathbf{U}$ *with orthonormal columns, a norm* $\|\cdot\|_X$ *is defined to be unitarily invariant if* $\|\mathbf{U}\mathbf{M}\|_X = \|\mathbf{M}\|_X$. *The Schatten-p norm is unitarily invariant for all* $p \geq 1$.

There exists a closed-form expression for the low-rank approximation problem under Schatten-p norms:

Fact 3.9 (Schatten-p Low-Rank Approximation). *Given a matrix* $\mathbf{A} \in \mathbb{R}^{n \times d}$ *and an integer* $k \in \mathbb{N}$,

$$\mathbf{A}_k = \arg\min_{rank(\mathbf{X}) \leq k} \|\mathbf{A} - \mathbf{X}\|_{\mathcal{S}_p},$$

where $\mathbf{A}_k$ *is the truncated SVD of* $\mathbf{A}$.

Fact 3.10 (Araki–Lieb–Thirring Inequality [1]). *Given PSD matrices* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d \times d}$, *for any* $r \geq 1$, *the following inequality holds:* $\text{Tr}\left((\mathbf{BAB})^r\right) \leq \text{Tr}\left(\mathbf{B}^r\mathbf{A}^r\mathbf{B}^r\right)$. *Further, for* $0 < r < 1$, *the reverse holds* $\text{Tr}\left((\mathbf{BAB})^r\right) \geq \text{Tr}\left(\mathbf{B}^r\mathbf{A}^r\mathbf{B}^r\right)$.

Fact 3.11 (Mahler's Orthogonal Operator Inequality, Theorem 1.7 in [25]). *Given $p \geq 2$, and matrices $\mathbf{P}$ and $\mathbf{Q}$ such that the row (column) span of $\mathbf{P}$ is orthogonal to the row (column) span of $\mathbf{Q}$, the following inequality holds:*

$$\|\mathbf{P}\|_{\mathcal{S}_p}^p + \|\mathbf{Q}\|_{\mathcal{S}_p}^p \leq \|\mathbf{P}+\mathbf{Q}\|_{\mathcal{S}_p}^p .$$

Fact 3.12 (Hölder's Inequality for Schatten-p Norms, Corollary 4.2.6 [7]). *Given matrices $\mathbf{A}, \mathbf{B}^\top \in \mathbb{R}^{n\times d}$ and $p \in [1,\infty)$, the following holds*

$$\|\mathbf{AB}\|_{\mathcal{S}_p} \leq \|\mathbf{A}\|_{\mathcal{S}_q} \cdot \|\mathbf{B}\|_{\mathcal{S}_r},$$

for any q, r such that $\frac{1}{p} = \frac{1}{q} + \frac{1}{r}$.

We also require *pinching inequalities* that were originally introduced to relate norms for partitioned operators over direct sums of Hilbert spaces. In our context, these inequalities simplify to norm inequalities for block matrices:

Fact 3.13 (Pinching Inequalities for Schatten-p Norms, [8]). *Let $\mathbf{M} \in \mathbb{R}^{td\times td}$ be the following block matrix*

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{(1,1)} & \mathbf{M}_{(1,2)} & \cdots & \mathbf{M}_{(1,t)} \\ \mathbf{M}_{(2,1)} & \mathbf{M}_{(2,2)} & \cdots & \mathbf{M}_{(1,t)} \\ \vdots & & \ddots & \vdots \\ \mathbf{M}_{(t,1)} & \mathbf{M}_{(t,2)} & \cdots & \mathbf{M}_{(t,t)} \end{bmatrix},$$

where for all $i, j \in [t]$, $\mathbf{M}_{(i,j)} \in \mathbb{R}^{d\times d}$. For all $p \geq 1$, the following inequality holds:

$$\left(\sum_{i\in[t]} \left\|\mathbf{M}_{(i,i)}\right\|_{\mathcal{S}_p}^p\right)^{1/p} \leq \|\mathbf{M}\|_{\mathcal{S}_p} .$$

We also require a norm compression inequality that is a special case of Conjecture 2.2 (and known to be true), when each block is aligned in the following sense:

Fact 3.14 (Aligned Norm Compression Inequality, Section 4.3 in [2]). *Let $\mathbf{M} = \begin{pmatrix} \mathbf{M}_1 & \mathbf{M}_2 \\ \mathbf{M}_3 & \mathbf{M}_4 \end{pmatrix}$ such that there exist scalars $\alpha_1, \alpha_2, \beta_1, \beta_2$ such that $\mathbf{M}_1 = \alpha_1 \mathbf{X}$, $\mathbf{M}_2 = \alpha_2 \mathbf{X}$, $\mathbf{M}_3 = \beta_1 \mathbf{Y}$ and $\mathbf{M}_4 = \beta_2 \mathbf{Y}$. Then, for any $p \geq 2$,*

$$\|\mathbf{M}\|_{\mathcal{S}_p} \leq \left\| \begin{pmatrix} \|\mathbf{M}_1\|_{\mathcal{S}_p} & \|\mathbf{M}_2\|_{\mathcal{S}_p} \\ \|\mathbf{M}_3\|_{\mathcal{S}_p} & \|\mathbf{M}_4\|_{\mathcal{S}_p} \end{pmatrix} \right\|_{\mathcal{S}_p} .$$

Random Matrix Theory. Next, we recall some basic facts for Wishart ensembles from random matrix theory (we refer the reader to [46] for a comprehensive overview).

Definition 3.15 (Wishart Ensemble)**.** An $n\times n$ matrix $\mathbf{W}$ is sampled from a Wishart Ensemble, Wishart(n), if $\mathbf{W} = \mathbf{X}\mathbf{X}^\top$ such that for all $i, j \in [n]$ $\mathbf{X}_{i,j} \sim \mathcal{N}\left(0, \frac{1}{n}\mathbf{I}\right)$.

Fact 3.16 (Norms of a Wishart Ensemble). *Let $\mathbf{W} \sim$ Wishart(n) such that $n = \Omega(1/\varepsilon^3)$. Then, with probability 99/100, $\|\mathbf{W}\|_{op} \leq 5$ and for any fixed constant p, $\left\|\mathbf{I} - \frac{1}{5}\mathbf{W}\right\|_{\mathcal{S}_p}^p = \Theta\left(\frac{1}{\varepsilon^{1/3}}\right)$.*

4 ALGORITHMS FOR SCHATTEN-p LRA

In this section, we focus on obtaining algorithms for low-rank approximation in Schatten-p norm, simultaneously for all real, not necessarily constant, $p \in [1, O(\log(d)/\epsilon)]$. For the special case of $p \in \{2, \infty\}$, Musco and Musco [30] showed an algorithm with matrix-vector query complexity $\tilde{O}(k/\epsilon^{1/2})$, given below as Algorithm 4.5. We show that the number of matrix-vector products we require scales proportional to $\tilde{O}\left(kp^{1/6}/\epsilon^{1/3}\right)$ instead. Finally, recall when $p > \log(d)/\epsilon$, it suffices to run Block Krylov for $p = \infty$, which requires $O(\log(d/\epsilon)k/\sqrt{\epsilon})$ matrix-vector products. We note that proofs of intermediate lemmas have been ommited and appear in the full version.

Algorithm 4.1 (Optimal Schatten-p Low-rank Approximation)**.**

Input: An $n \times d$ matrix $\mathbf{A}$, target rank $k \leq d$, accuracy parameter $0 < \varepsilon < 1$, and $p \geq 1$.

(1) Let $\gamma_1 = \varepsilon^{2/3}/p^{1/3}$. Run Block Krylov Iteration (Algorithm 4.5) on $\mathbf{A}$ with block size k, and number of iterations $q = O\left(\log(d/\gamma_1)/\sqrt{\gamma_1} + \log(d/\epsilon)\sqrt{p}\right)$. Let $\mathbf{Z}_1 \in \mathbb{R}^{d\times k}$ be the corresponding output with orthonormal columns.

(2) Run Block Krylov Iteration (Algorithm 4.5) on $\mathbf{A}^\top$ with block size k, and number of iterations $q = O\left(\log(d/\gamma_1)/\sqrt{\gamma_1}\right)$. Let $\mathbf{W}_1 \in \mathbb{R}^{n\times k}$ be the corresponding output with orthonormal columns.

(3) Let $\gamma_2 = \varepsilon$ and let $s = O\left(p^{-1/3}k/\varepsilon^{1/3}\right)$. Run Block Krylov Iteration (Algorithm 4.5) on $\mathbf{A}^\top$ with block size s, and number of iterations $q = O\left(\log(d/\gamma_2)\sqrt{p}\right)$. Let $\mathbf{W}_2 \in \mathbb{R}^{n\times k}$ be the corresponding output with orthonormal columns.

(4) Run Block Krylov on $\mathbf{A}$ with target rank $k+1$ and number of iterations $q = O\left((\log(dp) + \log(d/\epsilon))\sqrt{p}\right)$, and let $\hat{\mathbf{Z}}_1$ be the resulting $d \times (k+1)$ output matrix. Compute $\hat{\sigma}_1^2 = \left\|\mathbf{A}(\hat{\mathbf{Z}}_1)_{*,1}\right\|_2^2$ and $\hat{\sigma}_{k+1}^2 = \left\|\mathbf{A}(\hat{\mathbf{Z}}_1)_{*,k+1}\right\|_2^2$, rough estimates of the 1-st and $(k+1)$-st singular values of $\mathbf{A}$. Run Block Krylov on $\mathbf{A}$ with target rank s, where $s = O\left(p^{-1/3}k/\varepsilon^{1/3}\right)$ and iterations $q = O\left(\log(d/\epsilon)\sqrt{p}\right)$, and let $\hat{\mathbf{Z}}_2$ be the resulting $d\times s$ output matrix. Compute $\hat{\sigma}_s^2 = \left\|\mathbf{A}(\hat{\mathbf{Z}}_2)_{*,s}\right\|_2^2$, an estimate to the s-th singular value of $\mathbf{A}$.

(5) If $\hat{\sigma}_1^2 \geq (1+0.5/p)\hat{\sigma}_{k+1}^2$, set $\mathbf{Z} = \mathbf{Z}_1$. Else, if $\hat{\sigma}_s^2 \leq \hat{\sigma}_{k+1}^2/(1+0.5/p)$, set $\mathbf{Z}$ to be an orthonormal basis for $\mathbf{A}^\top\mathbf{W}_2\mathbf{W}_2^\top$ and otherwise set $\mathbf{Z}$ to be an orthonormal basis for $\mathbf{A}^\top\mathbf{W}_1\mathbf{W}_1^\top$.

Output: A matrix $\mathbf{Z} \in \mathbb{R}^{d\times k}$ with orthonormal columns such that

$$\left\|\mathbf{A}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p}^p \leq (1+\epsilon) \min_{\mathbf{U}:\ \mathbf{U}^\top\mathbf{U}=\mathbf{I}_k} \left\|\mathbf{A}\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\right\|_{\mathcal{S}_p}^p .$$

Theorem 4.2 (Optimal Schatten-p Low-Rank Approximation). *Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, a target rank $k \in [d]$, an accuracy parameter $\epsilon \in (0,1)$ and any $p \in [1, O(\log(d)/\epsilon)]$, Algorithm*

4.1 performs $O\Big(kp^{1/6}\log(d/\epsilon)/\epsilon^{1/3}+\log(d/\epsilon)k\sqrt{p}\Big)$ matrix-vector products and outputs a $d\times k$ matrix $\mathbf{Z}$ with orthonormal columns such that with probability at least 9/10,

$$\left\|\mathbf{A}\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p} \leq (1+\epsilon) \min_{\mathbf{U}:\ \mathbf{U}^\top\mathbf{U}=\mathbf{I}_k} \left\|\mathbf{A}\left(\mathbf{I}-\mathbf{U}\mathbf{U}^\top\right)\right\|_{\mathcal{S}_p}.$$

Further, in the RAM model, the algorithm runs in time

$$O(\mathsf{nnz}(\mathbf{A})p^{1/6}k\log(d/\epsilon)/\epsilon^{1/3}+np^{(\omega-1)/6}k^{\omega-1}/\epsilon^{(\omega-1)/3}).$$

We first introduce the following lemmas from Musco and Musco [30] that provide convergence bounds for the performance of Block Krylov Iteration (Algorithm 4.5) :

Lemma 4.3 (Gap Independent Block Krylov with Arbitrary Accuracy). *Let $\mathbf{A}$ be an $n\times d$ matrix, k be the target rank and $\gamma>0$ be an accuracy parameter. Then, initializing Algorithm 4.5 with block size k and running for $q=\Omega\left(\log(d/\gamma)/\sqrt{\gamma}\right)$ iterations outputs a $d\times k$ matrix $\mathbf{Z}$ such that with probability 99/100, for all $i\in[k]$,*

$$\left\|\mathbf{A}\mathbf{Z}_{*,i}\right\|_2^2=\sigma_i^2\pm\gamma\sigma_{k+1}^2.$$

Further, the total number of matrix-vector products is $O(kq)$ and the running time in the RAM model is $O\Big(\mathsf{nnz}(\mathbf{A})kq+n\,(kq)^2+(kq)^\omega\Big)$.

The aforementioned lemma follows directly from Theorem 1 in [30], using the per-vector error guarantee (3).

Lemma 4.4 (Gap Dependent Block Krylov, Theorem 13 [30]). *Let $\mathbf{A}$ be an $n\times d$ matrix and $\gamma>0$, be an accuracy parameter and $p,k\in\mathcal{N}$ be such that $b\geq k$. Let $\sigma_1,\sigma_2\ldots\sigma_d$ be the singular values of $\mathbf{A}$. Then, initializing Algorithm 4.5 with block size b and running for $q=\Omega\left(\log(n/\gamma)\sqrt{\sigma_k}/\sqrt{\sigma_k-\sigma_b}\right)$ iterations outputs a $d\times k$ matrix $\mathbf{Z}$ such that with probability 99/100, for all $i\in[k]$*

$$\left\|\mathbf{A}\mathbf{Z}_{*,i}\right\|_2^2=\sigma_i^2\pm\gamma\sigma_{k+1}^2.$$

Further, the total number of matrix-vector products is $O(pq)$ and the running time in the RAM model is $O\Big(\mathsf{nnz}(\mathbf{A})bq+n\,(bq)^2+(bq)^\omega\Big)$.

Algorithm 4.5 (Block Krylov Iteration, [30]).

Input: An $n\times d$ matrix $\mathbf{A}$, target rank k, iteration count q and a block size parameter s such that $k\leq s\leq d$.

(1) Let $\mathbf{U}$ be a $n\times s$ matrix such that each entry is drawn i.i.d. from $\mathcal{N}(0,1)$. Let
$$\mathbf{K}=\left[\mathbf{A}^\top\mathbf{U};(\mathbf{A}^\top\mathbf{A})\mathbf{A}^\top\mathbf{U};\ldots;(\mathbf{A}^\top\mathbf{A})^q\mathbf{A}^\top\mathbf{U}\right]$$
be the $d\times s(q+1)$ Krylov matrix obtained by concatenating the matrices $\mathbf{A}^\top\mathbf{U},\ldots,\left(\mathbf{A}^\top\mathbf{A}\right)^q\mathbf{A}^\top\mathbf{U}$.

(2) Compute an orthonomal basis $\mathbf{Q}$ for the column span of $\mathbf{K}$. Let $\mathbf{M}=\mathbf{Q}^\top\mathbf{A}^\top\mathbf{A}\mathbf{Q}$.

(3) Compute the top k left singular vectors of $\mathbf{M}$, and denote them by $\mathbf{Y}_k$.

Output: $\mathbf{Z}=\mathbf{Q}\mathbf{Y}_k$

Next, we prove the following key lemma relating the Schatten-p norm of row and column projections applied to a matrix $\mathbf{A}$ to the Schatten-p norm of the matrix itself. We can interpret this lemma as an extension of the Pythagorean Theorem to Schatten-p spaces and believe this lemma is of independent interest. We note that we appeal to *pinching inequality* for partitioned operators to obtain this lemma.

Lemma 4.6 (Schatten-p Norms for Orthogonal Projections). *Let $\mathbf{A}$ be an $n\times d$ matrix, let $\mathbf{P}$ be an $n\times n$ matrix, and let $\mathbf{Q}$ be a $d\times d$ matrix such that both $\mathbf{P}$ and $\mathbf{Q}$ are orthogonal projection matrices of rank k (see Definition 3.1). Then, the following inequality holds for all $p\geq 1$:*

$$\|\mathbf{A}\|_{\mathcal{S}_p}^p\geq\|\mathbf{P}\mathbf{A}\mathbf{Q}\|_{\mathcal{S}_p}^p+\|(\mathbf{I}-\mathbf{P})\,\mathbf{A}\,(\mathbf{I}-\mathbf{Q})\|_{\mathcal{S}_p}^p.$$

Note, despite establishing Lemma 4.6, it is not immediately apparent how to lower bound $\left\|\mathbf{A}\mathbf{Z}\mathbf{Z}^\top\right\|_{\mathcal{S}_p}^p$, where $\mathbf{Z}$ is a candidate solution. Next, we show how to translate a guarantee on the Euclidean norm of $\mathbf{A}$ times a column of $\mathbf{Z}$ to a lower bound on $\left\|\mathbf{A}\mathbf{Z}\mathbf{Z}^\top\right\|_{\mathcal{S}_p}^p$.

Lemma 4.7 (Per-Vector Guarantees to Schatten Norms). *Let $\mathbf{A}$ be an $n\times d$ matrix with singular values denoted by $\{\sigma_i\left(\mathbf{A}\right)\}_{i\in[d]}$. Let $\mathbf{Z}$ be a $d\times k$ matrix with orthonormal columns that is output by Algorithm 4.5, such that for all $i\in[k]$, with probability at least 99/100, $\left\|\mathbf{A}\mathbf{Z}_{*,i}\right\|_2^2\geq\sigma_i^2\left(\mathbf{A}\right)-\gamma\sigma_{k+1}^2\left(\mathbf{A}\right)$, for some $\gamma\in(0,1)$. Then, for any $p\geq 1$, we have $\left\|\mathbf{A}\mathbf{Z}\mathbf{Z}^\top\right\|_{\mathcal{S}_p}^p\geq\|\mathbf{A}_k\|_{\mathcal{S}_p}^p-O(\gamma p)\sum_{i\in[k]}\sigma_{k+1}^2\left(\mathbf{A}\right)\sigma_i^{p-2}\left(\mathbf{A}\right)$.*

Finally, we also need the following lemma:

Lemma 4.8 (Singular Values to Alignment of Singular Vectors). *Let $\mathbf{A}=\mathbf{U}\Sigma\mathbf{V}^\top$ be the SVD and let $\mathbf{Z}$ be a $d\times k$ orthonormal matrix such that for all $i\in[k]$, $\left\|\mathbf{A}\mathbf{Z}_{*,i}\right\|_2^2\geq\sigma_i^2\left(\mathbf{A}\right)-(\epsilon/d)^c\sigma_{k+1}^2$, for some fixed constant $c\geq 10$. Further, assume there exists a $j^*\in[k]$ such that for all $j\in[j^*]$, $\sigma_j^2(\mathbf{A})\geq(1+\epsilon/d)\,\sigma_{k+1}^2(\mathbf{A})$ and $\sigma_{j^*+1}^2(\mathbf{A})\leq(1-\epsilon/d)\,\sigma_{j^*}^2(\mathbf{A})$. Then, $\left\|\mathbf{V}_{j^*}^\top\mathbf{Z}\right\|_F^2\geq j^*-(\epsilon/d)^{c-4}$, where $\mathbf{V}_{j^*}^\top$ is the top-j^* rows of $\mathbf{V}^\top$.*

We now have all the ingredients we need to complete the proof of Theorem 4.2.

Proof of Theorem 4.2. Observe, using Lemma 4.3 with probability at least 97/100, Step 3 of Algorithm 4.1 outputs the following: $\hat{\sigma}_1^2=(1\pm 0.1/p)\,\sigma_1^2$, $\hat{\sigma}_{k+1}^2=(1\pm 0.1/p)\,\sigma_{k+1}^2$ and $\hat{\sigma}_s^2=(1\pm 0.1/p)\,\sigma_s^2$, for $s=O(kp^{-1/3}/\epsilon^{1/3})$. Condition on this event. Our proof proceeds via case analysis. The case where there is at least a constant gap between the first and $(k+1)$-st singular value is easy to handle since we can use gap-dependent guarantees to obtain highly accurate estimates of the top-k singular values. When there is no gap, either the Schatten-p norm of the tail is large compared to the $(k+1)$-st singular value, and we don't require a highly accurate solution, or the Schatten-p norm of the tail is small, and increasing the block size induces a gap. We formalize this intuition into a proof.

Let us first consider the case where there is a constant gap between the top and the $(k+1)$-st singular values, i.e., $\sigma_1>(1+1/p)\sigma_{k+1}$. Observe, since we have $(1+0.1/p)$-approximate estimates to σ_1 and σ_{k+1}, we can detect that we are in this case and Algorithm 4.1 outputs $\mathbf{Z}=\mathbf{Z}_1$. We further observe that the Algorithm 4.1 runs at least $\Omega(\log(d/\epsilon)\sqrt{p})$ iterations (since $p\leq\log(d)/\epsilon$) since $\mathbf{Z}=\mathbf{Z}_1$. We observe that in this case, there exists a gap of size p between σ_1 and σ_{k+1}, since $1-\sigma_{k+1}/\sigma_1\leq 1/p$. It follows from Lemma 4.4 that running Block Krylov Iteration for $O\big(\log(d/\epsilon)\sqrt{p}\big)$ iterations with block size $\geq k$ suffices to output a matrix $\mathbf{Z}$ such that with probability at least 99/100, for all $i \in [k]$,

$$\left\|\mathbf{A}\mathbf{Z}_{*,i}\right\|_2^2 \geq \sigma_i^2(\mathbf{A}) - \text{poly}\left(\frac{\epsilon}{d}\right)\sigma_{k+1}^2(\mathbf{A}). \tag{16}$$

We note that we cannot simply take $p/2$-th powers here (for large p) as this would introduce cross terms that scale proportional to $\sigma_i(\mathbf{A})$, which can be significantly larger than $\sigma_{k+1}(\mathbf{A})$. Instead, we require a finer analysis by splitting $\mathbf{A}$ into a head and tail term.

Let $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$ be the SVD of $\mathbf{A}$ and for all $j \in [d]$, let $v_j = \mathbf{V}_{j,*}^\top$ denote the j-th row of $\mathbf{V}^\top$. By the Pythagorean Theorem, we have

$$\begin{aligned}\|\mathbf{A}\mathbf{Z}\|_F^2 &= \|\mathbf{A}_k\mathbf{Z}\|_F^2 + \|(\mathbf{A} - \mathbf{A}_k)\mathbf{Z}\|_F^2 \\ &\leq \sum_{j\in[k]} \left(\sigma_j^2 - \sigma_{k+1}^2\right)\left\|v_j^\top \mathbf{Z}\right\|_2^2 + \sigma_{k+1}^2 k.\end{aligned} \tag{17}$$

Summing over $j \in [k]$ for the guarantee obtained in Equation 16, we have

$$\|\mathbf{A}\mathbf{Z}\|_F^2 = \sum_{j\in[k]} \left\|\mathbf{A}\mathbf{Z}_{*,j}\right\|_F^2 \geq \sum_{j\in[k]} \sigma_j^2 - O(\gamma k)\,\sigma_{k+1}^2. \tag{18}$$

where $\gamma = \text{poly}(\epsilon/d)$. Combining Equations (17) and (18), we can conclude

$$\sum_{j\in[k]} \left(\sigma_j^2 - \sigma_{k+1}^2\right) - O(\gamma k)\,\sigma_{k+1}^2 \leq \sum_{j\in[k]} \left(\sigma_j^2 - \sigma_{k+1}^2\right)\left\|v_j^\top \mathbf{Z}\right\|_2^2. \tag{19}$$

Let $j' \in [k]$ be the largest integer such that for all $j \leq j'$, $\sigma_j^2 \geq (1+\epsilon/d)\,\sigma_{k+1}^2$. Next, let $j^* \in [j', k]$ be such that $\sigma_{j^*+1} \leq (1-\epsilon/d)\sigma_{j^*}$. Observe, such a j^* is guaranteed to exist since there is a gap between σ_1 and σ_{k+1}. Since $\left\|v_j^\top \mathbf{Z}\right\|_2^2 \leq 1$, we can restate Equation (19), as follows:

$$\begin{aligned}&\sum_{j\in[k]} \left(\sigma_j^2 - \sigma_{k+1}^2\right) - O(\gamma k)\,\sigma_{k+1}^2 \\ &\leq \sum_{j\in[j^*]} \left(\sigma_j^2 - \sigma_{k+1}^2\right)\left\|v_j^\top \mathbf{Z}\right\|_2^2 + \sum_{j\in[j^*+1,k]} \left(\sigma_j^2 - \sigma_{k+1}^2\right).\end{aligned}$$

Subtracting $\sum_{j\in[j^*+1,k]}\left(\sigma_j^2 - \sigma_{k+1}^2\right)$ from both sides, and rearranging, we have

$$\begin{aligned}\sum_{j\in[j^*]} \left(\sigma_j^2 - \sigma_{k+1}^2\right) - \gamma k \sigma_{k+1}^2 + \sigma_{k+1}^2 \sum_{j\in[j^*]} \left\|v_j^\top \mathbf{Z}\right\|_2^2 \\ \leq \sum_{j\in[j^*]} \sigma_j^2 \left\|v_j^\top \mathbf{Z}\right\|_2^2\end{aligned} \tag{20}$$

We are now ready to bound $\left\|\mathbf{A}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p}$. By the triangle inequality,

$$\begin{aligned}\left\|\mathbf{A}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p} &\leq \left\|\mathbf{A}_{j^*}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p} \\ &\quad + \left\|\left(\mathbf{A} - \mathbf{A}_{j^*}\right)\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p}\end{aligned} \tag{21}$$

Observe, for any $p \geq 1$, $\left\|\mathbf{A}_{j^*}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p} \leq \sqrt{k}\left\|\mathbf{A}_{j^*}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_F$, since $\mathbf{A}_{j^*}$ has rank at most k, with $p = 1$ achieving the worst inequality. Therefore, using the Pythagorean theorem again, and plugging in the lower bound from Equation (20)

$$\left\|\mathbf{A}_{j^*}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p} \leq \sqrt{k}\sigma_{k+1} \cdot \left(j^* - \sum_{j\in[j^*]} \left\|v_j^\top \mathbf{Z}\right\|_2^2 + O(\gamma k)\right)^{1/2} \tag{22}$$

It therefore remains to lower bound $\sum_{j\in[j^*]} \left\|v_j^\top \mathbf{Z}\right\|_2^2$. Applying Lemma 4.8, we have,

$$\sum_{j\in[j^*]} \left\|v_j^\top \mathbf{Z}\right\|_2^2 = \left\|\mathbf{V}_{j^*}^\top \mathbf{Z}\right\|_F^2 \geq j^* - O\left((\epsilon/d)^4\right) \tag{23}$$

Plugging back into Equation (22), $\left\|\mathbf{A}_{j^*}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p} \leq O\left(\frac{\epsilon}{d}\sigma_{k+1}\right)$ and thus substituting into Equation (21),

$$\begin{aligned}\left\|\mathbf{A}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p} &\leq O\left(\frac{\epsilon}{d}\right)\|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p} \\ &\quad + \left\|\left(\mathbf{A} - \mathbf{A}_{j^*}\right)\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p}.\end{aligned} \tag{24}$$

It remains to bound term 24.1 above.

Applying Lemma 4.6 with $\mathbf{Q} = \mathbf{Z}\mathbf{Z}^\top$ and $\mathbf{P} = \mathbf{W}\mathbf{W}^\top$ being the projection on the column span of $\mathbf{A}\mathbf{Z}\mathbf{Z}^\top$, we have

$$\left\|\left(\mathbf{A} - \mathbf{A}_{j^*}\right)\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p}^p \leq \sum_{j\in[j^*+1,d]} \sigma_j^p - \sum_{j\in[k]} \sigma_j^p\left(\mathbf{W}^\top\left(\mathbf{A} - \mathbf{A}_{j^*}\right)\right)$$

Next, we show that for all $j \in [k]$, $\sigma_j(\mathbf{W}^\top(\mathbf{A}-\mathbf{A}_{j^*})) \geq \sigma_{j+j^*}(\mathbf{W}^\top\mathbf{A})$. Here, we invoke Fact 3.5 for $\mathbf{X} = \left(\mathbf{A} - \mathbf{A}_{j*}\right)$ and $\mathbf{Y} = \mathbf{A}_{j*}$, with $i = j$ and $j = j^*$. Note, the precondition on the indices i, j in Fact 3.5 is satisfied since $\mathbf{X}, \mathbf{Y}$ are $n \times k$ matrices, and $j \in [k]$ and $j^* < k$. Then, we have $\sigma_{j+j^*}\left(\mathbf{W}^\top\mathbf{A}\right) \leq \sigma_j\left(\mathbf{W}^\top\left(\mathbf{A} - \mathbf{A}_{j^*}\right)\right) + \sigma_{j^*+1}\left(\mathbf{W}^\top\mathbf{A}_{j*}\right)$. But $\mathbf{A}_{j*}\mathbf{Z}$ is a rank $\leq j^*$ matrix, and thus $\sigma_{j^*+1}\left(\mathbf{A}_{j*}\mathbf{Z}\right) = 0$. Therefore, we can conclude,

$$\left\|\left(\mathbf{A} - \mathbf{A}_{j^*}\right)\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p}^p \leq \sum_{j\in[j^*+1,d]} \sigma_j^p - \sum_{j\in[j^*,k+j^*]} \sigma_j^p\left(\mathbf{W}^\top\mathbf{A}\right) \tag{25}$$

Finally, we show that $\sigma_j^p(\mathbf{W}^\top\mathbf{A}) \geq \sigma_j^p(\mathbf{A}\mathbf{Z})$ (we defer the proof to the full version) and by definition, for $j \in [j^*+1, k+j^*]$, $\sigma_j \leq (1+\epsilon/d)\,\sigma_{k+1}$ and thus, it follows from Lemma 4.7 that for all $j \in [j^*+1, k]$,

$$\sigma_j^p(\mathbf{A}\mathbf{Z}) \geq \sigma_j^p - O(\gamma p)\,\sigma_{k+1}^p, \tag{26}$$

where the last inequality uses that $p = O(\log(d)/\epsilon)$. Substituting this back into Equation (25), we have

$$\left\|\left(\mathbf{A} - \mathbf{A}_{j^*}\right)\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p}^p \leq (1 + O(\gamma pk))\,\|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p}^p. \tag{27}$$

Taking the p-th root and substituting back into Equation (24),

$$\begin{aligned}&\left\|\mathbf{A}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p} \\ &\leq (1 + O(\gamma pk))^{1/p}\,\|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p} + O\left(\frac{\epsilon}{d}\right)\|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p},\end{aligned} \tag{28}$$

and since $\gamma = \text{poly}(\epsilon/d)$, we have $\left\|\mathbf{A}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p} \leq (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p}$, which completes the analysis for this case.

Next, we consider the case where the gap between the top and the $(k+1)$-st singular value is small, i.e., $\sigma_1 < (1+1/p)\,\sigma_{k+1}$. We yet again split into cases, and consider the case where the Schatten-p

norm of the tail is small, i.e. $\|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p}^p \le \frac{k}{p^{1/3}\epsilon^{1/3}} \cdot \sigma_{k+1}^p$. Observe, for any $t \in [1, d-k-1]$,

$$\frac{k}{p^{1/3}\epsilon^{1/3}} \cdot \sigma_{k+1}^p \ge \|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p}^p \ge \sum_{i=k+1}^{k+1+t} \sigma_i^p \ge t\sigma_{k+1+t}^p. \tag{29}$$

Then, setting $t = \frac{(1+1/p)^p k}{\epsilon^{1/3}p^{1/3}} = \Theta\left(\frac{k}{\epsilon^{1/3}p^{1/3}}\right)$, we have $\sigma_{k+1+t} \le \sigma_{k+1}/(1+1/p)$. It suffices to show that we can detect this gap for some $s \ge k+1+t$. Recall, we know that $\hat{\sigma}_{k+1} = (1 \pm 0.1/p)\sigma_{k+1}$ and $\hat{\sigma}_s = (1 \pm 0.1/p)\sigma_s$. Then, we have

$$\hat{\sigma}_s \le \left(1 + \frac{0.1}{p}\right)\sigma_s \le \left(1 + \frac{0.1}{p}\right)\sigma_{k+1+t} \le \frac{1}{(1+0.5/p)}\hat{\sigma}_{k+1}. \tag{30}$$

Therefore, Algorithm 4.1 outputs $\mathbf{Z}$, an orthonormal basis for $\mathbf{A}^\top\mathbf{W}_2$, where $\mathbf{W}_2$ is obtained by running Algorithm 4.5 on $\mathbf{A}^\top$, initialized with a block size of $\Theta\left(\frac{k}{\epsilon^{1/3}p^{1/3}}\right)$ and run for $O\left(\log(d/\epsilon)\sqrt{p}\right)$ iterations. Observe, since $\sigma_{k+1+t} \le \sigma_{k+1}/(1+1/p)$, this suffices to demonstrate a gap that depends on p as follows: $\frac{\sigma_k}{\sigma_k - \sigma_{k+t+1}} \le p$. Recall, we account for this gap by running $O\left(\log(d)\sqrt{p}\right)$ iterations. Using the gap dependent analysis (Lemma 4.4), we can conclude that with probability at least 99/100, for all $i \in [k]$,

$$\left\|\mathbf{A}^\top(\mathbf{W}_2)_{*,i}\right\|_2^2 \ge \sigma_i^2 - \text{poly}\left(\frac{\epsilon}{d}\right)\sigma_{k+1}^2. \tag{31}$$

Then, applying Lemma 4.7 with $\mathbf{W}_2\mathbf{W}_2^\top$ satisfying the guarantee in (31), we have

$$\left\|\mathbf{A}^\top\mathbf{W}_2\mathbf{W}_2^\top\right\|_{\mathcal{S}_p}^p \ge \|\mathbf{A}_k\|_{\mathcal{S}_p}^p - \text{poly}\left(\frac{\epsilon}{d}\right)\sigma_{k+1}^p. \tag{32}$$

Next, we use Lemma 4.6 to relate $\left\|\mathbf{A}^\top\mathbf{W}_2\mathbf{W}_2^\top\right\|_{\mathcal{S}_p}^p$ to $\left\|\mathbf{A}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p}^p$, where $\mathbf{Z}$ is an orthonormal basis for $\mathbf{A}^\top\mathbf{W}_2\mathbf{W}_2^\top$ as output by the algorithm. Setting $\mathbf{Q} = \mathbf{Z}\mathbf{Z}^\top$ and $\mathbf{P} = \mathbf{W}_2\mathbf{W}_2^\top$, we observe that $\|\mathbf{PAQ}\|_{\mathcal{S}_p}^p = \left\|\mathbf{W}_2\mathbf{W}_2^\top\mathbf{A}\right\|_{\mathcal{S}_p}^p$ and $\|(\mathbf{I}-\mathbf{P})\mathbf{A}(\mathbf{I}-\mathbf{Q})\|_{\mathcal{S}_p}^p = \|\mathbf{A}(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top)\|_{\mathcal{S}_p}^p$. Then, invoking Lemma 4.6 and plugging in Equation (32), we have

$$\begin{aligned}\left\|\mathbf{A}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p}^p &\le \|\mathbf{A}\|_{\mathcal{S}_p}^p - \left\|\mathbf{A}^\top\mathbf{W}_2\mathbf{W}_2^\top\right\|_{\mathcal{S}_p}^p \\ &\le \left(1 + \text{poly}\left(\frac{\epsilon}{d}\right)\right)\|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p}^p,\end{aligned} \tag{33}$$

which concludes the analysis in this case.

As shown in Equation 30, we can detect a gap between σ_{k+1+t} and σ_{k+1} by comparing $\hat{\sigma}_s$ and $\hat{\sigma}_{k+1}$. When 30 does not hold, we know that $\hat{\sigma}_s \ge (1+0.5/p)\,\hat{\sigma}_{k+1}$ and Algorithm 4.1 outputs $\mathbf{Z}$, an orthonormal basis for $\mathbf{A}^\top\mathbf{W}_1\mathbf{W}_1^\top$. Since we have $(1 \pm 0.1/p)$-approximate estimates to these quantities, we can conclude that $\sigma_s \ge (1+0.1/p)\,\sigma_{k+1}$. Then, we have

$$\|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p}^p \ge s \cdot \sigma_s^p = \Omega\left(\frac{k}{\epsilon^{1/3}p^{1/3}}\right)\sigma_{k+1}^p.$$

It therefore remains to consider the case where $\|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p}^p > \frac{ck}{p^{1/3}\epsilon^{1/3}} \cdot \sigma_{k+1}^p$, for a fixed universal constant c. Here, we note that the tail is large enough that an additive error of $O\left(\epsilon^{2/3}p^{1/3}\right)\sigma_{k+1}^2$ on each of the top-k singular values suffices. Formally, it follows from Lemma 4.3 (setting $\gamma = \epsilon^{2/3}p^{-1/3}$, and invoking it for $\mathbf{A}^\top$) that initializing Algorithm 4.5 with block size k and running for $O\left(\log(d/\epsilon)p^{1/6}/\epsilon^{1/3}\right)$ iterations suffices to output a $n \times k$ matrix $\mathbf{W}_1$ such that with probability at least 99/100, for all $i \in [k]$,

$$\left\|\mathbf{A}^\top(\mathbf{W}_1)_{*,i}\right\|_2^2 \ge \sigma_i^2 - \epsilon^{2/3}p^{-1/3}\sigma_{k+1}^2.$$

Then, invoking Lemma 4.7 with $\mathbf{A}^\top$ and $\mathbf{W}_1$ as defined above, we have

$$\begin{aligned}\left\|\mathbf{A}^\top\mathbf{W}_1\mathbf{W}_1^\top\right\|_{\mathcal{S}_p}^p &= \left\|\mathbf{W}_1\mathbf{W}_1^\top\mathbf{A}\right\|_{\mathcal{S}_p}^p \\ &\ge \|\mathbf{A}_k\|_{\mathcal{S}_p}^p - O\left(k\epsilon^{2/3}p^{2/3}\right)\sigma_{k+1}^p\end{aligned} \tag{34}$$

where the last inequality uses that $\sigma_1 < (1+1/p)\sigma_{k+1}$ and $(1+1/p)^p = O(1)$. Recall, in this case, Algorithm 4.1 outputs $\mathbf{Z}\mathbf{Z}^\top$ where $\mathbf{Z}$ is an orthonormal basis for $\mathbf{A}^\top\mathbf{W}_1\mathbf{W}_1^\top$. Next, we invoke Lemma 4.6 to relate $\left\|\mathbf{A}^\top\mathbf{W}_1\mathbf{W}_1^\top\right\|_{\mathcal{S}_p}^p$ to $\left\|\mathbf{A}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p}^p$. Setting $\mathbf{Q} = \mathbf{Z}\mathbf{Z}^\top$ and $\mathbf{P} = \mathbf{W}_1\mathbf{W}_1^\top$, we observe that $\|\mathbf{PAQ}\|_{\mathcal{S}_p}^p = \left\|\mathbf{W}_1\mathbf{W}_1^\top\mathbf{A}\right\|_{\mathcal{S}_p}^p$ and $\|(\mathbf{I}-\mathbf{P})\,\mathbf{A}\,(\mathbf{I}-\mathbf{Q})\|_{\mathcal{S}_p}^p = \left\|\mathbf{A}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p}^p$. Then, invoking Lemma 4.6 and plugging in Equation (34), we have

$$\|(\mathbf{I}-\mathbf{P})\,\mathbf{A}\,(\mathbf{I}-\mathbf{Q})\|_{\mathcal{S}_p}^p \le (1 + O(p\epsilon))\,\|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p}^p, \tag{35}$$

where the last inequality follows from our assumption on the Schatten-p norm of the tail, given the case we are in. Taking the $(1/p)$-th root, and recalling that $\epsilon < 1/2$, we obtain

$$\left\|\mathbf{A}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p} \le (1 + O(\epsilon))\,\|\mathbf{A} - \mathbf{A}_k\|_p, \tag{36}$$

which concludes the final case.

Next, we analyze the running time and matrix-vector products. Running Algorithm 4.5 with block size k for $q = O\left(\log(d)p^{1/6}/\epsilon^{1/3}\right)$ iterations requires $O\left(\frac{\text{nnz}(\mathbf{A})kp^{1/6}\log(d)}{\epsilon^{1/3}}\right)$ time and $O\left(\frac{kp^{1/6}\log(d)}{\epsilon^{1/3}}\right)$ matrix-vector products. Similarly, running with block size $O\left(\frac{k}{(\epsilon p)^{\frac{1}{3}}}\right)$ for $q = O\left(\log(d/\epsilon)\sqrt{p}\right)$ iterations requires $O\left(\frac{\text{nnz}(\mathbf{A})kp^{1/6}\log(d/\epsilon)}{\epsilon^{1/3}}\right)$ time and $O\left(\frac{kp^{1/6}\log(d)}{\epsilon^{1/3}}\right)$ matrix-vector products. Finally, we observe that to obtain a $(1+1/p)$-approximation to σ_1 and σ_{k+1}, we need $O\left(\log(d)\sqrt{p}\right)$ iterations with blocksize $k+1$ and this requires $O\left(\log(d)\sqrt{p}k\right)$ matrix-vector products. Note, our setting of the exponent of p and ϵ was chosen to balance the two cases, and this concludes the proof.

□

5 QUERY LOWER BOUNDS

Next, we show that the ϵ-dependence obtained by our algorithms for Schatten-p low-rank approximation is optimal in the restricted computation model of matrix-vector products. The matrix-vector product model is defined as follows: given a matrix $\mathbf{A}$, our algorithm is allowed to make adaptive matrix-vector queries to $\mathbf{A}$, where one matrix-vector query is of the form $\mathbf{A}v$, for any $v \in \mathbb{R}^d$. Our lower bounds are information-theoretic and rely on the hardness of estimating the smallest eigenvalue of a Wishart ensemble, as established in recent work of Braverman, Hazan, Simchowitz and Woodworth [9].

We split the lower bounds into the case of $p \in [1,2]$ and $p > 2$. For $p \in [1,2]$, we have a simple argument based on the Araki-Lieb-Thirring inequality (Fact 3.10), whereas for $p > 2$, our lower bounds require an involved argument using a norm compression inequality for partitioned operators (Fact 3.14).

5.1 Lower Bounds for $p \in [1,2]$

The main lower bound we prove in this sub-section is as follows:

Theorem 5.1 (Query Lower Bound for $p \in [1,2]$). *Given $\varepsilon > 0$, and $p \in [1,2]$, there exists a distribution $\mathcal{D}$ over $n \times n$ matrices such that for $\mathbf{A} \sim \mathcal{D}$, any randomized algorithm that with probability at least $9/10$ outputs a rank-1 matrix $\mathbf{B}$ such that $\|\mathbf{A}-\mathbf{B}\|_{\mathcal{S}_p}^p \leq (1+\varepsilon)\|\mathbf{A}-\mathbf{A}_1\|_{\mathcal{S}_p}^p$ must make $\Omega(1/\varepsilon^{1/3})$ matrix-vector queries to $\mathbf{A}$.*

We require the following theorem on the hardness of computing the minimum eigenvalue of a Wishart Matrix, introduced recently by Braverman, Hazan, Simchowitz and Woodworth [9]:

Theorem 5.2 (Computing Min Eigenvalue of Wishart, Theorem 3.1 [9]). *Given $\epsilon \in (0,1)$, there exists a function $\mathbf{d} : (0,1) \to \mathbb{N}$ such that for all $d \geq \mathbf{d}(\epsilon)$, the following holds. Let $\mathbf{W} \sim$ Wishart(d) be a Wishart matrix and $\{\lambda_i\}_{i \in [d]}$ be the eigenvalues of $\mathbf{W}$, in descending order. Then, there exists a universal constant c^* such that:*

(1) *Let ζ_1 be the event that $\lambda_d(\mathbf{W}) \leq c_1/d^2$, ζ_2 be the event that $\lambda_{d-1}(\mathbf{W})-\lambda_d(\mathbf{W}) \geq c_2/d^2$ and ζ_3 be the event that $\|\mathbf{W}\|_{op} \leq 5$, where c_1 and c_2 are constants that depend only on ϵ. Then, $\Pr_{\mathbf{W}}[\zeta_1 \cap \zeta_2 \cap \zeta_3] \geq 1 - \frac{c^*\sqrt{\epsilon}}{2}$.*
(2) *Any randomized algorithm that makes at most $(1-\epsilon)d$ adaptive matrix-vector queries and outputs an estimate $\hat{\lambda}_d$ must satisfy*
$$\Pr_{\mathbf{W}}\left[\left|\hat{\lambda}_d - \lambda_d\right| \geq \frac{1}{4d^2}\right] \geq c^*\sqrt{\epsilon}.$$

We also use the following lemma from [9] bounding the minimum eigenvalue of a Wishart ensemble:

Lemma 5.3 (Non-Asymptotic Spectra of Wishart Ensembles, Corollary 3.3 [9]). *Let $\mathbf{W} \sim$ Wishart(n) be such that $n = \Omega(1/\varepsilon^3)$. Then, there exists a universal constant $c_2 > 0$ such that*
$$\Pr\left[\lambda_n(\mathbf{W}) \geq \frac{1}{n^2}\right] \geq c_2, \qquad \text{and} \qquad \Pr\left[\lambda_n(\mathbf{W}) < \frac{1}{2n^2}\right] \geq \frac{c_2}{2}.$$

We are now ready to prove Theorem 5.1. Our high level approach is to show that we can take any solution that is a $(1+\varepsilon)$-relative-error Schatten-p low-rank approximation to the hard instance $\mathbf{I} - \frac{1}{5}\mathbf{W}$, where $\mathbf{W}$ is a Wishart ensemble, and extract from it an accurate estimate of the minimum eigenvalue of $\mathbf{W}$, thus appealing to the hardness stated in (2) of Theorem 5.2 above.

Proof of Theorem 5.1. Let $n = \Theta\left(1/\epsilon^{1/3}\right)$ and let $\mathbf{A} = \mathbf{I} - \frac{1}{5}\mathbf{W}$ be an $n \times n$ instance where $\mathbf{W} \sim$ Wishart(n). Let ζ_1 be the event that $\|\mathbf{W}\|_{\text{op}} \leq 5$. It follows from Fact 3.16 that ζ_1 holds with probability at least $99/100$, and we condition on this event. Let ζ_2 be the event that $\lambda_n(\mathbf{W}) \geq \frac{1}{n^2} = \frac{\varepsilon^{2/3}}{c^*}$ and ζ_3 be the event that $\lambda_n(\mathbf{W}) < \frac{1}{2n^2} = \frac{\varepsilon^{2/3}}{2c^*}$.

Then, conditioning on ζ_2, we have that $1 - \frac{1}{5}\lambda_n(\mathbf{W}) \leq 1 - \frac{\varepsilon^{2/3}}{5c^*}$ and conditioning on ζ_3, we have that $1 - \frac{1}{5}\lambda_n(\mathbf{W}) \geq 1 - \frac{\varepsilon^{2/3}}{10c^*}$. We observe that for $p \in [1,2]$, using Bernoulli's inequality (Fact 3.6) we have
$$\left(1 - \frac{1}{5}\lambda_n(\mathbf{W})\right)^p \geq 1 - \frac{p}{5}\lambda_n(\mathbf{W})$$
and since $(1-x)^p \leq (1-x)$ for any $x \in (0,1)$, we also have that,
$$\left(1 - \frac{1}{5}\lambda_n(\mathbf{W})\right)^p \leq 1 - \frac{1}{5}\lambda_n(\mathbf{W})$$
Therefore, we can conclude, $\left(1 - \frac{1}{5}\lambda_n(\mathbf{W})\right)^p = 1 - \Theta(\lambda_n(\mathbf{W}))$. Further, it follows from part (1) of Fact 3.16 that $0 \preceq \mathbf{I} - \frac{1}{5}\mathbf{W} \preceq \mathbf{I}$, and thus
$$\|\mathbf{A}\|_{\mathcal{S}_p}^p = \sum_{i \in [n]} \lambda_i^p\left(\mathbf{I} - \frac{1}{5}\mathbf{W}\right) \leq \sum_{i \in [n]} \lambda_i\left(\mathbf{I} - \frac{1}{5}\mathbf{W}\right) \leq O\left(\frac{1}{\epsilon^{1/3}}\right) \tag{37}$$
where the last inequality follows from the fact that $n = \sqrt{c^*}/\epsilon^{1/3}$. Let $\mathbf{A}_1$ denote the best rank-1 approximation to $\mathbf{A}$. Then, it follows from Equation (37) that
$$\epsilon\|\mathbf{A}-\mathbf{A}_1\|_{\mathcal{S}_p}^p \leq \epsilon\|\mathbf{A}\|_{\mathcal{S}_p}^p \leq O\left(\epsilon^{2/3}\right) \tag{38}$$
Observe, any $(1+\epsilon)$-approximate relative-error Schatten-p low-rank approximation algorithm for $k = 1$ outputs a matrix $vv^\top$ such that
$$\left\|\mathbf{A}\left(\mathbf{I} - vv^\top\right)\right\|_{\mathcal{S}_p}^p \leq \|\mathbf{A}\|_{\mathcal{S}_p}^p - \|\mathbf{A}\|_{\text{op}}^p + \Theta(\epsilon^{2/3}) \tag{39}$$
By definition of the Schatten-p norm we have:
$$\begin{aligned}\left\|\mathbf{A}\left(\mathbf{I} - vv^\top\right)\right\|_{\mathcal{S}_p}^p &\geq \mathrm{Tr}\left(\left(\mathbf{I} - vv^\top\right)^p \mathbf{A}^p \left(\mathbf{I} - vv^\top\right)^p\right) \\ &= \|\mathbf{A}\|_{\mathcal{S}_p}^p - \mathrm{Tr}\left(\left(vv^\top\right)^{p/2}\left(\mathbf{A}^2\right)^{p/2}\left(vv^\top\right)^{p/2}\right) \\ &\geq \|\mathbf{A}\|_{\mathcal{S}_p}^p - \mathrm{Tr}\left(\left(vv^\top\mathbf{A}^2 vv^\top\right)^{p/2}\right) \\ &= \|\mathbf{A}\|_{\mathcal{S}_p}^p - \left\|\mathbf{A}vv^\top\right\|_{\mathcal{S}_p}^p\end{aligned} \tag{40}$$
where the first and last inequality follows from the reverse Araki-Lieb-Thirring inequality (Fact 3.10). Combining equations (39) and (40), we have that
$$\|\mathbf{A}\|_{\text{op}}^p \geq \|\mathbf{A}v\|_2^p \geq \|\mathbf{A}\|_{\text{op}}^p - \Theta(\epsilon^{2/3}) \tag{41}$$
Next, we observe that $\mathbf{A}v = (\mathbf{I} - 1/5\mathbf{W})\,v$ can be computed with one additional matrix-vector product and
$$\|\mathbf{A}\|_{\text{op}}^p = \left(1 - \frac{1}{5}\lambda_n(\mathbf{W})\right)^p = 1 - \frac{p}{5}\lambda_n(\mathbf{W}) + O\left(\lambda_n^2(\mathbf{W})\right) \tag{42}$$
Consider the estimator $\hat{\lambda}(\mathbf{W}) = \frac{5}{p}\left(1 - \left\|\left(\mathbf{I} - \frac{1}{5}\mathbf{W}\right)v\right\|_2^p\right)$. Combining equations (41) and (42), we can conclude
$$\hat{\lambda}(\mathbf{W}) = \lambda_{\min}(\mathbf{W}) \pm \Theta(\epsilon^{2/3}).$$
obtaining an additive error estimate to the minimum eigenvalue of $\mathbf{W}$ by computing an additional matrix-vector product. It follows that we satisfy conditions (1) and (2) in Theorem 5.2 and thus any algorithm for computing a rank-1 approximation to the matrix

$\mathbf{A} = \mathbf{I} - \frac{1}{5}\mathbf{W}$ in Schatten p norm must make at least $\frac{1}{\epsilon^{1/3}}$ queries to the aforementioned matrix, completing the proof. The claim follows from Theorem 5.2. ◻

5.2 Lower Bound for $p > 2$

We now consider the case when $p > 2$. We note that the previous approach no longer works since we cannot lower bound the cost of $\| (\mathbf{I} - \mathbf{W}/5) (\mathbf{I} - vv^\top) \|_{\mathcal{S}_p}$, as the Araki-Lieb-Thirring inequality reverses (see application in Equation 40). Therefore, we require a new approach, and appeal to a special case of Conjecture 2.2 that is known to be true, i.e. the Aligned Norm Compression inequality (see Fact 3.14). The main theorem we prove in this sub-section is as follows:

Theorem 5.4 (Query Lower Bound for $p > 2$). *Given $\varepsilon > 0$, and $p \geq 2$ such that $p = O(1)$, there exists a distribution $\mathcal{D}$ over $n \times n$ matrices such that for $\mathbf{A} \sim \mathcal{D}$, any randomized algorithm that with probability at least 99/100 outputs a unit vector u such that $\|\mathbf{A} - \mathbf{A}uu^\top\|_{\mathcal{S}_p}^p \leq (1+\varepsilon) \|\mathbf{A} - \mathbf{A}_1\|_{\mathcal{S}_p}^p$ must make $\Omega\left(1/\varepsilon^{1/3}\right)$ matrix-vector queries to $\mathbf{A}$.*

We first introduce a sequence of key lemmas required for our proof.

Corollary 5.5 (Special Case of Lemma 4.3). *Given $\gamma \in [0,1]$, a vector $v \in \mathbb{R}^d$ and an $n \times d$ matrix $\mathbf{A}$, let $t = \log(n/\gamma)/(c\sqrt{\gamma})$, for a fixed universal constant c. Then, there exists an algorithm that computes t matrix-vector products with $\mathbf{A}$ and outputs a unit vector u such that with probability at least 99/100,*

$$\|\mathbf{A}\|_{op}^2 - \|\mathbf{A}u\|_2^2 \leq O\left(\gamma\sigma_2^2\right).$$

where σ_2 is the second largest singular value of $\mathbf{A}$.

Next, we prove a key lemma relating the norm of a matrix to norms of orthogonal projections applied to the matrix. We note that this lemma is straight forward and holds for arbitrary vectors unit u, v if Conjecture 2.2 holds. However, we show that we can transform our matrix to have structure such that we can apply Fact 3.14 instead.

Lemma 5.6 (Orthogonal Projectors to Block Matrices). *Given an $n \times d$ matrix $\mathbf{A}$, $p > 2$ and unit vectors $u \in \mathbb{R}^d, v \in \mathbb{R}^n$, such that $\left(\mathbf{I} - vv^\top\right)\mathbf{A}uu^\top = 0$. Then, we have*

$$\|\mathbf{A}\|_{\mathcal{S}_p} \leq \left\| \begin{pmatrix} \|vv^\top \mathbf{A}uu^\top\|_{\mathcal{S}_p} & \|vv^\top \mathbf{A}\left(\mathbf{I} - uu^\top\right)\|_{\mathcal{S}_p} \\ 0 & \|\left(\mathbf{I} - vv^\top\right)\mathbf{A}\left(\mathbf{I} - uu^\top\right)\|_{\mathcal{S}_p} \end{pmatrix} \right\|_{\mathcal{S}_p}.$$

Fact 5.7 (SVD of a 2 × 2 Matrix). *Given a 2×2 matrix $\mathbf{M} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ let $\mathbf{U\Sigma V}^\top$ be the SVD of $\mathbf{M}$. Then,*

$$\Sigma_{1,1} = \sqrt{\frac{a^2+b^2+c^2+d^2+\sqrt{\left(a^2+b^2-c^2-d^2\right)^2+4\left(ac+bd\right)^2}}{2}},$$

and

$$\Sigma_{2,2} = \sqrt{\frac{a^2+b^2+c^2+d^2-\sqrt{\left(a^2+b^2-c^2-d^2\right)^2+4\left(ac+bd\right)^2}}{2}}.$$

Now, we are ready to prove Theorem 5.4.

Proof of Theorem 5.4. Let $\mathbf{A} = \mathbf{I} - \frac{1}{5}\mathbf{W}$ where $\mathbf{W}$ is an $n \times n$ Wishart matrix as in the proof of Theorem 5.1 and we have by hypothesis that there is an algorithm that with probability at least 99/100, outputs a unit vector u such that $\|\mathbf{A}\left(\mathbf{I} - uu^\top\right)\|_{\mathcal{S}_p}^p \leq (1+\varepsilon) \|\mathbf{A} - \mathbf{A}_1\|_{\mathcal{S}_p}^p$. Let $v = \mathbf{A}u/\|\mathbf{A}u\|_2$ and observe, $\left(\mathbf{I} - vv^\top\right)\mathbf{A}uu^\top = 0$. Further, by the unitary invariance of the Schatten-p norm,

$$\|vv^\top \mathbf{A}uu^\top\|_{\mathcal{S}_p} = |v^\top \mathbf{A}u| = \frac{|u^\top \mathbf{A}^\top \mathbf{A}u|}{\|\mathbf{A}u\|_2} = \|\mathbf{A}u\|_2. \tag{43}$$

Similarly,

$$\|vv^\top \mathbf{A}\left(\mathbf{I} - uu^\top\right)\|_{\mathcal{S}_p} = \sqrt{\|v^\top \mathbf{A}\left(\mathbf{I} - uu^\top\right)\|_2^2} \leq \epsilon^{1/3}\sigma_2 \tag{44}$$

where we use sub-multiplicativity of the ℓ_2 norm and Corollary 5.5 with $\gamma = \epsilon^{2/3}$. Note that we can assume w.l.o.g. that Corollary 5.5 holds since we can just iterate Block Krylov $q = (1/c\epsilon^{1/3})$ times, for a sufficiently large constant c, starting the iterations with the vector u output by the algorithm hypothesized for the theorem, and pay only $(1/c\epsilon^{1/3})$ extra matrix-vector products. Since $vv^\top \mathbf{A} + \mathbf{A}uu^\top - vv^\top \mathbf{A}uu^\top$ has rank at most 3,

$$\|\left(\mathbf{I} - vv^\top\right)\mathbf{A}\left(\mathbf{I} - uu^\top\right)\|_{\mathcal{S}_p}^p = \Omega\left(1/\epsilon^{1/3}\right), \tag{45}$$

where the last inequality follows from Fact 3.16.

Let $\mathbf{M} = \begin{pmatrix} \|vv^\top \mathbf{A}uu^\top\|_{\mathcal{S}_p} & \|vv^\top \mathbf{A}\left(\mathbf{I} - uu^\top\right)\|_{\mathcal{S}_p} \\ \|\left(\mathbf{I} - vv^\top\right)\mathbf{A}uu^\top\|_{\mathcal{S}_p} & \|\left(\mathbf{I} - vv^\top\right)\mathbf{A}\left(\mathbf{I} - uu^\top\right)\|_{\mathcal{S}_p} \end{pmatrix}^\top$. Then, it follows from Fact 5.7 that

$$\Sigma_{1,1}(\mathbf{M}) = \sqrt{c^2 + d^2 + \Theta\left(\frac{a^2c^2}{c^2+d^2-a^2}\right)}, \tag{46}$$

where we use that $b = 0$, $c, a \leq 1$ and $1 \ll d$ and the Taylor expansion of $\sqrt{x+y}$ for $x, y \geq 0$. Similarly,

$$\Sigma_{2,2}(\mathbf{M}) = \sqrt{a^2 - \Theta\left(\frac{a^2c^2}{c^2+d^2-a^2}\right)}. \tag{47}$$

Then, using equations (46) and (47) we can bound the Schatten-p norm of $\mathbf{M}$ as follows:

$$\|\mathbf{M}\|_{\mathcal{S}_p}^p \leq \left(1 + O\left(\epsilon^{2p/3}\right)\right) \|\mathbf{A} - \mathbf{A}_1\|_{\mathcal{S}_p}^p + \|\mathbf{A}u\|_2^p \tag{48}$$

It follows from Lemma 5.6, that $\|\mathbf{M}\|_{\mathcal{S}_p}^p \geq \|\mathbf{A}\|_{\mathcal{S}_p}^p$ and thus

$$\begin{aligned} \|\mathbf{A}u\|_2^p &\geq \|\mathbf{A}\|_{\mathcal{S}_p}^p - \left(1 + O\left(\epsilon^{2p/3}\right)\right) \|\mathbf{A} - \mathbf{A}_1\|_{\mathcal{S}_p}^p \\ &\geq \|\mathbf{A}\|_{\text{op}}^p - O\left(\varepsilon^{2/3}\right) \end{aligned} \tag{49}$$

where the second to last inequality follows from recalling $p \geq 2$. The remainder of the proof is as in that following (41) in the proof of Theorem 5.1. ◻

ACKNOWLEDGMENTS

A. Bakshi and D. Woodruff would like to thank the National Science Foundation under Grant No. CCF-1815840, Office of Naval Research (ONR) grant N00014-18-1-2562, and a Simons Investigator Award. The authors also thank Praneeth Kacham for pointing out an error in a previous version, and anonymous reviewers for their careful reading of our manuscript and for several suggestions.

REFERENCES

[1] Huzihiro Araki. 1990. On an inequality of Lieb and Thirring. *LMaPh* 19, 2 (1990), 167–170.

[2] Koenraad MR Audenaert. 2008. On a norm compression inequality for 2× N partitioned block matrices. *Linear algebra and its applications* 428, 4 (2008), 781–795.

[3] Koenraad MR Audenaert and Fuad Kittaneh. 2012. Problems and conjectures in matrix and operator inequalities. *arXiv preprint arXiv:1201.5232* (2012).

[4] Haim Avron. 2010. Counting triangles in large graphs using randomized matrix trace estimation. In *Workshop on Large-scale Data Mining: Theory and Applications*, Vol. 10. 10–9.

[5] Haim Avron, Kenneth L. Clarkson, and David P. Woodruff. 2017. Sharper Bounds for Regularized Data Fitting. arXiv:1611.03225 [cs.DS]

[6] Zhaojun Bai, Gark Fahey, and Gene Golub. 1996. Some large-scale matrix computation problems. *J. Comput. Appl. Math.* 74, 1-2 (1996), 71–89.

[7] Rajendra Bhatia. 2013. *Matrix analysis*. Vol. 169. Springer Science & Business Media.

[8] Rajendra Bhatia, William Kahan, and Ren-Cang Li. 2002. Pinchings and norms of scaled triangular matrices. *Linear and Multilinear Algebra* 50, 1 (2002), 15–21.

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

[10] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. 2011. Robust principal component analysis? *Journal of the ACM (JACM)* 58, 3 (2011), 1–37.

[11] Emmanuel J Candes and Yaniv Plan. 2010. Matrix completion with noise. *Proc. IEEE* 98, 6 (2010), 925–936.

[12] Emmanuel J Candès and Benjamin Recht. 2009. Exact matrix completion via convex optimization. *Foundations of Computational mathematics* 9, 6 (2009), 717–772.

[13] Kenneth L Clarkson and David P Woodruff. 2009. Numerical linear algebra in the streaming model. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*. ACM, 205–214.

[14] Kenneth L Clarkson and David P Woodruff. 2013. Low rank approximation and regression in input sparsity time. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. ACM, 81–90.

[15] Michael B Cohen. 2016. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 278–287.

[16] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. 2019. An Investigation into Neural Net Optimization via Hessian Eigenvalue Density. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 2232–2241. http://proceedings.mlr.press/v97/ghorbani19b.html

[17] Shuhang Gu, Qi Xie, Deyu Meng, Wangmeng Zuo, Xiangchu Feng, and Lei Zhang. 2017. Weighted nuclear norm minimization and its applications to low level vision. *International journal of computer vision* 121, 2 (2017), 183–208.

[18] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. 2014. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2862–2869.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[20] Ravi Kannan and Santosh S. Vempala. 2009. Spectral Algorithms. *Found. Trends Theor. Comput. Sci.* 4, 3-4 (2009), 157–288.

[21] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).

[22] Yi Li, Huy L. Nguyen, and David P. Woodruff. 2014. Turnstile streaming algorithms might as well be linear sketches. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*. 174–183.

[23] Yi Li and David P. Woodruff. 2020. Input-Sparsity Low Rank Approximation in Schatten Norm. *CoRR* abs/2004.12646 (2020).

[24] Jörg Liesen and Zdenek Strakos. 2013. *Krylov subspace methods: principles and analysis*. Oxford University Press.

[25] Philip J Maher. 1990. Some operator inequalities concerning generalized inverses. *Illinois Journal of Mathematics* 34, 3 (1990), 503–514.

[26] Michael W. Mahoney. 2011. Randomized Algorithms for Matrices and Data. *Found. Trends Mach. Learn.* 3, 2 (2011), 123–224.

[27] John C Mason and David C Handscomb. 2002. *Chebyshev polynomials*. CRC press.

[28] Xiangrui Meng and Michael W Mahoney. 2013. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. ACM, 91–100.

[29] Raphael A. Meyer, Cameron Musco, Christopher Musco, and David P. Woodruff. 2021. Hutch++: Optimal Stochastic Trace Estimation. In *4th Symposium on Simplicity in Algorithms, SOSA 2021, Virtual Conference, January 11-12, 2021*. 142–155.

[30] Cameron Musco and Christopher Musco. 2015. Randomized block Krylov methods for stronger and faster approximate singular value decomposition. In *Advances in Neural Information Processing Systems*. 1396–1404.

[31] Cameron Musco and David Woodruff. 2017. Is input sparsity time possible for kernel low-rank approximation? *Advances in Neural Information Processing Systems* 30 (2017), 4435–4445.

[32] S. Muthukrishnan. 2005. Data Streams: Algorithms and Applications. *Found. Trends Theor. Comput. Sci.* 1, 2 (2005).

[33] Jelani Nelson and Huy L. Nguyen. 2013. OSNAP: Faster Numerical Linear Algebra Algorithms via Sparser Subspace Embeddings. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*. 117–126.

[34] Jelani Jelani Osei Nelson. 2011. *Sketching and streaming high-dimensional vectors*. Ph. D. Dissertation. Massachusetts Institute of Technology.

[35] Barak A. Pearlmutter. 1994. Fast Exact Multiplication by the Hessian. *Neural Computation* 6 (1994), 147–160.

[36] William Peebles, John Peebles, Jun-Yan Zhu, Alexei A. Efros, and Antonio Torralba. 2020. The Hessian Penalty: A Weak Prior for Unsupervised Disentanglement. In *Proceedings of European Conference on Computer Vision (ECCV)*.

[37] Cyrus Rashtchian, David P. Woodruff, Peng Ye, and Hanlin Zhu. 2021. Average-Case Communication Complexity of Statistical Problems.

[38] Cyrus Rashtchian, David P. Woodruff, and Hanlin Zhu. 2020. Vector-Matrix-Vector Queries for Solving Linear Algebra, Statistics, and Graph Problems. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2020, August 17-19, 2020, Virtual Conference*. 26:1–26:20. https://doi.org/10.4230/LIPIcs.APPROX/RANDOM.2020.26

[39] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. 2010. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review* 52, 3 (2010), 471–501.

[40] Theodore J Rivlin. 2020. *Chebyshev polynomials*. Courier Dover Publications.

[41] Sam T Roweis and Lawrence K Saul. 2000. Nonlinear dimensionality reduction by locally linear embedding. *science* 290, 5500 (2000), 2323–2326.

[42] Yousef Saad. 1981. Krylov subspace methods for solving large unsymmetric linear systems. *Mathematics of computation* 37, 155 (1981), 105–126.

[43] Robert Schatten. 1960. Norm ideals of completely continuous operators. (1960).

[44] Max Simchowitz, Ahmed El Alaoui, and Benjamin Recht. 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, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*. 1249–1259.

[45] Xiaoming Sun, David P. Woodruff, Guang Yang, and Jialin Zhang. 2019. Querying a Matrix Through Matrix-Vector Products. In *46th International Colloquium on Automata, Languages, and Programming, ICALP 2019, July 9-12, 2019, Patras, Greece*. 94:1–94:16. https://doi.org/10.4230/LIPIcs.ICALP.2019.94

[46] Terence Tao. 2012. *Topics in random matrix theory*. Vol. 132. American Mathematical Soc.

[47] Terence Tao. 2020. Notes 3a: Eigenvalues and sums of Hermitian matrices.

[48] Joshua B Tenenbaum, Vin De Silva, and John C Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *science* 290, 5500 (2000), 2319–2323.

[49] Charalampos E Tsourakakis. 2008. Fast counting of triangles in large real networks without counting: Algorithms and laws. In *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 608–617.

[50] John Von Neumann. 1937. *Some matrix-inequalities and metrization of matric space*.

[51] Karl Wimmer, Yi Wu, and Peng Zhang. 2014. Optimal Query Complexity for Estimating the Trace of a Matrix. In *Automata, Languages, and Programming - 41st International Colloquium, ICALP 2014, Copenhagen, Denmark, July 8-11, 2014, Proceedings, Part I*. 1051–1062.

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

[53] Huan Xu, Constantine Caramanis, and Sujay Sanghavi. 2010. Robust PCA via outlier pursuit. *arXiv preprint arXiv:1010.4237* (2010).

[54] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael Mahoney. 2020. PyHessian: Neural Networks Through the Lens of the Hessian. arXiv:1912.07145 [cs.LG]

[55] Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. 2016. Fast algorithms for robust PCA via gradient descent. In *Advances in neural information processing systems*. 4152–4160.

[56] Ming Yuan and Cun-Hui Zhang. 2016. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics* 16, 4 (2016), 1031–1068.