

Robust and Sample Optimal Algorithms for PSD Low-Rank Approximation

Ainesh Bakshi *
CMU
abakshi@cs.cmu.edu

Nadiia Chepurko
MIT
nadiia@mit.edu

David P. Woodruff*
CMU
dwoodruf@cs.cmu.edu

Abstract

Recently, Musco and Woodruff (FOCS, 2017) showed that given an $n \times n$ positive semidefinite (PSD) matrix \mathbf{A} , it is possible to compute a $(1 + \epsilon)$ -approximate relative-error low-rank approximation to \mathbf{A} by querying $\tilde{O}(nk/\epsilon^{2.5})$ entries of \mathbf{A} in time $\tilde{O}(nk/\epsilon^{2.5} + nk^{\omega-1}/\epsilon^{2(\omega-1)})$. They also showed that any relative-error low-rank approximation algorithm must query $\Omega(nk/\epsilon)$ entries of \mathbf{A} , this gap has since remained open. Our main result is to resolve this question by obtaining an optimal algorithm that queries $\tilde{O}(nk/\epsilon)$ entries of \mathbf{A} and outputs a relative-error low-rank approximation in $\tilde{O}(n \cdot (k/\epsilon)^{\omega-1})$ time. Note, our running time improves that of Musco and Woodruff, and matches the information-theoretic lower bound if the matrix-multiplication exponent ω is 2.

We then extend our techniques to negative-type distance matrices. Here, our input is a pair-wise distance matrix \mathbf{A} corresponding to a point set $\mathcal{P} = \{x_1, x_2, \dots, x_n\}$ such that $\mathbf{A}_{i,j} = \|x_i - x_j\|_2^2$. Bakshi and Woodruff (NeurIPS, 2018) showed a bi-criteria, relative-error low-rank approximation for negative-type metrics. Their algorithm queries $\tilde{O}(nk/\epsilon^{2.5})$ entries and outputs a rank- $(k + 4)$ matrix. We show that the bi-criteria guarantee is not necessary and obtain an $\tilde{O}(nk/\epsilon)$ query algorithm, which is optimal. Our algorithm applies to all distance matrices that arise from metrics satisfying negative-type inequalities, including ℓ_1 , ℓ_2 , spherical metrics, hypermetrics and effective resistances on a graph. We also obtain faster algorithms for ridge regression.

Next, we introduce a new robust low-rank approximation model which captures PSD matrices that have been corrupted with noise. We assume that the Frobenius norm of the corruption is bounded. Here, we relax the notion of approximation to additive-error, since it is information-theoretically impossible to obtain a relative-error approximation in this setting. While a sample complexity lower bound precludes sublinear algorithms for arbitrary PSD matrices, we provide the first sublinear time and query algorithms when the corruption on the diagonal entries is bounded. As a special case, we show sample-optimal sublinear time algorithms for low-rank approximation of correlation matrices corrupted by noise.

*Ainesh Bakshi and David Woodruff acknowledge support in part from NSF No. CCF-1815840.

Contents

1	Introduction	1
1.1	Our Results	2
2	Technical Overview	5
2.1	Sample-Optimal Low-Rank Approximation	5
2.2	Robust Low-Rank Approximation	11
3	Preliminaries and Notation	15
4	Relative Error PSD Low-Rank Approximation	16
4.1	Structured Projections to Low-Rank Approximation	20
4.2	Spectral Regression	25
4.3	Sample-Optimal Algorithm	35
4.4	Negative-Type Distances	39
4.5	Ridge Regression	41
5	Robust Low-Rank Approximation	45
5.1	Lower Bound for Robust PSD Low-Rank Approximation	45
5.2	Robust Sublinear Low-Rank Approximation Algorithms	48

1 Introduction

Low-rank approximation is one of the most common dimensionality reduction techniques, whereby one replaces a large matrix \mathbf{A} with a low-rank factorization $\mathbf{U} \cdot \mathbf{V} \approx \mathbf{A}$. Such a factorization provides a compact way of storing \mathbf{A} and allows one to multiply \mathbf{A} quickly by a vector. It is used as an algorithmic primitive in clustering [DFK⁺04, McS01], recommendation systems [DKR02], web search [AFKM01, Kle99], and learning mixtures of distributions [AM05, KSV05], and has numerous other applications.

A large body of recent work has looked at *relative-error* low-rank approximation, whereby given an $n \times n$ matrix \mathbf{A} , an accuracy parameter $\epsilon > 0$, and a rank parameter k , one seeks to output a rank- k matrix \mathbf{B} for which

$$\|\mathbf{A} - \mathbf{B}\|_F^2 \leq (1 + \epsilon) \|\mathbf{A} - \mathbf{A}_k\|_F^2, \quad (1.1)$$

where for a matrix \mathbf{C} , $\|\mathbf{C}\|_F^2 = \sum_{i,j} \mathbf{C}_{i,j}^2$, and \mathbf{A}_k denotes the best rank- k approximation to \mathbf{A} in Frobenius norm. \mathbf{A}_k can be computed exactly using the singular value decomposition, but takes time $O(n^\omega)$, where ω is the matrix multiplication constant. We refer the reader to the survey [Woo14] and references therein.

For worst-case matrices, it is not hard to see that any algorithm achieving (1.1) must spend at least $\tilde{\Omega}(\text{nnz}(\mathbf{A}))$ time, where $\text{nnz}(\mathbf{A})$ denotes the number of non-zero entries (sparsity) of \mathbf{A} . Indeed, without reading most of the non-zero entries of \mathbf{A} , one could fail to read a single large entry, thus making one's output matrix \mathbf{B} an arbitrarily bad approximation.

A flurry of recent work [KP16, MW17, BW18, CLW18, Tan19, RSML18, GLT18, IVWW19, SW19, GSLW19] has looked at the possibility of achieving *sublinear* time algorithms (classical and quantum) for low-rank approximation. In particular, Musco and Woodruff [MW17] consider the important case of positive-semidefinite (PSD) matrices. PSD matrices include as special cases covariance matrices, correlation matrices, graph Laplacians, kernel matrices and random dot product models. Further, the special case where the input itself is low-rank (PSD Matrix Completion) has applications in quantum state tomography [GLF⁺10]. Subsequently, Bakshi and Woodruff [BW18] considered low-rank approximation of the closely related family of Negative-type (Euclidean Squared) distance matrices. Negative-type metrics include as special cases ℓ_1 and ℓ_2 metrics, spherical metrics and hypermetrics, as well as effective resistances in graphs [DL09, TD87, CRR⁺96, CKM⁺11]. Negative-type metrics have found various applications in algorithm design and optimization [ALN08, SS11, KMP14, MST15].

Musco and Woodruff show that it is possible to output a low-rank matrix \mathbf{B} in factored form achieving (1.1) in $\tilde{O}(nk/\epsilon^{2.5} + nk^{\omega-1}/\epsilon^{2(\omega-1)})$ time, while reading only $\tilde{O}(nk/\epsilon^{2.5})$ entries of \mathbf{A} . They also showed a lower bound that any algorithm achieving (1.1) must read $\tilde{\Omega}(nk/\epsilon)$ entries, and closing the gap between these bounds has remained an open question. Similarly, Bakshi and Woodruff exploit the structure of Negative-type metrics to reduce to the PSD case and obtain a bi-criteria algorithm that requires $\tilde{O}(nk/\epsilon^{2.5})$ queries. The gap in the sample complexity and the requirement of a bi-criteria guarantee remained open. We resolve these questions here, and describe our novel technical contributions in Section 2.

Next we consider PSD matrices that have been corrupted by a small amount of noise. A drawback of algorithms achieving (1.1) is that they cannot tolerate any amount of unstructured noise. For instance, if one slightly corrupts a few off-diagonal entries, making the input matrix \mathbf{A} no longer PSD, then it is impossible to detect such corruptions in sublinear time, making the relative-error guarantee (1.1) information-theoretically impossible. Motivated by this, we also introduce a new framework where an adversary corrupts the input by adding a noise matrix \mathbf{N} to a psd matrix \mathbf{A} . We assume that the Frobenius norm of the corruption is bounded relative to the Frobenius norm of \mathbf{A} , i.e., $\|\mathbf{N}\|_F^2 \leq \eta \|\mathbf{A}\|_F^2$. We also assume the corruption is well-spread, i.e., each row of \mathbf{N} has ℓ_2^2 -norm at most a fixed constant factor larger than ℓ_2^2 -norm of the corresponding row of \mathbf{A} .

This model captures small perturbations to PSD matrices that we may observe in real-world datasets, as a consequence of round-off or numerical errors in tasks such as computing Laplacian pseudoinverses, and systematic measurement errors when computing a covariance matrix. One important application captured by our model is low-rank approximation of corrupted *correlation matrices*. Finding a low-rank approximation of such matrices occurs when measured correlations are asynchronous or incomplete, or when models are stress-tested by adjusting individual correlations. Low-rank approximation of correlation matrices also has many applications in finance [Hig02].

Given that it is information-theoretically impossible to obtain the relative-error guarantee (1.1) in the *robust model*, we relax our notion of approximation to the following well-studied additive-error guarantee:

$$\|\mathbf{A} - \mathbf{B}\|_F^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + (\epsilon + \eta) \|\mathbf{A}\|_F^2. \quad (1.2)$$

This additive-error guarantee was introduced by the seminal work of Frieze et. al. [FKV04], and triggered a long line of work on low-rank approximation from a computational perspective. Frieze et al. showed that it is possible to achieve (1.2) in $O(\text{nnz}(\mathbf{A}))$ time. Further, given access to an oracle for computing row norms of \mathbf{A} , 1.2 is achievable in sublinear time. More recently, the same notion of approximation was used to obtain sublinear sample complexity and running time algorithms for *distance matrices* [BW18, IVWW19], and a quantum algorithm for recommendation systems [KP16], which was subsequently dequantized [Tan19].

This raises the question of how robust are our sublinear low-rank approximation algorithms for structured matrices, if we relax to additive-error guarantees and allow for corruption. In particular, can we obtain additive-error low-rank approximation algorithms for PSD matrices that achieve sublinear time and sample complexity in the presence of noise? We characterize when such robust algorithms are achievable in sublinear time.

1.1 Our Results

We begin with stating our results for low-rank approximation for structured matrices. Our main result is an optimal algorithm for low-rank approximation of PSD matrices:

Theorem 4.1 (*Informal Sample-Optimal PSD LRA.*) *Given a PSD matrix \mathbf{A} , there exists an algorithm*

Problem	Prior Work		Our Results		Query Lower Bound
	Query	Run Time	Query	Run Time	
PSD LRA	$O\left(\frac{nk}{\epsilon^{2.5}}\right)$	$O\left(\frac{nk^{\omega-1}}{\epsilon^{2\omega-2}} + \frac{nk}{\epsilon^{2.5}}\right)$	$O^*\left(\frac{nk}{\epsilon}\right)$	$O^+\left(\frac{nk^{\omega-1}}{\epsilon^{\omega-1}}\right)$	$\Omega\left(\frac{nk}{\epsilon}\right)$
	[MW17]		Thm. 4.1		[MW17]
PSD LRA PSD Output	$O\left(nk\left(\frac{k}{\epsilon^2} + \frac{1}{\epsilon^3}\right)\right)$	$O\left(nk^{\omega-1}\left(\frac{k}{\epsilon^{\omega}} + \frac{1}{\epsilon^{3\omega-3}}\right)\right)$	$O^*\left(\frac{nk}{\epsilon}\right)$	$O^+\left(\frac{nk^{\omega-1}}{\epsilon^{\omega-1}}\right)$	$\Omega\left(\frac{nk}{\epsilon}\right)$
	[MW17]		Thm. 4.1		[MW17]
Negative-Type LRA	$O\left(\frac{nk}{\epsilon^{2.5}}\right)$	$O\left(\frac{nk^{\omega-1}}{\epsilon^{2\omega-2}} + \frac{nk}{\epsilon^{2.5}}\right)$	$O^*\left(\frac{nk}{\epsilon}\right)$	$O^+\left(\frac{nk^{\omega-1}}{\epsilon^{\omega-1}}\right)$	$\Omega\left(\frac{nk}{\epsilon}\right)$
	Bi-criteria, [BW18]		No Bi-criteria, Thm. 4.28		[BW18]
Coreset Ridge Regression	$O\left(\frac{ns_{\lambda}^2}{\epsilon^4}\right)$	$O\left(\frac{ns_{\lambda}^{\omega}}{\epsilon^{\omega}}\right)$	$O^*\left(\frac{ns_{\lambda}}{\epsilon^2}\right)$	$O^+\left(\frac{ns_{\lambda}^{\omega-1}}{\epsilon^{2\omega-2}}\right)$	$\Omega\left(\frac{ns_{\lambda}}{\epsilon^2}\right)$
	[MW17]		Thm. 4.31		Thm 4.33

Table 1: Comparison with prior work. The notation O^* and O^+ represent existence of matching lower bounds for query complexity and running time (assuming the fast matrix multiplication exponent ω is 2) respectively. The notation s_{λ} is used to denote the statistical dimension of ridge regression. All bounds are stated ignoring polylogarithmic factors in n, k and ϵ .

that queries $\tilde{O}(nk/\epsilon)$ entries in \mathbf{A} and outputs a rank k matrix \mathbf{B} such that with probability 99/100, $\|\mathbf{A} - \mathbf{B}\|_F^2 \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2$, and the algorithm runs in time $\tilde{O}(n \cdot (k/\epsilon)^{\omega-1})$.

Remark 1.1. Our algorithm matches the sample complexity lower bound of Musco and Woodruff, up to logarithmic factors, which shows that any randomized algorithm that outputs a $(1+\epsilon)$ -relative-error low-rank approximation for a PSD matrix \mathbf{A} must read $\tilde{\Omega}(nk/\epsilon)$ entries. Our running time also improves that of Musco and Woodruff and is optimal if the matrix multiplication exponent ω is 2.

Remark 1.2. We can extend our algorithm such that the low-rank matrix \mathbf{B} we output is also PSD with the same query complexity and running time. In comparison, the algorithm of Musco and Woodruff accesses $\tilde{O}(nk/\epsilon^3 + nk^2/\epsilon^2)$ entries in \mathbf{A} and runs in time $\tilde{O}(n(k/\epsilon)^{\omega} + nk^{\omega-1}/\epsilon^{3(\omega-1)})$.

At the core of our analysis is a sample optimal algorithm for Spectral Regression: $\min_{\mathbf{X}} \|\mathbf{D}\mathbf{X} - \mathbf{E}\|_2^2$. We show that when \mathbf{D} has orthonormal columns and \mathbf{E} is arbitrary, we can sketch the problem by sampling rows proportional to the leverage scores of \mathbf{D} and approximately preserve the minimum cost. This is particularly surprising since our sketch only computes sampling probabilities by reading entries in \mathbf{D} , while being completely agnostic to the entries in \mathbf{E} . Here, we also prove a spectral approximate matrix product guarantee for our one-sided leverage score sketch, which may be of independent interest. We note that such a guarantee for leverage score sampling does not appear in prior work, and we discuss the technical challenges we need to overcome in the subsequent section.

The techniques we develop for PSD low-rank approximation also extend to computing a low-rank approximation for distance matrices that arise from negative-type (Euclidean-squared) metrics. Here, our input is a pair-wise distance matrix \mathbf{A} corresponding to a point set $\mathcal{P} = \{x_1, x_2, \dots, x_n\} \in \mathbb{R}^d$ such that $\mathbf{A}_{i,j} = \|x_i - x_j\|_2^2$. We obtain an optimal algorithm for computing a low-rank approximation of such matrices:

Theorem 4.28 (Informal Sample-Optimal LRA for Negative-Type Metrics.) *Given a negative-type distance matrix \mathbf{A} , there exists an algorithm that queries $\tilde{O}(nk/\epsilon)$ entries in \mathbf{A} and outputs a rank k matrix \mathbf{B} such that with probability 99/100, $\|\mathbf{A} - \mathbf{B}\|_F^2 \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2$, and the algorithm runs in time $\tilde{O}(n \cdot (k/\epsilon)^{\omega-1})$.*

Remark 1.3. Prior work of Bakshi and Woodruff [BW18] obtains a $\tilde{O}(nk/\epsilon^{2.5})$ query algorithm that outputs a rank- $(k + 4)$ matrix \mathbf{B} such that $\|\mathbf{A} - \mathbf{B}\|_F^2 \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2$. We show that the bi-criteria guarantee is not necessary, thereby resolving an open question in their paper.

Structured Regression. The sample-optimal algorithm for PSD Low-Rank Approximation also leads to a faster algorithm for Ridge Regression, when the design matrix is PSD. Given a PSD matrix \mathbf{A} , a vector y and a regularization parameter λ , we consider the following optimization problem: $\min_{x \in \mathbb{R}^n} \|\mathbf{A}x - y\|_2^2 + \lambda \|x\|_2^2$. This problem is often referred to as Ridge Regression and has been the focus of numerous theoretical and practical works (see [Gru17] and references therein).

Theorem 4.31 (Informal Ridge Regression.) *Given a PSD matrix \mathbf{A} , a regularization parameter λ and statistical dimension $s_\lambda = \text{Tr}((\mathbf{A}^2 + \lambda \mathbf{I})^{-1} \mathbf{A}^2)$, there exists an algorithm that queries $\tilde{O}(ns_\lambda/\epsilon^2)$ entries of \mathbf{A} and with probability 99/100 outputs a $(1 + \epsilon)$ approximate solution to the Ridge Regression objective and runs in $\tilde{O}(n(s_\lambda/\epsilon^2)^{\omega-1})$ time.*

Remark 1.4. Our result improves on prior work by Musco and Woodruff [MW17], who obtain an algorithm that queries $\tilde{O}(ns_\lambda^2/\epsilon^4)$ entries in \mathbf{A} and runs in $\tilde{O}(n(s_\lambda/\epsilon^2)^\omega)$ time.

Remark 1.5. Since our algorithm works for all y simultaneously, we obtain a low-rank *coreset* of the design matrix (in factored form) that preserves the Ridge Regression cost up to a $(1 + \epsilon)$ factor. Further, in Theorem 4.33, we prove a matching lower bound on the query complexity for any coreset construction.

Robust Low-Rank Approximation. Next, we consider a robust form of low-rank approximation problem, where the input is a PSD matrix corrupted by noise. In this setting, we have query access to the corrupted matrix $\mathbf{A} + \mathbf{N}$, where \mathbf{A} is PSD and \mathbf{N} is such that $\|\mathbf{N}\|_F^2 \leq \eta \|\mathbf{A}\|_F^2$. Further, for all $i \in [n]$ $\|\mathbf{N}_{i,*}\|_2^2 \leq c \|\mathbf{A}_{i,*}\|_2^2$, for a fixed constant c . The diagonal of a PSD matrix carries crucial information since the largest diagonal entry upper bounds all off-diagonal entries. Therefore, a reasonable adversarial strategy is to corrupt the largest diagonal entries and make them close to the small diagonal entries, which enables the resulting matrix to have large off-diagonal entries that are hard to find. Capturing this intuition we parameterize our algorithms and lower bounds by the largest ratio between a diagonal entry of \mathbf{A} and $\mathbf{A} + \mathbf{N}$, denoted by $\phi_{\max} = \max_{j \in [n]} \mathbf{A}_{j,j} / |(\mathbf{A} + \mathbf{N})_{j,j}|$.

Theorem 5.4. (Informal lower bound.) *Let $\epsilon > \eta > 0$. Given $\mathbf{A} + \mathbf{N}$ such that \mathbf{A} is PSD and \mathbf{N} is a corruption matrix as defined above, any randomized algorithm that with probability at least 2/3 outputs a rank- k approximation up to additive error $(\epsilon + \eta)\|\mathbf{A}\|_F^2$ must read $\tilde{\Omega}(\phi_{\max}^2 nk/\epsilon)$ entries of $\mathbf{A} + \mathbf{N}$.*

Remark 1.6. Any algorithm must incur additive error $\eta \|\mathbf{A}\|_F^2$, since \mathbf{A} is not even identifiable below additive-error $\eta \|\mathbf{A}\|_F^2$.

Remark 1.7. In our hard instance, ϕ_{\max}^2 can be as large as $\epsilon n/k$, which implies a sample-complexity lower bound of $\tilde{\Omega}(n^2)$. While this lower bound precludes sublinear algorithms for arbitrary PSD matrices, we observe that in many applications ϕ_{\max} can be significantly smaller. For instance, if \mathbf{A} is a correlation matrix, we know that the true diagonal entries of $\mathbf{A} + \mathbf{N}$ are 1 and can ignore any corruption on them to bound ϕ_{\max} by 1.

Motivated by the aforementioned observation, we introduce algorithms for robust low-rank approximation, parameterized by the corruption on the diagonal entries. We obtain the following theorem:

Theorem 5.11 (*Informal Robust LRA.*) *Given $\mathbf{A} + \mathbf{N}$, which satisfies our noise model, there exists an algorithm that queries $\tilde{O}(\phi_{\max}^2 nk/\epsilon)$ entries in $\mathbf{A} + \mathbf{N}$ and computes a rank k matrix \mathbf{B} such that with probability at least $99/100$, $\|\mathbf{A} - \mathbf{B}\|_F^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + (\epsilon + \sqrt{\eta})\|\mathbf{A}\|_F^2$.*

Remark 1.8. While the sample complexity of this algorithm matches the sample complexity in the lower bound, it incurs additive-error $\sqrt{\eta}\|\mathbf{A}\|_F^2$ as opposed to $\eta\|\mathbf{A}\|_F^2$. An interesting open question here is whether we can achieve additive-error $o(\sqrt{\eta}\|\mathbf{A}\|_F^2)$, though we note that when $\eta^2 \leq \epsilon$, this just changes the additive error guarantee of our low-rank approximation by a constant factor.

Remark 1.9. Our techniques extend to low-rank approximation of correlation matrices, and we obtain a sample complexity of $\tilde{O}(nk/\epsilon)$, which is optimal. In fact, the hard instance in [MW17] implies an $\tilde{\Omega}(nk/\epsilon)$ lower bound on the sample complexity, even in the presence of no noise. Surprisingly, corrupting a correlation matrix does not increase the sample complexity and only incurs an additive error of $\sqrt{\eta}\|\mathbf{A}\|_F^2$ (see Corollary 5.18 for a formal statement).

2 Technical Overview

In this section, we provide an overview of our techniques and supply intuition for our proofs. As a first step, it is easy to see that the $\Omega(\text{nnz}(\mathbf{A}))$ lower bound for general matrices does not apply to PSD matrices, since it proceeds by hiding arbitrarily large entries. Observe that, reading the diagonal of a PSD matrix certifies an upper bound on all entries of the matrix and thus off-diagonals cannot be arbitrarily large. With this intuition in mind, we focus on sublinear algorithms.

2.1 Sample-Optimal Low-Rank Approximation

At a high level, our algorithm consists of two stages: first, we use the existing machinery developed by Musco and Woodruff [MW17] to obtain *weak projection-cost preserving sketches* for \mathbf{A} . Our sketches are smaller than those obtained by Musco and Woodruff, albeit satisfying a weaker guarantee. Recall that such sketches reduce the dimensionality of the column (row) space, while ensuring that the norm of *all* low-rank projections in the orthogonal complement of the column (row) space are simultaneously preserved. Constructing such sketches for both the column and row spaces of \mathbf{A} results in a much smaller matrix, which we can afford to query.

At this point our approach diverges from that of Musco and Woodruff, since it is not possible to follow their strategy and recover a $(1 + \epsilon)$ -relative-error low-rank approximation from the weaker

sketch we constructed above. However, we show that our sketch has enough information to extract a *structured subspace* (represented by an orthonormal basis) for \mathbf{A} such that the projection onto the orthogonal complement of this subspace is comparable to the optimal low-rank approximation cost, in spectral norm. Note, this guarantee is stronger than the span of the structured subspace containing a low-rank approximation comparable to the optimal in Frobenius norm, and indeed the latter does not suffice. In the second stage, we show that we can recover a rank- k matrix in the span of the structured subspace such that it is a $(1 + \epsilon)$ -relative-error low-rank approximation for \mathbf{A} . Further, we show that all these steps can be performed in sublinear time and by reading only $\tilde{O}(nk/\epsilon)$ entries in \mathbf{A} (see Theorem 4.1 for a precise statement).

We begin by providing a bird's eye view of the Musco-Woodruff algorithm and how to adapt parts of it to obtain *weak projection-cost preserving sketches (PCPs)*. For ease of exposition, we ignore polylogarithmic factors in the subsequent discussion. Their algorithm begins with computing the so-called *ridge leverage scores* (Definition 4.7) for $\mathbf{A}^{1/2}$, which approximate the *ridge leverage scores* of \mathbf{A} up to a $\sqrt{n/k}$ -factor. The ridge leverage scores of $\mathbf{A}^{1/2}$ can be approximated efficiently since we can compute the row norms of $\mathbf{A}^{1/2}$ by simply reading the diagonal of \mathbf{A} . It is well known [CMM17] that sampling k/ϵ_0^2 columns of \mathbf{A} proportional to its ridge leverage scores results in a sketch \mathbf{C} that preserves the cost of all rank- k projections \mathbf{P} :

$$\|\mathbf{C} - \mathbf{P}\mathbf{C}\|_F^2 = (1 \pm \epsilon_0)\|\mathbf{A} - \mathbf{P}\mathbf{A}\|_F^2 \quad (2.1)$$

In prior work, \mathbf{C} is referred to as a projection-cost preserving sketch (PCP). PCP constructions are useful since a low-rank approximation for \mathbf{C} translates to a low-rank approximation for \mathbf{A} , while \mathbf{C} has much smaller dimension. Observe that oversampling columns of \mathbf{A} proportional to the *ridge leverage scores* of $\mathbf{A}^{1/2}$, by a $\sqrt{n/k}$ factor, suffices to obtain the guarantee of 2.1 (see Lemma 4.11 for a precise statement). Note, \mathbf{C} may have $\Omega(n^{1.5}/\epsilon_0^2)$ non-zeros but the algorithm need not query any entries in \mathbf{C} at this stage. Musco and Woodruff then construct a row PCP for \mathbf{C} by sampling $\sqrt{nk}/\epsilon_0^{2.5}$ rows of \mathbf{C} proportional to the rank- (k/ϵ_0) *ridge leverage scores* of \mathbf{A} . The resulting matrix \mathbf{R} is a $\sqrt{nk}/\epsilon_0^{2.5} \times \sqrt{nk}/\epsilon_0^2$ matrix such that for any rank- k projection \mathbf{P} ,

$$\|\mathbf{R} - \mathbf{R}\mathbf{P}\|_F^2 + O(\|\mathbf{A} - \mathbf{A}_k\|_F^2) = (1 \pm \epsilon_0)\|\mathbf{C} - \mathbf{P}\mathbf{C}\|_F^2 \quad (2.2)$$

Since \mathbf{R} is a much smaller matrix, they run an input-sparsity time algorithm to compute a low-rank approximation for it [CW13]. Using standard regression techniques (described in [CMM17, MW17, BW18]) along with equations 2.1 and 2.2, setting $\epsilon_0 = \epsilon$ results in a $(1 + \epsilon)$ -low-rank approximation of \mathbf{A} by querying $O(nk/\epsilon^{4.5})$ entries. Musco and Woodruff instead use a more complicated algorithm to get a $1/\epsilon^{2.5}$ dependence.

Our starting point is to observe that the PCP construction above allows to preserve the projection of columns of \mathbf{A} on all (k/ϵ) -dimensional subspaces, albeit up to a constant factor. Therefore, a natural approach is to set the error parameter ϵ_0 in the PCP constructions to be a small fixed constant, say 0.1, and the rank parameter k to be k/ϵ , where ϵ is the desired input accuracy. Further, we observe that the guarantee obtained in Equation 2.1 can be strengthened to a mixed *Spectral-Frobenius PCP* guarantee (also introduced by [MW17]): for all rank- (k/ϵ) projection matrices \mathbf{P} , the column

PCP \mathbf{C} satisfies :

$$(1 - 0.1)\|\mathbf{A} - \mathbf{PA}\|_2^2 - \frac{\epsilon}{10k}\|\mathbf{A} - \mathbf{A}_{k/\epsilon}\|_F^2 \leq \|\mathbf{C} - \mathbf{PC}\|_2^2 \leq (1 + 0.1)\|\mathbf{A} - \mathbf{PA}\|_2^2 + \frac{\epsilon}{10k}\|\mathbf{A} - \mathbf{A}_{k/\epsilon}\|_F^2 \quad (2.3)$$

For a formal statement refer to Lemma 4.12. Sampling rows of \mathbf{C} proportional to the same distribution results in a row PCP for \mathbf{C} such that for all rank- (k/ϵ) projections \mathbf{P} ,

$$(1 - 0.1)\|\mathbf{C} - \mathbf{CP}\|_2^2 - \frac{\epsilon}{10k}\|\mathbf{A} - \mathbf{A}_{k/\epsilon}\|_F^2 \leq \|\mathbf{R} - \mathbf{RP}\|_2^2 \leq (1 + 0.1)\|\mathbf{C} - \mathbf{CP}\|_2^2 + \frac{\epsilon}{10k}\|\mathbf{A} - \mathbf{A}_{k/\epsilon}\|_F^2 \quad (2.4)$$

We then use an *input-sparsity* spectral-low-rank approximation algorithm by [CEM⁺15] (Lemma 4.6), to obtain a low-dimensional subspace, represented by a $\sqrt{nk/\epsilon} \times k/\epsilon$ matrix \mathbf{Z} with orthonormal columns such that

$$\|\mathbf{R} - \mathbf{RZZ}^\top\|_2^2 \leq \frac{\epsilon}{k}\|\mathbf{R} - \mathbf{R}_{k/\epsilon}\|_F^2 \quad (2.5)$$

Following the notation of Clarkson and Woodruff [CW17], we refer to the projection matrix \mathbf{ZZ}^\top as a Spectral-Frobenius (SF) projection (as in Definition 4.13). A key property of an SF projection is that it spans a $(1 + \epsilon)$ -relative-error low-rank approximation to \mathbf{R} (Lemma 4.14), i.e. $\|\mathbf{R} - \mathbf{ZZ}^\top \mathbf{R}_{k/\epsilon} \mathbf{ZZ}^\top\|_F^2 \leq (1 + \epsilon)\|\mathbf{R} - \mathbf{R}_{k/\epsilon}\|_F^2$. Now, using the fact that \mathbf{R} is a Spectral-Frobenius PCP, plugging in $\mathbf{P} = \mathbf{ZZ}^\top$ in Equation 2.4, we can bound $\|\mathbf{C} - \mathbf{CZZ}^\top\|_2^2$ as follows :

$$\begin{aligned} \|\mathbf{C} - \mathbf{CZZ}^\top\|_2^2 &\leq \frac{10}{9}\|\mathbf{R} - \mathbf{RZZ}^\top\|_2^2 + \frac{\epsilon}{9k}\|\mathbf{R} - \mathbf{R}_{k/\epsilon}\|_F^2 \leq \frac{\epsilon}{10k}\|\mathbf{R} - \mathbf{R}_k\|_F^2 + \frac{\epsilon}{9k}\|\mathbf{R} - \mathbf{R}_{k/\epsilon}\|_F^2 \\ &\leq O\left(\frac{\epsilon}{k}\right)\|\mathbf{C} - \mathbf{C}_k\|_F^2 \end{aligned} \quad (2.6)$$

where the second inequality follows from Equation 2.5 and the third follows from the fact that PCPs preserve Frobenius norm up to a constant factor, i.e., $\|\mathbf{R} - \mathbf{R}_{k/\epsilon}\|_F^2 = \Theta(\|\mathbf{C} - \mathbf{C}_{k/\epsilon}\|_F^2) = \Theta(\|\mathbf{A} - \mathbf{A}_{k/\epsilon}\|_F^2)$. Therefore, \mathbf{ZZ}^\top is also a Spectral-Frobenius projection for \mathbf{C} . Here, we are faced with a few challenges. First, the relative-error approximation spanned by the subspace has rank k/ϵ . Second, it is unclear how to obtain any reasonable result for \mathbf{A} from the above structural property, given that even the dimensions of \mathbf{ZZ}^\top do not match \mathbf{A} .

We begin by showing that a Spectral-Frobenius projection for \mathbf{A} suffices to obtain a low-rank approximation with $O(nk/\epsilon)$ queries. Assuming we are handed a (k/ϵ) -dimensional *structured subspace* that contains a relative-error low-rank approximation for \mathbf{A} itself. This is represented as an $n \times k/\epsilon$ matrix \mathbf{Q} with orthonormal columns such that $\|\mathbf{A} - \mathbf{QQ}^\top \mathbf{A}\|_2^2 \leq \epsilon/k \cdot \|\mathbf{A} - \mathbf{A}_{k/\epsilon}\|_F^2$. We prove that given such a structured subspace, we can extract a rank- k relative-error approximation by reading only nk/ϵ entries in \mathbf{A} (Theorem 4.16 in Section 4.1). We provide an overview of the proof here.

The SF projection property implies $\|\mathbf{A} - \mathbf{QQ}^\top \mathbf{A}_k \mathbf{QQ}^\top\|_F^2 \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2$. Therefore, it suffices to solve the following optimization problem:

$$\min_{\text{rank}(\mathbf{X}) \leq k} \|\mathbf{A} - \mathbf{QXQ}^\top\|_F^2 \quad (2.7)$$

since $\mathbf{X} = \mathbf{Q}^\top \mathbf{A}_k \mathbf{Q}$ is always feasible. While we are now optimizing over a $k/\epsilon \times k/\epsilon$ matrix \mathbf{X} , with rank at most k , the problem still seems intractable to solve *optimally* in sublinear time and

queries to \mathbf{A} . The key idea here is that \mathbf{Q} has orthonormal columns and thus the leverage scores are precomputed for us. We can then sample columns and rows proportional to the leverage scores of \mathbf{Q} and consider a significantly smaller sketched problem. Therefore, we create sampling matrices \mathbf{S} and \mathbf{T} that sample $\text{poly}(k/\epsilon)$ rows proportional to the leverage scores of \mathbf{Q} and consider the resulting optimization problem:

$$\min_{\text{rank}(X) \leq k} \|\mathbf{SAT} - \mathbf{SQXQ}^\top \mathbf{T}\|_F^2 \quad (2.8)$$

Here, we are faced with an intriguing phenomenon: our sketched optimization problem does not have the property that the minimum cost for Equation 2.8 is a $(1+\epsilon)$ -approximation to the minimum cost for Equation 2.7. The reason is that our sketch incurs a fixed additive shift term, which we cannot approximate in sublinear time. We note that this is the bottleneck in approximating the cost of the optimal low-rank approximation, and as mentioned in [MW17], it is open to estimate this cost in $o(n^{3/2})$ time.

However, we can apply the structural result in Lemma 4.5 twice, to show that the optimal solution to Equation 2.8, when plugged in to Equation 2.7 obtains a $(1+\epsilon)$ -approximation to the minimum cost. Formally, \mathbf{S} and \mathbf{T} have the property that if $\hat{\mathbf{X}} = \arg \min_{\mathbf{X}} \|\mathbf{SAT} - \mathbf{SQXQ}^\top \mathbf{T}\|_F^2$, then

$$\|\mathbf{A} - \mathbf{Q}\hat{\mathbf{X}}\mathbf{Q}^\top\|_F^2 \leq (1 + O(\epsilon)) \min_{\text{rank}(X) \leq k} \|\mathbf{A} - \mathbf{QXQ}^\top\|_F^2.$$

The optimization problem in Equation 2.8 is called *Generalized Low-Rank Approximation* and admits a closed form solution [FT07] (Theorem 4.15). Further, since the problem now has all dimensions independent of n , we can afford to explicitly compute \mathbf{SAT} by querying the corresponding entries in \mathbf{A} . The resulting closed-form solution can also be computed in $\text{poly}(k/\epsilon)$ time (and queries) which only contributes a lower order term. We obtain one factor for the low-rank approximation for \mathbf{A} by simply computing an orthonormal basis for $\mathbf{Q}\hat{\mathbf{X}}$. In order to compute the second factor, we set up and approximately solve a regression problem, the details of which can be found in Algorithm 1. Efficiently solving such a regression problem is now standard in low-rank approximation literature [CMM17, MW17, BW18]. Therefore, we can output a low-rank approximation to \mathbf{A} by querying only $\tilde{O}(nk/\epsilon)$ entries.

We have now reduced our problem to computing an SF projection for \mathbf{A} , while reading only nk/ϵ entries. Recall, in Equation 2.6, we obtained a Spectral-Frobenius projection for \mathbf{C} but stopped short since we did not see a natural way to proceed. Here, we observe that if we had a such a projection for the *column-space* of \mathbf{C} , by Equation 2.1, it would also work for \mathbf{A} and we would be done. To this end, we consider the following optimization problem:

$$\min_{\mathbf{W} \in \mathbb{R}^{n \times k'}} \|\mathbf{C} - \mathbf{WZ}^\top\|_2^2 \quad (2.9)$$

We show that an orthonormal basis \mathbf{Q} for an approximate minimizer to Equation 2.9 is an SF projection for \mathbf{C} and in turn \mathbf{A} (since \mathbf{C} is a column PCP for \mathbf{A}). Therefore, we focus on optimizing Equation 2.9 and refer to this problem as *Spectral Regression*. We note that unlike standard regression, here we minimize the Spectral (Operator) norm. While the corresponding problem for minimizing Frobenius norm is extensively studied and well understood, to the best of our

knowledge the only relevant related work on Spectral Regression is in the streaming model, by Clarkson and Woodruff [CW09]. They construct an oblivious sketch, consisting of random entries in $\{-1, 1\}$, for Equation 2.9 that preserves the optimal solution up to a $(1 \pm \epsilon)$ factor. Unfortunately, we cannot use oblivious sketching here, since \mathbf{C} may be a dense matrix and we cannot afford to read all of it.

Here, we emphasize that obtaining a sample-optimal algorithm for the aforementioned Spectral Regression problem is crucial for our main algorithmic result. Given that \mathbf{C} is an $n \times \sqrt{nk}/\epsilon$ matrix, we cannot query most of it and thus approximating its leverage scores is infeasible. A natural approach here would be to follow the *Affine Embedding idea* for Frobenius norm (refer to Lemma 4.5) and hope a similar guarantee holds for the spectral norm as well. Here, one might hope to obtain a small sketch that preserves the spectral norm cost of all \mathbf{W} up to a $(1 \pm \epsilon)$ factor. While such a guarantee would suffice, we note that \mathbf{Z} could have rank as large as k/ϵ and we can no longer afford a $(1 + \epsilon)$ -approximate affine embedding even for Frobenius norm, without incurring a larger dependence on ϵ . This precludes all known approaches for sketching Equation 2.9 to preserve the optimal cost.

Instead, we relax the notion of approximation for our sketch. We observe that it suffices to construct a sketch \mathbf{S} such that if $\widehat{\mathbf{W}} = \arg \min_{\mathbf{W}} \|\mathbf{C}\mathbf{S} - \mathbf{W}\mathbf{Z}^\top \mathbf{S}\|_2^2$, then

$$\|\mathbf{C} - \widehat{\mathbf{W}}\mathbf{Z}^\top\|_2^2 \leq O(1) \left(\min_{\mathbf{W} \in \mathbb{R}^{n \times k/\epsilon}} \|\mathbf{C} - \mathbf{W}\mathbf{Z}^\top\|_2^2 + \frac{\epsilon}{k} \|\mathbf{C} - \mathbf{C}_{k/\epsilon}\|_F^2 \right) \quad (2.10)$$

Note, this is a weaker guarantee for the sketch \mathbf{S} , since we only need to preserve the cost of the optimal solution up to a mixed relative and additive error. First, we observe such a guarantee suffices, since we can upper bound the cost from Equation 2.10 by $O(\epsilon/k) \cdot \|\mathbf{C} - \mathbf{C}_{k/\epsilon}\|_F^2$ and the Spectral-Frobenius PCP from Equation 2.1 incurs this term anyway. In Theorem 4.18, we show that we can construct such a sketch \mathbf{S} satisfying Equation 2.10 by sampling k/ϵ columns of \mathbf{C} proportional to the leverage scores of \mathbf{Z}^\top . This is surprising since we completely ignore all information about \mathbf{C} and our sketch is not an oblivious sketch.

The key technical lemma (Lemma 4.25) we prove here is a *weak approximate matrix product* for \mathbf{C}^* and \mathbf{Z}^\top where $\mathbf{C}^* = \mathbf{C}(\mathbf{I} - \mathbf{P}_{\mathbf{Z}^\top})$ is the projection onto the orthogonal complement of \mathbf{Z}^\top . While *approximate matrix product* has been extensively studied [DFK⁺04, Sar06, CW13], even for spectral norm [CNW15], it is important to emphasize here that all known constructions are either oblivious sketches or require sketches that are sampled proportional to both \mathbf{C}^* and \mathbf{Z}^\top . Since \mathbf{Z}^\top has no information about the spectrum of \mathbf{C}^* , the main challenge here is to control the spectrum of $\mathbf{C}^* \mathbf{S} \mathbf{S}^\top \mathbf{Z}^\top$.

In order to bound $\|\mathbf{C}^* \mathbf{S} \mathbf{S}^\top \mathbf{Z}^\top\|_2$, we analyze how sampling columns of \mathbf{C}^* proportional to the *leverage scores* of \mathbf{Z}^\top affects the spectrum of \mathbf{C}^* . An important tool in our analysis is the following result by Rudelson and Vershynin on how the spectral norm of a matrix degrades when we sample a uniformly random subset of rows [RV07]. They show that sampling q rows of a matrix \mathbf{M} uniformly at random, indexed by the set Q , results in a matrix $\mathbf{M}_{|Q}$ such that

$$\mathbb{E} [\|\mathbf{M}_{|Q}\|_2] = O \left(\sqrt{\frac{q}{n}} \|\mathbf{M}\|_2 + \sqrt{\log(q)} \|\mathbf{M}\|_{(n/q)} \right)$$

where $\|\mathbf{A}\|_{(n/q)}$ is the average of the largest n/q ℓ_2 -norms of columns of \mathbf{A} . Here, we prove that expected spectral norm of \mathbf{C} restricted to the columns sampled by \mathbf{S} proportional to the leverage scores of \mathbf{Z}^\top only exceeds that of a random subset by a polylogarithmic factor. This result may be of independent interest in applications where we would want to bound the spectrum of random submatrices, where the rows or columns are *not* sampled uniformly.

Intuitively, there are two technical challenges we overcome in order to apply the Rudelson-Vershynin result in our setting. First, a leverage score sampling matrix \mathbf{S} need not sample columns uniformly at random, since we have no control over the squared column norms of \mathbf{Z}^\top . Given that the squared column norms of \mathbf{Z}^\top may be lopsided, the subset of columns we select could be far from a uniform sample in the worst case. Second, the matrix we apply it to is not square and $\|\cdot\|_{(n/q)}$ norm only shrinks substantially when the columns of \mathbf{A} have the same ℓ_2^2 norm, up to a constant.

We therefore obtain a variant of Spectral norm decay for rectangular matrices, i.e. for any $n \times m$ matrix \mathbf{M} with roughly the same squared column norms, we show that

$$\mathbb{E} [\|\mathbf{M}_{|Q}\|_2] = O \left(\sqrt{\frac{q}{n}} \|\mathbf{M}\|_2 + \sqrt{\log(q)/b} \|\mathbf{M}\|_{(n/q)} \right) \quad (2.11)$$

where $b = \lceil n/m \rceil$. To apply the above result, we then partition the rows of \mathbf{C} (since \mathbf{S} samples columns of \mathbf{C} as opposed to rows) into $\log(n)$ groups such that within each group, all rows have roughly the same squared norm. We then analyze leverage score sampling proportional to the column norms of \mathbf{Z}^\top on each group independently. We show that we can obtain a coupling between the two random processes, namely uniform sampling and leverage score sampling, such that we obtain a decay bound similar to Equation 2.11, up to log factors. We describe our solution in more detail in Section 4. We note that our results extend to outputting a low-rank PSD matrix as well.

Negative-Type Matrices. We then use the techniques developed above to obtain an optimal relative-error low-rank approximation for Negative-Type distance matrices. While arbitrary metrics do not admit sublinear time algorithms for relative-error low-rank approximation (see Theorem 7.1 in [BW18]) Bakshi and Woodruff provided a sublinear time algorithm for metrics that satisfy negative-type inequalities. They obtain a $(1 + \epsilon)$ -relative-error approximation, that queries $\tilde{O}(nk/\epsilon^{2.5})$ entries in the input. However, this algorithm outputs a bi-criteria solution, i.e., given a negative-type matrix \mathbf{A} , it outputs a rank- $(k + 4)$ matrix \mathbf{M} such that $\|\mathbf{A} - \mathbf{M}\|_F^2 \leq (1 + \epsilon) \|\mathbf{A} - \mathbf{A}_k\|_F^2$.

The key observation they make is that negative-type metrics can be realized as the distances corresponding to a point set $\mathcal{P} = \{x_1, x_2, \dots, x_n\}$ such that $\mathbf{A}_{i,j} = \|x_i - x_j\|_2^2 = \|x_i\|_2^2 + \|x_j\|_2^2 - 2\langle x_i, x_j \rangle$. Therefore, \mathbf{A} admits the following decomposition: $\mathbf{A} = \mathbf{R}_1 + \mathbf{R}_2 - 2\mathbf{B}$, where for all $j \in [n]$, $(\mathbf{R}_1)_{i,j} = \|x_i\|_2^2$, $\mathbf{R}_2 = \mathbf{R}_1^\top$ and \mathbf{B} is PSD. Observe that query access to \mathbf{A} suffices to obtain query access to \mathbf{B} by simply assuming w.l.o.g. that x_1 is centered at the origin and the i -th entry in the first row corresponds to $\|x_i\|_2^2$. Therefore, any PSD low-rank approximation algorithm can be simulated on the matrix \mathbf{B} by only having query access to \mathbf{A} . Bakshi and Woodruff show that obtaining the low-rank approximation for \mathbf{B} and appending the column span of \mathbf{R}_1 and \mathbf{R}_2 to it results in a rank- $(k + 4)$ bi-criteria approximation to \mathbf{A} . The bi-criteria algorithm can be improved

to $k + 2$ using Cauchy's Interlacing Theorem [Fis05] and observing $\mathbf{R}_1, \mathbf{R}_2$ are rank-1 updates to \mathbf{B} , but this seems to be the limit of such approaches.

We show here that our SF projection framework can be used to obtain a sample-optimal algorithm for negative-type metrics, and the bi-criteria approximation is not necessary. Recall, from our discussion above, that an SF projection for \mathbf{A} suffices to obtain a low-rank approximation for \mathbf{A} . Our key observation is that we can use the techniques we developed for PSD matrices to obtain an SF projection, $\mathbf{Q}\mathbf{Q}^T$, for \mathbf{B} (in the decomposition above), to which we append the column span of $\mathbf{R}_1, \mathbf{R}_2$ to \mathbf{Q} , and the resulting projection (denoted by $\mathbf{\Omega}$) is an SF projection for \mathbf{A} . To see this, observe, $\|\mathbf{A} - \mathbf{\Omega}\mathbf{A}_k\mathbf{\Omega}\|_F^2 = \|\mathbf{A} - \mathbf{A}_k\|_F^2 + \|\mathbf{A}_k - \mathbf{\Omega}\mathbf{A}_k\mathbf{\Omega}\|_F^2 + 2\text{Tr}((\mathbf{A} - \mathbf{A}_k)(\mathbf{I} - \mathbf{\Omega})\mathbf{A}_k\mathbf{\Omega})$. A simple calculation using Von-Neumann's trace inequality bounds the deviation by $O(k\|\mathbf{A}(\mathbf{I} - \mathbf{\Omega})\|_2^2)$. Since $\mathbf{\Omega}$ spans \mathbf{R}_1 and \mathbf{R}_2 , and is an SF projection for \mathbf{B} , we can bound the above cost by $O(\epsilon/k)\|\mathbf{B} - \mathbf{B}_{k+2}\|_F$. It is easy to see that $\|\mathbf{A} - \mathbf{A}_k\|_F^2 = O(\|\mathbf{B} - \mathbf{B}_{k+2}\|_F^2)$ and therefore, we conclude $\|\mathbf{A} - \mathbf{\Omega}\mathbf{A}_k\mathbf{\Omega}\|_F^2 \leq (1 + O(\epsilon))\|\mathbf{A} - \mathbf{A}_k\|_F^2$ (see Lemma 4.29 for details). Subsequently, we use the sublinear algorithm we developed for PSD matrices to obtain a low-rank approximation for \mathbf{A} .

Ridge Regression. Our techniques also naturally extend to ridge regression, when the design matrix is PSD. This connection was originally outlined by Musco and Woodruff and they obtain sublinear time algorithms for solving ridge regression, parametrized by the statistical dimension s_λ . At a high level, we compute a rank- (s_λ/ϵ^2) spectral approximation to the input and solve ridge regression on the resulting matrix, i.e., given a PSD matrix \mathbf{A} , we compute a low-rank matrix \mathbf{B} such that $\|\mathbf{A} - \mathbf{B}\|_2^2 \leq O(\epsilon/k)\|\mathbf{A} - \mathbf{A}_k\|_F^2$. Further, we observe that the low-rank matrix is in fact a coresets for the input as it simultaneously preserves the cost of all x and y .

We then obtain a matching query lower bound for constructing coresets for ridge regression. Our lower bound proceeds by showing that a coresets can output a low-rank approximation on the instance of Musco and Woodruff with a stronger quadratic, rather than a linear dependence on ϵ . Intuitively, the hard instance has multiple principle submatrices of all 1s placed randomly over the matrix. Since a coresets simultaneously preserves the ridge regression cost for all x, y , it suffices to query the coresets on tuples of (scaled) eigenvectors and learn the positions of the blocks. However, a priori, we do not know what the eigenvectors of \mathbf{A} are. Instead, we query the coresets on every vector with a bounded support, and pick all vectors with small regression cost. We show that our resulting set only contains vectors which do not overlap much on the locations of the hidden blocks and we show this suffices.

2.2 Robust Low-Rank Approximation

The robustness model we consider is as follows: we begin with an $n \times n$ PSD matrix \mathbf{A} . An adversary is then allowed to arbitrarily corrupt \mathbf{A} by adding a perturbation matrix \mathbf{N} such that $\|\mathbf{N}\|_F^2 \leq \eta\|\mathbf{A}\|_F^2$ and for all $i \in [j]$, $\|\mathbf{N}_{i,*}\|_2^2 \leq c\|\mathbf{A}_{i,*}\|_2^2$, for a fixed constant c . Note, while the adversary is unrestricted in the entries of \mathbf{A} that it corrupts, the Frobenius norm of the corruption is bounded in terms of the Frobenius norm of \mathbf{A} and the corruption is well-spread. The motivation for considering such a model is that many matrices that we observe in practice might be close but not exactly PSD, for instance, small perturbations to PSD matrices.

It is impossible to obtain a relative-error low-rank approximation in this setting, since we cannot even identify \mathbf{A} after querying all n^2 entries of $\mathbf{A} + \mathbf{N}$. To see this, consider the case where \mathbf{A} is rank- k , and observe that a relative-error algorithm requires identifying \mathbf{A} exactly. However, by querying all entries of $\mathbf{A} + \mathbf{N}$, we can determine the row norms exactly. Therefore, we can run the algorithm of Frieze-Kannan-Vempala [FKV04] to obtain a rank- k matrix \mathbf{XY}^\top (in factored form) such that with probability at least $99/100$,

$$\begin{aligned}\|\mathbf{A} + \mathbf{N} - \mathbf{XY}^\top\|_F^2 &\leq \|\mathbf{A} + \mathbf{N} - (\mathbf{A} + \mathbf{N})_k\|_F^2 + \epsilon \|\mathbf{A} + \mathbf{N}\|_F^2 \\ &\leq \|\mathbf{A} + \mathbf{N} - \mathbf{A}_k\|_F^2 + \epsilon \|\mathbf{A} + \mathbf{N}\|_F^2 \\ &\leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + \|\mathbf{N}\|_F^2 + 2\langle \mathbf{A} - \mathbf{A}_k, \mathbf{N} \rangle + (3 + \eta)\epsilon \|\mathbf{A}\|_F^2 \\ &\leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + O(\epsilon + \sqrt{\eta})\|\mathbf{A}\|_F^2\end{aligned}\tag{2.12}$$

where the second inequality follows from $(\mathbf{A} + \mathbf{N})_k$ being the best rank- k approximation to $\mathbf{A} + \mathbf{N}$ and \mathbf{A}_k is any other rank- k matrix. The third inequality uses $\|\mathbf{A} + \mathbf{N}\|_F^2 \leq 2(\|\mathbf{A}\|_F^2 + \|\mathbf{N}\|_F^2)$, which follows from ℓ_2^2 distance satisfying triangle-inequality up to a factor of 2. The last inequality uses Cauchy-Schwarz on $2|\langle \mathbf{A} - \mathbf{A}_k, \mathbf{N} \rangle| \leq 2\|\mathbf{A} - \mathbf{A}_k\|_F \cdot \|\mathbf{N}\|_F \leq 2\sqrt{\eta}\|\mathbf{A}\|_F^2$, which follows from the assumption on \mathbf{N} . Additionally

$$\begin{aligned}\|\mathbf{A} + \mathbf{N} - \mathbf{XY}^\top\|_F^2 &= \|\mathbf{A} - \mathbf{XY}^\top\|_F^2 + \|\mathbf{N}\|_F^2 + 2\langle \mathbf{A} - \mathbf{XY}^\top, \mathbf{N} \rangle \\ &\geq \|\mathbf{A} - \mathbf{XY}^\top\|_F^2 - 2\sqrt{\eta}\|\mathbf{A}\|_F^2\end{aligned}\tag{2.13}$$

Combining Equations 2.12 and 2.13, we have $\|\mathbf{A} - \mathbf{XY}^\top\|_F^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + O(\epsilon + \sqrt{\eta})\|\mathbf{A}\|_F^2$. While this algorithm is far from optimal in terms of sample complexity, it indicates that relaxing our guarantees to additive-error is amenable to robust algorithms and indicates why we pick up a $\sqrt{\eta}$ term. The central question we focus on in this section is whether there exists a *robust sublinear time and query algorithm* to obtain an additive-error low-rank approximation for PSD matrices.

We begin by showing a sample complexity lower bound if \mathbf{A} is an arbitrary PSD matrix. The intuition from the relative-error setting still applies and the diagonal entries are crucial for sublinear algorithms. In tune with this intuition, the adversary corrupts large diagonal entries to decrease their magnitude and thus obfuscate rows that contain large off-diagonal entries. We therefore parameterize our lower bound and algorithms by the largest ratio between a diagonal entry of \mathbf{A} and $\mathbf{A} + \mathbf{N}$, denoted by $\phi_{\max} = \max_{j \in [n]} \mathbf{A}_{j,j} / |(\mathbf{A} + \mathbf{N})_{j,j}|$. Recall, we obtain the following lower bound:

Theorem 5.4. (*Informal lower bound.*) *Let $\epsilon > \eta > 0$. Given $\mathbf{A} + \mathbf{N}$ such that \mathbf{A} is PSD and \mathbf{N} is a corruption matrix as defined above, any randomized algorithm, that with probability at least $2/3$, outputs a rank- k approximation up to additive error $(\epsilon + \eta)\|\mathbf{A}\|_F^2$ must read $\Omega(\phi_{\max}^2 nk / \epsilon)$ entries of $\mathbf{A} + \mathbf{N}$.*

In our hard instance, we have a block matrix \mathbf{A} , where we place a random $\epsilon/\eta \times \epsilon/\eta$, rank-1, non-contiguous block \mathbf{B}_1 such that each entry in the block is $\sqrt{\eta^2 n / \epsilon}$ and the remaining matrix has 1s on the diagonals and 0s everywhere else. It is easy to see this matrix is PSD. We observe that the block \mathbf{B}_1 contributes an ϵ -fraction of the Frobenius norm of \mathbf{A} , and the ℓ_2^2 norm of the diagonals is an η -fraction of the Frobenius norm of \mathbf{A} . Therefore, the adversary can afford to corrupt all

the diagonal entries in \mathbf{B}_1 and set them to be 1. Such a perturbation is feasible in our model and successfully obfuscates the large off-diagonal entries. Note, for this perturbation $\phi_{\max}^2 = \eta^2 n / \epsilon$.

Let the resulting matrix be denoted by $\mathbf{A} + \mathbf{N}$. Here, we observe any ϵ -additive-error low-rank approximation cannot ignore the block \mathbf{B}_1 . Since the diagonals of $\mathbf{A} + \mathbf{N}$ now provide no information about the off-diagonal entries, any algorithm that correctly outputs a low-rank approximation for both $\mathbf{A} + \mathbf{N}$ and \mathbf{I} must detect at least one entry in \mathbf{B}_1 . Since \mathbf{B}_1 has ϵ^2 / η^2 non-zeros, any algorithm must query $\Omega(\eta^2 n^2 / \epsilon^2) = \Omega(\phi_{\max}^2 n / \epsilon)$ entries to detect one entry. To obtain a linear dependence on k , we simply create k independent copies of \mathbf{B}_1 .

Robust Algorithm. Next, we focus on a robust, additive-error low-rank approximation algorithm, where the sample complexity is parameterized by ϕ_{\max} . We begin by introducing a new sampling procedure to construct *projection-cost preserving sketches*. Our construction is simple to state: we sample each column proportional to the corresponding diagonal entry. Computing these sampling probabilities *exactly* requires reading only n entries in $\mathbf{A} + \mathbf{N}$. We show that sampling $\tilde{O}(\phi_{\max}^2 \sqrt{nk^2} / \epsilon^2)$ columns proportional to this distribution preserves the projection of the columns of \mathbf{A} onto the orthogonal complement of any rank- k subspace, up to additive error $(\epsilon + \sqrt{\eta}) \|\mathbf{A}\|_F^2$.

Theorem 5.7. (Informal Robust Column PCP.) *Let $\mathbf{A} + \mathbf{N}$ be an $n \times n$ matrix following the assumptions of our noise model. Let $k \in [n]$ and $\epsilon, \sqrt{\eta} > 0$. Let $q = \{q_1, q_2, \dots, q_n\}$ be a probability distribution over the columns of \mathbf{A} such that $q_j = (\mathbf{A} + \mathbf{N})_{j,j} / \text{Tr}(\mathbf{A} + \mathbf{N})$. Construct \mathbf{C} by sampling $\tilde{O}(\phi_{\max}^2 \sqrt{nk^2} / \epsilon^2)$ columns of $\mathbf{A} + \mathbf{N}$ proportional to q and rescaling appropriately. Then, with probability at least $1 - c$, for any rank- k orthogonal projection \mathbf{X} ,*

$$\|\mathbf{C} - \mathbf{X}\mathbf{C}\|_F^2 = \|\mathbf{A} - \mathbf{X}\mathbf{A}\|_F^2 \pm (\epsilon + \sqrt{\eta}) \|\mathbf{A}\|_F^2$$

We note that all prior PCP constructions work in the noiseless setting. As a comparison, the construction of Cohen et. al. [CMM17] works for arbitrary \mathbf{A} , but requires $n \text{nnz}(\mathbf{A})$ time and queries to compute the approximate ridge-leverage scores of \mathbf{A} . Musco and Musco [MM17] describe how to approximately compute the ridge leverage scores of $\mathbf{A}^{1/2}$ (if \mathbf{A} is PSD) using the Nystrom approximation, where $\mathbf{A} = \mathbf{A}^{1/2} \cdot \mathbf{A}^{1/2}$. Musco and Woodruff [MW17] use this method to compute the ridge leverage scores of $\mathbf{A}^{1/2}$ with $\Theta(nk)$ queries and show that this provides a $(\sqrt{n/k})$ -approximation to the ridge leverage scores of \mathbf{A} . We note that the guarantees obtained by [CMM17, MW17] are relative error, as opposed to the additive error guarantee in the theorem above. Finally, Bakshi and Woodruff [BW18] provide an additive-error sublinear time construction for *distance matrices* by sampling proportional to column norms. In all the aforementioned constructions, computing the sampling distribution is a non-trivial task, whereas we simply sample proportional to the diagonal entries.

We observe that we sample columns of $\mathbf{A} + \mathbf{N}$, to obtain \mathbf{C} which is an unbiased estimator for $\|\mathbf{A} + \mathbf{N}\|_F^2$. The main technical challenge in our construction is to relate the cost of rank- k projections for the column space of \mathbf{A} to that of \mathbf{C} , while obtaining an optimal dependence on n and k . Note, while we do not obtain the correct dependence on ϵ , we do not have to explicitly compute all of \mathbf{C} , only a subset of it.

We then extend the diagonal sampling algorithm to construct a robust row PCP for the matrix \mathbf{C} . We note that the construction for \mathbf{A} does not immediately give a row PCP for \mathbf{C} since \mathbf{C} is no longer

a corrupted PSD matrix or even a square matrix, and thus there is no notion of a diagonal. Here, all previous approaches to construct a PCP with a sublinear number of queries hit a roadblock, since the matrix \mathbf{C} need not have any well-defined structure apart from being a scaled subset of the columns of the original corrupted PSD matrix $\mathbf{A} + \mathbf{N}$. However, we show that sampling rows of \mathbf{C} proportional to the diagonal entries of $\mathbf{A} + \mathbf{N}$ results in a row PCP for \mathbf{C} .

Theorem 5.10. (*Informal Robust Row PCP.*) Let $\mathbf{A} + \mathbf{N}$ be an $n \times n$ matrix corresponding to our noise model and let \mathbf{C} be a column PCP for \mathbf{A} as defined above. Let $p = \{p_1, p_2, \dots, p_n\}$ be a probability distribution over the rows of \mathbf{C} such that $p_j = (\mathbf{A} + \mathbf{N})_{j,j} / \text{Tr}(\mathbf{A} + \mathbf{N})$. Construct \mathbf{R} by sampling $\tilde{O}(\phi_{\max} \sqrt{nk^2/\epsilon^2})$ rows of \mathbf{C} proportional to p and scaling appropriately. With probability at least $1 - c$, for any rank- k orthogonal projection \mathbf{X} ,

$$\|\mathbf{R} - \mathbf{R}\mathbf{X}\|_F^2 = \|\mathbf{C} - \mathbf{C}\mathbf{X}\|_F^2 \pm (\epsilon + \sqrt{\eta}) \|\mathbf{A}\|_F^2$$

For our algorithm, we begin by constructing column and row PCPs of $\mathbf{A} + \mathbf{N}$, to obtain a $t \times t$ matrix \mathbf{R} , where $t = \tilde{O}(\phi_{\max} \sqrt{nk^2/\epsilon^2})$. Instead of reading the entire matrix, we uniformly sample $\epsilon^3 t/k^3$ entries in each row of \mathbf{R} , and query these entries. Note, this corresponds to reading $\epsilon^3 t^2/k^3 = \tilde{O}(\phi_{\max}^2 nk/\epsilon)$ entries in $\mathbf{A} + \mathbf{N}$. Ideally we would want to estimate the ℓ_2^2 norms of each row of \mathbf{R} to then use a result of Frieze-Kannan-Vempala to obtain a low-rank approximation for \mathbf{R} [FKV04]. It is well known that to recover a low-rank approximation for \mathbf{R} , one can sample rows of \mathbf{R} proportional to row norm estimates, denoted by \mathcal{Y}_i [FKV04]. As shown in [IVWW19] the following two conditions are a relaxation of those required in [FKV04], and suffice to obtain an additive error low-rank approximation :

1. For all $i \in [t]$, the corresponding estimate over-estimates the row norm of $\mathbf{R}_{i,*}$, i.e., $\mathcal{Y}_i \geq \|\mathbf{R}_{i,*}\|_2^2$.
2. The sum of the over-estimates is not too much larger than the Frobenius norm of the matrix, i.e., $\sum_{i \in [t]} \mathcal{Y}_i \leq \phi_{\max}^2 n/t \|\mathbf{R}\|_F^2$

If the two conditions are satisfied, Frieze-Kannan-Vempala implies sampling s rows of \mathbf{R} proportional to \mathcal{Y}_i results in an $s \times t$ matrix \mathbf{S} such that the row space of \mathbf{S} contains a good rank- k approximation, where $s = O(\phi_{\max}^2 nk/\epsilon t)$, which matches our desired sample complexity. We show that, unfortunately, such a guarantee is not possible even in the uncorrupted case, where $\mathbf{N} = 0$ and $\phi_{\max} = 1$. Intuitively, \mathbf{R} may be a sparse matrix such that many rows have large norm, and uniform sampling cannot obtain concentration for all such rows, as required by the aforementioned conditions.

Instead, we settle for a weaker statement, where we show that the estimator obtained by uniform sampling in each row is accurate with $o(1)$ probability. At a high level, we show that we can design a sampling process that is statistically close to ℓ_2^2 sampling described by Frieze-Kannan-Vempala [FKV04]. We then open up the analyzes of Frieze-Kannan-Vempala and show that our sampling process suffices to recover the low-rank approximation guarantee. Given the flurry of recent work in quantum computing [KP16, CLW18, Tan19, RSML18, GLT18] that uses Frieze-Kannan-Vempala ℓ_2^2 sampling as a key algorithmic primitive, our analysis may be of independent interest.

Lemma 5.14. (*Informal Estimation of Row Norms.*) Let $\mathbf{R} \in \mathbb{R}^{t \times t}$ be the row PCP as defined above. For all $i \in [t]$ let $X_i = \sum_{j \in [\epsilon^3 t / k^3]} X_{i,j}$ such that $X_{i,j} = k^3 \mathbf{R}_{i,j}^2 / \epsilon^3$ with probability $1/t$, for all $j' \in [t]$. Then, for all $i \in [t]$, $X_i = (1 \pm 0.1) \|\mathbf{R}_{i,*}\|_2^2$ with probability at least $\min(\|\mathbf{R}_{i,*}\|_2^2 k / \epsilon n, 1)$.

We now face two major challenges: first, the probability with which the estimators are accurate is too small to even detect all rows with norm larger than $\phi_{\max}^2 n \|\mathbf{R}\|_F^2 / t^2$, and second, there is no small query certificate for when an estimator is accurate in estimating the row norms. Therefore, we cannot even identify the rows where we obtain an accurate estimate of norm.

To address the first issue, we make the crucial observation that while we cannot estimate the norm of each row accurately, we can hope to sample the row with the same probability as Frieze-Kannan-Vempala [FKV04]. Recall, their algorithm requires sampling row $\mathbf{R}_{i,*}$ with probability at least $\|\mathbf{R}_{i,*}\|_2^2 / \|\mathbf{R}\|_F^2$, which matches the probability in Lemma 5.14. Therefore, we can focus on designing a weaker notion of identifiability, that may potentially include extra rows.

We begin by partitioning the rows of \mathbf{R} into two sets. Let $\mathcal{H} = \{i \mid \|\mathbf{R}_{i,*}\|_2^2 \geq \phi_{\max}^2 n / t^2 \|\mathbf{R}\|_F^2\}$ be the set of heavy rows and $[t] \setminus \mathcal{H}$ be the remaining rows. Note, $|\mathcal{H}| = O(t^2 / \phi_{\max}^2 n) = O(k^4 \log^4(n) / \epsilon^4)$. We then condition on our estimator having norm at least $\phi_{\max}^2 n \|\mathbf{R}\|_F^2 / t^2$. Conditioned on this event, we sample the corresponding row of \mathbf{R} with probability 1. As before, we want to prevent sampling too many spurious rows, but we show only a subset of the rows in \mathcal{H} satisfy this condition. This ensures we identify rows in \mathcal{H} with the right probability. For all the remaining rows, we know the norm is at most $\phi_{\max}^2 n / t^2 \|\mathbf{R}\|_F^2$. We show that uniformly sampling $\phi_{\max}^2 n / t$ such rows suffices to simulate row norm sampling.

We then open up the analysis of Frieze-Kannan-Vempala to show that the above sampling procedure suffices to bound the overall variance, resulting in a relaxation of the conditions required to obtain an additive error low-rank approximation to \mathbf{R} . Once we compute a good low-rank approximation for \mathbf{R} we can follow the approach of [CMM17, MW17, BW18], where we set up two regression problems, and use the sketch and solve paradigm to compute a low-rank approximation for \mathbf{A} , culminating in Theorem 5.11.

For corrupted correlation matrices, we observe that the true uncorrupted matrix has all diagonal entries equal to 1. Therefore, we can discard the diagonal entries of $\mathbf{A} + \mathbf{N}$ and assume they are 1. In this case, no matter what the adversary does to the diagonal, $\phi_{\max} = 1$ and we obtain an $\tilde{O}(nk/\epsilon)$ query algorithm that satisfies the above guarantee. Further, we show a matching sample complexity lower bound of $\Omega(nk/\epsilon)$, to obtain ϵ -additive-error, even in the presence of no noise.

3 Preliminaries and Notation

Given an $m \times n$ matrix \mathbf{A} with rank r , we can compute its singular value decomposition, denoted by $\text{SVD}(\mathbf{A}) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, such that \mathbf{U} is an $m \times r$ matrix with orthonormal columns, \mathbf{V}^\top is an $r \times n$ matrix with orthonormal rows and $\mathbf{\Sigma}$ is an $r \times r$ diagonal matrix. The entries along the diagonal are the singular values of \mathbf{A} , denoted by $\sigma_1, \sigma_2 \dots \sigma_r$. 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., $\mathbf{A}_k = \mathbf{U}\mathbf{\Sigma}_k\mathbf{V}^\top$, where $\mathbf{\Sigma}_k$ 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 the Frobenius norm, i.e., $\mathbf{A}_k = \min_{\text{rank}(\mathbf{X}) \leq k} \|\mathbf{A} - \mathbf{X}\|_F$. 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. For an $n \times n$ PSD matrix \mathbf{A} , we denote the singular (eigenvalue) decomposition by $\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$. Further, since $\Sigma_{i,i} \geq 0$, let $\mathbf{A}^{1/2} = \mathbf{U}\mathbf{\Sigma}^{1/2}\mathbf{U}^\top$ be the square root of \mathbf{A} . Note that $\mathbf{A}_{i,j} = \langle \mathbf{A}_{i,*}^{1/2}, \mathbf{A}_{j,*}^{1/2} \rangle$. By Cauchy-Schwarz, for all $i, j \in [n]$, $\mathbf{A}_{i,j}^2 = \langle \mathbf{A}_{i,*}^{1/2}, \mathbf{A}_{j,*}^{1/2} \rangle^2 \leq \|\mathbf{A}_{i,*}^{1/2}\|_2^2 \cdot \|\mathbf{A}_{j,*}^{1/2}\|_2^2 = \mathbf{A}_{i,i} \cdot \mathbf{A}_{j,j}$. We use $\text{nnz}(\mathbf{A})$ to denote the number of non-zero entries (sparsity) of \mathbf{A} . We use operator and spectral norm interchangeably to denote $\|\mathbf{M}\|_2 = \max_{\|y\|_2=1} \|\mathbf{M}y\|_2$. We also use the notation \mathbf{M}^\dagger to denote the Moore-Penrose pseudoinverse.

4 Relative Error PSD Low-Rank Approximation

In this section, we describe our main algorithm for *relative-error* PSD Low-Rank Approximation, where we query only $\tilde{O}(nk/\epsilon)$ of the input matrix \mathbf{A} . This improves the best known algorithm by Musco and Woodruff that queries $\tilde{O}(nk/\epsilon^{2.5})$ and matches their query lower bound of $\Omega(nk/\epsilon)$ up to polylogarithmic factors [MW17]. Formally, we prove the following:

Theorem 4.1. (*Sample-Optimal PSD Low-Rank Approximation.*) *Given an $n \times n$ PSD matrix \mathbf{A} , an integer k , and $1 > \epsilon > 0$, Algorithm 3 samples $\tilde{O}(nk/\epsilon)$ entries in \mathbf{A} and outputs matrices $\mathbf{M}, \mathbf{N}^\top \in \mathbb{R}^{n \times k}$ such that with probability at least $9/10$,*

$$\|\mathbf{A} - \mathbf{M}\mathbf{N}\|_F^2 \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2$$

Further, the algorithm runs in $\tilde{O}(n(k/\epsilon)^{\omega-1} + (k/\epsilon^3)^\omega)$ time.

We begin by defining various statistical quantities associated with a given matrix, such as the leverage and ridge-leverage scores. The leverage score of a given row measures the importance of this row in composing the row span. Leverage scores have found numerous applications in regression, preconditioning, linear programming and graph sparsification [Sar06, SS11, LS15, CLM⁺15]. In the special case of graphs, they are referred to as *effective resistances*.

Definition 4.2. (*Leverage Scores.*) *Given a matrix $\mathbf{M} \in \mathbb{R}^{n \times m}$, let $\mathbf{m}_i = \mathbf{M}_{i,*}$ be the i -th row of \mathbf{M} . Then, for all $i \in [n]$ the i -th row leverage score of \mathbf{M} is given by*

$$\tau_i(\mathbf{M}) = \mathbf{m}_i(\mathbf{M}^\top \mathbf{M})^\dagger \mathbf{m}_i^\top$$

The column leverage scores can be defined analogously. Note, in the special case where \mathbf{M} has orthonormal columns, the row leverage scores of \mathbf{M} are simply the ℓ_2^2 norms of the rows i.e., $\tau_i(\mathbf{M}) = \|\mathbf{m}_i\|_2^2$. It is well-known that sampling rows of a matrix proportional to the leverage scores satisfies the subspace embedding property (Spectral Sparsification for Graphs) and leads to faster algorithms for ℓ_2 -norm Regression. Recall, for an $n \times m$ matrix \mathbf{A} , a leverage score sampling matrix $\mathbf{S} = \mathbf{D}\mathbf{\Omega}^\top$, where \mathbf{D} is a $t \times t$ diagonal matrix and $\mathbf{\Omega}$ is an $n \times t$ sampling matrix. For all $j \in [t]$, select row index $i \in [n]$ with probability $p_i = \tau_i(\mathbf{A}) / \sum_i \tau_i(\mathbf{A})$ and set $\Omega_{i,j} = 1$ and $\mathbf{D}_{j,j} = 1/\sqrt{tp_i}$.

Lemma 4.3. (*Subspace Embedding.*) Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$, $\epsilon > 0$, and a leverage score sampling matrix \mathbf{S} with $t = O(m \log(m)/\epsilon^2)$ rows, with probability at least $99/100$, for all $x \in \mathbb{R}^m$

$$\|\mathbf{S}\mathbf{A}x\|_2^2 = (1 \pm \epsilon)\|\mathbf{A}x\|_2^2$$

This simply follows from an application of the Matrix Chernoff bound. Observe that the sketch preserves all the singular values of \mathbf{A} up to a factor of $1 \pm \epsilon$. We refer the reader to a recent survey for more details [Woo14]. Next, we recall that leverage score sampling results in a fast algorithm for regression.

Lemma 4.4. (*Fast Regression, Theorem 38 [CW13].*) Given matrices $\mathbf{A} \in \mathbb{R}^{n \times m}$, $\mathbf{B} \in \mathbb{R}^{n \times d}$ such that $\text{rank}(\mathbf{A}) \leq r$ and $\epsilon > 0$, sample $O(r \log(r) + r/\epsilon)$ rows of \mathbf{A}, \mathbf{B} proportional to the leverage scores of \mathbf{A} to obtain a sketch \mathbf{S} such that $\mathbf{Y}^* = \arg \min_{\mathbf{Y}} \|\mathbf{S}\mathbf{A}\mathbf{Y} - \mathbf{S}\mathbf{B}\|_F^2$. Then, with probability at least $1 - c$,

$$\|\mathbf{A}\mathbf{Y}^* - \mathbf{B}\|_F^2 \leq (1 + \epsilon) \min_{\mathbf{Y}} \|\mathbf{A}\mathbf{Y} - \mathbf{B}\|_F^2$$

for a fixed small constant c . Further, the time to compute \mathbf{Y}^* is $O(n \text{nnz}(\mathbf{A}) \log(r/\epsilon) + (n + d)(r/\epsilon)^{\omega-1} + \text{poly}(r/\epsilon))$.

Note, the terms in the running time follow from using Cohen's construction for OSNAP [Coh16]. Leverage score sampling matrices also approximately preserve norms in affine spaces, which leads to faster algorithms for multi-response regression, i.e., $\min_{\mathbf{X}} \|\mathbf{A}\mathbf{X} - \mathbf{B}\|_F^2$, where \mathbf{B} now has a large number of columns.

Lemma 4.5. (*Affine Embeddings, Theorem 39 [CW13].*) Given matrices $\mathbf{A} \in \mathbb{R}^{n \times m}$, such that $\text{rank}(\mathbf{A}) = r$, and $\mathbf{B} \in \mathbb{R}^{n \times d}$, let \mathbf{S} be a leverage score sampling matrix with $t = O(r/\epsilon^2)$ rows. Further, let \mathbf{X}^* be the optimizer for $\min_{\mathbf{X}} \|\mathbf{A}\mathbf{X} - \mathbf{B}\|_F^2$ and let $\mathbf{B}^* = \mathbf{A}\mathbf{X}^* - \mathbf{B}$. Then, with probability at least $1 - c$, for all $\mathbf{X} \in \mathbb{R}^{m \times d}$

$$\|\mathbf{S}\mathbf{A}\mathbf{X} - \mathbf{S}\mathbf{B}\|_F^2 - \|\mathbf{S}\mathbf{B}^*\|_F^2 = (1 \pm \epsilon)\|\mathbf{A}\mathbf{X} - \mathbf{B}\|_F^2 - \|\mathbf{B}^*\|_F^2$$

for a fixed small constant c .

An important application of the above lemma (which we use extensively) is to sketch constrained regression problems, for example, when the matrix \mathbf{X} has a fixed small rank. Since affine embeddings approximately preserve the cost of all affine spaces up to a fixed shift, this guarantee in particular holds for \mathbf{X} with small rank. Recall, an important caveat here is that the cost of the sketched problem is not a relative-error approximation to the cost of the original problem since we cannot estimate $\|\mathbf{B}^*\|_F^2$ in general. However, the upshot here is that the aforementioned guarantee still suffices for optimization since the fixed shift does not change the optimizer.

The next tool we use is input-sparsity time low-rank approximation. This was achieved by Clarkson and Woodruff [CW13] and the exact dependence on k, ϵ was improved in subsequent works [MM13, NN13, BDN15, Coh16]. While the standard low-rank approximation guarantee achieves relative-error under Frobenius norm, here we will require a spectral norm bound, which follows from results of [CEM⁺15, CMM17].

Lemma 4.6. (Input-Sparsity Spectral LRA [CEM⁺15, CMM17].) Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\epsilon, \delta > 0$ and $k \in \mathbb{N}$, let $k' = k/\epsilon$. Then, there exists an algorithm that outputs a matrix $\mathbf{Z}^\top \in \mathbb{R}^{k' \times n}$ such that with probability at least $1 - \delta$,

$$\|\mathbf{A} - \mathbf{A}\mathbf{Z}\mathbf{Z}^\top\|_2^2 \leq O\left(\frac{\epsilon}{k}\right) \|\mathbf{A} - \mathbf{A}_{k/\epsilon}\|_F^2$$

in time and query complexity $\tilde{O}(nnz(\mathbf{A}) + (n + d)\text{poly}(k/\epsilon\delta))$.

Proof. By Lemma 18 from [CEM⁺15] it suffices to use any obvious subspace embedding matrix with $\epsilon = O(1)$ and $k = k/\epsilon$. Here, we use OSNAP in the regime that requires $\tilde{O}(k/\epsilon^2)$ rows and sparsity $\text{polylog}(k)/\epsilon$ [NN13]. Instantiating this OSNAP construction with $\epsilon = O(1)$ and $k = k/\epsilon$ results in \mathbf{Z}^\top with k/ϵ rows in the desired running time. \square

Next we define the ridge leverage scores of a matrix. The ridge leverage scores were used as sampling probabilities in the context of linear regression and spectral approximation [LMP13, KLM⁺17, AM15], and low-rank approximation [CMM17, MW17]. Intuitively, the ridge leverage scores can be thought of as adding a regularization term that attenuates the smaller singular directions such that they are sampled with proportionately lower probability.

Definition 4.7. (Ridge Leverage Scores.) Given a matrix $\mathbf{M} \in \mathbb{R}^{n \times m}$ and an integer k , let $\mathbf{m}_i = \mathbf{M}_{i,*}$ be the i -th row of \mathbf{M} . Then, for all $i \in [n]$, the i -th rank- k ridge leverage score of \mathbf{M} is

$$\rho_i^k(\mathbf{M}) = \mathbf{m}_i \left(\mathbf{M}^\top \mathbf{M} + \frac{\|\mathbf{M} - \mathbf{M}_k\|_F^2}{k} \mathbf{I} \right)^+ \mathbf{m}_i^\top$$

Since we typically use the row ridge leverage scores to define a probability distribution over the rows and sample according to this distribution, it is crucial that their sum is small as this controls the number of rows we would need to sample. This follows from a straightforward calculation:

Lemma 4.8. (Lemma 4 from [CMM17].) Let $\rho_i^k(\mathbf{M})$ be the i -th ridge leverage score of \mathbf{M} . Then,

$$\sum_{i \in [n]} \rho_i^k(\mathbf{M}) \leq 2k$$

Cohen et. al. [CMM17] show that the ridge leverage scores of a matrix can be approximated up to a small constant in $O(nnz(\mathbf{A}))$ time, however this involves reading the entire matrix \mathbf{A} . For the special case of \mathbf{A} being PSD, Musco and Musco [MM17] show that the ridge leverage scores of $\mathbf{A}^{\frac{1}{2}}$ can be approximated up to a small constant using a so-called Nyström approximation.

Lemma 4.9. (Lemma 4 of [MW17].) Given a PSD matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ and integer k , there exists an algorithm that accesses $O(nk \log(k/\delta))$ entries in \mathbf{A} and computes $\hat{\rho}_i^k(\mathbf{A}^{\frac{1}{2}})$ for all $i \in [n]$, such that with probability $1 - \delta$,

$$\rho_i^k(\mathbf{A}^{1/2}) \leq \hat{\rho}_i^k(\mathbf{A}^{1/2}) \leq 3\rho_i^k(\mathbf{A}^{1/2})$$

and runs in time $O(n(k \log(k/\delta))^{\omega-1})$, where ω is the matrix multiplication exponent.

Note, while it is not known how to compute ridge leverage scores of a PSD matrix in sublinear time, Musco and Woodruff [MW17] show that the ridge leverage scores of $\mathbf{A}^{\frac{1}{2}}$ are a coarse approximation to the ridge leverage scores of \mathbf{A} .

Lemma 4.10. (Lemma 5 in [MW17].) Given a PSD matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, for all $i \in [n]$,

$$\rho_i^k(\mathbf{A}) \leq 2\sqrt{\frac{n}{k}} \rho_i^k(\mathbf{A}^{\frac{1}{2}})$$

Musco and Woodruff then show that sampling columns of \mathbf{A} , according to the corresponding ridge leverage scores of $\mathbf{A}^{\frac{1}{2}}$, suffices to obtain a column projection-cost preserving sketch (PCP), if we oversample by a $\sqrt{n/k}$ factor. Projection-cost preserving sketches were introduced by Feldman et. al. [FSS13] and Cohen et. al. [CEM⁺15] and studied in the context of low-rank approximation in [CMM17, MW17, BW18].

Lemma 4.11. (Column PCP from [MW17].) Given a PSD matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, integer k and $\epsilon > 0$, for all $j \in [n]$ let $\bar{\rho}_j^k(\mathbf{A}^{\frac{1}{2}})$ be a constant approximation to the column-ridge leverage scores of $\mathbf{A}^{\frac{1}{2}}$. Let $q_j = \bar{\rho}_j^k(\mathbf{A}^{\frac{1}{2}}) / \sum_j \bar{\rho}_j^k(\mathbf{A}^{\frac{1}{2}})$ and let $t = O\left(\sqrt{\frac{n}{k}} \sum_j \bar{\rho}_j^k(\mathbf{A}^{\frac{1}{2}}) \log(k/\delta) / \epsilon^2\right) = O\left(\sqrt{nk} \log(k/\delta) / \epsilon^2\right)$. Construct $\mathbf{C} \in \mathbb{R}^{n \times t}$ by sampling t columns of \mathbf{A} and setting each one to be $\frac{1}{\sqrt{tq_j}} \mathbf{A}_{*,j}$ with probability q_j . Then, with probability $1 - \delta$, for any rank- k projection matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$,

$$(1 - \epsilon) \|\mathbf{A} - \mathbf{XA}\|_F^2 \leq \|\mathbf{C} - \mathbf{XC}\|_F^2 \leq (1 + \epsilon) \|\mathbf{A} - \mathbf{XA}\|_F^2$$

Further, such a \mathbf{C} can be computed by accessing $\tilde{O}(nk)$ entries in \mathbf{A} and in time $O(nk^{\omega-1})$.

This result also implies that the resulting matrix \mathbf{C} is a *Spectral-Frobenius PCP* for \mathbf{A} (Lemma 24 in [MW17]), i.e., for any rank- k projection matrix \mathbf{X} ,

$$(1 - \epsilon) \|\mathbf{A} - \mathbf{XA}\|_2^2 - \frac{\epsilon}{k} \|\mathbf{A} - \mathbf{A}_k\|_F^2 \leq \|\mathbf{C} - \mathbf{XC}\|_2^2 \leq (1 + \epsilon) \|\mathbf{A} - \mathbf{XA}\|_2^2 + \frac{\epsilon}{k} \|\mathbf{A} - \mathbf{A}_k\|_F^2 \quad (4.1)$$

As noted by Musco and Woodruff, the resulting matrix \mathbf{C} is not even square and thus is it unclear how to sample rows of \mathbf{C} to obtain a row-PCP in sublinear time and queries. In particular, the ridge-leverage scores of rows of \mathbf{C} can be an n/k -factor larger than the corresponding ridge-leverage scores of $\mathbf{A}^{\frac{1}{2}}$. Instead, Musco and Woodruff sample rows of \mathbf{C} proportional to the rank- k/ϵ^2 ridge leverage scores of $\mathbf{A}^{\frac{1}{2}}$. In addition, they show the stronger guarantee that a *Spectral-Frobenius PCP* holds (by Lemma 8 of [MW17]) for PSD Matrices.

Lemma 4.12. (Spectral-Frobenius PCP.) Given a PSD matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, an integer k and $\epsilon > 0$, let $\mathbf{C} \in \mathbb{R}^{n \times t}$ be a column PCP for \mathbf{A} , following Lemma 4.11. Let $k' = k/\epsilon^2$. For all $i \in [n]$, let $\bar{\rho}_i^{k'}(\mathbf{A}^{\frac{1}{2}})$ be a constant approximation to the rank- k' row ridge leverage scores of $\mathbf{A}^{\frac{1}{2}}$. Let $p_i = \bar{\rho}_i^{k'}(\mathbf{A}^{\frac{1}{2}}) / \sum_i \bar{\rho}_i^{k'}(\mathbf{A}^{\frac{1}{2}})$ and let $t = O\left(\sqrt{\frac{n}{k}} \sum_i \bar{\rho}_i^{k'}(\mathbf{A}^{\frac{1}{2}}) \log(n)/\epsilon\right) = O\left(\sqrt{nk} \log(n)/\epsilon^3\right)$. Then, with probability $1 - c$, for all rank- k' projection matrices \mathbf{X} ,

$$(1 - \epsilon) \|\mathbf{C} - \mathbf{CX}\|_2^2 - \frac{\epsilon}{k} \|\mathbf{A} - \mathbf{A}_k\|_F^2 \leq \|\mathbf{R} - \mathbf{RX}\|_2^2 \leq (1 + \epsilon) \|\mathbf{C} - \mathbf{CX}\|_2^2 + \frac{\epsilon}{k} \|\mathbf{A} - \mathbf{A}_k\|_F^2$$

We observe that we could compute a low-rank approximation to \mathbf{R} in input sparsity time, which already requires querying $\Omega(\text{nnz}(\mathbf{R})) = \Omega(nk/\epsilon^4)$ entries in \mathbf{A} and is far from optimal in terms of the dependence on ϵ . It is here that we digress from the approach of Musco and Woodruff. We observe that the dependence on n and k is optimal and thus we instantiate the aforementioned column and row PCPs with $\epsilon = O(1)$ and $k = k/\epsilon$. While this results in weaker PCP guarantees, the resulting matrix \mathbf{R} is a $\sqrt{nk/\epsilon} \times \sqrt{nk/\epsilon}$ matrix and we can now afford to read all of it and thus we can compute a rank- k low-rank approximation to \mathbf{R} using the input sparsity time algorithm of Clarkson and Woodruff [CW13].

However, the main technical challenge here is that we can no longer use the approach of [CMM17, MW17, BW18] to use the low-rank approximation for \mathbf{R} and solve regression problems to recover an ϵ -approximate low-rank matrix for \mathbf{A} . In particular, we can now only hope for an $O(1)$ approximation if we use the standard technique of iteratively solving regression problems. Our first insight is that computing a Spectral Low-Rank Approximation to \mathbf{R} results in a *structured* projection matrix for \mathbf{C} , from which we can compute a *structured* projection matrix for \mathbf{A} . Further, this structured projection can be computed with only $\tilde{O}(nk/\epsilon)$ queries. We first describe how this structured projection matrix for \mathbf{A} results in an efficient low-rank approximation algorithm.

4.1 Structured Projections to Low-Rank Approximation

Our starting point is a structural result based on the Spectral-Frobenius projection (SF) property introduced by Clarkson and Woodruff in the context of approximating arbitrary matrices with low-rank PSD matrices [CW17]. In this subsection, we show that if we are given a projection matrix that satisfies the SF property, we can obtain a query-optimal algorithm for Low-Rank Approximation. We begin by defining this property:

Definition 4.13. ((ϵ, k) -SF Projection.) Given any matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, integer k , and $\epsilon > 0$, a projection matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$ is (ϵ, k) -SF w.r.t. \mathbf{A} if

$$\|\mathbf{A} - \mathbf{A}\mathbf{P}\|_2^2 \leq \frac{\epsilon}{k} \|\mathbf{A} - \mathbf{A}_k\|_F^2$$

or

$$\|\mathbf{A} - \mathbf{P}\mathbf{A}\|_2^2 \leq \frac{\epsilon}{k} \|\mathbf{A} - \mathbf{A}_k\|_F^2$$

Intuitively, the following structural result of Clarkson and Woodruff relates an (ϵ, k) -SF projection to a relative-error low-rank approximation. We leverage this connection heavily in subsequent sections.

Lemma 4.14. (Structured Projections and Low-Rank Approximation [CW17].) Let $\mathbf{P} \in \mathbb{R}^{n \times n}$ be an (ϵ, k) -SF projection w.r.t \mathbf{A} , then

$$\|\mathbf{A} - \mathbf{P}\mathbf{A}_k\mathbf{P}\|_F^2 \leq (1 + \epsilon) \|\mathbf{A} - \mathbf{A}_k\|_F^2$$

Ignoring computational and query complexity constraints, suppose we were given a matrix $\mathbf{Q} \in \mathbb{R}^{n \times k'}$ with orthonormal columns such that $\mathbf{P} = \mathbf{Q}\mathbf{Q}^\top$ is an (ϵ, k) -SF Projection, where k' is

the dimension of the space \mathbf{P} projects onto. Note, for now it suffices to set $k' = \text{poly}(k/\epsilon)$. As a consequence of Lemma 4.14, we observe that solving the following constrained regression problem suffices to obtain a $(1 + \epsilon)$ -relative error solution to the Low-Rank Approximation problem:

$$\min_{\text{rank}(\mathbf{X}) \leq k} \|\mathbf{A} - \mathbf{Q}\mathbf{X}\mathbf{Q}^\top\|_F^2 \quad (4.2)$$

Algorithm 1 : Structured Projection to Low-Rank Approximation

Input: A PSD Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, integer $k, \epsilon > 0$, an orthonormal matrix $\mathbf{Q} \in \mathbb{R}^{n \times k'}$ such that the projection matrix $\mathbf{P} = \mathbf{Q}\mathbf{Q}^\top$ satisfies $\|\mathbf{A} - \mathbf{P}\mathbf{A}\|_2^2 \leq \frac{\epsilon}{k} \|\mathbf{A} - \mathbf{A}_k\|_F^2$

1. Consider the optimization problem:

$$\min_{\text{rank}(\mathbf{X}) \leq k} \|\mathbf{A} - \mathbf{Q}\mathbf{X}\mathbf{Q}^\top\|_F^2$$

2. For all $i \in [n]$, compute the leverage scores, $\tau_i(\mathbf{Q})$. Since \mathbf{Q} has orthonormal columns, $\tau_i(\mathbf{Q}) = \|\mathbf{Q}_{i,*}\|_2^2$ and can be computed exactly. Let $p = \{p_1, p_2, \dots, p_n\}$ denote a distribution over rows of \mathbf{A} for which $p_i = \tau_i(\mathbf{Q}) / \sum_{i'} \tau_{i'}(\mathbf{Q})$.
3. Let $t = k' / \epsilon^2$. Construct a *leverage score sampling* matrix \mathbf{S} by sampling t rows of \mathbf{A} , such that $\mathbf{S} = \mathbf{D}\mathbf{\Omega}^\top$, where \mathbf{D} is a $t \times t$ diagonal matrix and $\mathbf{\Omega}$ is an $n \times t$ sampling matrix. For all $j \in [t]$, select row index $i \in [n]$ with probability p_i and set $\Omega_{i,j} = 1$ and $D_{j,j} = 1/\sqrt{tp_i}$. Repeat this sampling process to construct another *leverage score sampling* matrix \mathbf{T} .
4. Consider the sketched optimization problem :

$$\min_{\text{rank}(\mathbf{X}) \leq k} \|\mathbf{S}\mathbf{A}\mathbf{T} - \mathbf{S}\mathbf{Q}\mathbf{X}\mathbf{Q}^\top\mathbf{T}\|_F^2$$

Compute $\mathbf{S}\mathbf{A}\mathbf{T}$, $\mathbf{P}_{\mathbf{S}\mathbf{Q}}$, $\mathbf{P}_{\mathbf{Q}^\top\mathbf{T}}(\mathbf{S}\mathbf{Q})^\dagger$ and $(\mathbf{Q}^\top\mathbf{T})^\dagger$, where $\mathbf{P}_{\mathbf{S}\mathbf{Q}}$ and $\mathbf{P}_{\mathbf{Q}^\top\mathbf{T}}$ are the projections onto $\mathbf{S}\mathbf{Q}$ and $\mathbf{Q}^\top\mathbf{T}$ respectively. Compute $\text{SVD}(\mathbf{P}_{\mathbf{S}\mathbf{Q}}\mathbf{S}\mathbf{A}\mathbf{T}\mathbf{P}_{\mathbf{Q}^\top\mathbf{T}})$. By Theorem 4.15 the sketched problem is minimized by $\mathbf{X}^* = (\mathbf{S}\mathbf{Q})^\dagger [\mathbf{P}_{\mathbf{S}\mathbf{Q}}\mathbf{S}\mathbf{A}\mathbf{T}\mathbf{P}_{\mathbf{Q}^\top\mathbf{T}}]_k (\mathbf{Q}^\top\mathbf{T})^\dagger$.

5. Let $\mathbf{U}^* \in \mathbb{R}^{k' \times k}$ be an orthonormal basis for the columns of \mathbf{X}^* . Compute an orthonormal basis \mathbf{M} for $\mathbf{Q}\mathbf{U}^*$. Consider the following regression problem: $\min_{\mathbf{Y} \in \mathbb{R}^{k \times n}} \|\mathbf{A} - \mathbf{M}\mathbf{Y}\|_F^2$. For all $i \in [n]$, compute $\tau_i(\mathbf{M}) = \|\mathbf{M}_{i,*}\|_2^2$. Let $q = \{q_1, q_2, \dots, q_n\}$ be a distribution over the rows of \mathbf{A} such that $q_i = \tau_i(\mathbf{M}) / \sum_{i' \in [n]} \tau_{i'}(\mathbf{M})$. Let \mathbf{W} be a *leverage score sampling* matrix with k/ϵ rows sampled proportional to q .
6. Consider the sketched regression problem: $\min_{\mathbf{Y} \in \mathbb{R}^{k \times n}} \|\mathbf{W}\mathbf{A} - \mathbf{W}\mathbf{M}\mathbf{Y}\|_F^2$. Let \mathbf{N} be the minimizer to this regression problem computed using the algorithm from Lemma 4.4.

Output: $\mathbf{M}, \mathbf{N}^\top \in \mathbb{R}^{n \times k}$ such that $\|\mathbf{A} - \mathbf{M}\mathbf{N}\|_F^2 \leq (1 + \epsilon) \|\mathbf{A} - \mathbf{A}_k\|_F^2$

However, there are several challenges pertaining to this approach. As noted above, it is not immediately clear how to obtain such a \mathbf{Q} with nk/ϵ queries to \mathbf{A} . Further, it is not immediately

clear how to solve Equation 4.2 efficiently. While we have reduced to optimizing over $k' \times k'$ sized matrices \mathbf{X} with rank at most k , the problem still seems intractable in sublinear time and queries.

We begin by describing how to solve the optimization problem in Equation 4.2 with $\tilde{O}(nk/\epsilon)$ queries given that we have access to \mathbf{Q} and $\mathbf{Q}\mathbf{Q}^\top$ is an (ϵ, k) -SF Projection. At a high level, our approach is to sketch the problem by sampling rows and columns proportional to the row leverage scores of \mathbf{Q} . We observe that since \mathbf{Q} has orthonormal columns, the row leverage scores of \mathbf{Q} are simply the ℓ_2^2 norms of corresponding rows. Therefore, we create sampling matrices \mathbf{S} and \mathbf{T} that sample $\text{poly}(k')$ rows proportional to the leverage scores of \mathbf{Q} and consider the resulting optimization problem:

$$\min_{\text{rank}(\mathbf{X}) \leq k} \|\mathbf{SAT} - \mathbf{SQXQ}^\top \mathbf{T}\|_F^2 \quad (4.3)$$

We then show that the minimizer for Equation 4.3 is an approximate minimizer for Equation 4.2. Further, the optimization problem in Equation 4.3 is referred to as Generalized Low-Rank Approximation and admits a closed form solution:

Theorem 4.15. (Generalized Low-Rank Approximation [FT07].) Let $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times k'}$ and $\mathbf{C} \in \mathbb{R}^{k' \times n}$ and $k \in \mathbb{N}$. Then, the Generalized Low-Rank Approximation problem

$$\min_{\text{rank}(\mathbf{X}) \leq k} \|\mathbf{A} - \mathbf{BXC}\|_F^2$$

is minimized by $\mathbf{X} = \mathbf{B}^\dagger [\mathbf{P}_B \mathbf{A} \mathbf{P}_C]_k \mathbf{C}^\dagger$, where $\mathbf{P}_B, \mathbf{P}_C$ are the projection matrices onto \mathbf{B} and \mathbf{C} respectively.

We apply the above theorem to Equation 4.3. Both the query complexity and running time here contribute a lower-order term and we can afford to compute the SVD for each term. Let \mathbf{X}^* be the solution to the sketched optimization problem in Equation 4.3. Then, we can compute \mathbf{U}^* , an orthonormal column basis for \mathbf{X}^* and consider \mathbf{M} , an orthonormal basis for $\mathbf{Q}\mathbf{U}^* \in \mathbb{R}^{n \times k}$ to be one of the low-rank factors for \mathbf{A} . To find the second factor, we set up the following regression problem:

$$\min_{\mathbf{N} \in \mathbb{R}^{k \times n}} \|\mathbf{A} - \mathbf{MN}\|_F^2 \quad (4.4)$$

Again, \mathbf{M} has orthonormal columns and thus we can efficiently compute the corresponding row leverage scores and sample k/ϵ rows. By Lemma 4.5 this achieves a $(1 + \epsilon)$ -approximation to the optimal cost in Equation 4.4 and obtains an \mathbf{N}^* with $\tilde{O}(nk/\epsilon)$ queries to \mathbf{A} . At this stage, we have obtained a $(1 + \epsilon)$ -approximate rank- k solution to Equation 4.2 and Lemma 4.14 implies that we are done. We now formalize this argument:

Theorem 4.16. (Structured Projection to Low-Rank Approximation.) Given a rank- k' projection matrix $\mathbf{P} = \mathbf{Q}\mathbf{Q}^\top$, such that \mathbf{P} is an (ϵ, k) -SF projection, Algorithm 1 queries $\tilde{O}(nk/\epsilon + k'^2/\epsilon^4)$ entries in \mathbf{A} and with probability 99/100 outputs $\mathbf{M}, \mathbf{N}^\top \in \mathbb{R}^{n \times k}$ such that

$$\|\mathbf{A} - \mathbf{MN}\|_F^2 \leq (1 + \epsilon) \|\mathbf{A} - \mathbf{A}_k\|_F^2$$

Further, Algorithm 1 runs in time $\tilde{O}(n(k/\epsilon)^{\omega-1} + nk'^{\omega-1} + (k'/\epsilon^2)^\omega)$.

Proof. Since \mathbf{P} is an (ϵ, k) -SF projection and \mathbf{A}_k is a feasible solution to $\min_{\text{rank}(\mathbf{Y}) \leq k} \|\mathbf{A} - \mathbf{PYP}\|_F^2$, from Lemma 4.14 we have

$$\min_{\text{rank}(\mathbf{Y}) \leq k} \|\mathbf{A} - \mathbf{PYP}\|_F^2 \leq (1 + \epsilon) \|\mathbf{A} - \mathbf{A}_k\|_F^2 \quad (4.5)$$

Since $\mathbf{P} = \mathbf{QQ}^\top$, we can substitute it in Equation 4.5 to get $\min_{\text{rank}(\mathbf{Y}) \leq k} \|\mathbf{A} - \mathbf{QQ}^\top \mathbf{YQQ}^\top\|_F^2$. We further relax this by optimizing over all rank- k matrices $\mathbf{X} \in \mathbb{R}^{k' \times k'}$ instead of matrices of the form \mathbf{QYQ}^\top . Therefore,

$$\min_{\text{rank}(\mathbf{X}) \leq k} \|\mathbf{A} - \mathbf{QXQ}^\top\|_F^2 \leq (1 + \epsilon) \|\mathbf{A} - \mathbf{A}_k\|_F^2 \quad (4.6)$$

where we are now optimizing over a $k' \times k'$ matrix \mathbf{X} , which is considerably smaller than \mathbf{Y} . Let $\mathbf{S}, \mathbf{T}^\top \in \mathbb{R}^{k'/\epsilon^2 \times n}$ be the *leverage score sampling* matrices as defined in Algorithm 1. Observe, from Lemma 4.5 we know \mathbf{S} has a sufficient number of rows to be an affine embedding for Equation 4.6. However, we cannot directly apply the affine embedding guarantee since $\mathbf{A} - \mathbf{QXQ}^\top$ is not an affine space. Let \mathbf{H} be $k' \times n$ matrix, let $\mathbf{H}^* = \arg \min_{\mathbf{H}} \|\mathbf{A} - \mathbf{QH}\|_F^2$ and let \mathbf{A}^* be $\mathbf{A} - \mathbf{QH}^*$. Then, with probability at least $1 - c_1$, for all \mathbf{H} ,

$$\|\mathbf{SA} - \mathbf{SQH}\|_F^2 - \|\mathbf{SA}^*\|_F^2 = (1 \pm \epsilon) \|\mathbf{A} - \mathbf{QH}\|_F^2 + \|\mathbf{A}^*\|_F^2 \quad (4.7)$$

Since Equation 4.7 holds for all \mathbf{H} , in particular it holds for all rank- k matrices \mathbf{X} such that $\mathbf{H} = \mathbf{XQ}^\top$. Therefore, with probability at least $1 - c_1$, for all rank k matrices \mathbf{X} ,

$$\|\mathbf{SA} - \mathbf{SQXQ}^\top\|_F^2 - \|\mathbf{SA}^*\|_F^2 = (1 \pm \epsilon) \|\mathbf{A} - \mathbf{QXQ}^\top\|_F^2 + \|\mathbf{A}^*\|_F^2 \quad (4.8)$$

Here, we observe that while we cannot estimate $\|\mathbf{A}^*\|_F^2$ accurately, it is a fixed matrix independent of \mathbf{X} and thus we can still approximately optimize. Let ζ_1 be the event that Equation 4.8 holds. We now use the sampling matrix \mathbf{T} to sketch $\|\mathbf{SA} - \mathbf{ZQ}^\top\|_F^2$. Let $\mathbf{Z}' = \arg \min_{\mathbf{Z}} \|\mathbf{SA} - \mathbf{ZQ}^\top\|_F^2$ and let $\mathbf{SA}' = \mathbf{SA} - \mathbf{Z}'\mathbf{Q}^\top$. Then, with probability at least $1 - c_2$, for all \mathbf{Z} ,

$$\|\mathbf{SAT} - \mathbf{ZQ}^\top \mathbf{T}\|_F^2 - \|\mathbf{SA}'\mathbf{T}\|_F^2 = (1 \pm \epsilon) \|\mathbf{SA} - \mathbf{ZQ}^\top\|_F^2 + \|\mathbf{SA}'\|_F^2 \quad (4.9)$$

In particular, the above equation holds for all rank- k matrices \mathbf{X} such that $\mathbf{Z} = \mathbf{SQ}^\top \mathbf{X}$. Let ζ_2 be the event that the aforementioned equation holds. Combining equations 4.8 and 4.9 and conditioning on ζ_1 and ζ_2 , for all rank- k matrices \mathbf{X} ,

$$\|\mathbf{SAT} - \mathbf{SQXQ}^\top \mathbf{T}\|_F^2 - \|\mathbf{SA}'\mathbf{T}\|_F^2 = (1 \pm \epsilon)^2 (\|\mathbf{A} - \mathbf{QXQ}^\top\|_F^2 + \|\mathbf{SA}^*\|_F^2 + \|\mathbf{A}^*\|_F^2) + \|\mathbf{SA}'\|_F^2 \quad (4.10)$$

Here, we observe that while the sketch does not preserve the cost of all \mathbf{X} up to relative error $(1 + \epsilon)$, the additive error $\Delta \leq (1 + \epsilon) (\|\mathbf{SA}^*\|_F^2 + \|\mathbf{A}^*\|_F^2 + \|\mathbf{SA}'\|_F^2 + \|\mathbf{SA}'\mathbf{T}\|_F^2)$ is fixed and is independent of \mathbf{X} . Let $\mathbf{X}^* = \arg \min_{\text{rank}(\mathbf{X}) \leq k} \|\mathbf{SAT} - \mathbf{SQXQ}^\top \mathbf{T}\|_F^2$. Then, union bounding over ζ_1 and ζ_2 , with probability $1 - c_1 - c_2$,

$$\|\mathbf{A} - \mathbf{QX}^*\mathbf{Q}^\top\|_F^2 \leq (1 + \epsilon) \min_{\text{rank}(\mathbf{X}) \leq k} \|\mathbf{A} - \mathbf{SQXQ}^\top\|_F^2 \quad (4.11)$$

Therefore, it suffices to efficiently compute \mathbf{X}^* . By Theorem 4.15, we know that the sketched optimization problem above is minimized by $\mathbf{X}^* = (\mathbf{SQ})^\dagger [\mathbf{P}_{\mathbf{SQ}} \mathbf{SATP}_{\mathbf{Q}^\top \mathbf{T}}]_k (\mathbf{Q}^\top \mathbf{T})^\dagger$, which can be

computed exactly as shown in Step 4 of Algorithm 1. We note that we can now explicitly compute \mathbf{SAT} by querying the relevant entries in \mathbf{A} . Further, we can compute \mathbf{SQ} and $\mathbf{Q}^\top \mathbf{T}$ without querying \mathbf{A} at all. Recalling equation 4.11 we can approximate the optimal low rank approximation cost:

$$\|\mathbf{A} - \mathbf{Q}\mathbf{X}^*\mathbf{Q}^\top\|_F^2 \leq (1 + O(\epsilon))\|\mathbf{A} - \mathbf{A}_k\|_F^2$$

While we have now approximately minimized the optimization problem from Equation 4.2, recall our goal was to obtain a rank- k approximation to \mathbf{A} in factored form i.e., outputting $n \times k$ matrices $\mathbf{M}, \mathbf{N}^\top$ such that the low rank approximation is given by \mathbf{MN} . Towards this end, we compute \mathbf{U}^* , an orthonormal column basis for \mathbf{X}^* such that $\mathbf{X}^* = \mathbf{U}^*\mathbf{V}^*$. Substituting this in the above equation we have

$$\|\mathbf{A} - \mathbf{Q}\mathbf{U}^*\mathbf{V}^*\mathbf{Q}^\top\|_F^2 \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2 \quad (4.12)$$

Let $\mathbf{M} = \mathbf{Q}\mathbf{U}^* \in \mathbb{R}^{n \times k}$ be one of the low-rank factors for \mathbf{A} . To find the second factor, we observe :

$$\min_{\mathbf{Y} \in \mathbb{R}^{k \times n}} \|\mathbf{A} - \mathbf{M}\mathbf{Y}\|_F^2 \leq \|\mathbf{A} - \mathbf{M}\mathbf{V}^*\mathbf{Q}^\top\|_F^2 \quad (4.13)$$

and therefore, approximately optimizing $\min_{\mathbf{Y} \in \mathbb{R}^{k \times n}} \|\mathbf{A} - \mathbf{M}\mathbf{Y}\|_F^2$ suffices. Again, \mathbf{M} has orthonormal columns and thus we can efficiently compute the corresponding leverage scores to create a sketch \mathbf{W} with $O(k/\epsilon)$ rows. From Lemma 4.4, with probability at least $1 - c_3$ for all \mathbf{Y} ,

$$\|\mathbf{W}\mathbf{A} - \mathbf{W}\mathbf{M}\mathbf{Y}\|_F^2 = (1 \pm \epsilon)\|\mathbf{A} - \mathbf{M}\mathbf{Y}\|_F^2$$

Let \mathbf{N} be the optimal solution for the sketched problem as defined in Algorithm 1. Then, with probability at least $1 - c_3$,

$$\|\mathbf{A} - \mathbf{MN}\|_F^2 \leq \left(\frac{1 + \epsilon}{1 - \epsilon}\right) \min_{\mathbf{Y} \in \mathbb{R}^{k \times n}} \|\mathbf{A} - \mathbf{M}\mathbf{Y}\|_F^2 \quad (4.14)$$

We conclude correctness by union bounding over the failure probabilities of all the sketches and observing that with probability at least 99/100,

$$\|\mathbf{A} - \mathbf{MN}\|_F^2 \leq (1 + O(\epsilon))\|\mathbf{A} - \mathbf{M}\mathbf{V}^*\mathbf{Q}^\top\|_F^2 \leq (1 + O(\epsilon))\|\mathbf{A} - \mathbf{A}_k\|_F^2$$

where the inequalities follow from Equations 4.4, 4.13 and 4.12.

Finally, we analyze the query complexity and running time of our algorithm. Since Algorithm 1 is given \mathbf{Q} as input, computing the leverage scores in Step 2 requires no queries to \mathbf{A} and requires $O(nk')$ time. Next, observe we do not have to explicitly compute \mathbf{SA} or \mathbf{AT} , since \mathbf{SAT} is simply a submatrix of \mathbf{A} with $(k'/\epsilon^2)^2$ entries appropriately scaled, it suffices to query them. \mathbf{SAT} can be computed in $O(k'^2/\epsilon^4)$ time. Next, we compute $\text{SVD}(\mathbf{SQ})$ and $\text{SVD}(\mathbf{Q}^\top \mathbf{T})$, which requires no queries to \mathbf{A} and time $O(k'^\omega/\epsilon^2)$. We can then compute $(\mathbf{SQ})^\dagger, (\mathbf{Q}^\top \mathbf{T})^\dagger, \mathbf{P}_{\mathbf{SQ}}$ and $\mathbf{P}_{\mathbf{Q}^\top \mathbf{T}}$ from the aforementioned SVDs. Next, we compute the matrix $\mathbf{P}_{\mathbf{SQ}}\mathbf{SAT}\mathbf{P}_{\mathbf{Q}^\top \mathbf{T}}$, which requires no extra queries to \mathbf{A} and time $O((k'/\epsilon^2)^\omega)$, which is also the time required to compute its SVD. We can then compute \mathbf{X}^* in Step 4 with a total of $O(k'^2/\epsilon^4)$ queries to \mathbf{A} in time $O(nk' + (k'/\epsilon^2)^\omega)$.

In Step 5, we can compute \mathbf{U}^* by computing the SVD of \mathbf{X}^* and compute \mathbf{M} in time $O(nk'^{\omega-1} + k'^\omega)$ and do not require any queries to \mathbf{A} . In Step 6, computing \mathbf{WA} requires $O(nk/\epsilon)$ queries to \mathbf{A} ,

since \mathbf{W} has $\tilde{O}(k/\epsilon)$ rows. Note, this step contributes the leading term to the query complexity and it is crucial \mathbf{W} does not have more rows. By Lemma 4.4, \mathbf{N} can be computed in time $\tilde{O}(nk/\epsilon + n(k/\epsilon)^{\omega-1} + k^3/\epsilon)$. Overall, Algorithm 1 requires $\tilde{O}(nk/\epsilon + k'^2/\epsilon^4)$ queries to \mathbf{A} and runs in time $\tilde{O}(n(k/\epsilon)^{\omega-1} + nk'^{\omega-1} + (k'/\epsilon^2)^\omega)$. \square

In light of Theorem 4.16, to obtain a low rank approximation for \mathbf{A} , it suffices to obtain an SF Projection. In particular, it suffices to obtain a matrix $\mathbf{Q} \in \mathbb{R}^{n \times k'}$, for $k' = \text{poly}(k, 1/\epsilon)$ such that $\mathbf{P} = \mathbf{Q}\mathbf{Q}^\top$ is an (ϵ, k) -SF projection, by querying $\tilde{O}(nk/\epsilon)$ entries in \mathbf{A} . One possible approach to computing such a \mathbf{Q} is to use the following result by Musco and Woodruff:

Theorem 4.17. (Theorem 25, [MW17].) *Given a PSD matrix \mathbf{A} , integer $k, \epsilon > 0$, there exists an algorithm that reads $\tilde{O}(nk/\epsilon^6 + nk^2/\epsilon^2)$ entries of \mathbf{A} and with probability at least 99/100, outputs $\mathbf{M}, \mathbf{N}^\top \in \mathbb{R}^{n \times k}$ such that*

$$\|\mathbf{A} - \mathbf{M}\mathbf{N}\|_2^2 \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{A}_k\|_2^2 + \frac{\epsilon}{k}\|\mathbf{A} - \mathbf{A}_k\|_F^2$$

Instantiating this theorem with $\epsilon = O(1)$ and $k = k/\epsilon$, we obtain a matrix $\mathbf{M}, \mathbf{N}^\top \in \mathbb{R}^{n \times k/\epsilon}$ such that

$$\begin{aligned} \|\mathbf{A} - \mathbf{M}\mathbf{N}^\top\|_2^2 &\leq O(1)\|\mathbf{A} - \mathbf{A}_{k/\epsilon}\|_2^2 + O\left(\frac{\epsilon}{k}\right)\|\mathbf{A} - \mathbf{A}_{k/\epsilon}\|_F^2 \\ &\leq O\left(\frac{\epsilon}{k}\right)\|\mathbf{A} - \mathbf{A}_k\|_F^2 \end{aligned}$$

where the last inequality follows from observing $\|\mathbf{A} - \mathbf{A}_{k/\epsilon}\|_F^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_F^2$ and

$$\|\mathbf{A} - \mathbf{A}_k\|_F^2 = \sum_{j=k+1}^n \sigma_j^2(\mathbf{A}) \geq \left(\frac{k}{\epsilon} - k\right) \sigma_{k/\epsilon}^2 \geq \left(\frac{k}{\epsilon} - k\right) \|\mathbf{A} - \mathbf{A}_{k/\epsilon}\|_2^2$$

We can then compute an orthonormal basis for \mathbf{M} and denote it by \mathbf{Q} . Here, we observe $\mathbf{P} = \mathbf{Q}\mathbf{Q}^\top$ is an (ϵ, k) -SF projection matrix. Further, the algorithm of Musco and Woodruff instantiated with the above parameters queries $\tilde{O}(nk^2/\epsilon^2)$ entries in \mathbf{A} . As a corollary of Theorem 4.16, providing the rank- k/ϵ projection matrix \mathbf{Q} as input to Algorithm 1, implies an algorithm for low rank approximation which queries $\tilde{O}(nk^2/\epsilon^2)$ entries in \mathbf{A} . This already improves the ϵ -dependence in the query complexity of best known algorithm for PSD low-rank approximation, since the algorithm of Musco and Woodruff requires $\tilde{O}(nk/\epsilon^{2.5})$ queries [MW17]. Note, this algorithm has worse dependence on k . However, our goal is to obtain linear dependence on both k and $1/\epsilon$. Towards this end, we focus on obtaining an SF projection with fewer queries to \mathbf{A} .

4.2 Spectral Regression

In this subsection, we consider the *Spectral Regression* problem. This problem is a natural generalization of least-squares regression, when the response variable is a matrix. *Spectral Regression* arises in the context of Regularized Least Squares Classification, for instance [CLL⁺10]. Given matrices $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\mathbf{X} \in \mathbb{R}^{d \times m}$ and $\mathbf{B} \in \mathbb{R}^{n \times m}$, the *Spectral Regression* problem considers the following optimization problem:

$$\min_{\mathbf{X}} \|\mathbf{A}\mathbf{X} - \mathbf{B}\|_2$$

We note that this is natural variant of multi-response regression where we minimize the difference between \mathbf{AX} and \mathbf{B} in spectral norm as opposed to the extensively studied and well-understood Frobenius norm. To the best of our knowledge the only relevant related work on *Spectral Regression* is by Clarkson and Woodruff [CW09] and Cohen et. al. [CNW15]. Both these works provide oblivious sketches to reduce the dimension of the problem, which unfortunately do not suffice for our application. Instead of *Spectral Regression* in its full generality we focus on the following special case:

Given an $n \times n$ PSD matrix \mathbf{A} , a rank parameter k , and an accuracy parameter ϵ , let \mathbf{C} be a $n \times \sqrt{nk/\epsilon}$ matrix such that it is a *column PCP* for \mathbf{A} , satisfying the guarantees of Lemma 4.11, instantiated with $k = k/\epsilon$, and $\epsilon = O(1)$. Let \mathbf{Z}^\top be a $k/\epsilon \times \sqrt{nk/\epsilon}$ matrix with orthonormal rows such that the corresponding projection matrix \mathbf{ZZ}^\top is an $(O(1), k/\epsilon)$ -SF Projection for \mathbf{C} . Then, we consider the following *Spectral Regression* problem:

$$\min_{\mathbf{W} \in \mathbb{R}^{n \times k/\epsilon}} \|\mathbf{C} - \mathbf{WZ}^\top\|_2 \quad (4.15)$$

Our main technical contribution here is to obtain a new algorithm to solve this optimization problem. We subsequently show how understanding this special case is crucial to obtaining *optimal* algorithms for low rank approximation of PSD matrices. The techniques we develop here may be of independent interest and find applications to other problems. Formally, we prove the following:

Algorithm 2 : Approximate Spectral Regression

Input: A PSD Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, integer k , and $\epsilon > 0$. $\mathbf{C} \in \mathbb{R}^{n \times \sqrt{nk/\epsilon}}$, a *column PCP* for \mathbf{A} satisfying the guarantees of Lemma 4.11 instantiated with $k = k/\epsilon$ and $\epsilon = O(1)$. $\mathbf{Z} \in \mathbb{R}^{\sqrt{nk/\epsilon} \times k/\epsilon}$ be an orthonormal matrix such that \mathbf{ZZ}^\top is an $(O(1), k/\epsilon)$ -SF projection for \mathbf{C} .

1. Consider the *Spectral Regression* problem:

$$\min_{\mathbf{W}} \|\mathbf{C} - \mathbf{WZ}^\top\|_2^2$$

Let $t = \sqrt{nk/\epsilon}$. For all $j \in [t]$, compute $\tau_j(\mathbf{Z}^\top) = \|\mathbf{Z}_{j,*}\|_2^2$. Let $q = \{q_1, q_2, \dots, q_t\}$ be a distribution of columns of \mathbf{C} such that for all $j \in [t]$, $q_j = \min(\tau_j(\mathbf{Z}^\top), 1)$.

2. Construct a sampling matrix \mathbf{S} such that \mathbf{CS} selects each column $\mathbf{C}_{*,j}$ independently with probability q_j and scales it by $1/\sqrt{q_j}$. Similarly, construct $\mathbf{Z}^\top \mathbf{S}$. Consider the sketched optimization problem :

$$\min_{\mathbf{W}} \|\mathbf{CS} - \mathbf{WZ}^\top \mathbf{S}\|_2^2$$

3. Compute $(\mathbf{Z}^\top \mathbf{S})^\dagger = \mathbf{S}^\top \mathbf{Z}(\mathbf{Z}^\top \mathbf{S} \mathbf{S}^\top \mathbf{Z})^{-1}$. Let $\widehat{\mathbf{W}} = \mathbf{CS}(\mathbf{Z}^\top \mathbf{S})^\dagger$ be the solution to the sketched optimization problem.

Output: $\widehat{\mathbf{W}} \in \mathbb{R}^{n \times k/\epsilon}$ such that $\|\mathbf{C} - \widehat{\mathbf{W}}\mathbf{Z}^\top\|_2^2 \leq \widetilde{O}(1) \min_{\mathbf{W}} \|\mathbf{C} - \mathbf{WZ}^\top\|_F^2 + \widetilde{O}(\epsilon/k) \|\mathbf{C} - \mathbf{C}_{k/\epsilon}\|_F^2$

Theorem 4.18. (Approximate Spectral Regression.) Let $\mathbf{C} \in \mathbb{R}^{n \times \sqrt{nk/\epsilon}}$ be a column PCP for \mathbf{A} satisfying the guarantees of Lemma 4.11 instantiated with $k = k/\epsilon$ and $\epsilon = O(1)$. Let $\mathbf{Z} \in \mathbb{R}^{\sqrt{nk/\epsilon} \times k/\epsilon}$ be an orthonormal matrix such that $\mathbf{Z}\mathbf{Z}^\top$ is an $(O(1), k/\epsilon)$ -SF projection for \mathbf{C} . Then, Algorithm 2 queries $\tilde{O}(nk/\epsilon)$ entries in \mathbf{A} and with probability 99/100 computes $\widehat{\mathbf{W}}$ such that

$$\|\mathbf{C} - \widehat{\mathbf{W}}\mathbf{Z}^\top\|_2^2 \leq \tilde{O}(1) \left(\min_{\mathbf{W} \in \mathbb{R}^{n \times k/\epsilon}} \|\mathbf{C} - \mathbf{W}\mathbf{Z}^\top\|_2^2 + \frac{\epsilon}{k} \|\mathbf{C} - \mathbf{C}_{k/\epsilon}\|_F^2 \right)$$

Further, the algorithm runs in time $\tilde{O}(nk/\epsilon + (k/\epsilon)^\omega)$.

We begin by characterizing the optimal solution and optimal cost for the *Spectral Regression* problem. We prove a structural result that shows the optimal solution for *Spectral Regression* is given by projecting \mathbf{C} away from the span of \mathbf{Z}^\top . This matches the characterization of the optimal solution to regression under the Frobenius norm, given by the well-known normal equations. Recall, by definition of the Moore-Penrose pseudoinverse, this projection matrix is $(\mathbf{Z}^\top)^\dagger \mathbf{Z}^\top$.

Then, the optimal cost for Equation 4.15 is $\|\mathbf{C} - \mathbf{C}(\mathbf{Z}^\top)^\dagger \mathbf{Z}^\top\|_2^2$ and is achieved by $\mathbf{W}^* = \mathbf{C}(\mathbf{Z}^\top)^\dagger$. Intuitively, we show that any feasible \mathbf{W} must incur the above cost by analyzing $\|y^\top (\mathbf{C} - \mathbf{C}(\mathbf{Z}^\top)^\dagger \mathbf{Z}^\top)\|_2^2$ for a fixed vector y . This enables us to exploit the geometry of Euclidean space and instantiate y as needed to relate it back to the spectral norm.

Lemma 4.19. (Characterizing Opt for Spectral Regression.) Let \mathbf{C} and \mathbf{Z} be matrices as defined in Theorem 4.18. Let $\mathbf{W}^* = \mathbf{C}(\mathbf{Z}^\top)^\dagger = \mathbf{C}\mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}$, such that $\mathbf{W}^* \mathbf{Z}^\top$ is the projection of \mathbf{C} on the $\text{colspan}(\mathbf{Z})$. Let $\mathbf{C}^* = \mathbf{C} - \mathbf{W}^* \mathbf{Z}^\top$ be the projection of \mathbf{C} orthogonal to $\text{colspan}(\mathbf{Z})$. Then,

$$\|\mathbf{C}^*\|_2^2 = \min_{\mathbf{W}} \|\mathbf{C} - \mathbf{W}\mathbf{Z}^\top\|_2^2$$

and the corresponding minimizer is \mathbf{W}^* .

Proof. Note, by definition $\|\mathbf{C} - \mathbf{W}^* \mathbf{Z}^\top\|_2^2 = \|\mathbf{C}^*\|_2^2$ and since \mathbf{W}^* is feasible,

$$\min_{\mathbf{W}} \|\mathbf{C} - \mathbf{W}\mathbf{Z}^\top\|_2^2 \leq \|\mathbf{C}^*\|_2^2.$$

Therefore, it suffices to show any \mathbf{W} must incur cost at least $\|\mathbf{C}^*\|_2^2$. By definition, we have

$$\mathbf{C} = \mathbf{C}(\mathbf{Z}^\top)^\dagger \mathbf{Z}^\top + \mathbf{C}(\mathbf{I} - (\mathbf{Z}^\top)^\dagger \mathbf{Z}^\top) = \mathbf{C}(\mathbf{Z}^\top)^\dagger \mathbf{Z}^\top + \mathbf{C}^*$$

By definition of spectral norm, $\|\mathbf{C} - \mathbf{W}\mathbf{Z}^\top\|_2^2 \geq \|y^\top \mathbf{C} - y^\top \mathbf{W}\mathbf{Z}^\top\|_2^2$, for all y such that $\|y\|_2 = 1$. Next, for any unit vector $y \in \mathbb{R}^n$,

$$\begin{aligned} \|y^\top \mathbf{C} - y^\top \mathbf{W}\mathbf{Z}^\top\|_2^2 &= \|y^\top (\mathbf{C}(\mathbf{Z}^\top)^\dagger \mathbf{Z}^\top + \mathbf{C}^*) - y^\top \mathbf{W}\mathbf{Z}^\top\|_2^2 \\ &= \|y^\top \mathbf{C}^* - y^\top (\mathbf{W} - \mathbf{W}^*) \mathbf{Z}^\top\|_2^2 \\ &= \|y^\top \mathbf{C}^*\|_2^2 + \|y^\top (\mathbf{W} - \mathbf{W}^*) \mathbf{Z}^\top\|_2^2 + 2\langle y^\top \mathbf{C}^*, y^\top (\mathbf{W} - \mathbf{W}^*) \mathbf{Z}^\top \rangle \end{aligned} \tag{4.16}$$

We observe that $\mathbf{W}\mathbf{Z}^\top = \mathbf{C}(\mathbf{Z}^\top)^\dagger \mathbf{Z}^\top$ is the projection of \mathbf{C} on the rowspan of \mathbf{Z}^\top and \mathbf{C}^* is the projection of \mathbf{C} on the orthogonal complement of rowspan of \mathbf{Z}^\top . Therefore, $\langle \mathbf{C}(\mathbf{Z}^\top)^\dagger \mathbf{Z}^\top, \mathbf{C}^* \rangle = 0$.

Further, for any y , $y^\top(\mathbf{W} - \mathbf{W}^*)\mathbf{Z}^\top$ is in the row span of \mathbf{Z}^\top and is thus perpendicular to $y^\top\mathbf{C}^*$. Plugging this back in to Equation 4.16, we have

$$\|y^\top\mathbf{C} - y^\top\mathbf{W}\mathbf{Z}^\top\|_2^2 = \|y^\top\mathbf{C}^*\|_2^2 + \|y^\top(\mathbf{W} - \mathbf{W}^*)\mathbf{Z}^\top\|_2^2 \geq \|y^\top\mathbf{C}^*\|_2^2 \quad (4.17)$$

where the inequality follows from non-negativity of norms. Since Equation 4.17 holds for all y , we can pick y such that $\|y^\top\mathbf{C}^*\|_2^2 = \|\mathbf{C}^*\|_2^2$. Therefore, $\|\mathbf{C} - \mathbf{W}\mathbf{Z}^\top\|_2^2 \geq \|y^\top\mathbf{C} - y^\top\mathbf{W}\mathbf{Z}^\top\|_2^2 \geq \|\mathbf{C}^*\|_2^2$. This completes the proof. \square

Next, we sketch the *Spectral Regression* problem from Equation 4.15 such that we approximately preserve the spectral norm cost of all $\mathbf{W} \in \mathbb{R}^{n \times k/\epsilon}$. A natural approach here would be to follow the Affine Embedding idea for Frobenius norm and hope a similar guarantee holds for spectral norm as well. However, since \mathbf{Z} could have rank as large as k/ϵ and we can no longer obtain a relative-error $(1 + \epsilon)$ -approximate Affine Embedding even for Frobenius norm without incurring a larger dependence on ϵ . Instead, we relax the notion of approximation for our sketch. We note that it suffices to construct a sketch \mathbf{S} such that if

$$\widehat{\mathbf{W}} = \arg \min_{\mathbf{W}} \|\mathbf{C}\mathbf{S} - \mathbf{W}\mathbf{Z}^\top\mathbf{S}\|_2^2$$

then

$$\|\mathbf{C} - \widehat{\mathbf{W}}\mathbf{Z}^\top\|_2^2 \leq \widetilde{O}(1) \left(\min_{\mathbf{W} \in \mathbb{R}^{n \times k/\epsilon}} \|\mathbf{C} - \mathbf{W}\mathbf{Z}^\top\|_2^2 + \frac{\epsilon}{k} \|\mathbf{C} - \mathbf{C}_{k/\epsilon}\|_F^2 \right)$$

as stated in Theorem 4.18. Note, here we only need to weakly preserve the cost of the optimal \mathbf{W} for the sketched problem as opposed to preserving the cost of all matrices \mathbf{W} . At a high level, this comes down to analyzing the spectrum of $\|\mathbf{C}^*\mathbf{S}\mathbf{S}^\top\|_2$. We begin with the definition of the Approximate Matrix Multiplication (AMM) guarantee and discuss its application in approximately minimizing *Spectral Regression*.

Definition 4.20. ((ϵ, k) -Spectral AMM.) Given matrices $\mathbf{A} \in \mathbb{R}^{n \times m}$ and $\mathbf{B} \in \mathbb{R}^{m \times d}$, a sketch $\mathbf{\Pi} \in \mathbb{R}^{m \times t}$ satisfies (ϵ, k) -Spectral AMM if with probability at least $1 - \delta$,

$$\|\mathbf{A}\mathbf{\Pi}\mathbf{\Pi}^\top\mathbf{B} - \mathbf{A}\mathbf{B}\|_2 \leq \epsilon \sqrt{\left(\|\mathbf{A}\|_2^2 + \frac{\|\mathbf{A}\|_F^2}{k} \right) \cdot \left(\|\mathbf{B}\|_2^2 + \frac{\|\mathbf{B}\|_F^2}{k} \right)}$$

Approximate Matrix Multiplication was introduced by Drineas et al. [DKM06] with respect to the Frobenius norm, as opposed to the spectral norm above. Subsequent work by Cohen et al. [CNW15] studied the spectral norm bound and showed that any sketch $\mathbf{\Pi}$ that is an oblivious subspace embedding (i.e., satisfies Lemma 4.3 with $\mathbf{\Pi}$ being an oblivious sketch) implies an AMM guarantee, as long as $\mathbf{\Pi}$ has $\Theta(k + \log(1/\delta)/\epsilon^2)$ columns. The *Spectral AMM* property combined with an $O(1)$ -Subspace Embedding suffice to approximately minimize the Spectral Regression problem :

Theorem 4.21. (Theorem 3, [CNW15]) Let \mathbf{A} , \mathbf{B} and $\mathbf{\Pi}$ be as defined above. If $\mathbf{\Pi}$ is an $(\sqrt{\epsilon}, \text{rank}(\mathbf{A}))$ -Spectral AMM for \mathbf{U}_A and $(\mathbf{I} - \mathbf{P}_A)\mathbf{B}$, and an $O(1)$ -Subspace Embedding for \mathbf{A} , and $\hat{\mathbf{X}} = \arg \min_{\mathbf{X}} \|\mathbf{\Pi A X} - \mathbf{\Pi B}\|_2^2$, then with probability $99/100$,

$$\|\mathbf{A}\hat{\mathbf{X}} - \mathbf{B}\|_2^2 \leq (1 + \epsilon)\|\mathbf{P}_A\mathbf{B} - \mathbf{B}\|_2^2 + \frac{\epsilon}{k}\|\mathbf{P}_A\mathbf{B} - \mathbf{B}\|_F^2$$

where \mathbf{U}_A is an orthonormal basis for \mathbf{A} and \mathbf{P}_A is the projection onto the span of \mathbf{A} .

However, all the constructions presented for the sketch in [CEM⁺15] are either oblivious sketches or require sampling proportional to both \mathbf{A} and \mathbf{B} . Applying an oblivious sketch \mathbf{S} in our problem requires computing \mathbf{CS} which would query $\Omega(\text{nnz}(\mathbf{C})) = \Omega(n^{1.5}\sqrt{k/\epsilon})$ entries in \mathbf{A} . Therefore, the main challenge here is to construct a sampling matrix \mathbf{S} while reading $\tilde{O}(nk/\epsilon)$ entries in \mathbf{A} such that \mathbf{S} is an $(\tilde{O}(1), k/\epsilon)$ -Spectral AMM and an $O(1)$ -Subspace Embedding. We construct \mathbf{S} by sampling $\tilde{O}(k/\epsilon)$ columns of \mathbf{Z}^\top proportional to the leverage scores of \mathbf{Z}^\top . While it is easy to show \mathbf{S} is a Subspace Embedding, observe that our sampling probabilities are computed without reading \mathbf{C} .

Proof of Theorem 4.18. As a starting point, we observe that yet again, since \mathbf{Z}^\top has orthonormal rows, the leverage scores are simply the ℓ_2^2 norms of the columns of \mathbf{Z}^\top . Therefore, one possible approach is to construct a *leverage score sampling* sketch \mathbf{S} for \mathbf{C} , by sampling columns proportional to the leverage scores of \mathbf{Z}^\top . We note we can afford to sample at most $\tilde{O}(k/\epsilon)$ columns, since our algorithm queries all entries in the resulting sketched matrix \mathbf{CS} .

Further, for reasons to be discussed later, it is crucial that we sample columns of \mathbf{C} independently, as opposed to the standard way of sampling with replacement we have used thus far. The independent sampling process can be described as follows: for all $j \in [\sqrt{nk/\epsilon}]$, we sample $\mathbf{C}_{*,j}^*$ with probability $\min(\|\mathbf{Z}_{*,j}^\top\|_2^2, 1)$. We use the following lemma from [CMM17] to show that independently sampling columns satisfies some desirable properties.

Lemma 4.22. (Lemma 21, [CMM17].) Given a matrix $\mathbf{M} \in \mathbb{R}^{n \times m}$, for all $j \in [m]$ let $\bar{\rho}_j^k(\mathbf{M}) = \Theta(\rho_j^k(\mathbf{M}))$ be estimates of the rank- k column ridge-leverage scores of \mathbf{M} and let $q_j = \min(\bar{\rho}_j^k(\mathbf{M}) \log(k/\delta)/\epsilon^2, 1)$. Then, construct \mathbf{MS} by selecting each column $\mathbf{M}_{*,j}$ with probability q_j and scale it by $1/\sqrt{q_j}$. Then, with probability at least $1 - \delta$, \mathbf{MS} has $\sum_{j \in [m]} \bar{\rho}_j^k \cdot \log(k/\delta)/\epsilon^2$ columns and

$$(1 - \epsilon)\mathbf{MSS}^\top \mathbf{M}^\top - \frac{\epsilon}{k}\|\mathbf{M} - \mathbf{M}_k\|_F^2 \mathbf{I} \leq \mathbf{MM}^\top \leq (1 + \epsilon)\mathbf{MSS}^\top \mathbf{M}^\top + \frac{\epsilon}{k}\|\mathbf{M} - \mathbf{M}_k\|_F^2 \mathbf{I}$$

The above lemma independently samples columns proportional to the *ridge leverage scores*. In our setting, we can set the ridge parameter $\lambda = 0$, and sample according to the exact leverage scores of \mathbf{Z}^\top . Formally, let $q = \{q_1, q_2, \dots, q_m\}$ be the corresponding distribution over columns of \mathbf{Z}^\top such that $q_j = \min(\|\mathbf{Z}^\top\|_2^2 \log(k), 1)$. Since \mathbf{Z}^\top has k/ϵ orthonormal rows, the leverage scores sum up to $\text{rank}(\mathbf{Z}^\top) \leq k/\epsilon$. We then use Lemma 4.22 by setting $\epsilon = 1/10$, $\delta = 0.01$ and thus with probability at least $99/100$, $\mathbf{Z}^\top \mathbf{S}$ has $\sum_{j \in [\sqrt{nk/\epsilon}]} \tau_j(\mathbf{Z}^\top) \log(n) = \tilde{O}(k/\epsilon)$ rows and

$$\frac{9}{10}\mathbf{Z}^\top \mathbf{SS}^\top \mathbf{Z} \leq \mathbf{Z}^\top \mathbf{Z} \leq \frac{11}{10}\mathbf{Z}^\top \mathbf{SS}^\top \mathbf{Z} \quad (4.18)$$

If the guarantee in Equation 4.18 holds for a sketch \mathbf{S} , we refer to \mathbf{S} as satisfying an $O(1)$ -Subspace Embedding property. Observe, this is equivalent to \mathbf{S} preserving all singular values of \mathbf{Z}^\top up to a constant.

We can now obtain a closed form solution for the *Spectral Regression* problem in the sketched space. By Lemma 4.19, the optimal solution to the optimization problem in Step 2 of Algorithm 2 is given by $\widehat{\mathbf{W}} = \mathbf{C}\mathbf{S}(\mathbf{Z}^\top\mathbf{S})^\dagger$. Since \mathbf{S} satisfies the $O(1)$ -Subspace Embedding property in Equation 4.18, it preserves the rank of \mathbf{Z}^\top . Therefore, $\mathbf{Z}^\top\mathbf{S}$ has full row rank and $(\mathbf{Z}^\top\mathbf{S})^\dagger = \mathbf{S}^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z})^{-1}$ and thus $\widehat{\mathbf{W}} = \mathbf{C}\mathbf{S}\mathbf{S}^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z})^{-1}$ is the optimal solution. Next, we bound the cost of $\widehat{\mathbf{W}}$ in the original problem. Let $\mathbf{P}_{\mathbf{Z}^\top} = \mathbf{Z}^\top\mathbf{Z}^\top$ be the orthogonal projection matrix onto \mathbf{Z}^\top . Using the fact that $\|\mathbf{M}\|_2^2 = \max_{\|y\|_2=1} \|y^\top\mathbf{M}\|_2^2$ and the Pythagorean Theorem for Euclidean space we have

$$\begin{aligned} \|\mathbf{C} - \widehat{\mathbf{W}}\mathbf{Z}^\top\|_2^2 &= \|\mathbf{C} - \mathbf{C}\mathbf{S}\mathbf{S}^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\|_2^2 \\ &= \max_{\|y\|_2^2=1} \|y^\top\mathbf{C}\mathbf{P}_{\mathbf{Z}^\top} - y^\top\mathbf{C}\mathbf{S}\mathbf{S}^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{P}_{\mathbf{Z}^\top}\|_2^2 + \\ &\quad \|y^\top\mathbf{C}(\mathbf{I} - \mathbf{P}_{\mathbf{Z}^\top}) - y^\top\mathbf{C}\mathbf{S}\mathbf{S}^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top(\mathbf{I} - \mathbf{P}_{\mathbf{Z}^\top})\|_2^2 \end{aligned} \quad (4.19)$$

Here, we observe $y^\top\mathbf{C}\mathbf{S}\mathbf{S}^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top$ is a vector in the row space of \mathbf{Z}^\top and $(\mathbf{I} - \mathbf{P}_{\mathbf{Z}^\top})$ is the projection on the orthogonal complement of $\text{rowspan}(\mathbf{Z}^\top)$, thus this evaluates to 0. Since $\mathbf{C}(\mathbf{I} - \mathbf{P}_{\mathbf{Z}^\top}) = \mathbf{C}^*$, we can upper bound $\|y^\top\mathbf{C}(\mathbf{I} - \mathbf{P}_{\mathbf{Z}^\top})\|_2$ by $\|\mathbf{C}^*\|_2^2$. Similarly, we can upper bound the first term by its spectral norm. Therefore, plugging this back into Equation 4.19,

$$\begin{aligned} \|\mathbf{C} - \widehat{\mathbf{W}}\mathbf{Z}^\top\|_2^2 &\leq \|\mathbf{C}(\mathbf{Z}^\top)^\dagger\mathbf{Z}^\top - \mathbf{C}\mathbf{S}\mathbf{S}^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\|_2^2 + \|\mathbf{C}^*\|_2^2 \\ &= \left\| \left(\mathbf{C}(\mathbf{Z}^\top)^\dagger\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z} - \mathbf{C}\mathbf{S}\mathbf{S}^\top\mathbf{Z} \right) (\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z})^{-1} \right\|_2^2 + \|\mathbf{C}^*\|_2^2 \\ &\leq \|\mathbf{C}(\mathbf{Z}^\top)^\dagger\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z} - \mathbf{C}\mathbf{S}\mathbf{S}^\top\mathbf{Z}\|_2^2 \|(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z})^{-1}\|_2^2 + \|\mathbf{C}^*\|_2^2 \end{aligned} \quad (4.20)$$

where we use that \mathbf{Z}^\top has orthonormal columns and the sub-multiplicativity of the spectral norm. From Equation 4.18, it follows that for all $i \in [k/\epsilon]$, $\sigma_i^2(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z}) = (1 \pm 0.1)^2 \sigma_i^2(\mathbf{Z}^\top\mathbf{Z}) = (1 \pm 0.1)^2$. Therefore, $\|(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z})^{-1}\|_2^2 = 1/\sigma_{\min}^2(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z}) \leq 100/81$. Substituting this back into Equation 4.20, we have

$$\begin{aligned} \|\mathbf{C} - \widehat{\mathbf{W}}\mathbf{Z}^\top\|_2^2 &\leq O(1) \|(\mathbf{C}(\mathbf{Z}^\top)^\dagger\mathbf{Z}^\top - \mathbf{C})\mathbf{S}\mathbf{S}^\top\mathbf{Z}\|_2^2 + \|\mathbf{C}^*\|_2^2 \\ &\leq O(1) \|\mathbf{C}^*\mathbf{S}\mathbf{S}^\top\mathbf{Z}\|_2^2 + \|\mathbf{C}^*\|_2^2 \end{aligned} \quad (4.21)$$

where the last inequality follows from the definition of \mathbf{C}^* . In order to bound the cost above, we focus on analyzing $\|\mathbf{C}^*\mathbf{S}\mathbf{S}^\top\mathbf{Z}\|_2^2$. Since we want to compare $\|\mathbf{C}^*\mathbf{S}\mathbf{S}^\top\mathbf{Z}\|_2^2$ to $\|\mathbf{C}^*\mathbf{Z}\|_2^2$, a natural way to proceed would be to interpret this term as an instance of *Approximate Matrix Product*. Therefore, we next show that the leverage score sampling matrix \mathbf{S} satisfies the *Spectral AMM* property for \mathbf{C}^* and \mathbf{Z}^\top . Here, we want to analyze how sampling columns of \mathbf{C}^* proportional to the *leverage scores* of \mathbf{Z}^\top affects the spectrum of \mathbf{C}^* . An important tool in this analysis is the following result by Rudelson and Vershynin on how the spectral norm of a matrix degrades when we sample a uniformly random subset of rows of a matrix:

Theorem 4.23. (Theorem 1.8 in [RV07]) Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, let \mathbf{Q} be a uniformly random subset of $[n]$ s.t. $\mathbb{E}[\mathbf{Q}] = q$. Let $\mathbf{A}_{|\mathbf{Q}}$ denote the submatrix restricted to the rows indexed by \mathbf{Q} . Then,

$$\mathbb{E} [\|\mathbf{A}_{|\mathbf{Q}}\|_2] = O \left(\sqrt{\frac{q}{n}} \|\mathbf{A}\|_2 + \sqrt{\log(q)} \|\mathbf{A}\|_{(n/q)} \right)$$

where $\|\mathbf{A}\|_{(n/q)}$ is the average of the largest n/q ℓ_2 -norms of columns of \mathbf{A} .

We extend the above statement to rectangular matrices:

Corollary 4.24 (Spectral Decay for Rectangular Matrices). Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$, s.t. for all $j, j' \in [m]$, $\|\mathbf{A}_{*,j}\|_2^2 = \Theta(\|\mathbf{A}_{*,j'}\|_2^2)$. Let \mathbf{Q} be a uniformly random subset of $[n]$ s.t. $\mathbb{E}[\mathbf{Q}] = q$. Let $b = \lceil n/m \rceil$ and $\mathbf{A}_{|\mathbf{Q}}$ denote the submatrix restricted to the rows indexed by \mathbf{Q} . Then,

$$\mathbb{E} [\|\mathbf{A}_{|\mathbf{Q}}\|_2] = O \left(\sqrt{\frac{q}{n}} \|\mathbf{A}\|_2 + \sqrt{\log(q)/b} \|\mathbf{A}\|_{(n/q)} \right)$$

Proof. First, consider the case when $m \geq n$. To see this, let $\text{SVD}(\mathbf{A}) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ where $\mathbf{U}\mathbf{\Sigma}$ is an $n \times n$ matrix. Now, $\|\mathbf{A}\|_2 = \|\mathbf{U}\mathbf{\Sigma}\|_2$ and applying Theorem 4.23 to $\mathbf{U}\mathbf{\Sigma}$, we have

$$\begin{aligned} \mathbb{E} [\|\mathbf{A}_{|\mathbf{Q}}\|_2] &= \mathbb{E} [\|(\mathbf{U}\mathbf{\Sigma})_{|\mathbf{Q}}\|_2] = O \left(\sqrt{\frac{q}{n}} \|\mathbf{U}\mathbf{\Sigma}\|_2 + \sqrt{\log(q)} \|\mathbf{U}\mathbf{\Sigma}\|_{(n/q)} \right) \\ &= O \left(\sqrt{\frac{q}{n}} \|\mathbf{A}\|_2 + \sqrt{\log(q)} \|\mathbf{A}\|_{(n/q)} \right) \end{aligned} \quad (4.22)$$

where we repeatedly use that \mathbf{V}^\top has orthonormal rows. Here, we note that since the columns of \mathbf{A} have the same squared norm up to a constant, $\|\mathbf{A}\|_{(n/q)} = \Theta(\|\mathbf{A}\|_{1 \rightarrow 2})$, i.e. the max column norm of \mathbf{A} .

Next, consider the case where $m < n$. Let $b = \lceil n/m \rceil$. In order to analyze the spectral norm of $\mathbf{A}_{|\mathbf{Q}}$, we create b copies of \mathbf{A} and concatenate them such that the resulting matrix \mathbf{A}^* has more columns than rows. Applying Equation (4.22) to \mathbf{A}^* and substituting the average with max, we have

$$\mathbb{E} [\|\mathbf{A}_{|\mathbf{Q}}^*\|_2] = O \left(\sqrt{\frac{q}{n}} \|\mathbf{A}^*\|_2 + \sqrt{\log(q)} \|\mathbf{A}^*\|_{1 \rightarrow 2} \right) \quad (4.23)$$

Observe, $\mathbf{A}_{|\mathbf{Q}}^*$ selects uniformly random rows of \mathbf{A}^* and $\|\mathbf{A}_{|\mathbf{Q}}^*\|_2 = \max_{\|x\|_2=1} \|x^\top \mathbf{A}_{|\mathbf{Q}}^*\|_2$ and for any vector x , $\|x^\top \mathbf{A}_{|\mathbf{Q}}^*\|_2 = \sqrt{b} \|x^\top \mathbf{A}_{|\mathbf{Q}}\|_2$. Therefore, $\mathbb{E} [\|\mathbf{A}_{|\mathbf{Q}}^*\|_2] = \sqrt{b} \cdot \mathbb{E} [\|\mathbf{A}_{|\mathbf{Q}}\|_2]$ and $\|\mathbf{A}^*\|_2 = \sqrt{b} \cdot \|\mathbf{A}\|_2$. Finally, it is easy to see that since the columns of \mathbf{A}^* are copies of columns of \mathbf{A} , the max column norm does not change. Therefore, (4.22) to \mathbf{A}^* , we have

$$\mathbb{E} [\|\mathbf{A}_{|\mathbf{Q}}\|_2] = O \left(\sqrt{\frac{q}{n}} \|\mathbf{A}\|_2 + \sqrt{\log(q)/b} \|\mathbf{A}\|_{1 \rightarrow 2} \right) = O \left(\sqrt{\frac{q}{n}} \|\mathbf{A}^*\|_2 + \sqrt{\log(q)} \|\mathbf{A}^*\|_{(n/q)} \right) \quad (4.24)$$

and the claim follows. \square

Intuitively, there are two technical challenges in applying Corollary 4.24. First, a *leverage score sampling* matrix need not sample columns uniformly at random, since we have no control over the column norms of \mathbf{Z}^\top . Second, the $\|\cdot\|_{(n/q)}$ norm only shrinks when all columns of \mathbf{A} have roughly the same squared norm. We overcome these challenges by partitioning the matrix, first according to row norms, such that each partition does indeed have the same row norm, up to a factor of 2. Next, we further partition each such matrix according to the sampling probabilities, such that within each partition, the sampling process is *close* to uniform sampling. Formally,

Lemma 4.25. (*Weak Spectral Approximate Matrix Product.*) Let \mathbf{Z} , \mathbf{C}^* and \mathbf{S} be as defined in Lemma 4.19. Then, with probability at least 99/100, \mathbf{S} satisfies $(\tilde{O}(1), k/\epsilon)$ -Spectral AMM, i.e.,

$$\|\mathbf{C}^* \mathbf{S} \mathbf{S}^\top \mathbf{Z}\|_2^2 \leq \tilde{O}(1) \left(\frac{\epsilon}{k} \|\mathbf{C}^*\|_F^2 + \|\mathbf{C}^*\|_2^2 \right)$$

Proof. By sub-multiplicativity of the spectral norm and \mathbf{S} being an $O(1)$ -subspace embedding for \mathbf{Z}^\top , we have

$$\begin{aligned} \|\mathbf{C}^* \mathbf{S} \mathbf{S}^\top \mathbf{Z}\|_2^2 &\leq \|\mathbf{C}^* \mathbf{S}\|_2^2 \cdot \|\mathbf{S}^\top \mathbf{Z}\|_2^2 \\ &\leq O(1) \|\mathbf{C}^* \mathbf{S}\|_2^2 \end{aligned} \tag{4.25}$$

where the second inequality follows from \mathbf{Z}^\top having orthonormal rows.

We begin by observing that Corollary 4.24 requires the squared row norms of \mathbf{C}^* to be roughly the same, which need not be the case in general. Note, here the sampling matrix subsamples columns of \mathbf{C}^* , as opposed to rows in Corollary 4.24. Thus, we partition the rows of \mathbf{C}^* into $O(\log(n))$ blocks such that either the squared column norms are the same up to a factor of 2 or they are at most $\|\mathbf{C}^*\|_F^2 / \text{poly}(n)$. Formally, for all $\ell \in [c \log(n)]$, let

$$\mathcal{B}_\ell = \left\{ i \in [n] : \frac{\|\mathbf{C}^*\|_F^2}{2^{\ell+1}} \leq \|\mathbf{C}^*_{i,*}\|_2^2 \leq \frac{\|\mathbf{C}^*\|_F^2}{2^\ell} \right\}$$

represent the blocks for rows with large squared norm. Let $\mathcal{B}_r = [n] \setminus \bigcup_{\ell \in [\log(n)]} \mathcal{B}_\ell$ be the remaining rows, which have norm at most $\|\mathbf{C}^*\|_F^2 / \text{poly}(n)$. Since the set of indices in the blocks form a partition of the rows of \mathbf{C}^* , we can write $\|\mathbf{C}^*\|_F^2 = \sum_{\ell \in [\log(n)]} \|\mathbf{C}^*_{\mathcal{B}_\ell}\|_F^2 + \|\mathbf{C}^*_{\mathcal{B}_r}\|_F^2$. Similarly, we can bound the spectral norm as follows:

$$\begin{aligned} \|\mathbf{C}^* \mathbf{S}\|_2^2 &= \max_{\|y\|_2=1} \|\mathbf{C}^* \mathbf{S} y\|_2^2 \leq O \left(\sum_{\ell \in [\log(n)]} \|\mathbf{C}^*_{\mathcal{B}_\ell} \mathbf{S} y\|_2^2 + \|\mathbf{C}^*_{\mathcal{B}_r} \mathbf{S} y\|_2^2 \right) \\ &\leq O \left(\sum_{\ell \in [\log(n)]} \|\mathbf{C}^*_{\mathcal{B}_\ell} \mathbf{S}\|_2^2 + \|\mathbf{C}^*_{\mathcal{B}_r} \mathbf{S}\|_2^2 \right) \end{aligned} \tag{4.26}$$

We now handle the two separately. Since \mathbf{S} is an unbiased estimator of the squared Frobenius norm of \mathbf{Z}^\top , it is an unbiased estimator of the squared Frobenius norm of \mathbf{C}^* . Therefore, with probability at least 99/100,

$$\|\mathbf{C}^*_{\mathcal{B}_r} \mathbf{S}\|_2^2 \leq \|\mathbf{C}^*_{\mathcal{B}_r}\|_F^2 = O(\|\mathbf{C}^*_{\mathcal{B}_r}\|_F^2) \leq \frac{\|\mathbf{C}^*\|_F^2}{\text{poly}(n)} < \frac{\epsilon}{k} \|\mathbf{C}^*\|_F^2 \tag{4.27}$$

For the remaining terms, we cannot use this naïve analysis as this would only leave us with an upper bound of $\|\mathbf{C}^*\|_F^2$, which is too large.

If instead of a leverage score sampling matrix, \mathbf{S} were a uniform sampling sketch that samples k/ϵ columns of \mathbf{C}^* in expectation, we could apply Corollary 4.24 for each ℓ , with $q = k/\epsilon$ and $n = \sqrt{nk/\epsilon}$ and $b = \left\lceil \sqrt{nk}/(\sqrt{\epsilon}|\mathcal{B}_\ell|) \right\rceil$, to obtain

$$\begin{aligned} \mathbb{E} \left[\|\mathbf{C}_{|\mathcal{B}_\ell}^* \mathbf{S}\|_2^2 \right] &= \sqrt{\frac{n\epsilon}{k}} \mathbb{E} \left[\|(\mathbf{C}_{|\mathcal{B}_\ell}^*)_Q\|_2^2 \right] \leq O \left(\|\mathbf{C}_{|\mathcal{B}_\ell}^*\|_2^2 + \frac{\epsilon \log(k/\epsilon)|\mathcal{B}_\ell|}{k} \|(\mathbf{C}^*)_{|\mathcal{B}_\ell}^\top\|_2^2 \sqrt{\epsilon n/k} \right) \\ &\leq O \left(\|\mathbf{C}_{|\mathcal{B}_\ell}^*\|_2^2 + \frac{\epsilon \log(k/\epsilon)}{k} \|\mathbf{C}_{|\mathcal{B}_\ell}^*\|_F^2 \right) \end{aligned} \quad (4.28)$$

where Q is the subset of columns selected by \mathbf{S} and the second inequality follows from observing that the all the row norms of $(\mathbf{C}^*)_{|\mathcal{B}_\ell}$ are within a factor of 2 of each other and thus the max squared row norm times the size of the set is the squared Frobenius norm.

Using Equations 4.27 and 4.28 to upper bound the two terms in Equation 4.26 suffices to finish the proof. Unfortunately, a similar analysis does not immediately go through when we replace a uniform sampling matrix with a leverage score sketch. Instead, we partition the sketch \mathbf{S} into buckets such that each bucket corresponds to rows in \mathbf{S} that scale columns of \mathbf{C}^* within a factor of 2. For notational convenience, let $m = \sqrt{nk/\epsilon}$ and $t = k/\epsilon$. Recall, we construct \mathbf{S} by sampling the j -th column of \mathbf{Z}^\top independently with probability $q_j = \min(\|\mathbf{Z}_{j,*}\|_2^2 \log(k), 1)$ and scale this column by $1/\sqrt{q_j}$. We group the scaling factors into buckets. Note, if for some j , $q_j < 1/n^3$, we can ignore the corresponding column.

Let ζ_j be the indicator for a column of \mathbf{Z}^\top to be sampled by \mathbf{S} . Then, $\Pr[\zeta_j = 1] = q_j = \min(\|\mathbf{Z}_{j,*}^\top\|_2^2 \log(k), 1)$. Since $q_j \leq 1/n^3$, we can union bound over at most m such events and conclude with probability at least $1 - 1/n^2$, for all $j \in m$, no column $\mathbf{Z}_{*,j}^\top$ is sampled such that $q_j \leq 1/n^3$. Further, since $q_j \leq 1$, $1/\sqrt{q_j} \in [1, n^{1.5}]$. Therefore, it suffices to bucket values in the range $[1, n^{1.5}]$. For all $h \in [c \log(n)]$, let \mathcal{S} denote the set of column indices from \mathbf{Z}^\top that were sampled by the sketch \mathbf{S} . Then,

$$\mathcal{T}_h = \left\{ j \in \mathcal{S} : 2^h \leq \frac{1}{\sqrt{q_j}} \leq 2^{h+1} \right\}$$

Let $\mathbf{S}_{\mathcal{T}_h}$ be the subset of rows of \mathbf{S} which are indexed by the set \mathcal{T}_h . Since t is fixed and the scaling factors in \mathcal{T}_h differ by at most a factor of 2, the corresponding sampling probabilities in \mathbf{D} differ by at most $\sqrt{2}$, which is still not uniform. To fix this, we change the sampling process and independently sample each column indexed by $j \in \mathcal{T}_h$ with probability 2^{h+1} , while still scaling it by $1/\sqrt{tq_j}$. Let this new distribution be denoted by q' . Under the new sampling process, we now sample rows independently and therefore, we are at least as likely to see all the rows sampled by $\mathbf{S}_{\mathcal{T}_h}$ in our new sampling process. Therefore, it now holds that

$$\mathbb{E}_q \left[\|\mathbf{C}_{|\mathcal{B}_\ell}^* \mathbf{S}_{\mathcal{T}_h}\|_2^2 \right] \leq \mathbb{E}_{q'} \left[\|\mathbf{C}_{|\mathcal{B}_\ell}^* \mathbf{S}_{\mathcal{T}_h}\|_2^2 \right]$$

Further, in the new sampling process, each row restricted to the set \mathcal{T}_h is uniformly sampled with probability $1/2^{(h+1)/2}$ and thus we can apply Corollary 4.24 to $\mathbf{C}_{|\mathcal{B}_\ell}^* \mathbf{S}_{\mathcal{T}_h}$.

$$\begin{aligned} \mathbb{E}_{q'} \left[\|\mathbf{C}_{|\mathcal{B}_\ell}^* \mathbf{S}_{\mathcal{T}_h}\|_2^2 \right] &\leq O \left(\|\mathbf{C}_{|\mathcal{B}_\ell}^*\|_2^2 + \frac{\epsilon \log(k/\epsilon) |\mathcal{B}_\ell|}{k} \|(\mathbf{C}_{|\mathcal{B}_\ell}^*)^\top\|_{(\sqrt{\epsilon n/k})} \right) \\ &\leq O \left(\|\mathbf{C}_{|\mathcal{B}_\ell}^*\|_2^2 + \frac{\epsilon \log(k/\epsilon)}{k} \|\mathbf{C}_{|\mathcal{B}_\ell}^*\|_F^2 \right) \end{aligned} \quad (4.29)$$

where the second inequality follows from squared row norms in $\mathbf{C}_{|\mathcal{B}_\ell}^*$ being equal up to a factor of 2. Therefore, with probability at least $1 - 1/c' \log(n)$,

$$\|\mathbf{C}_{|\mathcal{B}_\ell}^* \mathbf{S}_{\mathcal{T}_h}\|_2^2 \leq \tilde{O} \left(\|\mathbf{C}_{|\mathcal{B}_\ell}^*\|_2^2 + \frac{\epsilon \log(k/\epsilon)}{k} \|\mathbf{C}_{|\mathcal{B}_\ell}^*\|_F^2 \right) \quad (4.30)$$

Let η_h be the event that the above bound holds. Then, union bounding over all $c \log(n)$ such events, with probability at least 99/100, simultaneously for all h ,

$$\begin{aligned} \|\mathbf{C}_{|\mathcal{B}_\ell}^* \mathbf{S}\|_2^2 &\leq O \left(\sum_{h \in [c \log(n)]} \|\mathbf{C}_{|\mathcal{B}_\ell}^* \mathbf{S}_{\mathcal{T}_h}\|_2^2 \right) \\ &\leq \tilde{O} \left(\|\mathbf{C}_{|\mathcal{B}_\ell}^*\|_2^2 + \frac{\epsilon}{k} \|\mathbf{C}_{|\mathcal{B}_\ell}^*\|_F^2 \right) \end{aligned} \quad (4.31)$$

which follows from Equation 4.30. Substituting this back into Equation 4.26,

$$\begin{aligned} \|\mathbf{C}^* \mathbf{S}\|_2^2 &\leq O \left(\sum_{\ell \in [\log(n)]} \|\mathbf{C}_{|\mathcal{B}_\ell}^* \mathbf{S}\|_2^2 + \|\mathbf{C}_{|\mathcal{B}_r}^* \mathbf{S}\|_2^2 \right) \\ &\leq \tilde{O} \left(\|\mathbf{C}^*\|_2^2 + \frac{\epsilon}{k} \|\mathbf{C}^*\|_F^2 \right) \end{aligned} \quad (4.32)$$

where the second inequality follows from Equation 4.31 and observing that $\|\mathbf{C}_{\mathcal{B}_\ell}^*\|_2^2 \leq \|\mathbf{C}^*\|_2^2$ and $\sum_\ell \|\mathbf{C}_{\mathcal{B}_\ell}^*\|_F^2 = \|\mathbf{C}^*\|_F^2$, which completes the proof. \square

Combining the above lemma with 4.19, and observing that $\|\mathbf{C}^*\|_F^2 \leq \|\mathbf{C} - \mathbf{C}_{k/\epsilon}\|_F^2$, we can bound the cost of $\widehat{\mathbf{W}}$

$$\|\mathbf{C} - \widehat{\mathbf{W}} \mathbf{Z}^\top\|_2^2 \leq \tilde{O} \left(\min_{\mathbf{W}} \|\mathbf{C} - \mathbf{W} \mathbf{Z}^\top\|_2^2 + \frac{\epsilon}{k} \|\mathbf{C} - \mathbf{C}_{k/\epsilon}\|_F^2 \right)$$

which completes the correctness proof of Theorem 4.18. Next, we analyze the running time. In Step 1 of Algorithm 2, we compute a distribution over the columns of \mathbf{Z}^\top , which does not require reading any entries in \mathbf{A} and takes time $\sqrt{n k/\epsilon} \cdot k/\epsilon = \sqrt{n} (k/\epsilon)^{1.5}$. Step 2 requires computing $\mathbf{C} \mathbf{S}$ and $\mathbf{Z}^\top \mathbf{S}$. Note since \mathbf{S} samples $\tilde{O}(k/\epsilon)$ columns in \mathbf{C} , we have to query $n \cdot \tilde{O}(k/\epsilon)$ entries in \mathbf{A} to explicitly compute $\mathbf{C} \mathbf{S}$ and can be computed in as much time. Since \mathbf{Z}^\top has fewer rows the running time is dominated by computing $\mathbf{C} \mathbf{S}$. For Step 3, we compute $(\mathbf{Z}^\top \mathbf{S} \mathbf{S}^\top \mathbf{Z})^{-1}$, which requires no queries to \mathbf{A} and runs in time $\tilde{O}((k/\epsilon)^\omega)$ and thus $(\mathbf{Z}^\top \mathbf{S})^\dagger$ can be computed in the same time. Therefore, the total query complexity of Algorithm 2 is $\tilde{O}(n k/\epsilon)$ and the running time is $\tilde{O}(n k/\epsilon + (k/\epsilon)^\omega)$, which concludes the proof.

4.3 Sample-Optimal Algorithm

In this subsection, we describe our main algorithm for PSD Low-Rank Approximation. Given a PSD matrix \mathbf{A} , our algorithm queries $\tilde{O}(nk/\epsilon)$ entries in \mathbf{A} and runs in time $\tilde{O}(n(k/\epsilon)^{\omega-1} + (k/\epsilon^3)^\omega)$. This resolves an open question on the ϵ -dependence of the query complexity and matches the lower bound of $\Omega(nk/\epsilon)$ up to polylog factors from [MW17]. At a high level, our algorithm consists of two stages: first, we use the existing machinery developed by Musco and Woodruff to obtain weak PCPs by setting ϵ to be a constant. By observing that their algorithms have linear dependence on the rank, we can afford to rank- (k/ϵ) PCPs instead. This enables us to find a structured subspace that contains a spectral low-rank approximation for the PCP.

Algorithm 3 : Sample Optimal PSD Low-Rank Approximation

Input: A PSD Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, integer k , and $\epsilon > 0$.

1. Let $t = c\sqrt{\frac{nk}{\epsilon}} \log(n)$, for some constant c and let $k' = \tilde{O}(k/\epsilon)$. For all $j \in [n]$, let $\bar{\rho}_j^{k'}(\mathbf{A}^{1/2})$ be the approximate column ridge-leverage scores that satisfy Lemma 4.9. Let $q = \{q_1, q_2, \dots, q_n\}$ denote a distribution over columns of \mathbf{A} such that $q_j = \bar{\rho}_j^{k'}(\mathbf{A}^{1/2}) / \sum_j \bar{\rho}_j^{k'}(\mathbf{A}^{1/2})$.
2. Construct a *column PCP* for \mathbf{A} by sampling t columns of \mathbf{A} such that each column is set to $\frac{\mathbf{A}_{*,j}}{\sqrt{tq_j}}$ with probability q_j , for all $j \in [n]$. Let \mathbf{C} be the resulting $n \times t$ matrix that satisfies the guarantee of Lemma 4.11 instantiated with $k = k'$ and $\epsilon = O(1)$.
3. Construct a *row PCP* for \mathbf{C} by sampling t rows of \mathbf{C} such that each row is set to $\frac{\mathbf{C}_{i,*}}{\sqrt{tq_i}}$ with probability q_i , for all $i \in [n]$. Let \mathbf{R} be the resulting $t \times t$ matrix that satisfies the guarantee of Lemma 4.12 instantiated with $k = k/\epsilon$ and $\epsilon = O(1)$.
4. Run the *input-sparsity* algorithm from Lemma 4.6 to compute a rank- k/ϵ matrix \mathbf{Z} with orthonormal columns such that $\|\mathbf{R} - \mathbf{R}\mathbf{Z}\mathbf{Z}^\top\|_2^2 \leq O\left(\frac{\epsilon}{k}\right) \|\mathbf{R} - \mathbf{R}_{k/\epsilon}\|_F^2$.
5. Run Algorithm 2 with parameters k, ϵ on the *Spectral Regression* problem

$$\min_{\mathbf{W}} \|\mathbf{C} - \mathbf{W}\mathbf{Z}^\top\|_2$$

Let $\widehat{\mathbf{W}}$ be the output of Algorithm 2. Compute an orthonormal basis \mathbf{Q} for $\widehat{\mathbf{W}}$. Note, $\mathbf{Q}\mathbf{Q}^\top$ is an $(O(1), k/\epsilon)$ -SF projection for \mathbf{A} .

6. Run Algorithm 1 with input $\mathbf{A}, \mathbf{Q}, k$ and ϵ to approximately minimize $\|\mathbf{A} - \mathbf{Q}\mathbf{X}\mathbf{Q}^\top\|_F^2$ over rank k matrices \mathbf{X} . Let \mathbf{M}, \mathbf{N} be the output of Algorithm 1.

Output: $\mathbf{M}, \mathbf{N}^\top \in \mathbb{R}^{n \times k}$ such that $\|\mathbf{A} - \mathbf{M}\mathbf{N}\|_F^2 \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2$

Since our PCPs are accurate only up to $O(1)$ -error, we cannot directly extract a $(1 + \epsilon)$ relative error approximation for \mathbf{A} . However, we show that the PCPs have enough structure to obtain a

structured subspace that spans a $(1 + \epsilon)$ -approximate solution for \mathbf{A} . A key ingredient to recover this structured subspace is an efficient algorithm for *Spectral Regression*.

Following the approach of Musco and Woodruff we use the ridge leverage scores of $\mathbf{A}^{1/2}$ to compute \mathbf{C} , a column PCP for \mathbf{A} and \mathbf{R} a row PCP for \mathbf{C} , with a minor tweak: we instantiate their theorems (Lemmas 4.11 and 4.12) with $k = k/\epsilon$ and $\epsilon = O(1)$. While the precise guarantees satisfied by our PCPs are weaker than the PCPs used by Musco and Woodruff, the dimensions of our PCPs are smaller.

In particular, we obtain a row PCP \mathbf{R} , which is a $\sqrt{nk/\epsilon} \times \sqrt{nk/\epsilon}$ matrix (ignoring polylogarithmic factors) and we can afford to read all of it. The input-sparsity time algorithm from Lemma 4.6 queries $\text{nnz}(\mathbf{R}) = \tilde{O}(nk/\epsilon)$ entries to obtain a rank- (k/ϵ) matrix \mathbf{Z} with orthonormal columns such that

$$\|\mathbf{R} - \mathbf{RZZ}^\top\|_2^2 \leq \frac{\epsilon}{k} \|\mathbf{R} - \mathbf{R}_k\|_F^2 \quad (4.33)$$

Since \mathbf{R} is a *Spectral-Frobenius PCP* for \mathbf{C} , \mathbf{ZZ}^\top satisfies $\|\mathbf{C} - \mathbf{CZZ}^\top\|_2^2 \leq \frac{\epsilon}{k} \|\mathbf{C} - \mathbf{C}_k\|_F^2$. Since \mathbf{C} is a *Spectral-Frobenius PCP* for \mathbf{A} , it suffices to obtain a projection for the column space of \mathbf{C} that also satisfies the above guarantee. Therefore, we solve the following *Spectral Regression* problem: $\min_{\mathbf{W}} \|\mathbf{C} - \mathbf{WZ}^\top\|_2^2$. Recall, we can approximately optimize this using Algorithm 2. Let $\hat{\mathbf{W}}$ be the resulting solution. We can then compute an orthonormal basis for $\hat{\mathbf{W}}$ (denoted by \mathbf{Q}) and show that \mathbf{QQ}^\top is an $(O(1), k/\epsilon)$ -SF projection for \mathbf{A} . Then, we can obtain a low rank approximation for \mathbf{A} by simply running Algorithm 1.

Proof of Theorem 4.1. Let $k' = \tilde{O}(k/\epsilon)$. It follows from Lemma 4.9 that we can compute the rank- k' ridge leverage scores of $\mathbf{A}^{1/2}$, up to a constant factor using the algorithm of Musco and Musco [MM17]. By Lemma 4.10, the ridge leverage scores of $\mathbf{A}^{1/2}$ are a $\sqrt{\epsilon n/k'}$ -approximation to the ridge leverage scores of \mathbf{A} . Let q be a distribution over rows and columns of \mathbf{A} as defined in Algorithm 3. Since we sample $t = O(\sqrt{nk/\epsilon} \log(n))$ columns of \mathbf{A} proportional to q , instantiating Lemma 4.12 with $k = k'$ and $\epsilon = 0.1$, we obtain a mixed Spectral-Frobenius column PCP \mathbf{C} such that with probability at least $1 - c_1$, for all rank- k' projections \mathbf{X} ,

$$\frac{9}{10} \|\mathbf{A} - \mathbf{XA}\|_2^2 - \frac{1}{10k'} \|\mathbf{A} - \mathbf{A}_{k'}\|_F^2 \leq \|\mathbf{C} - \mathbf{XC}\|_2^2 \leq \frac{11}{10} \|\mathbf{A} - \mathbf{XA}\|_2^2 + \frac{1}{10k'} \|\mathbf{A} - \mathbf{A}_{k'}\|_F^2 \quad (4.34)$$

Let ζ_1 be the indicator for \mathbf{C} satisfying the above guarantee. Similarly, sampling t rows of \mathbf{C} proportional to q , results in a mixed Spectral-Frobenius row PCP for \mathbf{R} such that with probability at least $1 - c_2$, for all rank- k' projection matrices \mathbf{X} ,

$$\frac{9}{10} \|\mathbf{C} - \mathbf{CX}\|_2^2 - \frac{1}{10k'} \|\mathbf{C} - \mathbf{C}_{k'}\|_F^2 \leq \|\mathbf{R} - \mathbf{RX}\|_2^2 \leq \frac{11}{10} \|\mathbf{C} - \mathbf{CX}\|_2^2 + \frac{1}{10k'} \|\mathbf{C} - \mathbf{C}_{k'}\|_F^2 \quad (4.35)$$

Further, it is well-known that with the same probability $\|\mathbf{R} - \mathbf{R}_{k'}\|_F^2 = \|\mathbf{C} - \mathbf{C}_{k'}\|_F^2$. Let ζ_2 be the event that \mathbf{R} satisfies the above guarantee. Next, we compute a Spectral Low-Rank Approximation for \mathbf{R} , using the algorithm from Lemma 4.6, with $k = k'$ and $\epsilon = 0.1$. As a result, with probability at least $1 - c_3$, we obtain a rank- k' matrix $\mathbf{Z} \in \mathbb{R}^{t \times k}$, such that \mathbf{ZZ}^\top is a $(0.1, k')$ -SF projection for \mathbf{R} , i.e.,

$$\|\mathbf{R} - \mathbf{RZZ}^\top\|_2^2 \leq \frac{1}{10k'} \|\mathbf{R} - \mathbf{R}_{k'}\|_F^2 \quad (4.36)$$

Let ζ_3 be the event that \mathbf{Z} satisfies the above guarantee. Union bounding over $\zeta_1, \zeta_2, \zeta_3$, we know that all of them hold with probability at least $1 - (c_1 + c_2 + c_3)$. Since \mathbf{R} is a Spectral-Frobenius row PCP for \mathbf{C} and $\mathbf{Z}\mathbf{Z}^\top$ is a rank- k' projection matrix, it follows from Equation 4.35

$$\begin{aligned}\|\mathbf{C} - \mathbf{C}\mathbf{Z}\mathbf{Z}^\top\|_2^2 &\leq \frac{10}{9}\|\mathbf{R} - \mathbf{R}\mathbf{Z}\mathbf{Z}^\top\|_2^2 + \frac{1}{9k'}\|\mathbf{R} - \mathbf{R}_{k'}\|_F^2 \\ &\leq \frac{1}{10k'}\|\mathbf{R} - \mathbf{R}_k\|_F^2 + \frac{1}{9k'}\|\mathbf{R} - \mathbf{R}_{k'}\|_F^2 \\ &\leq \tilde{O}\left(\frac{\epsilon}{k}\right)\|\mathbf{C} - \mathbf{C}_{k'}\|_F^2\end{aligned}\tag{4.37}$$

where the second inequality follows from Equation 4.33 and the third follows from the fact that PCPs preserve Frobenius norm up to a constant factor. While conditioning on ζ_3 , it follows from Equation 4.37 that $\mathbf{Z}\mathbf{Z}^\top$ is an $(\tilde{O}(1), k/\epsilon)$ -SF projection for \mathbf{C} , our goal is to compute an SF projection for \mathbf{A} . Since $\mathbf{Z}\mathbf{Z}^\top$ is a $t \times t$ matrix, it does not even match the dimensions of \mathbf{A} . Therefore, we set up the following *Spectral Regression* problem:

$$\min_{\mathbf{W} \in \mathbb{R}^{n \times k'}} \|\mathbf{C} - \mathbf{W}\mathbf{Z}^\top\|_2^2\tag{4.38}$$

Let $\widehat{\mathbf{W}}$ be the approximate minimizer of the above problem obtained by running Algorithm 2. Then, it follows from Theorem 4.18 that with probability at least 99/100,

$$\begin{aligned}\|\mathbf{C} - \widehat{\mathbf{W}}\mathbf{Z}^\top\|_2^2 &\leq \tilde{O}(1) \left(\min_{\mathbf{W}} \|\mathbf{C} - \mathbf{W}\mathbf{Z}^\top\|_2^2 + \frac{\epsilon}{k} \|\mathbf{C} - \mathbf{C}_{k'}\|_F^2 \right) \\ &\leq \tilde{O}(1) \left(\|\mathbf{C} - \mathbf{C}_{k'}\|_2^2 + \frac{\epsilon}{k} \|\mathbf{C} - \mathbf{C}_{k'}\|_F^2 \right) \\ &\leq \tilde{O}(1) \left(\frac{\epsilon}{k} \|\mathbf{C} - \mathbf{C}_{k'}\|_F^2 \right)\end{aligned}\tag{4.39}$$

where the second inequality follows from $\|\mathbf{C} - \mathbf{C}\mathbf{Z}\mathbf{Z}^\top\|_2^2 \leq \tilde{O}\left(\frac{\epsilon}{k}\right)\|\mathbf{C} - \mathbf{C}_{k'}\|_F^2$ (by definition of an SF projection) and observing that $\mathbf{W} = \mathbf{C}\mathbf{Z}^\top$ is a feasible solution to Equation 4.38. Let ζ_4 be the event that Equation 4.39 holds. Next, let \mathbf{Q} be an orthonormal basis for \mathbf{W} . We observe that $\mathbf{Q}\mathbf{Q}^\top\mathbf{C}$ is the orthogonal projection of \mathbf{C} onto the subspace spanned by \mathbf{Q} and the matrix $\widehat{\mathbf{W}}\mathbf{Z}^\top$ also lies in the subspace. Therefore, by the Pythagorean Theorem, for any fixed unit vector y ,

$$\|\mathbf{C}y - \mathbf{Q}\mathbf{Q}^\top\mathbf{C}y\|_2^2 \leq \|\mathbf{C}y - \widehat{\mathbf{W}}\mathbf{Z}^\top y\|_2^2 \leq \|\mathbf{C} - \widehat{\mathbf{W}}\mathbf{Z}^\top\|_2^2$$

Picking y such that $\|\mathbf{C}y - \mathbf{Q}\mathbf{Q}^\top\mathbf{C}y\|_2^2 = \|\mathbf{C} - \mathbf{Q}\mathbf{Q}^\top\mathbf{C}\|_2^2$, and combining it with Equation 4.39 we have

$$\|\mathbf{C} - \mathbf{Q}\mathbf{Q}^\top\mathbf{C}\|_2^2 \leq \tilde{O}(1) \left(\frac{\epsilon}{k} \|\mathbf{C} - \mathbf{C}_{k'}\|_F^2 \right)\tag{4.40}$$

Conditioning on event ζ_1 , we know that $\|\mathbf{C} - \mathbf{C}_{k'}\|_F^2 = \|\mathbf{A} - \mathbf{A}_{k'}\|_F^2$. Since $\mathbf{Q}\mathbf{Q}^\top$ is a rank- k' projection matrix and \mathbf{C} is a mixed Spectral-Frobenius column PCP for \mathbf{A} , it follows from Equation 4.34,

$$\begin{aligned}\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\top\mathbf{A}\|_2^2 &\leq \frac{10}{9}\|\mathbf{C} - \mathbf{Q}\mathbf{Q}^\top\mathbf{C}\|_2^2 + \frac{1}{9k'}\|\mathbf{A} - \mathbf{A}_{k'}\|_F^2 \\ &\leq \tilde{O}(1) \left(\frac{\epsilon}{k} \|\mathbf{A} - \mathbf{A}_{k'}\|_F^2 \right)\end{aligned}\tag{4.41}$$

where the last inequality follows from Equation 4.40. Therefore, $\mathbf{Q}\mathbf{Q}^\top$ is a $(0.1, k')$ -SF projection for \mathbf{A} . Finally, we run Algorithm 1 on $\min_{\text{rank}(\mathbf{X})=k} \|\mathbf{A} - \mathbf{Q}\mathbf{X}\mathbf{Q}^\top\|_F^2$. Here, we note that for Algorithm 1, a $(0.1, k')$ -SF projection is equivalent to an (ϵ, k) -SF projection, up to polylogarithmic factors. Therefore, Theorem 4.16 holds as is. Then, by Theorem 4.16, we know that with probability at least 99/100 Algorithm 1 outputs matrices \mathbf{M}, \mathbf{N} such that $\|\mathbf{A} - \mathbf{M}\mathbf{N}^\top\|_F^2 \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2$. Let ζ_5 be the event that the aforementioned algorithm succeeds. Then, union bounding over $\zeta_1, \zeta_2, \zeta_3, \zeta_4$ and ζ_5 , with probability at least 9/10, \mathbf{M}, \mathbf{N} is a relative-error Low-Rank Approximation for \mathbf{A} , which concludes correctness.

Next, we analyze the query complexity and running time of Algorithm 3. Step 1 computes the rank- k' ridge leverage scores of $\mathbf{A}^{1/2}$ and by Lemma 4.9, requires reading $O(nk' \log(k')) = \tilde{O}(nk/\epsilon)$ entries in \mathbf{A} and runs in time $\tilde{O}(n(k/\epsilon)^{\omega-1})$. Steps 2 and 3 require no queries to \mathbf{A} and the sampling can be performed in $O(n)$ time. In step 4, the *input sparsity* algorithm from Lemma 4.6 queries $\text{nnz}(\mathbf{R}) = t^2 = \tilde{O}(nk/\epsilon)$ entries in \mathbf{A} and runs in $\tilde{O}(nk/\epsilon + \sqrt{n}\text{poly}(k/\epsilon)) = \tilde{O}(nk/\epsilon)$ time. We know from Theorem 4.16 that Step 5 requires $\tilde{O}(nk/\epsilon)$ queries to \mathbf{A} and runs in $\tilde{O}(nk/\epsilon + (k/\epsilon)^\omega)$ time. Finally, in Step 6, we run Algorithm 1 such that \mathbf{Q} is a $n \times k'$ matrix. Therefore, it follows from Theorem 4.16, that the total number of queries to \mathbf{A} is $\tilde{O}(nk/\epsilon + k^2/\epsilon^6)$ and the running time is $\tilde{O}(n(k/\epsilon)^{\omega-1} + (k/\epsilon^3)^\omega)$. The final query complexity and running time follows, and this concludes the proof.

Outputting a PSD Low-Rank Approximation. Here, we extend our algorithm to show that we can obtain a relative-error low-rank approximation matrix \mathbf{B} such that \mathbf{B} itself is a PSD matrix, using the same sample complexity and running time as in Theorem 4.1. Outputting a PSD low-rank approximation was first considered by Clarkson and Woodruff [CW17], who obtain an input-sparsity algorithm for arbitrary \mathbf{A} . When \mathbf{A} is PSD, Musco and Woodruff show that this problem can be solved with $\tilde{O}(nk/\epsilon^3 + nk^2/\epsilon^2)$ queries, in time $\tilde{O}(n(k/\epsilon)^\omega + nk^{\omega-1}/\epsilon^{3(\omega-1)})$.

We run Algorithm 3 till Step 5, i.e., we recover \mathbf{Q} such that $\mathbf{Q}\mathbf{Q}^\top$ is a SF projection for \mathbf{A} . We then modify Algorithm 1 by considering the following optimization problem instead:

$$\min_{\substack{\text{rank}(\mathbf{X}) \leq k \\ \mathbf{X} \geq 0}} \|\mathbf{A} - \mathbf{Q}\mathbf{X}\mathbf{Q}^\top\|_F^2 \quad (4.42)$$

As before, we sketch on both sides by sampling proportional to the leverage scores of \mathbf{Q} . Let the resulting sampling matrices be denoted by \mathbf{S}, \mathbf{T} . Then, we have the following sketched optimization problem:

$$\min_{\substack{\text{rank}(\mathbf{X}) \leq k \\ \mathbf{X} \geq 0}} \|\mathbf{SAT} - \mathbf{S}\mathbf{Q}\mathbf{X}\mathbf{Q}^\top\mathbf{T}\|_F^2 \quad (4.43)$$

Following Step 4 in Algorithm 1, we can compute $\mathbf{SAT}, \mathbf{P}_{\mathbf{S}\mathbf{Q}}, \mathbf{P}_{\mathbf{Q}^\top\mathbf{T}}$. We then compute $\hat{\mathbf{X}} = (\mathbf{S}\mathbf{Q})^\dagger \mathbf{P}_{\mathbf{S}\mathbf{Q}} \mathbf{SAT} \mathbf{P}_{\mathbf{Q}^\top\mathbf{T}} (\mathbf{Q}^\top\mathbf{T})^\dagger$ and $\mathbf{X}^* = [(\hat{\mathbf{X}} + \hat{\mathbf{X}}^\top)/2]_{k+}$, where for any matrix \mathbf{M} , $[\mathbf{M}]_{k+}$ is defined by setting all but the top- k positive eigenvalues to 0. Finally, we output $\mathbf{N}\mathbf{N}^\top$ where $\mathbf{N} = \mathbf{Q}(\mathbf{X}^*)^{1/2}$.

Corollary 4.26. (*Outputting a PSD Low-Rank Approximation.*) *Given an $n \times n$ PSD matrix \mathbf{A} , an integer k , and $1 > \epsilon > 0$, there exists an algorithm that samples $\tilde{O}(nk/\epsilon)$ entries in \mathbf{A} and outputs a rank- k $\mathbf{M}\mathbf{M}^\top$*

such that with probability at least $9/10$,

$$\|\mathbf{A} - \mathbf{M}\mathbf{M}^\top\|_F^2 \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2$$

Further, the algorithm runs in $\tilde{O}(n(k/\epsilon)^{\omega-1} + (k/\epsilon^3)^\omega)$ time.

Proof. We first note that an extension of Lemma 4.14 holds for outputting a PSD matrix as well. As a consequence of the following lemma, obtaining an approximate solution to the optimization problem in Equation 4.42 suffices.

Lemma 4.27. (Structured Projections and PSD LRA [CW17].) *Let $\mathbf{P} \in \mathbb{R}^{n \times n}$ be an (ϵ, k) -SF projection w.r.t \mathbf{A} , then*

$$\|\mathbf{A} - \mathbf{P}\mathbf{A}_{k+\mathbf{P}}\mathbf{P}\|_F^2 \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{A}_{k+}\|_F^2$$

We then use the analysis of Lemma 15 from [CW17] to conclude that \mathbf{X}^* is the minimizer for Equation 4.43. Finally, we note that the running time and query complexity is dominated by computing \mathbf{Q} and thus is the same as Theorem 4.1. Computing \mathbf{X}^* requires no additional queries to \mathbf{A} and only contributes a lower order term to the running time. \square

4.4 Negative-Type Distances

In this subsection, we consider the problem of computing low-rank approximation for distance matrices. Here, the input matrix \mathbf{A} is formed by the pairwise distances between a set of points $\mathbf{P} = \{p_1, \dots, p_n\}$ in an underlying metric space d , i.e., $A_{i,j} = d(p_i, p_j)$. Low-rank approximation for distance matrices was introduced by Bakshi and Woodruff [BW18] who obtained sublinear time *additive-error* algorithms for arbitrary metrics. Subsequently, Indyk et. al. [IVWW19] provided sample-optimal algorithms for additive-error low-rank approximation. For arbitrary distance matrices, it is known that relative-error algorithms require $\Omega(\text{nnz}(\mathbf{A}))$ queries [BW18].

Here, we focus on the special case of negative-type (Euclidean Squared) metrics [Sch38]. Negative-type metrics have numerous applications in algorithm design since it is possible to optimize over them using a semidefinite program (SDP). One significant algorithmic application of negative-type metrics appears in the Arora-Rao-Vazirani algorithm for the Sparsest Cut problem [ARV09]. We refer the reader to extensive subsequent work on embeddability of such metrics and the references therein [ALN08, ALN07, CGR05]. It is well-known that negative-type metrics include ℓ_1 and ℓ_2 metrics, spherical metrics and hyper metrics [DL09, TD87]. Therefore, our algorithms extend to distance matrices that arise from all such metrics.

For negative-type metrics, Bakshi and Woodruff obtain a bi-criteria relative-error low-rank approximation algorithm that queries $\tilde{O}(nk/\epsilon^{2.5})$ entries in \mathbf{A} and output a rank $k + 4$ matrix. In contrast, we obtain a sample-optimal algorithm that does not require a bi-criteria guarantee. As noted above, our algorithm works for any distance matrix where the distance can be realized as a negative-type metric.

Theorem 4.28 (Sample-Optimal Negative-Type LRA). *Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a negative-type distance matrix. Given $\epsilon > 0$ and $k \in [n]$, there exists an algorithm that queries $\tilde{O}(nk/\epsilon)$ entries in \mathbf{A} and outputs matrices*

$\mathbf{M}, \mathbf{N}^\top \in \mathbb{R}^{n \times k}$ such that with probability 99/100,

$$\|\mathbf{A} - \mathbf{MN}\|_F^2 \leq (1 + \epsilon) \|\mathbf{A} - \mathbf{A}_k\|_F^2$$

Further, the algorithm runs in time $\tilde{O}(n(k/\epsilon)^{\omega-1})$.

To demonstrate the connection between negative-type metrics and PSD matrices, we observe that a negative-type distance matrix \mathbf{A} can be realized as the distances corresponding to a point set $\mathcal{P} = \{x_1, x_2, \dots, x_n\}$ such that $\mathbf{A}_{i,j} = \|x_i - x_j\|_2^2 = \|x_i\|_2^2 + \|x_j\|_2^2 - 2\langle x_i, x_j \rangle$. Therefore, we can rewrite \mathbf{A} as $\mathbf{R}_1 + \mathbf{R}_2 - 2\mathbf{B}$, where for all $j \in [n]$, $(\mathbf{R}_1)_{i,j} = \|x_i\|_2^2$, $\mathbf{R}_2 = \mathbf{R}_1^\top$ and \mathbf{B} is PSD. Further, we can obtain query access to \mathbf{B} by simply assuming w.l.o.g. that x_1 is centered at the origin and the i -th entry in the first row corresponds to $\|x_i\|_2^2$. Therefore, we can simulate our PSD low-rank approximation algorithms on the matrix \mathbf{B} by only having query access to \mathbf{A} .

Our main contribution here is to show that if $\mathbf{P} = \mathbf{Q}\mathbf{Q}^\top$ is an $(O(1), k/\epsilon)$ -SF projection matrix for \mathbf{B} , then adjoining \mathbf{Q}^\top with the row span of \mathbf{R}_1 and \mathbf{R}_2 results in an SF-projection matrix for \mathbf{A} . Here, the row span of \mathbf{R}_1 is $\mathbf{1}^\top / \sqrt{n}$ and \mathbf{R}_2 is v such that for all $i \in [n]$ $v_i = \|x_i\|_2^2 / \sum_i \|x_i\|_2^2$. We note that once we obtain an SF projection for \mathbf{A} , we can run Algorithm 1 to output a $(1 + \epsilon)$ relative-error low-rank approximation.

Lemma 4.29 (Structured Projections for Distance Matrices). *Let \mathbf{A} be a negative-type matrix such that $\mathbf{A} = \mathbf{R}_1 + \mathbf{R}_2 - 2\mathbf{B}$, as defined above and let $\epsilon > 0$. Given an $(O(1), k/\epsilon)$ -SF projection $\mathbf{P} = \mathbf{Q}\mathbf{Q}^\top$ for \mathbf{B} , let Ω^\top be a basis for \mathbf{Q}^\top appended with the basis vectors for $\text{rowspan}(\mathbf{R}_1)$ and $\text{rowspan}(\mathbf{R}_2)$. Then, with probability at least 99/100,*

$$\min_{\text{rank}(\mathbf{X}) \leq k} \|\mathbf{A} - \Omega\mathbf{X}\Omega^\top\|_2^2 \leq (1 + \epsilon) \|\mathbf{A} - \mathbf{A}_k\|_F^2$$

Proof. By Lemma 7 in [CW17], for any symmetric matrices \mathbf{Y}, \mathbf{Z} such that $(\mathbf{Y} - \mathbf{Z})\mathbf{Z} = 0$ and projection matrix \mathbf{P} , the following holds:

$$\|\mathbf{Y} - \mathbf{PZP}\|_F^2 = \|\mathbf{Y} - \mathbf{Z}\|_F^2 + \|\mathbf{Z} - \mathbf{PZP}\|_F^2 + 2\text{Tr}((\mathbf{Y} - \mathbf{Z})(\mathbf{I} - \mathbf{P})\mathbf{ZP}) \quad (4.44)$$

Applying Equation 4.44 with $\mathbf{Y} = \mathbf{A}$ and $\mathbf{Z} = \mathbf{A}_k$, for any projection matrix \mathbf{P} , we have

$$\|\mathbf{A} - \mathbf{PA}_k\mathbf{P}\|_F^2 = \|\mathbf{A} - \mathbf{A}_k\|_F^2 + \|\mathbf{A}_k - \mathbf{PA}_k\mathbf{P}\|_F^2 + 2\text{Tr}((\mathbf{A} - \mathbf{A}_k)(\mathbf{I} - \mathbf{P})\mathbf{A}_k\mathbf{P}) \quad (4.45)$$

Next, we bound the $\|\mathbf{A}_k - \mathbf{PA}_k\mathbf{P}\|_F^2$ as follows:

$$\begin{aligned} \|\mathbf{A}_k - \mathbf{PA}_k\mathbf{P}\|_F^2 &\leq 2\|\mathbf{A}_k(\mathbf{I} - \mathbf{P})\|_F^2 \\ &\leq 2k\|\mathbf{A}_k(\mathbf{I} - \mathbf{P})\|_2^2 \\ &\leq 2k\|\mathbf{A}(\mathbf{I} - \mathbf{P})\|_2^2 \end{aligned} \quad (4.46)$$

To bound the trace, we use the Von Neuman trace inequality,

$$\begin{aligned} 2\text{Tr}((\mathbf{A} - \mathbf{A}_k)(\mathbf{I} - \mathbf{P})\mathbf{A}_k\mathbf{P}) &= 2\text{Tr}((\mathbf{A} - \mathbf{A}_k)(\mathbf{I} - \mathbf{P})^2\mathbf{A}_k\mathbf{P}) \\ &\leq 2 \sum_{i \in [n]} \sigma_i((\mathbf{A} - \mathbf{A}_k)(\mathbf{I} - \mathbf{P})) \sigma_i((\mathbf{I} - \mathbf{P})\mathbf{A}_k\mathbf{P}) \\ &\leq 2k\|(\mathbf{A} - \mathbf{A}_k)(\mathbf{I} - \mathbf{P})\|_2 \|(\mathbf{I} - \mathbf{P})\mathbf{A}_k\mathbf{P}\|_2 \\ &\leq 2k\|\mathbf{A}(\mathbf{I} - \mathbf{P})\|_2^2 \end{aligned} \quad (4.47)$$

It suffices to bound $\|\mathbf{A}(\mathbf{I} - \mathbf{P})\|_2^2$ for $\mathbf{P} = \mathbf{\Omega}\mathbf{\Omega}^\top$. Since $\mathbf{A} = \mathbf{R}_1 + \mathbf{R}_2 - 2\mathbf{B}$ and $(\mathbf{R}_1 + \mathbf{R}_2)(\mathbf{I} - \mathbf{P}) = 0$, we have

$$\begin{aligned}\|\mathbf{A}(\mathbf{I} - \mathbf{P})\|_2^2 &\leq 2\|\mathbf{B}(\mathbf{I} - \mathbf{P})\|_2^2 \\ &\leq 2\|\mathbf{B}(\mathbf{I} - \mathbf{Q}\mathbf{Q}^\top)\|_2^2 \\ &\leq O\left(\frac{\epsilon}{k}\right) \|\mathbf{B} - \mathbf{B}_{k+2}\|_F^2\end{aligned}$$

To relate $\|\mathbf{B} - \mathbf{B}_k\|_F^2$ back to \mathbf{A} , observe

$$\begin{aligned}\|\mathbf{A} - \mathbf{A}_k\|_F^2 &= \|\mathbf{R}_1 + \mathbf{R}_2 - 2\mathbf{B} - \mathbf{A}_k\|_F^2 = 4\|\mathbf{B} - (\mathbf{R}_1 + \mathbf{R}_2 - \mathbf{A}_k)/2\|_F^2 \\ &\geq 4\|\mathbf{B} - \mathbf{B}_{k+2}\|_F^2\end{aligned}$$

Therefore, $\|\mathbf{A}(\mathbf{I} - \mathbf{P})\|_2^2 \leq O(\epsilon/k)\|\mathbf{A} - \mathbf{A}_k\|_F^2$. We can thus bound Equations 4.47 and 4.46 with $O(\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2$. Substituting this into Equation 4.45, we conclude that $\|\mathbf{A} - \mathbf{P}\mathbf{A}_k\mathbf{P}\|_F^2 \leq (1 + O(\epsilon))\|\mathbf{A} - \mathbf{A}_k\|_F^2$ for $\mathbf{P} = \mathbf{\Omega}\mathbf{\Omega}^\top$ and the claim follows. \square

Recall, we can compute an SF projection for the PSD matrix \mathbf{B} efficiently using Algorithm 3 and then solve the optimization problem in Lemma 4.29 using Algorithm 1. We can therefore reduce low-rank approximation of negative-type matrices to PSD low-rank approximation with only $O(n)$ additional queries and Theorem 4.28 follows.

4.5 Ridge Regression

We consider the following regression problem: given a PSD matrix \mathbf{A} , a vector y and a ridge parameter λ ,

$$\min_x \|\mathbf{A}x - y\|_2^2 + \lambda\|x\|_2^2.$$

As a corollary of Theorem 4.1, we obtain a faster algorithm for the aforementioned problem. We begin with the following simple lemma from [MW17]:

Lemma 4.30 (Lemma 26 in [MW17]). *Given a PSD matrix \mathbf{A} , vector y , and $\lambda > 0$, let \mathbf{B} be a matrix such that $\|\mathbf{A} - \mathbf{B}\|_2^2 \leq \epsilon^2\lambda$. Then, for any vector \tilde{x} such that*

$$\|\mathbf{B}\tilde{x} - y\|_2^2 + \lambda\|\tilde{x}\|_2^2 \leq (1 + \epsilon') \left(\min_x \|\mathbf{B}x - y\|_2^2 + \lambda\|x\|_2^2 \right)$$

we have

$$\|\mathbf{A}\tilde{x} - y\|_2^2 + \lambda\|\tilde{x}\|_2^2 \leq (1 + \epsilon')(1 + 5\epsilon) \left(\min_x \|\mathbf{A}x - y\|_2^2 + \lambda\|x\|_2^2 \right)$$

Therefore, it suffices to find a rank- k matrix \mathbf{B} such that $\|\mathbf{A} - \mathbf{B}\|_2^2 \leq \epsilon^2\lambda$. Let \tilde{s}_λ be an upper bound on the statistical dimension $s_\lambda = \text{Tr}((\mathbf{A}^2 + \lambda\mathbf{I})^{-1}\mathbf{A}^2)$. Setting $k = \tilde{s}_\lambda/\epsilon^2$, we can bound $\|\mathbf{A} - \mathbf{A}_k\|_F^2$ as follows:

$$\begin{aligned}\frac{\epsilon^2}{\tilde{s}_\lambda} \|\mathbf{A} - \mathbf{A}_k\|_F^2 &\leq \epsilon^2 \frac{\sum_{i=k+1}^n \lambda_i^2(\mathbf{A})}{\sum_{i=1}^n \lambda_i^2(\mathbf{A})/(\lambda_i^2(\mathbf{A}) + \lambda)} \\ &\leq \epsilon^2 \frac{\sum_{i=k+1}^n \lambda_i^2(\mathbf{A})}{\sum_{i=k+1}^n \lambda_i^2(\mathbf{A})/(\lambda_i^2(\mathbf{A}) + \lambda)} \\ &\leq c\epsilon^2\lambda\end{aligned}$$

We can then solve the regression problem $\min_x \|\mathbf{B}x - y\|_2^2 + \lambda \|x\|_2^2$ exactly in time $O(nk^{\omega-1})$ and obtain the following result:

Theorem 4.31 (Ridge Regression). *Given a PSD matrix \mathbf{A} , a regularization parameter λ and an upper bound \tilde{s}_λ on the statistical dimension $s_\lambda = \text{Tr}((\mathbf{A}^2 + \lambda\mathbf{I})^{-1}\mathbf{A}^2)$, there exists an algorithm that queries $\tilde{O}(n\tilde{s}_\lambda/\epsilon^2)$ entries of \mathbf{A} and with probability 99/100 outputs \hat{x} such that for all $y \in \mathbb{R}^d$,*

$$\|\mathbf{A}\hat{x} - y\|_2^2 + \lambda \|\hat{x}\|_2^2 \leq (1 + \epsilon) \left(\min_x \|\mathbf{A}x - y\|_2^2 + \lambda \|x\|_2^2 \right)$$

Further, the algorithm runs in $\tilde{O}(n(\tilde{s}_\lambda/\epsilon^2)^{\omega-1})$ time.

Remark 4.32. Observe that we can derive a data structure from our algorithm that preserves the objective cost (up to $1 + \epsilon$) for all x and y simultaneously and thus we obtain a coreset for Ridge Regression.

To complement the above algorithmic result, we present a new lower bound for coreset constructions for ridge regression, which matches our upper bound in all parameters. At a high level, our hard instance for constant s_λ consists of $1/\epsilon^2$ blocks of all 1s, each of size $\epsilon\sqrt{n} \times \epsilon\sqrt{n}$, placed randomly across the matrix. Since any coreset construction must preserve the cost of all x, y , we pick pairs (x, y) to be the eigenvectors of \mathbf{A} (scaled appropriately) and show that in order to preserve the cost of all pairs, the coreset algorithm must find all the blocks, which requires $\Omega(n/\epsilon^2)$ queries to \mathbf{A} . Repeating the above construction s_λ -times suffices to obtain a linear lower bound in terms of s_λ . Formally,

Theorem 4.33 (Coreset Lower Bound for Ridge Regression). *Given a PSD matrix \mathbf{A} and $\epsilon, \lambda > 0$ let $s_\lambda = \text{Tr}((\mathbf{A}^2 + \lambda\mathbf{I})^{-1}\mathbf{A}^2)$ denote the statistical dimension of \mathbf{A} . Then, any coreset construction \mathbf{C} that with constant probability, preserves the ridge regression cost up to $(1 + \epsilon)$ simultaneously for all x, y , must read $\Omega(ns_\lambda/\epsilon^2)$ entries in \mathbf{A} .*

We recall the lower bound instance for low rank approximation of PSD matrices shown by Musco and Woodruff :

Theorem 4.34 (Lower Bound for PSD LRA ([MW17])). *Given an $n \times n$ PSD matrix \mathbf{A} , $\epsilon_0 > 0$ and $k_0 \in [n]$, any randomized algorithm that outputs a rank k_0 matrix \mathbf{B} such that with probability at least 9/10,*

$$\|\mathbf{A} - \mathbf{B}\|_F^2 \leq (1 + \epsilon_0) \|\mathbf{A} - \mathbf{A}_{k_0}\|_F^2$$

must query $\Omega(nk_0/\epsilon_0)$ entries in \mathbf{A} .

We consider the hard distribution defined by Musco and Woodruff, and show that we can obtain a low rank approximation to this instance with strengthened parameters by using a coreset for ridge regression.

Definition 4.35 (Hard Input Distribution for LRA ([MW17])). Let \mathbf{M} be an $n \times n$ matrix and let $\epsilon_0 > 0, k_0 \in [n]$. Let $\gamma(n, \epsilon_0, k_0)$ be a distribution over \mathbf{M} such that $\mathcal{S} \subset [n]$ is a uniformly random subset of size $n/2$, which is further partitioned into subsets $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_{k_0}$ such that for all $\ell \in [k]$,

\mathcal{S}_ℓ is picked uniformly at random and $|\mathcal{S}_\ell| = n/(2k_0)$. For each subset \mathcal{S}_ℓ , let $\mathbf{A}_{\mathcal{S}_\ell}$ denote the principle submatrix of \mathbf{A} indexed by the set \mathcal{S}_ℓ . Then, with probability $1/2$, $\mathbf{A}_{\mathcal{S}_\ell}$ is such that all the diagonal entries are set to 1 and a uniformly random principle submatrix of $\mathbf{A}_{\mathcal{S}_\ell}$, indexed by the set \mathcal{T}_ℓ , such that $|\mathcal{T}_\ell| = c\sqrt{\epsilon_0|\mathcal{S}_\ell|}$ is set to all 1s. With the remaining probability $\mathbf{A}_{\mathcal{S}_\ell}$ is set to the \mathbf{I} .

We show that we can derive a low-rank matrix \mathbf{B} that satisfies the relative-error guarantee above from a coresets for ridge regression.

Proof of Theorem 4.33. We show a proof by contradiction, where the high level idea is that a coresets for ridge regression can be used to derive a low-rank approximation to \mathbf{A} , when \mathbf{A} is picked from $\gamma(n, O(1), s_\lambda/\epsilon^2)$ (the hard distribution defined in 4.35). First, we observe that with probability at least $99/100$, the input distribution has $\Omega(s_\lambda/\epsilon^2)$ blocks that contain a principle submatrix with all 1s. To see this let $X_1, \dots, X_{s_\lambda/\epsilon^2}$ be indicators for the corresponding blocks $\mathbf{A}_{\mathcal{S}_\ell}$ having a principle submatrix of all 1s. Then,

$$\mathbb{E} \left[\sum_{\ell \in [s_\lambda/\epsilon^2]} X_\ell \right] = \frac{\epsilon^2 n}{2s_\lambda} \quad (4.48)$$

Since the X_ℓ 's are independent, by a Chernoff bound we have

$$\Pr \left[\sum_{\ell \in [s_\lambda/\epsilon^2]} X_\ell \leq (1 - \delta) \frac{\epsilon^2 n}{2s_\lambda} \right] \leq \exp \left(-\frac{c\delta\epsilon^2 n}{s_\lambda} \right) \quad (4.49)$$

For $n \geq \Omega(s_\lambda/\epsilon^2)$, we can bound the above probability by $1/100$. We begin by showing that for our input instance, $s_\lambda = \Theta(n/\lambda)$ and thus the aforementioned equations differ by $O(\epsilon n/s_\lambda)$. To see this observe

$$s_\lambda = \sum_{i \in [n]} \frac{\sigma_i^2(\mathbf{A})}{\sigma_i^2(\mathbf{A}) + \lambda} \quad (4.50)$$

Then, there are $s_\lambda/2\epsilon^2$ large eigenvalues, each of magnitude $\epsilon\sqrt{n/s_\lambda}$ and thus the total contribution is

$$\frac{s_\lambda}{2\epsilon^2} \cdot \frac{(\epsilon^2 n/s_\lambda)}{(\epsilon^2 n/s_\lambda) + \lambda} = \frac{n}{\epsilon^2 n/s_\lambda + \lambda}$$

The remaining eigenvalues simply contribute $1/(1 + \lambda)$ to the sum and since there are at most n of them, the total contribution is $n/(1 + \lambda)$. Therefore, we can conclude $s_\lambda = \Theta(n/\lambda)$.

For a block in \mathbf{A} indexed by ℓ , let \hat{x}_ℓ be the eigenvector supported on indices in \mathcal{S}_ℓ and let $\hat{y}_\ell = \sqrt{n/s_\lambda} \hat{x}_\ell$. For non-identity blocks, $\mathbf{A} \hat{x}_\ell = |\mathcal{T}_\ell| \hat{x}_\ell = \sqrt{\epsilon^2 n/s_\lambda} \hat{x}_\ell$ and the regression cost is

$$\|(1 - \epsilon)\sqrt{n/s_\lambda} \hat{x}_\ell\|_2^2 + \lambda = (1 - 2\epsilon)n/s_\lambda + cn/s_\lambda \quad (4.51)$$

When the block indexed by ℓ is the identity block, we get $\mathbf{A} \hat{x}_\ell = \hat{x}$ and the regression cost is

$$\|(\sqrt{n/s_\lambda} - 1)\hat{x}_\ell\|_2^2 + \lambda = (n/s_\lambda + 1 - 2\sqrt{n/s_\lambda}) + cn/s_\lambda \quad (4.52)$$

Instead consider a vector that intersects an eigenvector \tilde{x}_ℓ on a $(1 - \gamma)$ -fraction of the support and the rest is arbitrary. Then, when an all 1s block exists, $\mathbf{A}\tilde{x}_\ell \geq (1 - \gamma)^2 |\mathcal{T}_\ell| \hat{x}_\ell = (1 - \gamma)^2 \sqrt{\epsilon^2 n / s_\lambda} \hat{x}_\ell$ and thus the regression cost is at most

$$(1 - \epsilon(1 - 2\gamma))^2 n / s_\lambda + cn / s_\lambda$$

Further, when the block is simply the identity, a similar calculation shows that the regression cost is at least $(1 - 2\epsilon\gamma)^2 n / s_\lambda + cn / s_\lambda$. Therefore, the ridge regression cost determines the existence of a $(1 - 2\gamma)$ -fraction of an all 1s principle submatrix even when \hat{x}_ℓ intersects with an eigenvector on a $(1 - \gamma)$ -fraction of coordinates.

Consider a coresset C for the above instance. Since this coresset preserves the ridge regression objective upto a $(1 + \epsilon/1000)$ factor for all x, y , as per our above discussion we can query the coresset on the tuples $(\hat{x}_\ell, \hat{y}_\ell)$, which represent the eigenvectors of each block, to determine if a block contains a principle submatrix with all 1s. However, a priori we do not know the support of the eigenvector within each block \mathbf{A}_{S_ℓ} .

Instead we query the coresset on all possible supports and show we can determine the right one as follows: let \tilde{x} be supported on a set that intersects with a principle submatrix of all 1s on at most a γ fraction. Observe that $\mathbf{A}\tilde{x} \leq \frac{1}{\gamma} (\gamma^2 \epsilon \tilde{x})$ and thus the ridge regression cost can be lower bounded as follows:

$$(1 - \epsilon\gamma)^2 n / s_\lambda + cn / s_\lambda \tag{4.53}$$

We therefore take the set of all vectors on which the coresset cost is less than the above cost and let the resulting list be \mathcal{L} . Note, this list must include the eigenvectors and further, only includes vectors which intersect an all 1's submatrix on a $1 - \gamma$ -fraction. Therefore, picking a set of $\epsilon^2 n / s_\lambda$ vectors that have maximum support suffices.

Since we detect a $(1 - \gamma)$ -fraction of all principle submatrices in \mathbf{A} , it follows that we can output a $1 + c'$ -approximate low-rank approximation for \mathbf{A} , for a fixed small constant c' . To see this, observe that the optimal low-rank approximation to \mathbf{A} is given by the matrix that selects all the principle submatrices with all 1s and thus $\|\mathbf{A} - \mathbf{A}_{k_0}\|_F^2 = n - k_0 = n - s_\lambda / \epsilon^2$. Further, our approximation to \mathbf{A}_k , denoted by \mathbf{B} , matches \mathbf{A}_k on a $(1 - \gamma)$ -fraction of each principle submatrix of all 1s and thus we match \mathbf{A}_k on these entries. Subsequently we bound the additional cost that \mathbf{B} incurs which \mathbf{A}_k does not. This includes entries that are 1 in \mathbf{A}_k and 0 in \mathbf{B} and vice versa.

To bound the cost of the entries that exist in \mathbf{A}_k but do not exist in \mathbf{B} , observe on each principle submatrix, \mathbf{B} and \mathbf{A} intersect in at least $(1 - \gamma)^2$ -fraction of the entries and thus the remaining entries are at most $(1 - (1 - \gamma)^2) \cdot \frac{\epsilon^2 n}{cs_\lambda} \cdot \frac{s_\lambda}{\epsilon^2} = 4\gamma n / c$, since the size of each block is $\frac{\epsilon^2 n}{cs_\lambda}$ and the number of blocks are at most $\frac{s_\lambda}{\epsilon^2}$. Finally, observe that since we do not pick exact eigenvectors we can have non-zero off diagonal entries in \mathbf{B} that do not exist in \mathbf{A}_k . However, we have at most γ -fraction of the support on each indicator vector remaining and contributing to two rectangular blocks of 1s, each of size $\gamma \cdot \frac{\epsilon^2 n}{cs_\lambda} \cdot \frac{s_\lambda}{\epsilon^2} = \gamma n / c$. Therefore, the additional non-zero entries in \mathbf{B} that do not appear in \mathbf{A} are $2\gamma n / c$ in number.

Therefore, the overall cost $\|\mathbf{A} - \mathbf{B}\|_F^2 \leq (1 + 6\gamma/c)n$. By setting the constants γ and c and observing that $s_\lambda / \epsilon^2 \ll n$, we obtain a $(1 + c')$ -low-rank approximation to \mathbf{A} , for any arbitrary

small constant c' . Therefore, our reduction suffices to solve the hard instance above and a lower bound of $\Omega(nk_0/\epsilon_0) = \Omega(ns_\lambda/\epsilon^2)$ queries follows. \square

5 Robust Low-Rank Approximation

One drawback of relative-error guarantees is that the corresponding algorithms cannot tolerate any amount of noise. Therefore, we introduce a robust model for low-rank approximation by relaxing the requirement from relative-error guarantees to additive-error guarantees. In the robustness model we consider, we begin with an $n \times n$ PSD matrix \mathbf{A} . An adversary is then allowed to arbitrarily corrupt \mathbf{A} by adding a corruption matrix \mathbf{N} such that the corruption in each row is an fixed constant times the ℓ_2^2 row norm of the row and the total corruption is an η -fraction of squared Frobenius norm of \mathbf{A} . While the adversary may corrupt any number of entries of \mathbf{A} , the norm of the corruption matrix is bounded and the algorithm has query access to $\mathbf{A} + \mathbf{N}$. We parameterize our lower bound and algorithms by the largest ratio between a diagonal entry of \mathbf{A} and $\mathbf{A} + \mathbf{N}$, denoted by $\phi_{\max} = \max_{j \in [n]} \mathbf{A}_{j,j} / |(\mathbf{A} + \mathbf{N})_{j,j}|$. This captures the intuition that the diagonal is crucial for sublinear time low rank approximation and the sample complexity degrades as we corrupt larger diagonals entries.

5.1 Lower Bound for Robust PSD Low-Rank Approximation

In this subsection, we show a query lower bound of $\Omega(\eta^2 n^2 k / \epsilon^2) = \Omega(\phi_{\max}^2 nk / \epsilon)$ for any algorithm that outputs a low rank approximation up to additive-error $(\epsilon + \eta)\|\mathbf{A}\|_F^2$. Note, obtaining error smaller than $\eta\|\mathbf{A}\|_F^2$ is information-theoretically impossible and reflected in the query lower bound.

Our lower bound holds for randomized algorithms, and uses Yao's minimax principle [Yao77]. The overall strategy is to demonstrate a lower bound for deterministic algorithms on a carefully chosen input distribution. We construct our input distributions as follows: let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a block diagonal matrix with such that \mathbf{B}_1 is $5\epsilon/\eta \times 5\epsilon/\eta$ randomly positioned, non-contiguous block with all entries $\sqrt{n\eta^2/5\epsilon}$ and \mathbf{B}_2 is the identity matrix on the remaining indices. \mathbf{A} is clearly a PSD matrix since each principle submatrix is PSD. Observe $\|\mathbf{A}\|_F^2 = (25\epsilon/\eta^2) \cdot (n\eta^2/5\epsilon) + (n - 5\epsilon/\eta) = (1 + 5\epsilon)n - o(n)$. Further, the dense block \mathbf{B}_1 contributes a total squared Frobenius norm of at least $4\epsilon n$ and the diagonal entry contributes an η/ϵ fraction of each row. Since $\epsilon \geq \eta$, the diagonal contributes at most the entire ℓ_2^2 norm. The corresponding diagonal of \mathbf{B}_1 also has ℓ_2^2 squared norm $5\epsilon/\eta \cdot n\eta^2/5\epsilon = \eta n$.

At a high level, the adversary can then corrupt the diagonal and set each diagonal entry to be 1, making it hard for the algorithm to find rows corresponding to \mathbf{B}_1 . We show that any ϵ -additive-error low-rank approximation must detect at least one entry in \mathbf{B}_1 to adaptively sample the corresponding row and column, but the diagonals no longer provide any useful information. Thus any algorithm must query most entries in \mathbf{A} . Further, in our construction, note $\phi_{\max} = \sqrt{n\eta^2/5\epsilon}$.

We first describe intuitively why a low rank approximation needs to recover many rows from the block \mathbf{B}_1 . Since \mathbf{A} has this block structure, the best rank-1 approximation satisfies $\|\mathbf{A} - \mathbf{A}_k\|_F^2 = n - |\mathbf{B}_1|$. Therefore, assuming the cardinality of \mathbf{B}_1 is negligible, in order to obtain an overall error bound of $\epsilon\|\mathbf{A}\|_F^2 \geq \epsilon n$, the algorithm must find a constant fraction of off-diagonal entries in \mathbf{B}_1 .

This is because \mathbf{B}_1 contributes at least $5\epsilon n$ norm. However, since the diagonals no longer convey any information about the off-diagonals, and the block \mathbf{B}_1 is placed on a random subset of indices, any deterministic algorithm must read arbitrary off-diagonal entries until it finds a non-zero entry. Since there are only $25\epsilon^2/\eta^2$ non-zeros in \mathbf{B}_1 , to find one in expectation (over the input distribution) requires sampling $\epsilon^2 n^2/\eta^2$ entries. While the above serves well as intuition, a rigorous proof requires many additional steps. We begin by defining a distribution over the input matrices :

Definition 5.1. Given $n \in \mathbb{N}, \epsilon > \eta > 0$, let $\mathcal{S} \subset [n]$ be a uniformly random subset of size $\lceil 5\epsilon/\eta \rceil$. Let $\mu(n, \epsilon, \eta)$ be a distribution over matrices $\mathbf{M} \in \mathbb{R}^{n \times n}$ such that $\forall i \in [n], \mathbf{M}_{i,i} = 1$ and $\forall i, j \in \mathcal{S}, \mathbf{M}_{i,j} = \sqrt{\eta^2 n/5\epsilon}$. All remaining entries in \mathbf{M} are 0.

Next, we show that any \mathbf{M} sampled from $\mu(n, \epsilon, \eta)$ can be decomposed into $\mathbf{A} + \mathbf{N}$ such that \mathbf{A} is PSD and $\|\mathbf{N}\|_F^2 \leq \eta \|\mathbf{A}\|_F^2$. To see this, let \mathbf{N} be a diagonal matrix such that for all $i \in \mathcal{S}$, $\mathbf{N}_{i,i} = -\sqrt{\eta^2 n/5\epsilon}$ and let $\mathbf{A} = \mathbf{M} - \mathbf{N}$. We give an algebraic proof that \mathbf{A} is PSD, but \mathbf{A} can also be decomposed into a rank-1 block of all $\sqrt{\eta^2 n/5\epsilon}$ -s corresponding to all $i, j \in \mathcal{S}$ and identity on the remaining indices. For all $x \in \mathbb{R}^n$,

$$\begin{aligned} x^T \mathbf{A} x &= \sum_{i,j \in \mathcal{S}} \mathbf{A}_{i,j} x_i x_j + \sum_{i,j \notin \mathcal{S}} \mathbf{A}_{i,j} x_i x_j \\ &= \sqrt{\eta^2 n/5\epsilon} \sum_{i,j \in \mathcal{S}} x_i x_j + \sum_{i \notin \mathcal{S}} x_i^2 \\ &= \sqrt{\eta^2 n/5\epsilon} \left(\sum_{i \in \mathcal{S}} x_i \right)^2 + \sum_{i \notin \mathcal{S}} x_i^2 \\ &\geq 0 \end{aligned} \tag{5.1}$$

and thus \mathbf{A} is PSD. Further, $\|\mathbf{A}\|_F^2 = (25\epsilon^2/\eta^2) \cdot (n\eta^2/5\epsilon) + (n - 5\epsilon/\eta) = (1 + 5\epsilon)n - 5\epsilon/\eta$. Then, $\|\mathbf{N}\|_F^2 = 5\epsilon/\eta \cdot \eta^2 n/\epsilon = \eta \|\mathbf{A}\|_F^2$, as desired. Intuitively, we show that if \mathbf{B} is a rank- k matrix that is a good low-rank approximation for \mathbf{M} sampled from μ , then it cannot be a good low-rank approximation for \mathbf{I} . To this end, we consider a distribution where \mathbf{M} is drawn from $\mu(n, \epsilon, \eta)$ with probability $1/2$ and is $\mathbf{I}_{n \times n}$ with probability $1/2$.

Definition 5.2. (Hard Distribution) Given $n \in \mathbb{N}, \epsilon > \eta > 0$, let $\nu(n, \epsilon, \eta)$ be a distribution over $\mathbf{M} \in \mathbb{R}^{n \times n}$ such that with probability $1/2$, \mathbf{M} is sampled from $\mu(n, \epsilon, \eta)$ and with probability $1/2$, $\mathbf{M} = \mathbf{I}_{n \times n}$.

We now show that a low-rank approximation to \mathbf{M} can be used as a certificate to separate the mixture $\nu(n, \epsilon, \eta)$ since it can distinguish between the input being identity or far from it. Thus if the distributions are close in a statistical sense, any algorithm to distinguish between the two would require querying many entries in \mathbf{M} . Formally,

Lemma 5.3. (LRA as a Distinguisher.) Let \mathbf{M} be a matrix drawn from $\mu(n, \epsilon, \eta)$ and let \mathbf{B} be a rank- k matrix that is the candidate low-rank approximation to \mathbf{M} such that $\|\mathbf{M} - \mathbf{B}\|_F^2 \leq \epsilon n$. Then, $\|\mathbf{M} - \mathbf{I}\|_F^2 > 1.1\epsilon n$.

Proof. Since $\|\mathbf{M} - \mathbf{B}\|_F^2 \leq \epsilon n$, \mathbf{B} must have at least $4\epsilon n$ mass on the off-diagonal entries of \mathbf{M} . So, \mathbf{B} must have at least $10\epsilon^2/\eta^2$ non-zero off-diagonal entries. Therefore, it must have at least $5\epsilon^2/\eta^2$ entries with squared mass $\epsilon n/2$. To see why, assume there is a subset of at least $12\epsilon^2/\eta^2$ entries, each being at most $\sqrt{n\eta^2/10\epsilon}$. Restricted to only these entries, the squared Frobenius norm difference between \mathbf{M} and \mathbf{B} is already at least $1.2\epsilon n$, contradicting our assumption. Given that there exists a subset of $5\epsilon^2/\eta^2$ off-diagonal entries having squared mass $1.2\epsilon n$, $\|\mathbf{B} - \mathbf{I}\|_F^2 > 1.2\epsilon n$, and thus \mathbf{B} is not an additive error low-rank approximation for \mathbf{I} . \square

Theorem 5.4. (Lower bound for PSD Matrices.) Let \mathbf{A} be a PSD matrix, $k \in \mathbb{Z}$ and $\epsilon > 0$ be any constant. Let \mathbf{N} be an arbitrary matrix such that $\|\mathbf{N}\|_F^2 \leq \eta \|\mathbf{A}\|_F^2$. Any randomized algorithm \mathcal{A} that only has query access to $\mathbf{A} + \mathbf{N}$, with probability at least $2/3$, computes a rank- k matrix \mathbf{B} such that

$$\|\mathbf{A} - \mathbf{B}\|_F^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + \epsilon \|\mathbf{A}\|_F^2$$

must read $\Omega(\phi_{\max}^2 nk/\epsilon)$ entries of $\mathbf{A} + \mathbf{N}$ on some input, possibly adaptively, in expectation.

Proof. Let Algorithm \mathcal{A} be a deterministic algorithm that outputs a rank- k matrix \mathbf{B} such that it is an additive-error low-rank approximation \mathbf{M} . Let $T \subset [n^2]$ be the subset of entries read by \mathcal{A} . Let $L(\mu)$ denote the distribution of T conditioned on $\mathbf{M} \sim \mu(n, \epsilon, \eta)$ and $L(i)$ be the distribution of T conditioned on $\mathbf{M} = \mathbf{I}$. By Lemma 5.3, since the output of \mathcal{A} can be used to distinguish between the two distributions, it is well-known that the success probability over the randomness in T is at most $1/2 + D_{TV}(L(\mu), L(i))/2$ [BY02]. Since we assume \mathcal{A} succeeds with probability at least $2/3$,

$$D_{TV}(L(\mu), L(i)) \geq 1/3 \tag{5.2}$$

It remains to upper bound D_{TV} in terms of $|T|$. Recall, \mathcal{S} is the random set of indices where $\mu(n, \epsilon, \eta)$ is non-zero. Let $\tilde{\mathcal{S}}$ be the subset of \mathcal{S} restricted to the off-diagonal entries of \mathbf{M} . When $\mathbf{M} \sim \mu(n, \epsilon, \eta)$, $\forall i, j \in \tilde{\mathcal{S}}$, $\mathbf{M}_{i,j}$ is non-zero and when $\mathbf{M} = \mathbf{I}$, the same entries are 0. Observe, for all $i, j \notin \tilde{\mathcal{S}}$, $\mathbf{M}_{i,j}$ are fixed. Further, \mathcal{S} is a uniform subset of $[n]$. Therefore,

$$\Pr[(i, j) \in T \mid (i, j) \in \mathcal{S}] = \frac{|T|\epsilon^2}{\eta^2 n^2} \tag{5.3}$$

Then, with probability at least $1 - |T|\epsilon^2/\eta^2 n^2$, \mathcal{A} queries the same entries for both $L(\mu)$ and $L(i)$. Therefore

$$D_{TV}(L(\mu), L(i)) \leq |T|\epsilon^2/\eta^2 n^2.$$

Combined with Equation 5.2, if \mathcal{A} succeeds with probability at least $2/3$, $|T|\epsilon^2/\eta^2 n^2 \geq 1/3$ and thus $|T| = \Omega(\eta^2 n^2/\epsilon^2)$. Given that any deterministic algorithm must query $\Omega(\eta^2 n^2/\epsilon^2) = \Omega(\phi_{\max}^2 nk/\epsilon)$ entries for $\nu(n, \epsilon, \eta)$, to now obtain a linear dependence on the rank k , we can use the standard approach of creating k disjoint copies of the block \mathbf{B}_1 in the hard distribution, as shown in [MW17]. The theorem follows from Yao's minimax principle. \square

5.2 Robust Sublinear Low-Rank Approximation Algorithms

In this subsection, we provide a robust algorithm for the model discussed above. We parameterize our algorithms and lower bound by the largest ratio between a diagonal entry of \mathbf{A} and $\mathbf{A} + \mathbf{N}$, denoted by $\phi_{\max} = \max_{j \in [n]} \mathbf{A}_{j,j} / |(\mathbf{A} + \mathbf{N})_{j,j}|$. In addition, we provide robust PCP constructions, by introducing a new sampling procedure to construct projection-cost preserving sketches. Our sampling procedure is straightforward: we sample each column proportional to the diagonal entry in that column. This sampling requires n queries to the matrix \mathbf{A} to obtain an additive-error projection cost preservation guarantee. Further, for the special case of correlation matrices, we can uniformly sample columns of \mathbf{A} to obtain a smaller matrix such that all rank k projections in the column and row space are preserved.

For our algorithms, we assume we know ϕ_{\max} . In practice, this assumption may not hold, but we can query as many entries in $\mathbf{A} + \mathbf{N}$ as our budget allows, given that correctness holds only when the queries are at least $\tilde{O}(\phi_{\max}^2 nk / \epsilon)$. Since we read the diagonals of $\mathbf{A} + \mathbf{N}$ and we know ϕ_{\max} , we can obtain an upper bound on $\mathbf{A}_{i,i}$ and $\mathbf{A}_{j,j}$. Therefore, whenever we query an off-diagonal entry in $\mathbf{A} + \mathbf{N}$, we can truncate it to $\phi_{\max} \sqrt{|(\mathbf{A} + \mathbf{N})_{i,i}| \cdot |(\mathbf{A} + \mathbf{N})_{j,j}|}$ without increasing the corruption in our input.

Robust Projection-Cost Preserving Sketches. Here, we show that diagonal sampling is a robust sampling procedure to create *projection-cost preserving sketches*. We begin by relating the ℓ_2^2 row (or column) norms of a PSD matrix to its spectral norm. Let \mathbf{A} be a PSD matrix and let $\mathbf{U}\Sigma\mathbf{U}^T$ be the SVD for \mathbf{A} .

Lemma 5.5. *Given an $n \times n$ PSD matrix \mathbf{A} , for all $i \in [n]$, $\|\mathbf{A}_{i,*}\|_2^2 \leq \|\mathbf{A}\|_2 \cdot \mathbf{A}_{i,i}$.*

Proof. Observe, $\mathbf{A}_{i,*} = \mathbf{U}_i \Sigma \mathbf{U}^T$ and $\mathbf{A}_{i,i} = (\mathbf{U}_{i,*} \Sigma \mathbf{U}^T)_i = \sum_{j=1}^n \sigma_j(\mathbf{A}) \mathbf{U}_{i,j}^2$. Then,

$$\begin{aligned} \|\mathbf{A}_{i,*}\|_2^2 &= \mathbf{A}_{i,*} \mathbf{A}_{i,*}^T = \mathbf{U}_{i,*} \Sigma \mathbf{U}^T \mathbf{U} \Sigma \mathbf{U}^T \mathbf{U}_{i,*} = \sum_{j=1}^n \sigma_j^2(\mathbf{A}) \mathbf{U}_{i,j}^2 \\ &\leq \|\mathbf{A}\|_2 \sum_{j=1}^n \sigma_j \mathbf{U}_{i,j}^2 \\ &= \|\mathbf{A}\|_2 \cdot \mathbf{A}_{i,i} \end{aligned}$$

□

An immediate consequence of Lemma 5.5 is that the ℓ_2^2 norm of a row or column of a PSD matrix is at most $\frac{\|\mathbf{A}\|_F}{\mathbf{A}_{i,i}}$. Note, this precludes matrices where most of the mass is concentrated on a small number of rows or columns. Recall, we observe as input the matrix $\mathbf{A} + \mathbf{N}$ and our goal is to obtain a PCP for this in sublinear time and queries.

Musco and Musco [MM17] describe how to approximately compute the ridge leverage scores of $\mathbf{A}^{\frac{1}{2}}$ (if \mathbf{A} is PSD) using a Nystrom approximation. [MW17] use this method to compute the ridge leverage scores of $\mathbf{A}^{\frac{1}{2}}$ with $O(nk)$ queries, where $\mathbf{A} = \mathbf{A}^{\frac{1}{2}} \cdot \mathbf{A}^{\frac{1}{2}}$. However, these approaches do not apply when we perturb the input and it may no longer be PSD. Therefore, the best known

construction by Cohen et. al. [CMM17] would require $\Omega(\text{nnz}(\mathbf{A}))$ time to compute approximate ridge-leverage scores of \mathbf{A} . Note, this does not use the structure that \mathbf{A} has.

In contrast, we show that sampling columns proportional to the diagonal entries suffices to obtain a PCP. Note, we only need to query the diagonal of \mathbf{A} to compute the distribution over columns exactly. The main technical challenge here is to obtain the correct dependence on n and k and account for the perturbation to the input, given that our sampling probabilities are straightforward to compute and do not rely on spectral properties of $\mathbf{A} + \mathbf{N}$. Note, the following is a structural result and while we do not know \mathbf{A} , we can still show the following :

Theorem 5.6. (*Robust Spectral Bound.*) Let \mathbf{A} be an $n \times n$ PSD matrix and \mathbf{N} be an arbitrary matrix such that $\|\mathbf{N}\|_F^2 \leq \eta \|\mathbf{A}\|_F^2$ and for all $j \in [n]$, $\|\mathbf{N}_{*,j}\|_2^2 \leq c \|\mathbf{A}_{*,j}\|_2^2$, for any fixed constant c . Let $\phi_{\max} = \max_j \mathbf{A}_{j,j}/(\mathbf{A} + \mathbf{N})_{j,j}$ and let $q = \{q_1, q_2, \dots, q_n\}$ be a distribution over the columns of $\mathbf{A} + \mathbf{N}$ such that for all j , $q_j = (\mathbf{A} + \mathbf{N})_{j,j}/\text{Tr}(\mathbf{A} + \mathbf{N})$ and let $t = O(\phi_{\max} \sqrt{n} k^2 \log(n/\delta)/\epsilon^2)$. Then, construct a sampling matrix \mathbf{T} that samples t columns of $\mathbf{A} + \mathbf{N}$ such that it samples column $(\mathbf{A} + \mathbf{N})_{*,j}$ with probability q_j and scales it by $1/\sqrt{tq_j}$. With probability at least $1 - \delta$, for any rank- k orthogonal projection \mathbf{X} ,

$$\mathbf{A}\mathbf{A}^T - \left(\frac{\epsilon}{k}\right) \|\mathbf{A}\|_F^2 \mathbf{I} \leq \mathbf{A}\mathbf{T}(\mathbf{A}\mathbf{T})^T \leq \mathbf{A}\mathbf{A}^T + \left(\frac{\epsilon}{k}\right) \|\mathbf{A}\|_F^2 \mathbf{I}$$

Proof. First, we note that we cannot explicitly compute $\mathbf{A}\mathbf{T}$, but we can show that the sampling probabilities we have access to result in a PCP for \mathbf{A} . Let $\mathbf{Y} = \mathbf{A}\mathbf{T}(\mathbf{A}\mathbf{T})^T - \mathbf{A}\mathbf{A}^T$. For notational convenience let $\mathbf{A}_j = \mathbf{A}_{*,j}$. We can then write $\mathbf{Y} = \sum_{j \in [t]} \left(\mathbf{C}_{*,j} \mathbf{C}_{*,j}^T - \frac{1}{t} \mathbf{A}\mathbf{A}^T \right) = \sum_{j \in [t]} \mathbf{X}_j$, where $\mathbf{X}_j = \frac{1}{t} \left(\frac{1}{q_j} \mathbf{A}_j \mathbf{A}_j^T - \mathbf{A}\mathbf{A}^T \right)$ with probability q_j . We observe that $\mathbb{E}[\mathbf{X}_j] = \mathbb{E}[\mathbf{C}_{*,j} \mathbf{C}_{*,j}^T - \frac{1}{t} \mathbf{A}\mathbf{A}^T] = 0$, and therefore, $\mathbb{E}[\mathbf{Y}] = 0$. Next, we bound the operator norm of \mathbf{Y} . To this end, we use the Matrix Bernstein inequality, which in turn requires a bound on the operator norm of \mathbf{X}_j and variance of \mathbf{Y} . Recall,

$$\begin{aligned} \|\mathbf{X}_j\|_2 &= \left\| \frac{1}{tq_j} \mathbf{A}_j \mathbf{A}_j^T - \frac{1}{t} \mathbf{A}\mathbf{A}^T \right\|_2 \\ &\leq \frac{\text{Tr}(\mathbf{A} + \mathbf{N})}{t(\mathbf{A} + \mathbf{N})_{j,j}} \|\mathbf{A}_j\|_2^2 + \frac{1}{t} \|\mathbf{A}\|_2^2 \\ &\leq \frac{2\text{Tr}(\mathbf{A}) + |\text{Tr}(\mathbf{N})|}{t(\mathbf{A} + \mathbf{N})_{j,j}} ((1 + \eta) \|\mathbf{A}\|_2 \mathbf{A}_{j,j}) \\ &\leq \frac{c(\text{Tr}(\mathbf{A}) + |\text{Tr}(\mathbf{N})|) \|\mathbf{A}\|_2 \phi_{\max}}{t} \\ &\leq \frac{c\phi_{\max} \sqrt{n} \|\mathbf{A}\|_F \|\mathbf{A}\|_2}{t} \end{aligned} \tag{5.4}$$

where we use triangle inequality for operator norm to obtain the first inequality, triangle inequality up to a factor of 2 for ℓ_2^2 norms for the second inequality, $\|\mathbf{N}_j\|_2^2 \leq \eta \|\mathbf{A}_{*,j}\|_2^2$ and $\|\mathbf{A}_j\|_2^2 \leq \|\mathbf{A}\|_2 \cdot \mathbf{A}_{j,j}$ (from Lemma 5.5) for the third inequality and definition of ϕ_{\max} and $\eta = O(1)$ for the fourth. Finally, we relate the trace of \mathbf{A} and \mathbf{N} to their respective Frobenius norm using Cauchy-Schwarz:

$$\text{Tr}(\mathbf{A}) = \sum_{i=1}^n \sigma_i(\mathbf{A}) \leq \sqrt{\sum_{i=1}^n \sigma_i^2(\mathbf{A}) \cdot n} = \sqrt{n \|\mathbf{A}\|_F^2}$$

and

$$|\text{Tr}(\mathbf{N})| = \left| \sum_{i=1}^n \sigma_i(\mathbf{A}) \right| \leq \sqrt{\sum_{i=1}^n \sigma_i^2(\mathbf{N}) \cdot n} = \sqrt{n \|\mathbf{N}\|_F^2} \leq \sqrt{n\eta} \|\mathbf{A}\|_F$$

where the last inequality follows from $\|\mathbf{N}\|_F \leq \sqrt{\eta} \|\mathbf{A}\|_F$. Next, we bound $\mathbf{Var}[\mathbf{Y}] \leq \mathbb{E}[\mathbf{Y}^2]$.

$$\begin{aligned} \mathbb{E}[\mathbf{Y}^2] &= t \mathbb{E} \left[\left((\mathbf{A}\mathbf{T})_{*,j} (\mathbf{A}\mathbf{T})_{*,j}^T - \frac{1}{t} \mathbf{A}\mathbf{A}^T \right)^2 \right] \\ &= t \mathbb{E} \left[\left((\mathbf{A}\mathbf{T})_{*,j} (\mathbf{A}\mathbf{T})_{*,j}^T \right)^2 + \frac{1}{t^2} (\mathbf{A}\mathbf{A}^T)^2 - \frac{2}{t} (\mathbf{A}\mathbf{T})_{*,j} (\mathbf{A}\mathbf{T})_{*,j}^T \mathbf{A}\mathbf{A}^T \right] \\ &= \frac{1}{t} \left(\sum_{j \in [n]} \frac{(\mathbf{A}_j \mathbf{A}_j^T)^2}{q_j} + (\mathbf{A}\mathbf{A}^T)^2 - \sum_{j \in [n]} 2 \mathbf{A}_j \mathbf{A}_j^T \mathbf{A}\mathbf{A}^T \right) \\ &\leq \frac{\text{Tr}(\mathbf{A} + \mathbf{N})}{t \mathbf{A}_{j,j}} \left(\sum_{j \in [n]} (\mathbf{A}_j \mathbf{A}_j^T)^2 \right) \\ &\leq \frac{c \phi_{\max} \sqrt{n} \|\mathbf{A}\|_F \|\mathbf{A}\|_2}{t} \mathbf{A}\mathbf{A}^T \end{aligned} \tag{5.5}$$

where we use linearity of expectation, $(\mathbf{A}\mathbf{A}^T)^2 \geq 0$ and $\|\mathbf{A}_{j,*}\|_2^2 \leq \|\mathbf{A}\|_2 \cdot \mathbf{A}_{j,j}$. Applying the Matrix Bernstein inequality,

$$\begin{aligned} \Pr[\|\mathbf{Y}\|_2 \geq \epsilon \|\mathbf{A}\|_F^2] &\leq 2n \exp \left(- \frac{\epsilon^2 \|\mathbf{A}\|_F^4}{\frac{c \phi_{\max} \sqrt{n} \|\mathbf{A}\|_F \|\mathbf{A}\|_2^3}{t} + \frac{2 \phi_{\max} \sqrt{n} \|\mathbf{A}\|_F \|\mathbf{A}\|_2 (\epsilon \|\mathbf{A}\|_F^2)}{3t}} \right) \\ &\leq 2n \exp \left(- \frac{\epsilon^2 t}{c \phi_{\max} \sqrt{n} \|\mathbf{A}\|_F \|\mathbf{A}\|_2} \left(\frac{\|\mathbf{A}\|_F^4}{\|\mathbf{A}\|_2^2 + \epsilon \|\mathbf{A}\|_F^2} \right) \right) \\ &\leq 2n \exp \left(- \frac{\epsilon^2 t}{c' \phi_{\max} \sqrt{n}} \right) \\ &\leq \delta/2 \end{aligned}$$

where the last inequality follows from setting $t = O(\phi_{\max} \sqrt{n} \log(n/\delta)/\epsilon^2)$. To yield the claim, we set $\epsilon = \epsilon/k$. \square

We use the above spectral bound to show that sampling proportional to diagonal entries preserves the projection cost of the columns of \mathbf{A} on to any k -dimensional subspace up to an additive $(\epsilon + \sqrt{\eta}) \|\mathbf{A}\|_F^2$.

Theorem 5.7. (Column Projection-Cost Preservation.) *Given $\mathbf{A} + \mathbf{N}$, where \mathbf{A} is an $n \times n$ PSD matrix and \mathbf{N} is an arbitrary noise matrix as defined above, $k \in \mathbb{Z}$ and $\epsilon > \eta > 0$, let $q = \{q_1, q_2 \dots q_n\}$ be a probability distribution over the columns of $\mathbf{A} + \mathbf{N}$ such that $q_j = \frac{(\mathbf{A} + \mathbf{N})_{j,j}}{\text{Tr}(\mathbf{A} + \mathbf{N})}$. Let $t = O(\phi_{\max} \sqrt{n} k^2 \log(\frac{n}{\delta})/\epsilon^2)$. Then,*

construct \mathbf{C} using t columns of $\mathbf{A} + \mathbf{N}$ and set each one to $\frac{(\mathbf{A} + \mathbf{N})_{*,j}}{\sqrt{tq_j}}$ with probability q_j . With probability at least $1 - c$, for any rank- k orthogonal projection \mathbf{X} ,

$$\|\mathbf{C} - \mathbf{X}\mathbf{C}\|_F^2 = \|\mathbf{A} - \mathbf{X}\mathbf{A}\|_F^2 \pm (\epsilon + \sqrt{\eta})\|\mathbf{A}\|_F^2$$

for a fixed constant c .

Proof. Here, the matrix \mathbf{C} is actually a matrix we can compute. Observe that we can relate \mathbf{C} to the sampling matrix \mathbf{T} as defined in Theorem 5.6 as $\mathbf{C} = (\mathbf{A} + \mathbf{N})\mathbf{T}$. We follow the proof strategy of the relative error guarantees in [CMM17] and additive error guarantees in [BW18] but note, our spectral bounds from Theorem 5.6 apply to matrices that we do not actually compute. Observe, $\|\mathbf{A} - \mathbf{X}\mathbf{A}\|_F^2 = \text{Tr}((\mathbf{I} - \mathbf{X})\mathbf{A}\mathbf{A}^T(\mathbf{I} - \mathbf{X}))$. Then,

$$\begin{aligned} \text{Tr}((\mathbf{I} - \mathbf{X})\mathbf{A}\mathbf{A}^T(\mathbf{I} - \mathbf{X})) &= \text{Tr}(\mathbf{A}\mathbf{A}^T) + \text{Tr}(\mathbf{X}\mathbf{A}\mathbf{A}^T\mathbf{X}) - \text{Tr}(\mathbf{A}\mathbf{A}^T\mathbf{X}) - \text{Tr}(\mathbf{X}\mathbf{A}\mathbf{A}^T) \\ &= \text{Tr}(\mathbf{A}\mathbf{A}^T) + \text{Tr}(\mathbf{X}\mathbf{A}\mathbf{A}^T\mathbf{X}) - \text{Tr}(\mathbf{A}\mathbf{A}^T\mathbf{X}\mathbf{X}) - \text{Tr}(\mathbf{X}\mathbf{X}\mathbf{A}\mathbf{A}^T) \\ &= \text{Tr}(\mathbf{A}\mathbf{A}^T) + \text{Tr}(\mathbf{X}\mathbf{A}\mathbf{A}^T\mathbf{X}) - \text{Tr}(\mathbf{X}\mathbf{A}\mathbf{A}^T\mathbf{X}) - \text{Tr}(\mathbf{X}\mathbf{A}\mathbf{A}^T\mathbf{X}) \\ &= \text{Tr}(\mathbf{A}\mathbf{A}^T) - \text{Tr}(\mathbf{X}\mathbf{A}\mathbf{A}^T\mathbf{X}) \end{aligned} \quad (5.6)$$

where we used the fact that for any projection matrix $\mathbf{X} = \mathbf{X}^2$ in addition to the cyclic property of the trace. Similarly,

$$\|\mathbf{C} - \mathbf{X}\mathbf{C}\|_F^2 = \text{Tr}((\mathbf{I} - \mathbf{X})\mathbf{C}\mathbf{C}^T(\mathbf{I} - \mathbf{X})) = \text{Tr}(\mathbf{C}\mathbf{C}^T) - \text{Tr}(\mathbf{X}\mathbf{C}\mathbf{C}^T\mathbf{X}) \quad (5.7)$$

We first relate $\text{Tr}(\mathbf{A}\mathbf{A}^T)$ and $\text{Tr}(\mathbf{C}\mathbf{C}^T)$. Recall,

$$\mathbb{E}[\text{Tr}(\mathbf{C}\mathbf{C}^T)] = \mathbb{E}[\|\mathbf{C}\|_F^2] = \|\mathbf{A} + \mathbf{N}\|_F^2 \leq \text{Tr}(\mathbf{A}\mathbf{A}^T) + 2\sqrt{\eta}\|\mathbf{A}\|_F^2$$

Using a scalar Chernoff bound, we show that with probability at least $1 - 1/\text{poly}(n)$, $\|\mathbf{C}\|_F^2 = (1 \pm \epsilon)\|\mathbf{A} + \mathbf{N}\|_F^2$. This is equivalent to $|\|\mathbf{C}\|_F^2 - \|\mathbf{A} + \mathbf{N}\|_F^2| \leq \epsilon\|\mathbf{A} + \mathbf{N}\|_F^2$. Observe, for all $j \in [t]$, $\mathbf{C}_{*,j} = \frac{1}{\sqrt{q_j t}}(\mathbf{A} + \mathbf{N})_{*,j'}$ for some $j' \in [n]$. Then,

$$\begin{aligned} \|\mathbf{C}_{*,j}\|_2^2 &= \frac{1}{q_j t} \|(\mathbf{A} + \mathbf{N})_{*,j'}\|_2^2 = \frac{\text{Tr}(\mathbf{A} + \mathbf{N}) \epsilon^2}{\phi_{\max} \sqrt{n} k \log(n/\delta) (\mathbf{A} + \mathbf{N})_{j',j'}} \|(\mathbf{A} + \mathbf{N})_{*,j'}\|_2^2 \\ &\leq \frac{c\sqrt{n}\|\mathbf{A}\|_F \epsilon^2}{\sqrt{n} \log(n/\delta)} \|\mathbf{A}\|_2 \\ &\leq \frac{c\epsilon^2}{k \log(n/\delta)} \|\mathbf{A} + \mathbf{N}\|_F^2 \end{aligned} \quad (5.8)$$

where we use $\text{Tr}(\mathbf{A}) \leq \sqrt{n}\|\mathbf{A}\|_F$, $\text{Tr}(\mathbf{N}) \leq \sqrt{\eta n}\|\mathbf{A}\|_F$ and $t = O(\phi_{\max} \sqrt{n} k \log(n/\delta)/\epsilon^2)$. Therefore, $\frac{k \log(n/\delta)}{\epsilon^2 \|\mathbf{A}\|_F^2} \|\mathbf{C}_{*,j}\|_2^2 \in [0, 1]$. By a Chernoff bound,

$$\begin{aligned}
\Pr \left[\|\mathbf{C}\|_F^2 \geq (1 + 2\epsilon) \|\mathbf{A} + \mathbf{N}\|_F^2 \right] &= \Pr \left[\frac{k \log(n/\delta)}{\epsilon^2 \|\mathbf{A} + \mathbf{N}\|_F^2} \|\mathbf{C}\|_F^2 \geq \frac{k \log(n/\delta)}{\epsilon^2} (1 + \epsilon) \right] \\
&\leq \exp \left(-\frac{k \epsilon^2 \log(n/\delta)}{\epsilon^2} \right) \\
&\leq \frac{\delta}{2}
\end{aligned} \tag{5.9}$$

We can repeat the above argument to lower bound $\|\mathbf{C}\|_F^2$. Therefore, with probability $1 - \delta$, we have

$$| \|\mathbf{C}\|_F^2 - \|\mathbf{A} + \mathbf{N}\|_F^2 | \leq \epsilon \|\mathbf{A} + \mathbf{N}\|_F^2$$

Here, we can upper bound this by observing $\|\mathbf{A} + \mathbf{N}\|_F^2 \leq \|\mathbf{A}\|_F^2 + \|\mathbf{N}\|_F^2 + 2\langle \mathbf{A}, \mathbf{N} \rangle \leq \|\mathbf{A}\|_F^2 + 3\sqrt{\eta} \|\mathbf{A}\|_F^2$. Therefore,

$$| \|\mathbf{C}\|_F^2 - \|\mathbf{A}\|_F^2 | \leq \epsilon \|\mathbf{A}\|_F^2 + (1 + \epsilon) \sqrt{\eta} \|\mathbf{A}\|_F^2 \leq (\epsilon + 2\sqrt{\eta}) \|\mathbf{A}\|_F^2 \tag{5.10}$$

Next, we relate $\text{Tr}(\mathbf{XCC}^T\mathbf{X})$ and $\text{Tr}(\mathbf{XAA}^T\mathbf{X})$. First, we observe

$$\mathbf{CC}^T = (\mathbf{AT} + \mathbf{NT})(\mathbf{AT} + \mathbf{NT})^T = (\mathbf{AT})(\mathbf{AT})^T + (\mathbf{AT})(\mathbf{NT})^T + (\mathbf{NT})(\mathbf{AT})^T + (\mathbf{NT})(\mathbf{NT})^T \tag{5.11}$$

We begin by first bounding $\text{Tr}(\mathbf{X(AT)(AT)}^T\mathbf{X})$. Observe, \mathbf{X} is a rank k projection matrix and we can represent it as \mathbf{ZZ}^T , where $\mathbf{Z} \in \mathbb{R}^{n \times k}$ and has orthonormal columns. By the cyclic property of the trace, we have

$$\text{Tr}(\mathbf{ZZ}^T(\mathbf{AT})(\mathbf{AT})^T\mathbf{ZZ}^T) = \text{Tr}(\mathbf{Z}^T(\mathbf{AT})(\mathbf{AT})^T\mathbf{Z}) = \sum_{j \in [k]} \mathbf{Z}_{*,j}^T (\mathbf{AT})(\mathbf{AT})^T \mathbf{Z}_{*,j}$$

Similarly, $\text{Tr}(\mathbf{ZZ}^T\mathbf{AA}^T\mathbf{ZZ}^T) = \sum_{j \in [k]} \mathbf{Z}_{*,j}^T \mathbf{AA}^T \mathbf{Z}_{*,j}$. By Theorem 5.6, we have

$$\begin{aligned}
\sum_{j \in [k]} \left(\mathbf{Z}_{*,j}^T \mathbf{AA}^T \mathbf{Z}_{*,j} - \left(\frac{\epsilon}{k} \right) \|\mathbf{A}\|_F^2 \mathbf{Z}_{*,j}^T \mathbf{I} \mathbf{Z}_{*,j} \right) &\leq \sum_{j \in [k]} \left(\mathbf{Z}_{*,j}^T (\mathbf{AT})(\mathbf{AT})^T \mathbf{Z}_{*,j} \right) \\
&\leq \sum_{j \in [k]} \left(\mathbf{Z}_{*,j}^T \mathbf{AA}^T \mathbf{Z}_{*,j} + \left(\frac{\epsilon}{k} \right) \|\mathbf{A}\|_F^2 \mathbf{Z}_{*,j}^T \mathbf{I} \mathbf{Z}_{*,j} \right)
\end{aligned} \tag{5.12}$$

Since $\mathbf{Z}_{*,j}^T \mathbf{Z}_{*,j} = 1$ and $\text{Tr}(\mathbf{Z}^T \mathbf{AA}^T \mathbf{Z}) = \text{Tr}(\mathbf{XAA}^T\mathbf{X})$, we have

$$\text{Tr}(\mathbf{XAA}^T\mathbf{X}) - \epsilon \|\mathbf{A}\|_F^2 \leq \text{Tr}(\mathbf{X(AT)(AT)}^T\mathbf{X}) \leq \text{Tr}(\mathbf{XAA}^T\mathbf{X}) + \epsilon \|\mathbf{A}\|_F^2 \tag{5.13}$$

Next, we focus on $\text{Tr}(\mathbf{X(NT)(NT)}^T\mathbf{X}) = \|\mathbf{XNT}\|_F^2$. Observe, since \mathbf{T} is an unbiased estimator of Frobenius norm, by Markov's inequality we can show with probability at least $1 - c$, $\|\mathbf{XNT}\|_F = c \|\mathbf{N}\|_F = O(\sqrt{\eta}) \|\mathbf{A}\|_F$. Therefore, we can upper bound $\text{Tr}(\mathbf{X(NT)(NT)}^T\mathbf{X})$ by $O(\eta) \|\mathbf{A}\|_F^2$. Now, we focus on the cross terms. By Cauchy-Schwartz, and a Markov bound, with probability at least $1 - c$,

$$\text{Tr}(\mathbf{X(AT)(NT)}^T\mathbf{X}) \leq \|\mathbf{AT}\|_F \cdot \|\mathbf{NT}\|_F \leq O(\sqrt{\eta}) \|\mathbf{A}\|_F^2 \tag{5.14}$$

Combining equations 5.10, 5.13, 5.14 and union bounding over the success of the random events, with probability $1 - c$,

$$\|\mathbf{A} - \mathbf{XA}\|_F^2 - O(\epsilon + \sqrt{\eta})\|\mathbf{A}\|_F^2 \leq \|\mathbf{C} - \mathbf{XC}\|_F^2 \leq \|\mathbf{A} - \mathbf{XA}\|_F^2 + O(\epsilon + \sqrt{\eta})\|\mathbf{A}\|_F^2$$

□

Robust Row Projection Cost Preserving Sketches. We now extend the diagonal sampling algorithm to construct a row projection cost preserving sketch for the matrix \mathbf{C} . We note that following the construction for \mathbf{A} does not immediately give a row PCP for \mathbf{C} since \mathbf{C} is no longer PSD or even square matrix. Here, all previous approaches to construct a PCP with sublinear queries hit a roadblock, since the matrix \mathbf{C} need not have any well defined structure apart from being a scaled subset of the columns of a PSD matrix. However, we show that sampling rows of \mathbf{C} proportional to the diagonal entries of \mathbf{A} results in a row PCP.

We begin by relating the row norms of \mathbf{C} to the row norms of \mathbf{A} . Note, we do not expect to obtain concentration here, since such a sampling procedure would then help us estimate row norms of \mathbf{A} up to a constant and we would be done by using [FKV04]. Therefore, we obtain the following one-sided guarantee:

Lemma 5.8. *Let $\mathbf{AT} \in \mathbb{R}^{n \times t}$ be a column projection-cost preserving sketch for \mathbf{A} as described in Theorem 5.6. For all $i \in [n]$, with probability at least $1 - 1/n^c$,*

$$\|(\mathbf{AT})_{i,*}\|_2^2 \leq O\left(\log(n) \max\left\{\|\mathbf{A}_{i,*}\|_2^2, \frac{\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F\mathbf{A}_{i,i}}{t}\right\}\right)$$

where c is a fixed constant.

Proof. Observe that $\|(\mathbf{AT})_{i,*}\|_2^2 = \sum_{j \in [t]} (\mathbf{AT})_{i,j}^2$, where $(\mathbf{AT})_{i,j}^2 = \frac{\text{Tr}(\mathbf{A} + \mathbf{N})}{t(\mathbf{A} + \mathbf{N})_{j,j}} \mathbf{A}_{i,j}^2$ with probability $\frac{(\mathbf{A} + \mathbf{N})_{i,j}}{\text{Tr}(\mathbf{A} + \mathbf{N})}$. Then, $\mathbb{E}[\|(\mathbf{AT})_{i,*}\|_2^2] = \sum_{i=1}^n \mathbf{A}_{i,i}^2 = \|\mathbf{A}_{i,*}\|_2^2$. Next, we compute the variance of $\|(\mathbf{AT})_{i,*}\|_2^2$. $\text{Var}[\|(\mathbf{AT})_{i,*}\|_2^2] = t \text{Var}[(\mathbf{AT})_{i,j}^2] \leq \mathbb{E}[(\mathbf{AT})_{i,j}^4]$. Then,

$$\begin{aligned} t \mathbb{E}[(\mathbf{AT})_{i,j}^4] &= \sum_{j \in [n]} \frac{1}{t q_j} \mathbf{A}_{i,j}^4 \leq \sum_{j \in [n]} \frac{\text{Tr}(\mathbf{A} + \mathbf{N})}{t(\mathbf{A} + \mathbf{N})_{j,j}} \mathbf{A}_{i,j}^2 \mathbf{A}_{i,i} \mathbf{A}_{j,j} \\ &\leq \frac{\text{Tr}(\mathbf{A} + \mathbf{N}) \phi_{\max} \mathbf{A}_{i,i}}{t} \|\mathbf{A}_{i,*}\|_2^2 \\ &\leq \left(\frac{2\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F\mathbf{A}_{i,i}}{t}\right)^2 + \|\mathbf{A}_{i,*}\|_2^4 \quad [\text{AM-GM}] \end{aligned}$$

where we use $\mathbf{A}_{i,j}^2 \leq \mathbf{A}_{i,i} \mathbf{A}_{j,j}$, which follows from applying Cauchy-Schwarz to $\langle \mathbf{A}_{i,*}^{1/2}, \mathbf{A}_{j,*}^{1/2} \rangle$, i.e.,

$$\mathbf{A}_{i,j}^2 = \langle \mathbf{A}_{i,*}^{1/2}, \mathbf{A}_{j,*}^{1/2} \rangle^2 \leq \|\mathbf{A}_{i,*}^{1/2}\|_2^2 \|\mathbf{A}_{j,*}^{1/2}\|_2^2 = \mathbf{A}_{i,i} \mathbf{A}_{j,j}$$

Similarly, we bound

$$(\mathbf{AT})_{i,j}^2 = \frac{\text{Tr}(\mathbf{A} + \mathbf{N})}{t(\mathbf{A} + \mathbf{N})_{j,j}} \mathbf{A}_{i,j}^2 \leq \frac{2\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F}{t} \mathbf{A}_{i,i}$$

Applying Bernstein's inequality,

$$\begin{aligned} & \Pr \left[\left| \|(\mathbf{AT})_{i,*}\|_2^2 - \|\mathbf{A}_{i,*}\|_2^2 \right| \geq \gamma \max \left\{ \|\mathbf{A}_{i,*}\|_2^2, \frac{\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F \mathbf{A}_{i,i}}{t} \right\} \right] \\ & \leq 2 \exp \left(- \frac{\gamma^2 \max \left\{ \|\mathbf{A}_{i,*}\|_2^4, \left(\frac{\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F \mathbf{A}_{i,i}}{t} \right)^2 \right\}}{\left(\frac{2\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F \mathbf{A}_{i,i}}{t} \right)^2 + \|\mathbf{A}_{i,*}\|_2^4 + \gamma \max \left\{ \frac{\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F \mathbf{A}_{i,i} \|\mathbf{A}_{i,*}\|_2^2}{t}, \left(\frac{\text{Tr}(\mathbf{A}) \mathbf{A}_{i,i}}{t} \right)^2 \right\}} \right) \\ & \leq 2 \exp \left(- \frac{\gamma \max \left\{ \|\mathbf{A}_{i,*}\|_2^4, \left(\frac{\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F \mathbf{A}_{i,i}}{t} \right)^2 \right\}}{c' \left(\frac{\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F \mathbf{A}_{i,i}}{t} \right)^2 + c' \|\mathbf{A}_{i,*}\|_2^4} \right) \end{aligned}$$

where $\frac{\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F \mathbf{A}_{i,i}}{t} \|\mathbf{A}_{i,*}\|_2^2 \leq \left(\frac{\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F \mathbf{A}_{i,i}}{t} \right)^2 + \|\mathbf{A}_{i,*}\|_2^4$ follows from the AM-GM inequality. Setting $\gamma = \Omega(\log(n))$ completes the proof. \square

To construct a row projection cost preserving sketch of \mathbf{C} , we sample t rows of \mathbf{C} proportional to the corresponding diagonal entries of \mathbf{A} . Formally, we consider a probability distribution, $p = \{p_1, p_2, \dots, p_n\}$, over the rows of \mathbf{C} such that $p_i = \frac{\mathbf{A}_{i,i}}{\text{Tr}(\mathbf{A})}$. Let \mathbf{R} be a $t \times t$ matrix where each row of \mathbf{R} is set to $\frac{1}{\sqrt{tp_i}} \mathbf{C}_{i,*}$ with probability p_i . As before, \mathbf{R} can be represented as $\mathbf{S}\mathbf{C} = \mathbf{S}(\mathbf{AT} + \mathbf{NT})$. We first obtain a spectral guarantee for \mathbf{SAT} , while we cannot actually compute this.

Theorem 5.9. (Spectral Bounds.) Let \mathbf{AT} be an $n \times t$ matrix constructed as shown in Theorem 5.6. Let $p = \{p_1, p_2, \dots, p_n\}$ be a probability distribution over the rows of \mathbf{AT} such that $p_i = \frac{(\mathbf{A} + \mathbf{N})_{i,i}}{\text{Tr}(\mathbf{A} + \mathbf{N})}$. Let $t = O\left(\frac{\sqrt{n}k^2}{\epsilon^2} \log\left(\frac{n}{\delta}\right)\right)$. Construct a sampling matrix \mathbf{S} that samples t rows of \mathbf{AT} such that row $(\mathbf{AT})_{i,*}$ is picked with probability p_i and scaled by $\frac{1}{\sqrt{tp_i}}$. Then, with probability at least $1 - \delta$,

$$(\mathbf{AT})^T(\mathbf{AT}) - \frac{\epsilon}{k} \|\mathbf{A}\|_F^2 \mathbf{I} \leq (\mathbf{SAT})^T(\mathbf{SAT}) \leq (\mathbf{AT})^T(\mathbf{AT}) + \frac{\epsilon}{k} \|\mathbf{A}\|_F^2 \mathbf{I}$$

Proof. Let $\mathbf{Y} = (\mathbf{SAT})^T(\mathbf{SAT}) - (\mathbf{AT})^T(\mathbf{AT})$. For notational convenience let $(\mathbf{AT})_i = (\mathbf{AT})_{i,*}$ and $(\mathbf{SAT})_i = (\mathbf{SAT})_{i,*}$. We can then write $\mathbf{Y} = \sum_{i \in [t]} \left((\mathbf{SAT})_i^T (\mathbf{SAT})_i - \frac{1}{t} (\mathbf{AT})^T (\mathbf{AT}) \right) = \sum_{i \in [t]} \mathbf{X}_i$, where $\mathbf{X}_i = \frac{1}{t} \left(\frac{1}{p_i} (\mathbf{AT})_i^T (\mathbf{AT})_i - (\mathbf{AT})^T (\mathbf{AT}) \right)$ with probability p_i . We observe that $\mathbb{E}[\mathbf{X}_i] = \mathbb{E}[(\mathbf{SAT})_i^T (\mathbf{SAT})_i - \frac{1}{t} (\mathbf{AT})^T (\mathbf{AT})] = \sum_i \frac{p_i}{p_i} (\mathbf{AT})_i^T (\mathbf{AT})_i - (\mathbf{AT})^T (\mathbf{AT}) = 0$, and therefore, $\mathbb{E}[\mathbf{Y}] = 0$. Next, we bound the operator norm of \mathbf{Y} . To this end, we use the Matrix Bernstein inequality, which in turn requires a

bound on the operator norm of \mathbf{X}_i and variance of \mathbf{Y} . Recall, for some $i' \in [n]$

$$\begin{aligned}
\|\mathbf{X}_i\|_2 &= \left\| \frac{1}{tp_{i'}} (\mathbf{AT})_{i'}^T (\mathbf{AT})_{i'} - \frac{1}{t} (\mathbf{AT})^T (\mathbf{AT}) \right\|_2 \\
&\leq \frac{1}{tp_{i'}} \|(\mathbf{AT})_{i'}^T (\mathbf{AT})_{i'}\|_2 + \frac{1}{t} \|(\mathbf{AT})^T (\mathbf{AT})\|_2 \\
&= \frac{\|(\mathbf{AT})_{i'}\|_2^2}{tp_{i'}} + \frac{\|(\mathbf{AT})\|_2^2}{t} \\
&\leq \frac{\log(n)}{tp_{i'}} \max \left\{ \|\mathbf{A}_{i',*}\|_2^2, \frac{\phi_{\max} \sqrt{n} \|\mathbf{A}\|_F \mathbf{A}_{i,i}}{t} \right\} + \frac{\|(\mathbf{AT})\|_2^2}{t} \quad [\text{by Lemma 5.8}] \\
&\leq \frac{\log(n)}{t} \max \left\{ \phi_{\max} \sqrt{n} \|\mathbf{A}\|_F \|\mathbf{A}\|_2, \frac{(\phi_{\max} \sqrt{n} \|\mathbf{A}\|_F)^2}{t}, \|(\mathbf{AT})\|_2^2 \right\} \quad [\text{by Lemma 5.5}] \\
&\leq \frac{\phi_{\max} \sqrt{n} \log(n) \|\mathbf{A}\|_F^2}{t} \left(1 + \frac{\sqrt{n}}{t} \right) \leq \frac{2\phi_{\max} \sqrt{n} \log(n) \|\mathbf{A}\|_F^2}{t}
\end{aligned} \tag{5.15}$$

where the last inequality uses that $t = \Omega(\sqrt{n})$. Next, we bound $\mathbf{Var}[\mathbf{Y}] \leq \mathbb{E}[\mathbf{Y}^2]$ as follows

$$\begin{aligned}
\mathbb{E}[\mathbf{Y}^2] &= t \left(\sum_{i \in [n]} \frac{p_i}{t^2 p_i^2} ((\mathbf{AT})_i^T (\mathbf{AT})_i)^2 + \frac{1}{t^2} ((\mathbf{AT})^T (\mathbf{AT}))^2 - \sum_{i \in [n]} \frac{2p_i}{p_i t^2} (\mathbf{AT})_i^T (\mathbf{AT})_i (\mathbf{AT})^T (\mathbf{AT}) \right) \\
&= \frac{1}{t} \left(\sum_{i \in [n]} \frac{((\mathbf{AT})_i^T (\mathbf{AT})_i)^2}{p_i} + ((\mathbf{AT})^T (\mathbf{AT}))^2 - \sum_{i \in [n]} 2(\mathbf{AT})_i^T (\mathbf{AT})_i (\mathbf{AT})^T (\mathbf{AT}) \right) \\
&\leq \frac{1}{t} \left(\sum_{i \in [n]} \frac{((\mathbf{AT})_i^T (\mathbf{AT})_i)^2}{p_i} \right) \\
&\leq \frac{\log(n)}{t} \max \left\{ \phi_{\max} \sqrt{n} \|\mathbf{A}\|_F \|\mathbf{A}\|_2, \frac{(\phi_{\max} \sqrt{n} \|\mathbf{A}\|_F)^2}{t} \right\} \left(\sum_{i \in [n]} (\mathbf{AT})_i^T (\mathbf{AT})_i \right) \\
&\leq \frac{c \log(n) \sqrt{n} \|\mathbf{A}\|_F^2 \|(\mathbf{AT})\|_2^2}{t} \mathbf{I}_{n \times n}
\end{aligned} \tag{5.16}$$

where we again use Lemma 5.5 and Theorem 5.7. Observe,

Applying the Matrix Bernstein inequality with equations 5.15 and 5.16

$$\begin{aligned}
\Pr \left[\|\mathbf{Y}\|_2 \geq \epsilon \|\mathbf{A}\|_F^2 \right] &\leq 2n \exp \left(- \frac{\epsilon^2 \|\mathbf{A}\|_F^4}{\frac{c \log(n) \sqrt{n} \phi_{\max} \|\mathbf{A}\|_F^2 \|(\mathbf{AT})\|_2^2}{t} + \frac{\epsilon \sqrt{n} \log(n) \phi_{\max}}{3t} \|\mathbf{A}\|_F^4} \right) \\
&\leq 2n \exp \left(- \frac{\epsilon^2 t}{c' \phi_{\max} \sqrt{n} \log(n)} \right)
\end{aligned} \tag{5.17}$$

where the second inequality uses Theorem 5.6, to conclude that with probability at least $1 - \delta/2$ $\|\mathbf{AT}\|_2^2 \leq \|\mathbf{A}\|_2^2 + \epsilon/k \|\mathbf{A}\|_F^2 \leq O(\|\mathbf{A}\|_F^2)$. Therefore, it suffices to set $t = \frac{c' \phi_{\max} \sqrt{n} \log(n)}{\epsilon^2} \log(n/\delta)$, to bound the above probability by $\delta/2$. Union bounding over the error for both PCPs, and setting $\epsilon = \epsilon/k$, we can conclude that with probability at least $1 - \delta$,

$$(\mathbf{AT})^T(\mathbf{AT}) - \frac{\epsilon}{k} \|\mathbf{A}\|_F^2 \mathbf{I} \leq (\mathbf{SAT})^T(\mathbf{SAT}) \leq (\mathbf{AT})^T(\mathbf{AT}) + \frac{\epsilon}{k} \|\mathbf{A}\|_F^2 \mathbf{I}$$

when $t = \Omega\left(\frac{\phi_{\max} \sqrt{n} k^2}{\epsilon^2}\right)$.

□

We use the spectral bound from Theorem 5.9 to obtain a row projection-cost preservation guarantee. We follow the same proof strategy as Theorem 5.7, while requiring modified version of the scalar Chernoff bound. We do away with the head-tail split from [CMM17], [MW17] and [BW18] and analyze the projection-cost guarantee directly. This enables us to obtain a better ϵ dependence than [MW17] and [BW18]. Note, our ϵ dependence matches that of [CMM17] but our row projection cost preserving sketch can be computed in sub-linear time, albeit for PSD matrices.

Theorem 5.10. (Row Projection-Cost Preservation.) *Given as input $\mathbf{A} + \mathbf{N}$ let \mathbf{C} be an $n \times t$ matrix as defined in Theorem 5.7 such that $\mathbf{C} = \mathbf{AT} + \mathbf{NT}$. Let $p = \{p_1, p_2 \dots p_n\}$ be a probability distribution over the rows of \mathbf{C} such that $p_j = \frac{(\mathbf{A} + \mathbf{N})_{j,j}}{\text{Tr}(\mathbf{A} + \mathbf{N})}$. Let $t = O\left(\frac{\phi_{\max} \sqrt{n} k^2 \log^2(n)}{\epsilon^2}\right)$. Then, construct \mathbf{R} using t rows of \mathbf{C} and set each one to $\frac{\mathbf{C}_{i,*}}{\sqrt{tp_i}}$ with probability p_i . With probability at least $1 - c$, for any rank- k orthogonal projection \mathbf{X} ,*

$$\|\mathbf{R} - \mathbf{RX}\|_F^2 = \|\mathbf{C} - \mathbf{CX}\|_F^2 \pm O(\epsilon + \sqrt{\eta}) \|\mathbf{A}\|_F^2$$

for a fixed constant c .

Proof. Note, $\mathbf{R} = \mathbf{SC} = \mathbf{SAT} + \mathbf{SNT}$, where \mathbf{S} and \mathbf{T} are the corresponding sampling matrices. Observe, $\|\mathbf{C} - \mathbf{CX}\|_F^2 = \text{Tr}((\mathbf{I} - \mathbf{X})\mathbf{C}^T\mathbf{C}(\mathbf{I} - \mathbf{X}))$. Then,

$$\begin{aligned} \text{Tr}((\mathbf{I} - \mathbf{X})\mathbf{C}^T\mathbf{C}(\mathbf{I} - \mathbf{X})) &= \text{Tr}(\mathbf{C}^T\mathbf{C}) + \text{Tr}(\mathbf{XC}^T\mathbf{CX}) - \text{Tr}(\mathbf{C}^T\mathbf{CX}) - \text{Tr}(\mathbf{XC}^T\mathbf{C}) \\ &= \text{Tr}(\mathbf{C}^T\mathbf{C}^T) + \text{Tr}(\mathbf{X}^T\mathbf{C}^T\mathbf{CX}) - \text{Tr}(\mathbf{C}^T\mathbf{CXX}) - \text{Tr}(\mathbf{XXC}^T\mathbf{C}) \\ &= \text{Tr}(\mathbf{C}^T\mathbf{C}) + \text{Tr}(\mathbf{XC}^T\mathbf{CX}) - \text{Tr}(\mathbf{XC}^T\mathbf{CX}) - \text{Tr}(\mathbf{XC}^T\mathbf{CX}) \\ &= \text{Tr}(\mathbf{C}^T\mathbf{C}) - \text{Tr}(\mathbf{XC}^T\mathbf{CX}) \\ &= \text{Tr}(\mathbf{C}^T\mathbf{C}) - \text{Tr}(\mathbf{X}(\mathbf{AT} + \mathbf{NT})^T(\mathbf{AT} + \mathbf{NT})\mathbf{X}) \end{aligned} \quad (5.18)$$

where we used the fact that for any projection matrix $\mathbf{X} = \mathbf{X}^2$ in addition to the cyclic property of the trace. Here, for analyzing the cross and tail terms, we observe that with probability $1 - c$, $\|\mathbf{X}(\mathbf{AT})^T\|_F \leq O(1)\|\mathbf{A}\|_F$ and $\|\mathbf{X}(\mathbf{NT})^T\|_F^2 \leq O(\eta)\|\mathbf{A}\|_F^2$. Therefore,

$$\text{Tr}(\mathbf{X}(\mathbf{AT} + \mathbf{NT})^T(\mathbf{AT} + \mathbf{NT})\mathbf{X}) = \text{Tr}(\mathbf{X}(\mathbf{AT})^T\mathbf{ATX}) \pm O(\sqrt{\eta})\|\mathbf{A}\|_F^2 \quad (5.19)$$

Similarly,

$$\begin{aligned} \|\mathbf{R} - \mathbf{RX}\|_F^2 &= \text{Tr}((\mathbf{I} - \mathbf{X})\mathbf{R}^T\mathbf{R}(\mathbf{I} - \mathbf{X})) = \text{Tr}(\mathbf{R}^T\mathbf{R}) - \text{Tr}(\mathbf{XR}^T\mathbf{RX}) \\ &= \text{Tr}(\mathbf{R}^T\mathbf{R}) - \text{Tr}(\mathbf{X}(\mathbf{SAT})^T(\mathbf{SAT})\mathbf{X}) \pm O(\sqrt{\eta})\|\mathbf{A}\|_F^2 \end{aligned} \quad (5.20)$$

Here, we observe $\|\mathbf{SAT}\|_F^2$ is an unbiased estimator for $\|\mathbf{A}\|_F^2$ and $\|\mathbf{SNT}\|_F^2$ is an unbiased estimator for $\|\mathbf{N}\|_F^2$. Using the same idea as above, we can bound the cross and tail terms by $O(\sqrt{\eta})\|\mathbf{A}\|_F^2$. Our goal is show that Equations 5.18 and 5.20 are related up to additive error $O(\epsilon + \sqrt{\eta})\|\mathbf{A}\|_F^2$. We first relate $\text{Tr}(\mathbf{C}^T \mathbf{C})$ and $\text{Tr}(\mathbf{R}^T \mathbf{R})$. Recall, $\mathbb{E}[\text{Tr}(\mathbf{R}^T \mathbf{R})] = \mathbb{E}[\|\mathbf{R}\|_F^2] = \|\mathbf{C}\|_F^2 = \text{Tr}(\mathbf{C}^T \mathbf{C})$. Using a scalar Chernoff bound, we show that with probability at least $1 - 1/\text{poly}(n)$, $|\|\mathbf{R}\|_F^2 - \|\mathbf{C}\|_F^2| \leq \epsilon \|\mathbf{A}\|_F^2$. Observe, for all $i \in [t]$, $\mathbf{R}_{*,i} = \frac{1}{\sqrt{p_{i'}t}} \mathbf{C}_{i',*}$ for some $i' \in [n]$. Then,

$$\begin{aligned} \|\mathbf{R}_{i,*}\|_2^2 &= \frac{1}{p_{i'}t} \|\mathbf{C}_{i',*}\|_2^2 = \frac{\phi_{\max} \sqrt{n} \|\mathbf{A}\|_F \epsilon^2}{\phi_{\max} \sqrt{nk} \log(n) \log(n/\delta) \mathbf{A}_{i',i'}} \|\mathbf{C}_{i',*}\|_2^2 \\ &\leq \frac{c(1+\eta) \|\mathbf{A}\|_F \epsilon^2}{\sqrt{k} \log(n/\delta) \mathbf{A}_{i',i'}} \max \left\{ \|\mathbf{A}_{i',*}\|_2^2, \frac{\phi_{\max} \sqrt{n} \|\mathbf{A}\|_F \mathbf{A}_{i',i'}}{t} \right\} \\ &\leq \frac{c\epsilon^2}{\sqrt{k} \log(n/\delta)} \max \left\{ \|\mathbf{A}\|_2 \|\mathbf{A}\|_F, \frac{\|\mathbf{A}\|_F^2 \epsilon^2}{\sqrt{k} \log(n/\delta)} \right\} \\ &\leq \frac{c\epsilon^2}{\sqrt{k} \log(n/\delta)} \|\mathbf{A} + \mathbf{N}\|_F^2 \end{aligned} \quad (5.21)$$

where we use $\mathbf{C}_{i',*} = (\mathbf{AT})_{i',*} + (\mathbf{NT})_{i',*}$, $\|(\mathbf{NT})_{i,*}\|_2^2 \leq (\eta) \|(\mathbf{AT})_{i,*}\|_2^2$ for all i and Lemma 5.8 to bound $\|(\mathbf{AT})_{i,*}\|_2^2$. Therefore, $\frac{\sqrt{k} \log(n/\delta)}{c\epsilon^2 \|\mathbf{A}\|_F^2} \|\mathbf{R}_{i,*}\|_2^2 \in [0, 1]$. Note, $\|\mathbf{R}\|_F^2$ is an unbiased estimator for $\|\mathbf{A} + \mathbf{N}\|_F^2$. Using a Chernoff bound,

$$\begin{aligned} \Pr \left[\|\mathbf{R}\|_F^2 \geq (1+\epsilon) \|\mathbf{A} + \mathbf{N}\|_F^2 \right] &= \Pr \left[\frac{\sqrt{k} \log(n/\delta)}{\epsilon^2 \|\mathbf{A}\|_F^2} \|\mathbf{R}\|_F^2 \geq \frac{\sqrt{k} \log(n/\delta)}{\epsilon^2} (1+\epsilon) \right] \\ &\leq \exp \left(-\frac{\sqrt{k} \epsilon^2 \log(n/\delta)}{\epsilon^2} \right) \leq \frac{\delta}{10} \end{aligned} \quad (5.22)$$

Therefore, with probability at least $1 - \delta/10$, $|\|\mathbf{R}\|_F^2 - \|\mathbf{A} + \mathbf{N}\|_F^2| \leq \epsilon \|\mathbf{A} + \mathbf{N}\|_F^2$. Note, we can then bound $\|\mathbf{A} + \mathbf{N}\|_F^2 \leq \|\mathbf{A}\|_F^2 + 2\sqrt{\eta} \|\mathbf{A}\|_F^2$. Therefore,

$$|\|\mathbf{R}\|_F^2 - \|\mathbf{A}\|_F^2| \leq O(\epsilon + \sqrt{\eta}) \|\mathbf{A}\|_F^2$$

Recall, by equation 5.10, with probability $\delta/10$, $\|\mathbf{A}\|_F^2 = (1 \pm (\epsilon + 2\sqrt{\eta})) \|\mathbf{C}\|_F^2$ and thus we have that $\|\mathbf{R}\|_F^2 - \|\mathbf{C}\|_F^2 \leq 3\epsilon \|\mathbf{A}\|_F^2$. We can repeat the above argument to lower bound $\|\mathbf{R}\|_F^2$. Therefore, with probability $1 - \delta$, we have

$$|\|\mathbf{R}\|_F^2 - \|\mathbf{C}\|_F^2| \leq O(\epsilon + \sqrt{\eta}) \|\mathbf{A}\|_F^2 \quad (5.23)$$

Next, we relate $\text{Tr}(\mathbf{X}(\mathbf{SAT})^T(\mathbf{SAT})\mathbf{X})$ and $\text{Tr}(\mathbf{X}(\mathbf{AT})^T\mathbf{ATX})$. Observe, \mathbf{X} is a rank k projection matrix and we can represent it as \mathbf{ZZ}^T , where $\mathbf{Z} \in \mathbb{R}^{n \times k}$ and has orthonormal columns. By the cyclic property of the trace, we have

$$\text{Tr}(\mathbf{ZZ}^T(\mathbf{SAT})^T(\mathbf{SAT})\mathbf{ZZ}^T) = \text{Tr}(\mathbf{Z}^T(\mathbf{SAT})^T(\mathbf{SAT})\mathbf{Z}) = \sum_{j \in [k]} \mathbf{Z}_{*,j}^T (\mathbf{SAT})^T (\mathbf{SAT}) \mathbf{Z}_{*,j}$$

Similarly, $\text{Tr}(\mathbf{Z}\mathbf{Z}^T(\mathbf{A}\mathbf{T})^T\mathbf{A}\mathbf{T}\mathbf{Z}\mathbf{Z}^T) = \sum_{j \in [k]} \mathbf{Z}_{*,j}^T(\mathbf{A}\mathbf{T})^T\mathbf{A}\mathbf{T}\mathbf{Z}_{*,j}$. By Theorem 5.9, we have

$$\sum_{j \in [k]} \left(\mathbf{Z}_{*,j}^T(\mathbf{A}\mathbf{T})^T\mathbf{A}\mathbf{T}\mathbf{Z}_{*,j} \right) = \sum_{j \in [k]} \left(\mathbf{Z}_{*,j}^T(\mathbf{S}\mathbf{A}\mathbf{T})^T(\mathbf{S}\mathbf{A}\mathbf{T})\mathbf{Z}_{*,j} \pm \frac{\epsilon}{k} \|\mathbf{A}\|_F^2 \mathbf{Z}_{*,j}^T \mathbf{I} \mathbf{Z}_{*,j} \right) \quad (5.24)$$

Since $\mathbf{Z}_{*,j}^T \mathbf{Z}_{*,j} = 1$ and $\text{Tr}(\mathbf{Z}^T(\mathbf{S}\mathbf{A}\mathbf{T})^T(\mathbf{S}\mathbf{A}\mathbf{T})\mathbf{Z}) = \text{Tr}(\mathbf{X}(\mathbf{A}\mathbf{T})^T\mathbf{A}\mathbf{T}\mathbf{X})$, we obtain

$$\text{Tr}(\mathbf{X}(\mathbf{A}\mathbf{T})^T\mathbf{A}\mathbf{T}\mathbf{X}) - \epsilon \|\mathbf{A}\|_F^2 \leq \text{Tr}(\mathbf{X}(\mathbf{S}\mathbf{A}\mathbf{T})^T(\mathbf{S}\mathbf{A}\mathbf{T})\mathbf{X}) \leq \text{Tr}(\mathbf{X}(\mathbf{A}\mathbf{T})^T\mathbf{A}\mathbf{T}\mathbf{X}) + \epsilon \|\mathbf{A}\|_F^2 \quad (5.25)$$

Combining equations 5.25, 5.23, 5.19 and 5.20 with probability $1 - c$,

$$\|\mathbf{C} - \mathbf{C}\mathbf{X}\|_F^2 - O(\epsilon + \sqrt{\eta})\|\mathbf{A}\|_F^2 \leq \|\mathbf{R} - \mathbf{R}\mathbf{X}\|_F^2 \leq \|\mathbf{C} - \mathbf{C}\mathbf{X}\|_F^2 + O(\epsilon + \sqrt{\eta})\|\mathbf{A}\|_F^2$$

□

Full Algorithm. Next, we describe a sublinear time and query robust algorithm for low-rank approximation of PSD matrices. We show that querying $\tilde{O}(\phi_{\max}^2 nk/\epsilon)$ entries of \mathbf{A} suffices. While we assume we know ϕ_{\max} , in practice this need not be the case. Therefore, given a budget for the total number of queries, denoted by β , we can run the algorithm by querying a $\sqrt{\beta} \times \sqrt{\beta}$ submatrix (as described in Algorithm 4), but correctness only holds when $\beta \geq \tilde{\Theta}(\phi_{\max}^2 nk/\epsilon)$. Recall, whenever we read an entry in $(\mathbf{A} + \mathbf{N})_{i,j}$, we can truncate it to $\phi_{\max} \sqrt{|(\mathbf{A} + \mathbf{N})_{i,i}| \cdot |(\mathbf{A} + \mathbf{N})_{j,j}|}$. We can compute these thresholds by simply reading the diagonal of $\mathbf{A} + \mathbf{N}$.

We proceed by constructing column and row projection-cost preserving sketches of $\mathbf{A} + \mathbf{N}$, to obtain a $t \times t$ matrix \mathbf{R} , where $t = \tilde{O}(\phi_{\max} \sqrt{nk^2/\epsilon^2})$. Instead of reading the entire matrix, we sample $\epsilon^3 t/k^3$ entries in each row of \mathbf{R} , and read these entries. Ideally we would want to estimate ℓ_2^2 norms of rows of \mathbf{R} to then use a result of Frieze-Kannan-Vempala [FKV04] to show that there exists an $s \times t$ matrix \mathbf{S} such that the row space of \mathbf{S} contains a good rank- k approximation, where $s = c\phi_{\max}^2 nk/\epsilon t$, for some constant c . However, we show that it is not possible to obtain accurate estimates of the row norms of each row of \mathbf{R} with high probability.

Instead, we describe a new sampling procedure that ends up sampling rows of \mathbf{R} with the same probability as Frieze-Kannan-Vempala. Once we compute a good low-rank approximation for \mathbf{R} we can follow the approach of [CMM17], [MW17] and [BW18], where we set up two regression problems, and use fast approximate regression to compute a low rank approximation for \mathbf{A} . The main theorem we prove in this section is as follows:

Theorem 5.11. (Robust PSD LRA.) Let k be an integer and $\epsilon > \eta > 0$. Given a matrix $\mathbf{A} + \mathbf{N}$, where \mathbf{A} is PSD and \mathbf{N} is a corruption term such that $\|\mathbf{N}\|_F^2 \leq \sqrt{\eta} \|\mathbf{A}\|_F^2$ and for all $i \in [n]$ $\|\mathbf{N}_{i,*}\|_2^2 \leq c \|\mathbf{A}_{i,*}\|_2^2$, for a fixed constant c , Algorithm 4 samples $\tilde{O}(\phi_{\max}^2 nk/\epsilon)$ entries in $\mathbf{A} + \mathbf{N}$ and computes matrices $\mathbf{M}, \mathbf{N}^T \in \mathbb{R}^{n \times k}$ such that with probability at least $99/100$,

$$\|\mathbf{A} - \mathbf{M}\mathbf{N}^T\|_F^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + (\epsilon + \sqrt{\eta})\|\mathbf{A}\|_F^2$$

Algorithm 4 : Robust PSD Low-Rank Approximation

Input: A Matrix $\mathbf{A} + \mathbf{N}$, integer $k, \epsilon > 0$ and $\phi_{\max} = \max_j \mathbf{A}_{j,j}/(\mathbf{A} + \mathbf{N})_{j,j}$

1. Let $t = \frac{c\phi_{\max}^2 \sqrt{nk} \log^2(n)}{\epsilon^2}$, for some constant c . Let $q = \{q_1, q_2 \dots q_n\}$ denote a distribution over columns of $\mathbf{A} + \mathbf{N}$ such that $q_j = \frac{(\mathbf{A} + \mathbf{N})_{j,j}}{\text{Tr}(\mathbf{A} + \mathbf{N})}$. Construct a column PCP for $\mathbf{A} + \mathbf{N}$ by sampling t columns of $\mathbf{A} + \mathbf{N}$ such that each column is set to $\frac{(\mathbf{A} + \mathbf{N})_{*,j}}{\sqrt{tq_j}}$ with probability q_j . Let \mathbf{C} be the resulting $n \times t$ matrix that satisfies the guarantee of Theorem 5.7.
2. Let $p = \{p_1, p_2 \dots p_n\}$ denote a distribution over rows of \mathbf{C} such that $p_i = \frac{(\mathbf{A} + \mathbf{N})_{i,i}}{\text{Tr}(\mathbf{A} + \mathbf{N})}$. Construct a row PCP for \mathbf{C} by sampling t rows of \mathbf{C} such that each row is set to $\frac{\mathbf{C}_{i,*}}{\sqrt{tp_i}}$ with probability p_i . Let \mathbf{R} be the resulting $t \times t$ matrix that satisfies the guarantee of Theorem 5.10. Sample $\Theta(n)$ entries uniformly at random from \mathbf{A} and rescale such that $\tilde{v}^2 = \Theta(\|\mathbf{A}\|_F^2)$.
3. Let $\mu = \phi_{\max} \sqrt{[(\mathbf{A} + \mathbf{N})_{i,i}] \cdot [(\mathbf{A} + \mathbf{N})_{i',i'}]}$. For all $i \in [t]$, let $\mathcal{X}_i = \sum_{j \in [\epsilon^3 t / k^3]} \mathcal{X}_{i,j}$ such that $\mathcal{X}_{i,j} = (k^3 / \epsilon^3) \mathbf{R}_{i,i'}^2$, with probability $1/t$, for all $i' \in [t]$. Here, we query the entry corresponding to $\mathbf{R}_{i,i'}$ in $\mathbf{A} + \mathbf{N}$ and truncate it to μ . Let $\tau = \phi_{\max}^2 n \tilde{v}^2 / t^2$. If $\mathcal{X}_i > \tau$, sample row $\mathbf{R}_{i,*}$ with probability 1. For the remaining rows, sample $nk/(\epsilon t)$ rows uniformly at random.
4. Run the sampling algorithm from Frieze-Kannan-Vempala [FKV04] to compute a $t \times k$ matrix \mathbf{S} such that $\|\mathbf{R} - \mathbf{R}\mathbf{S}\mathbf{S}^T\|_F^2 \leq \|\mathbf{R} - \mathbf{R}_k\|_F^2 + \epsilon \|\mathbf{R}\|_F^2$. Consider the regression problem $\min_{\mathbf{X} \in \mathbb{R}^{n \times k}} \|\mathbf{C} - \mathbf{X}\mathbf{S}^T\|_F^2$. Sketch the problem using the leverage scores of \mathbf{S}^T , as shown in Lemma 4.4, to obtain a sampling matrix \mathbf{E} with $O(\frac{k}{\epsilon})$ columns. Compute $\mathbf{X}_\mathbf{C} = \arg\min_{\mathbf{X} \in \mathbb{R}^{n \times k}} \|\mathbf{C}\mathbf{E} - \mathbf{X}\mathbf{S}^T\mathbf{E}\|_F^2$. Let $\mathbf{X}_\mathbf{C}\mathbf{S}^T = \mathbf{U}\mathbf{V}^T$ be such that $\mathbf{U} \in \mathbb{R}^{n \times k}$ has orthonormal columns.
5. Consider the regression problem $\min_{\mathbf{X} \in \mathbb{R}^{k \times n}} \|\mathbf{A} - \mathbf{U}\mathbf{X}\|_F^2$. Sketch the problem as above, following Lemma 4.4 to obtain a sampling matrix \mathbf{E}' with $O(\frac{k}{\epsilon})$ rows. Compute $\mathbf{X}_\mathbf{A} = \arg\min_{\mathbf{X}} \|\mathbf{E}'\mathbf{A} - \mathbf{E}'\mathbf{U}\mathbf{X}\|_F^2$.

Output: $\mathbf{M} = \mathbf{U}, \mathbf{N}^T = \mathbf{X}_\mathbf{A}$

We begin with the following simple lemma for approximating the Frobenius norm :

Lemma 5.12. (Approximating Frobenius Norm.) *Given as input an $n \times n$ matrix $\mathbf{A} + \mathbf{N}$, there exists an algorithm that reads $O(\phi_{\max}^2 n)$ entries in \mathbf{A} and outputs an estimator \tilde{v} such that with probability at least $1 - \frac{1}{n^c}$, $\tilde{v} = \Theta(\|\mathbf{A}\|_F^2)$.*

Proof. There are multiple ways to see this. Observe, in Theorem 5.10, we show that sampling $\frac{\phi_{\max}^2 n \log(n)}{\epsilon^2}$ entries results in row projection-cost preserving sketch \mathbf{R} such that $\|\mathbf{R}\|_F^2 = (1 \pm \epsilon) \|\mathbf{A} + \mathbf{N}\|_F^2$. Setting ϵ to be a small constant suffices. \square

Next, we provide intuition for why uniformly sampling columns of \mathbf{R} does not suffice for obtaining a sketch that spans a good low rank approximation. For simplicity, we assume there is no noise ($\eta = 0$ and $\phi_{\max} = 1$) and show that our techniques to bound the column norms of \mathbf{R} results in an estimate that is too large. We note that this lemma is not required for proving our result, and is just present for intuition.

Lemma 5.13. *Let $\eta = 0$. Let $\mathbf{R} \in \mathbb{R}^{t \times t}$ be a row projection-cost preserving sketch for \mathbf{C} as described in Theorem 5.10. For all $j \in [t]$, with probability at least $1 - 1/n^c$,*

$$\|\mathbf{R}_{*,j}\|_2^2 \leq O\left(\log(n) \max\left\{\|\mathbf{C}_{*,j}\|_2^2, \frac{n\|\mathbf{A}\|_F^2}{t^2}\right\}\right) = O\left(\log(n) \max\left\{\frac{\sqrt{n}\|\mathbf{A}\|_F^2}{t}, \frac{n\|\mathbf{A}\|_F^2}{t^2}\right\}\right)$$

where c is a fixed constant.

Proof. Observe, $\|\mathbf{R}_{*,j}\|_2^2 = \sum_{i \in [t]} \mathbf{R}_{i,j}^2$, where $\mathbf{R}_{i,j}^2 = \frac{\text{Tr}(\mathbf{A})}{t\mathbf{A}_{i',i'}} \mathbf{C}_{i',j}^2$ with probability $\frac{\mathbf{A}_{i',i'}}{\text{Tr}(\mathbf{A})}$ for all $i' \in [n]$. Then, $\mathbb{E}[\|\mathbf{R}_{*,j}\|_2^2] = \sum_{i=1}^n \mathbf{C}_{i,j}^2 = \|\mathbf{C}_{*,j}\|_2^2$. Next, we compute the variance of $\|\mathbf{R}_{*,j}\|_2^2$. $\text{Var}[\|\mathbf{R}_{*,j}\|_2^2] = t \text{Var}[\mathbf{R}_{i,j}^2] \leq t \mathbb{E}[\mathbf{R}_{i,j}^4]$. Then,

$$\begin{aligned} t \mathbb{E}[\mathbf{R}_{i,j}^4] &= \sum_{i' \in [n]} \frac{1}{tp_{i'}} \mathbf{C}_{i',j}^4 \leq \sum_{i' \in [n]} \frac{\text{Tr}(\mathbf{A})^2}{t^2 \mathbf{A}_{i',i'} \mathbf{A}_{j,j}} \mathbf{A}_{i',j}^2 \mathbf{A}_{i',i'} \mathbf{A}_{j,j} \\ &\leq \frac{\text{Tr}(\mathbf{A})^2}{t^2} \|\mathbf{A}_{*,j}\|_2^2 \\ &= \frac{\text{Tr}(\mathbf{A})}{t} \|\mathbf{C}_{*,j}\|_2^2 \\ &\leq \left(\frac{\text{Tr}(\mathbf{A})}{t}\right)^2 + \|\mathbf{C}_{*,j}\|_2^4 \quad [\text{AM-GM}] \end{aligned}$$

where we use $\mathbf{A}_{i,j}^2 \leq \mathbf{A}_{i,i} \mathbf{A}_{j,j}$, which follows from applying Cauchy-Schwarz to $\langle \mathbf{A}_{i,*}^{1/2}, \mathbf{A}_{j,*}^{1/2} \rangle$. Similarly, we bound $\mathbf{R}_{i,j}^2 \leq \frac{\text{Tr}(\mathbf{A})}{t}$. Applying Bernstein's inequality,

$$\begin{aligned} &\Pr\left[\left|\|\mathbf{R}_{*,j}\|_2^2 - \|\mathbf{C}_{*,j}\|_2^2\right| \geq \eta \max\left\{\|\mathbf{C}_{*,j}\|_2^2, \frac{\text{Tr}(\mathbf{A})}{t}\right\}\right] \\ &\leq 2 \exp\left(-\frac{\eta^2 \max\left\{\|\mathbf{C}_{*,j}\|_2^4, \left(\frac{\text{Tr}(\mathbf{A})}{t}\right)^2\right\}}{\left(\frac{\text{Tr}(\mathbf{A})}{t}\right)^2 + \|\mathbf{C}_{*,j}\|_2^4 + \eta \max\left\{\frac{\text{Tr}(\mathbf{A})\|\mathbf{C}_{*,j}\|_2^2}{t}, \left(\frac{\text{Tr}(\mathbf{A})}{t}\right)^2\right\}}\right) \\ &\leq 2 \exp\left(-\frac{\eta \max\left\{\|\mathbf{C}_{*,j}\|_2^4, \left(\frac{\text{Tr}(\mathbf{A})}{t}\right)^2\right\}}{c' \left(\frac{\text{Tr}(\mathbf{A})}{t}\right)^2 + c' \|\mathbf{C}_{*,j}\|_2^4}\right) \end{aligned}$$

where we use the AM-GM inequality on $\frac{\text{Tr}(\mathbf{A})}{t} \|\mathbf{C}_{*,j}\|_2^2$ repeatedly. Setting $\eta = \Omega(\log(n))$ completes the proof. Finally, observe, for any $j \in [t]$, $\|\mathbf{C}_{*,j}\|_2^2 = \frac{\text{Tr}(\mathbf{A})}{t\mathbf{A}_{j',j'}} \|\mathbf{A}_{*,j'}\|_2^2$ for some $j' \in [n]$. We then use $\text{Tr}(\mathbf{A}) \leq \sqrt{n}\|\mathbf{A}\|_F$. \square

It is well-known that to recover a low-rank approximation for \mathbf{R} , one can sample rows of \mathbf{R} proportional to row norm estimates, denoted by \mathbf{Y}_i [FKV04]. As shown in [IVWW19] the following two conditions are a relaxation of those required in [FKV04], and suffice to obtain an additive error low-rank approximation :

1. For all $i \in [t]$, $\mathcal{Y}_i \geq \|\mathbf{R}_{i,*}\|_2^2$.
2. $\sum_{i \in [t]} \mathcal{Y}_i \leq \frac{n}{t} \|\mathbf{R}\|_F^2$

To satisfy the first condition, we need to obtain overestimates for each row. Since it is not immediately clear how to obtain overestimates for row norms of \mathbf{R} , a naïve approach would be to bias the estimate for each row by an upper bound on the row norm. However, by Lemma 5.13, a row norm could be as large as $\sqrt{n} \|\mathbf{A}\|_F^2 / t$. Observe, we cannot afford to bias the estimator of each row, \mathcal{Y}_i , by this amount since $\sum_{i \in [t]} \mathcal{Y}_i \geq \sqrt{n} \|\mathbf{A}\|_F^2 \geq \sqrt{n} \|\mathbf{R}\|_F^2$. Therefore, we would have to sample $\sqrt{n}k/\epsilon$ rows of \mathbf{R} , resulting in us querying $\Omega(nk^3/\epsilon^3)$ entries in \mathbf{A} , even when $\eta = 0$.

An alternative strategy would be to bias the estimator for each row by $n \|\mathbf{R}\|_F^2 / t^2$, as this would satisfy condition 2 above. We can now hope to detect rows of \mathbf{R} that have norm larger than $n \|\mathbf{R}\|_F^2 / t^2$ by sampling $\epsilon^3 t / k^3$ entries in each row of \mathbf{R} , uniformly at random. Note, this way we can construct an unbiased estimator for the ℓ_2^2 norm of each row. Ideally, we would want to show a high probability statement for concentration of our row norm estimates around the expectation. We could then union bound, and obtain concentration for all i simultaneously.

However, this is not possible since it may be the case that a row of \mathbf{R} is $\log(n)$ -sparse with each entry being large in magnitude. In this case, uniformly querying the row would not observe any non-zero with good probability and thus cannot distinguish between such a row and an empty row. Instead, we settle for a weaker statement, that shows our estimate is accurate with $o(1)$ probability. All subsequent statements hold for $\eta > 0$.

Lemma 5.14. (Estimating large row norms.) Let $\mathbf{R} \in \mathbb{R}^{t \times t}$ be the row PCP output by Step 2 of Algorithm 4. For all $i \in [t]$ let $\mathcal{X}_i = \sum_{j \in [\epsilon^3 t / k^3]} \mathcal{X}_{i,j}$ such that $\mathcal{X}_{i,j} = \frac{k^3 \mathbf{R}_{i,j'}^2}{\epsilon^3}$ with probability $\frac{1}{t}$, for all $j' \in [t]$. Then, for all $i \in [t]$, $\mathcal{X}_i = (1 \pm \frac{1}{10}) \|\mathbf{R}_{i,*}\|_2^2$ with probability at least $\frac{\|\mathbf{R}_{i,*}\|_2^2 k}{\epsilon n}$.

Proof. Observe, \mathcal{X}_i is an unbiased estimator of $\|\mathbf{R}_{i,*}\|_2^2$:

$$\mathbb{E}[\mathcal{X}_i] = \frac{\epsilon^3 t}{k^3} \mathbb{E}[\mathcal{X}_{i,j}] = \frac{\epsilon^3 t}{k^3} \sum_{j' \in [t]} \frac{k^3}{\epsilon^3 n} \mathbf{R}_{i,j'}^2 = \|\mathbf{R}_{i,*}\|_2^2$$

Next, we compute the variance of \mathcal{X}_i .

$$\begin{aligned}
\text{Var}[\mathcal{X}_i] &= \frac{\epsilon^3 t}{k^3} \text{Var}[\mathcal{X}_{i,j}] \leq \sum_{j \in [t]} \frac{1}{\epsilon^3} \mathbf{R}_{i,j}^4 \\
&\leq \sum_{j \in [t]} \frac{k^3}{\epsilon^3} \mathbf{R}_{i,j}^2 \left(\frac{\text{Tr}(\mathbf{A} + \mathbf{N})^2}{t^2 (\mathbf{A} + \mathbf{N})_{i,i} (\mathbf{A} + \mathbf{N})_{j,j}} (\mathbf{A} + \mathbf{N})_{i,j}^2 \right) \\
&\leq \sum_{j \in [t]} \frac{k^3}{\epsilon^3} \mathbf{R}_{i,j}^2 \left(\frac{\text{Tr}(\mathbf{A} + \mathbf{N})^2}{t^2 (\mathbf{A} + \mathbf{N})_{i,i} (\mathbf{A} + \mathbf{N})_{j,j}} (\mathbf{A}_{i,i} \mathbf{A}_{j,j} + \mathbf{N}_{i,j}^2) \right) \\
&\leq \sum_{j \in [t]} \frac{k^3}{\epsilon^3} \mathbf{R}_{i,j}^2 \left(\frac{\text{Tr}(\mathbf{A} + \mathbf{N})^2 \phi_{\max}^2}{t^2} + \frac{\text{Tr}(\mathbf{A} + \mathbf{N})^2 \phi_{\max}^2 (\mathbf{A} + \mathbf{N})_{i,i} (\mathbf{A} + \mathbf{N})_{j,j}}{t^2 (\mathbf{A} + \mathbf{N})_{i,i} (\mathbf{A} + \mathbf{N})_{j,j}} \right) \\
&\leq \sum_{j \in [t]} \frac{k^3}{\epsilon^3} \mathbf{R}_{i,j}^2 \left(\frac{\text{Tr}(\mathbf{A} + \mathbf{N})^2 \phi_{\max}^2}{t^2} \right) \leq O \left(\frac{\epsilon \|\mathbf{A}\|_F^2}{k} \|\mathbf{R}_{i,*}\|_2^2 \right)
\end{aligned} \tag{5.26}$$

Here, we use that $\mathbf{N}_{i,j}^2 \leq \phi_{\max}^2 (\mathbf{A} + \mathbf{N})_{i,i} (\mathbf{A} + \mathbf{N})_{j,j}$, which follows from our truncation procedure. Further, using $t = \phi_{\max} \sqrt{n} k^2 / \epsilon^2$ and $\text{Tr}(\mathbf{A} + \mathbf{N}) \leq \sqrt{n} \|\mathbf{A}\|_F + \sqrt{\eta n} \|\mathbf{A}\|_F$, we can bound

$$\frac{\text{Tr}(\mathbf{A} + \mathbf{N})^2 \phi_{\max}^2}{t^2} \leq O \left(\frac{\epsilon \|\mathbf{A}\|_F^2}{k} \right)$$

Further, using the same argument as above

$$\mathcal{X}_{i,j} \leq \frac{k^3}{\epsilon^3} \mathbf{R}_{i,j}^2 \leq O \left(\frac{\epsilon \|\mathbf{A}\|_F^2}{k} \right)$$

Using Equations 5.26 and 5.27 in Bernstein's inequality,

$$\begin{aligned}
\Pr[|\mathcal{X}_i - \mathbb{E}[\mathcal{X}_i]| \geq \delta \mathbb{E}[\mathcal{X}_i]] &\leq \exp \left(- \frac{\delta^2 \|\mathbf{R}_{i,*}\|_2^4}{\frac{\epsilon \|\mathbf{A}\|_F^2}{k} \|\mathbf{R}_{i,*}\|_2^2 + \frac{\delta \epsilon \|\mathbf{A}\|_F^2}{3k} \|\mathbf{R}_{i,*}\|_2^2} \right) \\
&\leq \exp \left(- \frac{\delta^2 \|\mathbf{R}_{i,*}\|_2^2 k \log^2(n)}{\epsilon \|\mathbf{R}\|_F^2} \right)
\end{aligned}$$

where we use that $\|\mathbf{A}\|_F^2 = \Theta(\|\mathbf{R}\|_F^2)$. Setting $\eta = \frac{1}{10}$, $\mathcal{X}_i = (1 \pm \frac{1}{10}) \|\mathbf{R}_{i,*}\|_2^2$ with probability at least $1 - \exp \left(- \frac{\|\mathbf{R}_{i,*}\|_2^2 k \log^2(n)}{\epsilon \|\mathbf{R}\|_F^2} \right)$. Let ξ_i be the event that $\mathcal{X}_i = (1 \pm \frac{1}{10}) \|\mathbf{R}_{i,*}\|_2^2$. Then, union bounding over $t \leq n$ such events ξ_i , simultaneously for all i , ξ_i is true with probability at least

$$1 - \exp \left(- \frac{\|\mathbf{R}_{i,*}\|_2^2 k \log(n)}{\epsilon \|\mathbf{R}\|_F^2} \right) \geq \frac{\|\mathbf{R}_{i,*}\|_2^2 k \log(n)}{\epsilon \|\mathbf{R}\|_F^2}$$

□

We now have two major challenges: first, the probability with which the estimators are accurate is too small to even detect all rows with norm larger than $\phi_{\max}^2 n \|\mathbf{R}\|_F^2 / t^2$, and second, there is no small query certificate for when an estimator is accurate in estimating the row norms. Therefore, we cannot even identify the rows where we obtain an accurate estimate of their norm.

To address the first issue, we make the crucial observation that while we cannot estimate the norm of each row accurately, we can hope to sample the row with the same probability as Frieze-Kannan-Vempala [FKV04]. Recall, their algorithm samples row $\mathbf{R}_{i,*}$ with probability at least $\|\mathbf{R}_{i,*}\|_2^2 / \|\mathbf{R}\|_F^2$, which matches the probability in Lemma 5.14. Therefore, we can focus on designing a weaker notion of identifiability, that may potentially include extra rows.

We begin by partitioning rows of \mathbf{R} into two sets. Let $\mathcal{H} = \{i \mid \|\mathbf{R}_{i,*}\|_2^2 \geq \phi_{\max}^2 n \bar{v}^2 / t^2\}$ be the set of heavy rows and $[t] \setminus \mathcal{H}$ be the remaining rows. Since with probability at least $1 - \frac{1}{\text{poly}(n)}$, $\|\mathbf{R}\|_F^2 = \Theta(\|\mathbf{A}\|_F^2) = \Theta(\bar{v}^2)$,

$$|\mathcal{H}| = O(t^2 / \phi_{\max}^2 n) = O(k^4 \log^4(n) / \epsilon^4)$$

Observe, every row in \mathcal{H} can potentially satisfy the threshold $\tau = \phi_{\max}^2 n \bar{v}^2 / t^2$. Therefore, even if our estimator X_i is $\Theta(\|\mathbf{R}_{i,*}\|_2^2)$ for all $i \in \mathcal{H}$, we include at most $\tilde{O}(k^4 / \epsilon^4)$ extra rows in \mathbf{S} , which is well within our budget. Observe, we can then sample a row with probability 1 whenever the corresponding estimate is larger than τ . This sampling process ensures that we identify rows in \mathcal{H} with the right probability and also does not query more than $O(\phi_{\max}^2 n k / \epsilon)$ entries in $\mathbf{A} + \mathbf{N}$. For all the remaining rows, we know the norm is at most $O(\phi_{\max}^2 n \bar{v}^2 / t^2)$. We then modify the analysis of [FKV04] to show that we can handle both cases separately.

Theorem 5.15. (Existence [FKV04].) *Let \mathbf{R} be a row projection-cost preserving sketch output by Step 2 of Algorithm 4. For all $i \in [t]$, let X_i be estimate for $\|\mathbf{R}_{i,*}\|_2^2$ as described in Step 3 of Algorithm 4. Let \mathbf{S} be a subset of $s = O(\phi_{\max}^2 n k / \epsilon t)$ columns of \mathbf{R} sampled according to distribution $r = \{r_1, r_2, \dots, r_t\}$ such that r_i is the probability of sampling the i -th row. Then, with probability at least 99/100, there exists a $t \times k$ matrix \mathbf{U} in the column span of \mathbf{S} such that*

$$\|\mathbf{R} - \mathbf{U}\mathbf{U}^T \mathbf{R}\|_F^2 \leq \|\mathbf{R} - \mathbf{R}_k\|_F^2 + \epsilon \|\mathbf{R}\|_F^2$$

Proof. We follow the proof strategy of [FKV04] and show how to directly bound the variance in our setting as opposed to reducing to the two conditions above. Let $\mathbf{R} = \mathbf{P}\Sigma\mathbf{Q}^T = \sum_{\ell \in [t]} \Sigma_{\ell,\ell} \mathbf{P}_{\ell,*} \mathbf{Q}_{\ell,*}^T = \sum_{\ell \in [t]} \sigma_{\ell} \mathbf{P}_{\ell} \mathbf{Q}_{\ell}^T$. Recall, $\mathbf{R}_k = \sum_{\ell \in [k]} \mathbf{A} \mathbf{Q}_{\ell} \mathbf{Q}_{\ell}^T$. For $\ell \in [t]$, let $\mathbf{W}_{\ell} = \frac{1}{s} \sum_{i' \in [s]} \mathbf{Y}_{i'}$ where $\mathbf{Y}_{i'} = \frac{\mathbf{P}_{i',\ell}}{r_i} \mathbf{R}_{i,*}$ with probability r_i , for all $i \in [t]$. Then,

$$\mathbb{E}[\mathbf{W}_{\ell}] = \mathbb{E}[\mathbf{Y}_{i'}] = \sum_{i \in [t]} \frac{\mathbf{P}_{i,\ell}}{r_i} \mathbf{R}_{i,*} r_i = \sigma_{\ell} \mathbf{Q}_{\ell} \quad (5.28)$$

Therefore, in expectation the span of the rows contain a good low-rank solution. Next, we bound the variance. Recall, here we consider the rows in \mathcal{H} and its complement separately. From Lemma For all $i \in \mathcal{H}$, we know that $X_i = \Theta(\|\mathbf{R}_{i,*}\|_2^2)$ with probability at least $\|\mathbf{R}_{i,*}\|_2^2 k \log(n) / \epsilon \|\mathbf{R}\|_F^2$. Since for all such i , $\|\mathbf{R}_{i,*}\|_2^2 \geq \tau$, the corresponding $r_i \geq \frac{\|\mathbf{R}_{i,*}\|_2^2 k \log(n)}{\epsilon \|\mathbf{R}\|_F^2}$, since every time we pass the

threshold we sample the row. For all $i \notin \mathcal{H}$, $r_i \geq 1/t$ since there can be at most t such i , and we sample each such row with uniform probability. Once we have a lower bound on r_i in both cases, we open up the analysis of the variance bound in [FKV04] and show that our lower bounds suffice.

$$\begin{aligned}
\mathbb{E} [\|\mathbf{W}_\ell - \sigma_\ell \mathbf{Q}_\ell\|_2^2] &= \frac{1}{s} \left(\sum_{j \in [t]} \frac{\mathbf{P}_{i,\ell}^2}{r_i} \|\mathbf{R}_{i,*}\|_2^2 \right) - \frac{\sigma_\ell^2}{s} \leq \frac{1}{s} \left(\sum_{j \in \mathcal{H}} \frac{\mathbf{P}_{i,\ell}^2}{r_i} \|\mathbf{R}_{i,*}\|_2^2 + \sum_{i \in [t] \setminus \mathcal{H}} \frac{\mathbf{P}_{i,\ell}^2}{r_i} \|\mathbf{R}_{i,*}\|_2^2 \right) \\
&\leq \frac{1}{s} \left(\sum_{j \in \mathcal{H}} \frac{k \log(n) \mathbf{P}_{i,\ell}^2 \|\mathbf{R}\|_F^2}{\epsilon} + \sum_{j \in [t] \setminus \mathcal{H}} t \mathbf{P}_{i,\ell}^2 \|\mathbf{R}_{i,*}\|_2^2 \right) \\
&\leq \frac{1}{s} \left(\sum_{j \in \mathcal{H}} \frac{k \log(n) \mathbf{P}_{i,\ell}^2 \|\mathbf{R}\|_F^2}{\epsilon} + \sum_{j \in [t] \setminus \mathcal{H}} \frac{n}{t} \mathbf{P}_{i,\ell}^2 \tilde{v}^2 \right) \\
&\leq \frac{1}{s} \left(\frac{k \log(n)}{\epsilon} + \frac{\phi_{\max}^2 n}{t} \right) \|\mathbf{R}\|_F^2
\end{aligned} \tag{5.29}$$

Now, we can repeat the argument from [FKV04] and it suffices to set $s = \left(\frac{\phi_{\max}^2 n}{t} + \frac{k}{\epsilon} \right) \frac{k}{\epsilon} = O(\frac{\phi_{\max}^2 n k}{\epsilon t})$. For completeness, we present the rest of the proof here. For all $\ell \in [t]$, let $\mathbf{Y}_\ell = \frac{1}{\sigma_\ell} \mathbf{W}_\ell$. Let $\mathcal{V} = \text{span}(\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_k)$. Let $\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_t$ be an orthonormal basis for \mathcal{R}^t such that $\mathcal{V} = \text{span}(\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_{k'})$, where $k' = \dim(\mathcal{V})$. Let $\mathbf{S} = \mathbf{R} \sum_{\ell \in [k]} \mathbf{Z}_\ell \mathbf{Z}_\ell^T$ be the candidate low-rank approximation. Then,

$$\begin{aligned}
\|\mathbf{R} - \mathbf{S}\|_F^2 &= \sum_{\ell \in [t]} \|(\mathbf{R} - \mathbf{S})\mathbf{Z}_\ell\|_2^2 \\
&= \sum_{\ell \in [k'+1, t]} \|\mathbf{R}\mathbf{Z}_\ell\|_2^2 \\
&= \sum_{\ell \in [k'+1, t]} \left\| \left(\mathbf{R} - \mathbf{R} \sum_{\ell' \in [k]} \mathbf{Q}_{\ell'} \mathbf{Y}_{\ell'}^T \right) \mathbf{Z}_\ell \right\|_2^2 \\
&\leq \left\| \mathbf{R} - \mathbf{R} \sum_{\ell' \in [k]} \mathbf{Y}_{\ell'} \mathbf{Y}_{\ell'}^T \right\|_F^2
\end{aligned} \tag{5.30}$$

where the first equality follows from $\|\mathbf{Z}_\ell\|_2^2 = 1$, the second follows from $\mathbf{Z}_{\ell'}^T \mathbf{Z}_\ell = 0$ for $\ell' \neq \ell$, the third follows from $\langle \mathbf{Y}_{\ell'}, \mathbf{Z}_\ell \rangle = 0$ for all $\ell' \leq k$ and $\ell > k'$. Let $\hat{\mathbf{S}} = \mathbf{R} \sum_{\ell' \in [k]} \mathbf{Y}_{\ell'} \mathbf{Y}_{\ell'}^T$. Since $\mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_t$ forms an orthonormal basis

$$\begin{aligned}
\|\mathbf{R} - \hat{\mathbf{S}}\|_F^2 &\leq \sum_{\ell \in [t]} \left\| \mathbf{P}_\ell (\mathbf{R} - \hat{\mathbf{S}}) \right\|_2^2 \\
&= \sum_{\ell \in [k]} \|\sigma_\ell \mathbf{Q}_\ell - \mathbf{W}_\ell\|_2^2 + \sum_{\ell \in [k+1, t]} \sigma_\ell^2
\end{aligned} \tag{5.31}$$

Taking expectations on both sides of equation 5.31, we have

$$\begin{aligned}\mathbb{E} \left[\left\| \mathbf{R} - \widehat{\mathbf{S}} \right\|_F^2 \right] &\leq \mathbb{E} \left[\sum_{\ell \in [k]} \left\| \sigma_\ell \mathbf{Q}_\ell - \mathbf{W}_\ell \right\|_2^2 \right] + \left\| \mathbf{R} - \mathbf{R}_k \right\|_F^2 \\ &\leq \frac{k}{s} \left(\frac{k \log(n)}{\epsilon} + \frac{\phi_{\max}^2 n}{t} \right) \left\| \mathbf{R} \right\|_F^2 + \left\| \mathbf{R} - \mathbf{R}_k \right\|_F^2\end{aligned}\tag{5.32}$$

Since $\widehat{\mathbf{S}}$ is a rank k matrix and \mathbf{R}_k is the best rank k approximation to \mathbf{R} , $\left\| \mathbf{R} - \widehat{\mathbf{S}} \right\|_F^2 - \left\| \mathbf{R} - \mathbf{R}_k \right\|_F^2$ is a non-negative random variable. Thus, using Markov's inequality and Equation 5.30,

$$\Pr \left[\left\| \mathbf{R} - \mathbf{S} \right\|_F^2 - \left\| \mathbf{R} - \mathbf{R}_k \right\|_F^2 \geq \frac{100nk}{st} \left\| \mathbf{R} \right\|_F^2 \right] \leq \frac{1}{100}$$

Therefore, it suffices to sample $s = O\left(\frac{\phi_{\max}^2 nk}{\epsilon t}\right)$ columns, read all of them and compute a low rank approximation for \mathbf{R} with probability at least $\frac{99}{100}$. Observe, the total entries read by this algorithm is $O\left(\frac{\phi_{\max}^2 nk}{\epsilon t} \cdot t\right) = O\left(\frac{\phi_{\max}^2 nk}{\epsilon}\right)$. \square

It remains to show that we can now recover a low-rank approximation for \mathbf{A} , in factored form, from the low-rank approximation for \mathbf{R} . Here, we follow the approach of [CMM17],[MW17] and [BW18], where we set up two regression problems, and use the sketch and solve paradigm to compute an approximate solution. We use the following Lemma from [BW18] that relates a good low-rank approximation of an additive error project-cost preserving sketch to a low-rank approximation of the original matrix. A similar guarantee for relative error appears in [CMM17] and [MW17].

Lemma 5.16. (Lemma 4.4 in [BW18].) *Let \mathbf{C} be a column PCP for \mathbf{A} satisfying the guarantee of Theorem 5.7. Let $\mathbf{P}_\mathbf{C}^* = \operatorname{argmin}_{\operatorname{rank}(\mathbf{X}) \leq k} \left\| \mathbf{C} - \mathbf{X}\mathbf{C} \right\|_F^2$ and $\mathbf{P}_\mathbf{A}^* = \operatorname{argmin}_{\operatorname{rank}(\mathbf{X}) \leq k} \left\| \mathbf{A} - \mathbf{X}\mathbf{A} \right\|_F^2$. Then, for any rank k projection matrix \mathbf{P} such that $\left\| \mathbf{C} - \mathbf{P}\mathbf{C} \right\|_F^2 \leq \left\| \mathbf{C} - \mathbf{P}_\mathbf{C}^* \mathbf{C} \right\|_F^2 + (\epsilon + \sqrt{\eta}) \left\| \mathbf{C} \right\|_F^2$, with probability at least 99/100,*

$$\left\| \mathbf{A} - \mathbf{P}\mathbf{A} \right\|_F^2 \leq \left\| \mathbf{A} - \mathbf{P}_\mathbf{A}^* \mathbf{A} \right\|_F^2 + (\epsilon + \sqrt{\eta}) \left\| \mathbf{A} \right\|_F^2$$

A similar guarantee holds for a row PCP of \mathbf{A} .

Note, while \mathbf{RSS}^T is an approximate rank- k solution for \mathbf{R} , it does not have the same dimensions as \mathbf{A} . If we do not consider running time, we could construct a low-rank approximation to \mathbf{A} as follows: since projecting \mathbf{R} onto \mathbf{S}^T is approximately optimal, it follows from Lemma 5.16 that with probability 99/100,

$$\left\| \mathbf{C} - \mathbf{C}\mathbf{S}\mathbf{S}^T \right\|_F^2 = \left\| \mathbf{C} - \mathbf{C}_k \right\|_F^2 \pm (\epsilon + \sqrt{\eta}) \left\| \mathbf{C} \right\|_F^2\tag{5.33}$$

Let $\mathbf{C}_k = \mathbf{U}'\mathbf{V}'^T$ be such that \mathbf{U}' has orthonormal columns. Then, $\left\| \mathbf{C} - \mathbf{U}'\mathbf{U}'^T \mathbf{C} \right\|_F^2 = \left\| \mathbf{C} - \mathbf{C}_k \right\|_F^2$ and by Lemma 5.16 it follows that with probability 98/100, $\left\| \mathbf{A} - \mathbf{U}'\mathbf{U}'^T \mathbf{A} \right\|_F^2 \leq \left\| \mathbf{A} - \mathbf{A}_k \right\|_F^2 + (\epsilon + \sqrt{\eta}) \left\| \mathbf{A} \right\|_F^2$. However, even approximately computing a column space \mathbf{U}' for \mathbf{C}_k using an input-sparsity time algorithm, such as [CW13], could require $\Omega(nt)$ queries. To get around this issue, we observe that

an approximate solution for \mathbf{R} lies in the row space of \mathbf{S}^T and therefore, an approximately optimal solution for \mathbf{C} lies in the row space of \mathbf{S}^T . We then set up the following regression problem:

$$\min_{\text{rank}(\mathbf{X}) \leq k} \|\mathbf{C} - \mathbf{X}\mathbf{S}^T\|_F^2 \quad (5.34)$$

Note, this regression problem is still too large to be solved in sublinear time. Therefore, we sketch it by sampling columns of \mathbf{C} to set up a smaller regression problem. Observe, since \mathbf{S} has orthonormal columns, the leverage scores are simply ℓ_2^2 norms of rows of \mathbf{S} . Now, using Lemma 4.4, approximately solving this regression problem requires sampling $\Omega(k/\epsilon)$ rows of \mathbf{C} , which in turn requires $\Omega(\frac{nk}{\epsilon})$ queries to $\mathbf{A} + \mathbf{N}$. Note, the above theorem applied to Equation 5.34 can take $O(nk + \text{poly}(k, \epsilon^{-1}))$ time and thus is a lower order term. Since \mathbf{S}^T has orthonormal rows, the leverage scores are precomputed. With probability at least 99/100, we can compute $\mathbf{X}_C = \arg\min_{\mathbf{X}} \|\mathbf{C}\mathbf{E} - \mathbf{X}\mathbf{S}^T\mathbf{E}\|_F^2$, where \mathbf{E} is a leverage score sketching matrix with $O(\frac{k}{\epsilon})$ columns, as shown in Lemma 4.4, and thus requires $O(\frac{nk}{\epsilon})$ queries to \mathbf{A} . Then,

$$\begin{aligned} \|\mathbf{C} - \mathbf{X}_C\mathbf{S}^T\|_F^2 &\leq (1 + \epsilon) \min_{\mathbf{X}} \|\mathbf{C} - \mathbf{X}\mathbf{S}^T\|_F^2 \\ &\leq (1 + \epsilon) \|\mathbf{C} - \mathbf{C}\mathbf{S}\mathbf{S}^T\|_F^2 \\ &= \|\mathbf{C} - \mathbf{C}_k\|_F^2 \pm (\epsilon + \sqrt{\eta}) \|\mathbf{C}\|_F^2 \end{aligned} \quad (5.35)$$

where the last two inequalities follow from equation 5.33. Let $\mathbf{X}_C\mathbf{S}^T = \mathbf{U}'\mathbf{V}'^T$ be such that \mathbf{U}' has orthonormal columns. Then, the column space of \mathbf{U}' contains an approximately optimal solution for \mathbf{A} , since $\|\mathbf{C} - \mathbf{U}'\mathbf{V}'^T\|_F^2 = \|\mathbf{C} - \mathbf{C}_k\|_F^2 \pm \epsilon \|\mathbf{C}\|_F^2$ and \mathbf{C} is a column PCP for \mathbf{A} . It follows from Lemma 5.16 that with probability at least 98/100,

$$\|\mathbf{A} - \mathbf{U}'\mathbf{U}'^T\mathbf{A}\|_F^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + (\epsilon + \sqrt{\eta}) \|\mathbf{A}\|_F^2 \quad (5.36)$$

Therefore, there exists a good solution for \mathbf{A} in the column space of \mathbf{U}' . Since we cannot compute this explicitly, we set up the following regression problem:

$$\min_{\mathbf{X}} \|\mathbf{A} - \mathbf{U}'\mathbf{X}\|_F^2 \quad (5.37)$$

Again, we sketch the regression problem above by sampling columns of \mathbf{A} and apply Lemma 4.4. We can then compute $\mathbf{X}_A = \arg\min_{\mathbf{X}} \|\mathbf{E}'\mathbf{A} - \mathbf{E}'\mathbf{U}'\mathbf{X}\|_F^2$ with probability at least 99/100, where \mathbf{E}' is a sketching matrix with $(\frac{k}{\epsilon})$ rows and $O(\frac{nk}{\epsilon})$ queries to \mathbf{A} . Then,

$$\begin{aligned} \|\mathbf{A} - \mathbf{U}'\mathbf{X}_A\|_F^2 &\leq (1 + \epsilon) \min_{\mathbf{X}} \|\mathbf{A} - \mathbf{U}'\mathbf{X}\|_F^2 \\ &\leq (1 + \epsilon) \|\mathbf{A} - \mathbf{U}'\mathbf{U}'^T\mathbf{A}\|_F^2 \\ &\leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + O(\epsilon + \sqrt{\eta}) \|\mathbf{A}\|_F^2 \end{aligned} \quad (5.38)$$

where the second inequality follows from \mathbf{X} being the minimizer and $\mathbf{U}'^T\mathbf{A}$ being some other matrix, and the last inequality follows from equation 5.36. Recall, \mathbf{U}' is an $n \times k$ matrix and the time taken to solve the regression problem is $O(nk + \text{poly}(k, \epsilon^{-1}))$.

Therefore, we observe that $\mathbf{U}'\mathbf{X}_\mathbf{A}$ suffices and we output it in factored form by setting $\mathbf{M} = \mathbf{U}'$ and $\mathbf{N} = \mathbf{X}_\mathbf{A}^T$. Union bounding over the probabilistic events, and rescaling ϵ , with probability at least 9/10, Algorithm 4 outputs $\mathbf{M} \in \mathbb{R}^{n \times k}$ and $\mathbf{N}^T \in \mathbb{R}^{n \times k}$ such that the total number of entries queried in \mathbf{A} are $\tilde{O}\left(\frac{\phi_{\max}^2 nk}{\epsilon}\right)$. This concludes the proof of Theorem 5.11.

Correlation Matrices. We introduce low-rank approximation of correlation Matrices, a special case of PSD matrices where the diagonal is all 1s. Correlation matrices are well studied in numerical linear algebra, statistics and finance since an important statistic of n random variables X_1, X_2, \dots, X_n is given by computing the pairwise correlation coefficient, $\text{corr}(X_i, X_j) = \text{cov}(X_i, X_j) / \sqrt{\text{var}(X_i) \cdot \text{var}(X_j)}$. A natural matrix representation of correlation coefficients results in a $n \times n$ correlation matrix \mathbf{A} such that $\mathbf{A}_{i,j} = \text{corr}(X_i, X_j)$.

Definition 5.17. (Correlation Matrices.) \mathbf{A} is an $n \times n$ correlation matrix if \mathbf{A} is PSD and $\mathbf{A}_{i,i} = 1$, for all $i \in [n]$.

Often, in practice the correlation matrices obtained are close to being PSD, but corrupted by noise in the form of missing or asynchronous observations, stress testing or aggregation. Here the goal is to query few entries of the corrupted matrix and recover a rank- k matrix close to the underlying correlation matrix, assuming that the underlying matrix is also close to low rank to begin with.

Here we observe that since correlation matrices have all diagonal entries equal to 1, we can compute ϕ_{\max} by simply reading the diagonal entries of $\mathbf{A} + \mathbf{N}$. However, we can do even better since we can discard the diagonal entries of $\mathbf{A} + \mathbf{N}$. The main insight here is that for correlation matrices, our algorithm simply uniformly samples columns and rows to construct our row and column PCPs, since we know what the true diagonals should be. In this case, no matter what the adversary does to the diagonal, $\phi_{\max} = 1$ and we obtain a $\tilde{O}(nk/\epsilon)$ query algorithm.

Corollary 5.18. (Robust LRA for Correlation Matrices.) Let k be an integer and $1 > \epsilon > \eta > 0$. Given $\mathbf{A} + \mathbf{N}$, where \mathbf{A} is a correlation matrix and \mathbf{N} is a corruption term such that $\|\mathbf{N}\|_F^2 \leq \eta \|\mathbf{A}\|_F^2$ and for all $i \in [n]$ $\|\mathbf{N}_{i,*}\|_2^2 \leq c \|\mathbf{A}_{i,*}\|_2^2$ for a fixed constant c , there exists an algorithm that samples $\tilde{O}(nk/\epsilon)$ entries in $\mathbf{A} + \mathbf{N}$ and with probability at least 99/100, computes a rank k matrix \mathbf{B} such that

$$\|\mathbf{A} - \mathbf{B}\|_F^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + (\epsilon + \sqrt{\eta}) \|\mathbf{A}\|_F^2$$

Note, the sample complexity of this algorithm is optimal, since there is an $\Omega(nk/\epsilon)$ query lower bound for additive-error low-rank approximation of correlation matrices, even when there is no corruption (see Corollary 5.20).

Additive-Error PSD Low-Rank Approximation. In the limit where $\eta = 0$, $\phi_{\max} = 1$, and we obtain an algorithm with query complexity $\tilde{O}(nk/\epsilon)$. While this guarantee is already implied by our algorithm for *relative-error* low-rank approximation, our additive-error algorithm is simpler to implement, since the sampling probabilities can be computed *exactly* by simply reading the diagonal.

Corollary 5.19. (*Sample-Optimal Additive-Error LRA.*) Given a PSD matrix \mathbf{A} , rank parameter k , and $\epsilon > 0$, there exists an algorithm that samples $\tilde{O}(nk/\epsilon)$ entries in \mathbf{A} and outputs a rank- k matrix \mathbf{B} such that with probability at least 99/100,

$$\|\mathbf{A} - \mathbf{B}\|_F^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + \epsilon \|\mathbf{A}\|_F^2$$

We show a matching lower bound on the query complexity of additive-error low-rank approximation of PSD matrices. Here, we simply observe that the lower bound construction introduced by [MW17] of $\Omega\left(\frac{nk}{\epsilon}\right)$ also holds for additive error. As a consequence our algorithm is optimal in the setting where there is no corruption.

Corollary 5.20. (*Correlation Matrix Lower Bound, Theorem 13 [MW17].*) Let \mathbf{A} be a PSD matrix, $k \in \mathbb{Z}$ and $\epsilon > 0$ be such that $\frac{nk}{\epsilon} = o(n^2)$. Any randomized algorithm, \mathcal{A} , that with probability at least 2/3, computes a rank k matrix \mathbf{B} such that

$$\|\mathbf{A} - \mathbf{B}\|_F^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + \epsilon \|\mathbf{A}\|_F^2$$

must read $\Omega\left(\frac{nk}{\epsilon}\right)$ entries of \mathbf{A} on some input, possibly adaptively, in expectation.

Proof. We observe that in the lower bound construction of [MW17], the matrix \mathbf{A} is binary, with all 1s on the diagonal, and k off-diagonal blocks of all 1s, each of size $\sqrt{\frac{2\epsilon n}{k}} \times \sqrt{\frac{2\epsilon n}{k}}$. Therefore, \mathbf{A} is a correlation matrix and $\|\mathbf{A}\|_F^2 = (1 + 2\epsilon)n$. Further, the optimal rank- k cost, $\|\mathbf{A} - \mathbf{A}_k\|_F^2 = \Theta(n)$. To compute an additive-error approximation, any algorithm must capture $\epsilon \|\mathbf{A}\|_F^2 = \epsilon n$ mass among the off-diagonal entries of \mathbf{A} . Note, the remaining proof is identical to Theorem 13 in [MW17]. \square

Acknowledgments. A. Bakshi and D. Woodruff would like to acknowledge support from the National Science Foundation under Grant No. CCF-1815840. Part of this work was also done while they were visiting the Simons Institute for the Theory of Computing.

References

- [AFKM01] Dimitris Achlioptas, Amos Fiat, Anna R Karlin, and Frank McSherry. Web search via hub synthesis. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 500–509. IEEE, 2001. [1](#)
- [ALN07] Sanjeev Arora, James R Lee, and Assaf Naor. Fréchet embeddings of negative type metrics. *Discrete & Computational Geometry*, 38(4):726–739, 2007. [39](#)
- [ALN08] Sanjeev Arora, James Lee, and Assaf Naor. Euclidean distortion and the sparsest cut. *Journal of the American Mathematical Society*, 21(1):1–21, 2008. [1](#), [39](#)
- [AM05] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory*, pages 458–469. Springer, 2005. [1](#)
- [AM15] Ahmed Alaoui and Michael W Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 775–783, 2015. [18](#)
- [ARV09] Sanjeev Arora, Satish Rao, and Umesh Vazirani. Expander flows, geometric embeddings and graph partitioning. *Journal of the ACM (JACM)*, 56(2):5, 2009. [39](#)
- [BDN15] Jean Bourgain, Sjoerd Dirksen, and Jelani Nelson. Toward a unified theory of sparse dimensionality reduction in euclidean space. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 499–508, 2015. [17](#)
- [BW18] Ainesh Bakshi and David Woodruff. Sublinear time low-rank approximation of distance matrices. In *Advances in Neural Information Processing Systems*, pages 3782–3792, 2018. [1](#), [2](#), [3](#), [4](#), [6](#), [8](#), [10](#), [13](#), [15](#), [19](#), [20](#), [39](#), [51](#), [56](#), [58](#), [65](#)
- [BY02] Ziv Bar-Yossef. *The complexity of massive data set computations*. PhD thesis, University of California, Berkeley, 2002. [47](#)
- [CEM⁺15] Michael B Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 163–172. ACM, 2015. [7](#), [17](#), [18](#), [19](#), [29](#)

- [CGR05] Shuchi Chawla, Anupam Gupta, and Harald Räcke. Embeddings of negative-type metrics and an improved approximation to generalized sparsest cut. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 102–111. Society for Industrial and Applied Mathematics, 2005. [39](#)
- [CKM⁺11] Paul Christiano, Jonathan A Kelner, Aleksander Madry, Daniel A Spielman, and Shang-Hua Teng. Electrical flows, laplacian systems, and faster approximation of maximum flow in undirected graphs. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 273–282. ACM, 2011. [1](#)
- [CLL⁺10] Pei-Chun Chen, Kuang-Yao Lee, Tsung-Ju Lee, Yuh-Jye Lee, and Su-Yun Huang. Multiclass support vector classification via coding and regression. *Neurocomputing*, 73(7-9):1501–1512, 2010. [25](#)
- [CLM⁺15] Michael B Cohen, Yin Tat Lee, Cameron Musco, Christopher Musco, Richard Peng, and Aaron Sidford. Uniform sampling for matrix approximation. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 181–190. ACM, 2015. [16](#)
- [CLW18] Nai-Hui Chia, Han-Hsuan Lin, and Chunhao Wang. Quantum-inspired sublinear classical algorithms for solving low-rank linear systems. *arXiv preprint arXiv:1811.04852*, 2018. [1](#), [14](#)
- [CMM17] Michael B. Cohen, Cameron Musco, and Christopher Musco. Input sparsity time low-rank approximation via ridge leverage score sampling. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19*, pages 1758–1777, 2017. [6](#), [8](#), [13](#), [15](#), [17](#), [18](#), [19](#), [20](#), [29](#), [49](#), [51](#), [56](#), [58](#), [65](#)
- [CNW15] Michael B Cohen, Jelani Nelson, and David P Woodruff. Optimal approximate matrix product in terms of stable rank. *arXiv preprint arXiv:1507.02268*, 2015. [9](#), [26](#), [28](#), [29](#)
- [Coh16] Michael B Cohen. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*, pages 278–287. SIAM, 2016. [17](#)
- [CRR⁺96] Ashok K Chandra, Prabhakar Raghavan, Walter L Ruzzo, Roman Smolensky, and Prasoon Tiwari. The electrical resistance of a graph captures its commute and cover times. *Computational Complexity*, 6(4):312–340, 1996. [1](#)
- [CW09] Kenneth L Clarkson and David P Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 205–214. ACM, 2009. [9](#), [26](#)
- [CW13] Kenneth L Clarkson and David P Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 81–90. ACM, 2013. [6](#), [9](#), [17](#), [20](#), [65](#)

- [CW17] Kenneth L Clarkson and David P Woodruff. Low-rank psd approximation in input-sparsity time. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2061–2072. Society for Industrial and Applied Mathematics, 2017. [7](#), [20](#), [38](#), [39](#), [40](#)
- [DFK⁺04] Petros Drineas, Alan Frieze, Ravi Kannan, Santosh Vempala, and V Vinay. Clustering large graphs via the singular value decomposition. *Machine learning*, 56(1-3):9–33, 2004. [1](#), [9](#)
- [DKM06] Petros Drineas, Ravi Kannan, and Michael W Mahoney. Fast monte carlo algorithms for matrices i: Approximating matrix multiplication. *SIAM Journal on Computing*, 36(1):132–157, 2006. [28](#)
- [DKR02] Petros Drineas, Iordanis Kerenidis, and Prabhakar Raghavan. Competitive recommendation systems. In *Proceedings of the thirty-fourth annual ACM symposium on Theory of computing*, pages 82–90. ACM, 2002. [1](#)
- [DL09] Michel Marie Deza and Monique Laurent. *Geometry of cuts and metrics*, volume 15. Springer, 2009. [1](#), [39](#)
- [Fis05] Steve Fisk. A very short proof of cauchy’s interlace theorem for eigenvalues of hermitian matrices. *arXiv preprint math/0502408*, 2005. [11](#)
- [FKV04] Alan M. Frieze, Ravi Kannan, and Santosh Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *J. ACM*, 51(6):1025–1041, 2004. [2](#), [12](#), [14](#), [15](#), [53](#), [58](#), [59](#), [61](#), [63](#), [64](#)
- [FSS13] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1434–1453. Society for Industrial and Applied Mathematics, 2013. [19](#)
- [FT07] Shmuel Friedland and Anatoli Torokhti. Generalized rank-constrained matrix approximations. *SIAM Journal on Matrix Analysis and Applications*, 29(2):656–659, 2007. [8](#), [22](#)
- [GLF⁺10] David Gross, Yi-Kai Liu, Steven T Flammia, Stephen Becker, and Jens Eisert. Quantum state tomography via compressed sensing. *Physical review letters*, 105(15):150401, 2010. [1](#)
- [GLT18] András Gilyén, Seth Lloyd, and Ewin Tang. Quantum-inspired low-rank stochastic regression with logarithmic dependence on the dimension. *arXiv preprint arXiv:1811.04909*, 2018. [1](#), [14](#)
- [Gru17] Marvin Gruber. *Improving Efficiency by Shrinkage: The James–Stein and Ridge Regression Estimators*. Routledge, 2017. [4](#)

- [GSLW19] András Gilyén, Yuan Su, Guang Hao Low, and Nathan Wiebe. Quantum singular value transformation and beyond: exponential improvements for quantum matrix arithmetics. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 193–204. ACM, 2019. [1](#)
- [Hig02] Nicholas J Higham. Computing the nearest correlation matrix—a problem from finance. *IMA journal of Numerical Analysis*, 22(3):329–343, 2002. [2](#)
- [IVWW19] Piotr Indyk, Ali Vakilian, Tal Wagner, and David Woodruff. Sample-optimal low-rank approximation of distance matrices. *arXiv preprint arXiv:1906.00339*, 2019. [1](#), [2](#), [14](#), [39](#), [61](#)
- [Kle99] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999. [1](#)
- [KLM⁺17] Michael Kapralov, Yin Tat Lee, CN Musco, Christopher Paul Musco, and Aaron Sidford. Single pass spectral sparsification in dynamic streams. *SIAM Journal on Computing*, 46(1):456–477, 2017. [18](#)
- [KMP14] Ioannis Koutis, Gary L Miller, and Richard Peng. Approaching optimality for solving sdd linear systems. *SIAM Journal on Computing*, 43(1):337–354, 2014. [1](#)
- [KP16] Iordanis Kerenidis and Anupam Prakash. Quantum recommendation systems. *arXiv preprint arXiv:1603.08675*, 2016. [1](#), [2](#), [14](#)
- [KSV05] Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. In *International Conference on Computational Learning Theory*, pages 444–457. Springer, 2005. [1](#)
- [LMP13] Mu Li, Gary L Miller, and Richard Peng. Iterative row sampling. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 127–136. IEEE, 2013. [18](#)
- [LS15] Yin Tat Lee and Aaron Sidford. Efficient inverse maintenance and faster algorithms for linear programming. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 230–249. IEEE, 2015. [16](#)
- [McS01] Frank McSherry. Spectral partitioning of random graphs. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 529–537. IEEE, 2001. [1](#)
- [MM13] Xiangrui Meng and Michael W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Symposium on Theory of Computing Conference, STOC’13, Palo Alto, CA, USA, June 1-4, 2013*, pages 91–100, 2013. [17](#)
- [MM17] Cameron Musco and Christopher Musco. Recursive sampling for the nystrom method. In *Advances in Neural Information Processing Systems*, pages 3833–3845, 2017. [13](#), [18](#), [36](#), [48](#)

- [MST15] Aleksander Madry, Damian Straszak, and Jakub Tarnawski. Fast generation of random spanning trees and the effective resistance metric. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 2019–2036. Society for Industrial and Applied Mathematics, 2015. [1](#)
- [MW17] Cameron Musco and David P. Woodruff. Sublinear time low-rank approximation of positive semidefinite matrices. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 672–683, 2017. [1](#), [3](#), [4](#), [5](#), [6](#), [8](#), [13](#), [15](#), [16](#), [18](#), [19](#), [20](#), [25](#), [35](#), [41](#), [42](#), [47](#), [48](#), [56](#), [58](#), [65](#), [68](#)
- [NN13] Jelani Nelson and Huy L. Nguyen. 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*, pages 117–126, 2013. [17](#), [18](#)
- [RSML18] Patrick Rebentrost, Adrian Steffens, Iman Marvian, and Seth Lloyd. Quantum singular-value decomposition of nonsparse low-rank matrices. *Physical review A*, 97(1):012327, 2018. [1](#), [14](#)
- [RV07] Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM (JACM)*, 54(4):21, 2007. [9](#), [31](#)
- [Sar06] Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *FOCS*, pages 143–152, 2006. [9](#), [16](#)
- [Sch38] Isaac J Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938. [39](#)
- [SS11] Daniel A Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011. [1](#), [16](#)
- [SW19] Xiaofei Shi and David P. Woodruff. Sublinear time numerical linear algebra for structured matrices. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019.*, pages 4918–4925, 2019. [1](#)
- [Tan19] Ewin Tang. A quantum-inspired classical algorithm for recommendation systems. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 217–228. ACM, 2019. [1](#), [2](#), [14](#)
- [TD87] Paul Terwilliger and Michel Deza. The classification of finite connected hypermetric spaces. *Graphs and Combinatorics*, 3(1):293–298, 1987. [1](#), [39](#)
- [Woo14] David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014. [1](#), [17](#)

- [Yao77] Andrew Chi-Chin Yao. Probabilistic computations: Toward a unified measure of complexity. In *Proceedings of the 18th Annual Symposium on Foundations of Computer Science, SFCS '77*, pages 222–227, Washington, DC, USA, 1977. IEEE Computer Society.
- [45](#)