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 $\widetilde{O}(nk/\epsilon^{2.5})$ entries of $\mathbf{A}$ in time $\widetilde{O}(nk/\epsilon^{2.5} + nk^{\omega-1}/\epsilon^{2(\omega-1)})$. 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 $\widetilde{O}(nk/\epsilon)$ entries of $\mathbf{A}$ and outputs a relative-error low-rank approximation in $\widetilde{O}(n \cdot (k/\epsilon)^{\omega-1})$ 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 $\mathcal{P} = \{x_1, x_2, \ldots, x_n\}$ 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 $\widetilde{O}(nk/\epsilon^{2.5})$ entries and outputs a rank-$(k+4)$ matrix. We show that the bi-criteria guarantee is not necessary and obtain an $\widetilde{O}(nk/\epsilon)$ 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.

***Keywords*-low rank approximation, psd matrices, regression, sublinear algorithms, randomized numerical linear algebra**

I. 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 [1], [2], recommendation systems [3], web search [4], [5], and learning mixtures of distributions [6], [7], 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

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

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 [8] and references therein.

For worst-case matrices, it is not hard to see that any algorithm achieving (1) must spend at least $\widetilde{\Omega}(\text{nnz}(\mathbf{A}))$ 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 [9]–[18] has looked at the possibility of achieving *sublinear* time algorithms (classical and quantum) for low-rank approximation. In particular, Musco and Woodruff [10] 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 [19]. Subsequently, Bakshi and Woodruff [11] 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 [20]–[23]. Negative-type metrics have found various applications in algorithm design and optimization [24]–[27].

Musco and Woodruff show that it is possible to output a low-rank matrix $\mathbf{B}$ in factored form achieving (1) in $\widetilde{O}(nk/\epsilon^{2.5} + nk^{\omega-1}/\epsilon^{2(\omega-1)})$ time, while reading only $\widetilde{O}(nk/\epsilon^{2.5})$ entries of $\mathbf{A}$. They also showed a lower bound that any algorithm achieving (1) must read $\widetilde{\Omega}(nk/\epsilon)$ entries, and closing the gap between these bounds has remained

2575-8454/20/$31.00 ©2020 IEEE
DOI 10.1109/FOCS46700.2020.00054

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 $\widetilde{O}(nk/\epsilon^{2.5})$ 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 II.

Next we consider PSD matrices that have been corrupted by a small amount of noise. A drawback of algorithms achieving (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) 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 [28].

Given that it is information-theoretically impossible to obtain the relative-error guarantee (1) in the *robust model*, we relax our notion of approximation to the following well-studied additive-error guarantee:

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

This additive-error guarantee was introduced by the seminal work of Frieze et. al. [29], 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 (2) in $O(\text{nnz}(\mathbf{A}))$ time. Further, given access to an oracle for computing row norms of $\mathbf{A}$, 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* [11], [16], and a quantum algorithm for recommendation systems [9], which was subsequently dequantized [13].

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.

A. 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 I.1 (Informal Sample-Optimal PSD LRA)**.** *Given a PSD matrix $\mathbf{A}$, there exists an algorithm that queries $\widetilde{O}(nk/\epsilon)$ 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 $\widetilde{O}(n \cdot (k/\epsilon)^{\omega-1})$.*

Remark I.2. *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 $\widetilde{\Omega}(nk/\epsilon)$ entries. Our running time also improves that of Musco and Woodruff and is optimal if the matrix multiplication exponent ω is 2.*

Remark I.3. *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 $\widetilde{O}(nk/\epsilon^3 + nk^2/\epsilon^2)$ entries in $\mathbf{A}$ and runs in time $\widetilde{O}(n(k/\epsilon)^\omega + nk^{\omega-1}/\epsilon^{3(\omega-1)})$.*

At the core of our analysis is a sample optimal algorithm for Spectral Regression: $\min_{\mathbf{X}} \|\mathbf{DX} - \mathbf{E}\|_2^2$. 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 $\mathcal{P} = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^d$ 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 I.4 (Informal Sample-Optimal LRA for Negative–Type Metrics)**.** *Given a negative-type distance matrix $\mathbf{A}$,*

Problem	Prior Work		Our Results		Query Lower Bound
	Query	Run Time	Query	Run Time	
PSD LRA	$O\left(\frac{nk}{\epsilon^{2.5}}\right)$	$O\left(\frac{nk^{\omega-1}}{\epsilon^{2\omega-2}}+\frac{nk}{\epsilon^{2.5}}\right)$	$O^*\left(\frac{nk}{\epsilon}\right)$	$O^\dagger\left(\frac{nk^{\omega-1}}{\epsilon^{\omega-1}}\right)$	$\Omega\left(\frac{nk}{\epsilon}\right)$
	[10]		Thm. I.1		[10]
PSD LRA PSD Output	$O\left(nk\left(\frac{k}{\epsilon^2}+\frac{1}{\epsilon^3}\right)\right)$	$O\left(nk^{\omega-1}\left(\frac{k}{\epsilon^\omega}+\frac{1}{\epsilon^{3\omega-3}}\right)\right)$	$O^*\left(\frac{nk}{\epsilon}\right)$	$O^\dagger\left(\frac{nk^{\omega-1}}{\epsilon^{\omega-1}}\right)$	$\Omega\left(\frac{nk}{\epsilon}\right)$
	[10]		Thm. I.1		[10]
Negative-Type LRA	$O\left(\frac{nk}{\epsilon^{2.5}}\right)$	$O\left(\frac{nk^{\omega-1}}{\epsilon^{2\omega-2}}+\frac{nk}{\epsilon^{2.5}}\right)$	$O^*\left(\frac{nk}{\epsilon}\right)$	$O^\dagger\left(\frac{nk^{\omega-1}}{\epsilon^{\omega-1}}\right)$	$\Omega\left(\frac{nk}{\epsilon}\right)$
	Bi-criteria, [11]		No Bi-criteria, Thm. I.4		[11]
Coreset Ridge Regression	$O\left(\frac{ns_\lambda^2}{\epsilon^4}\right)$	$O\left(\frac{ns_\lambda^\omega}{\epsilon^\omega}\right)$	$O^*\left(\frac{ns_\lambda}{\epsilon^2}\right)$	$O^\dagger\left(\frac{ns_\lambda^{\omega-1}}{\epsilon^{2\omega-2}}\right)$	$\Omega\left(\frac{ns_\lambda}{\epsilon^2}\right)$
	[10]		Thm. I.6		See full version

Table I

COMPARISON WITH PRIOR WORK. THE NOTATION O^* AND $O^\dagger$ 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.

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

Remark I.5. *Prior work of Bakshi and Woodruff [11] obtains a $\widetilde{O}(nk/\epsilon^{2.5})$ query algorithm that outputs a rank-$(k+4)$ matrix $\mathbf{B}$ such that $\|\mathbf{A}-\mathbf{B}\|_F^2 \le (1+\epsilon)\|\mathbf{A}-\mathbf{A}_k\|_F^2$. 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 [30] and references therein).

Theorem I.6 (Informal Ridge Regression.)**.** *Given a PSD matrix $\mathbf{A}$, a regularization parameter λ and statistical dimension $s_\lambda = \mathrm{Tr}\left((\mathbf{A}^2+\lambda\mathbb{I})^{-1}\mathbf{A}^2\right)$, there exists an algorithm that queries $\widetilde{O}(ns_\lambda/\epsilon^2)$ entries of $\mathbf{A}$ and with probability $99/100$ outputs a $(1+\epsilon)$ approximate solution to the Ridge Regression objective and runs in $\widetilde{O}(n(s_\lambda/\epsilon^2)^{\omega-1})$ time.*

Remark I.7. *Our result improves on prior work by Musco and Woodruff [10], who obtain an algorithm that queries $\widetilde{O}(ns_\lambda^2/\epsilon^4)$ entries in $\mathbf{A}$ and runs in $\widetilde{O}(n(s_\lambda/\epsilon^2)^\omega)$ time.*

Remark I.8. *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, 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 $\|\mathbf{N}\|_F^2 \le \eta\|\mathbf{A}\|_F^2$. Further, for all $i\in[n]$ $\|\mathbf{N}_{i,*}\|_2^2 \le c\|\mathbf{A}_{i,*}\|_2^2$, 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 I.9 (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 $\widetilde{\Omega}\left(\phi_{\max}^2 nk/\epsilon\right)$ entries of $\mathbf{A}+\mathbf{N}$.*

Remark I.10. *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 I.11. *In our hard instance, $\phi_{\max}^2$ can be as large as $\epsilon n/k$, which implies a sample-complexity lower bound of $\widetilde{\Omega}(n^2)$. While this lower bound precludes sublinear algorithms for arbitrary PSD matrices, we observe that in many applications $\phi_{\max}$ 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 $\phi_{\max}$ 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 I.12 (Informal Robust LRA)**.** *Given $\mathbf{A}+\mathbf{N}$, which satisfies our noise model, there exists an algorithm that queries $\widetilde{O}(\phi_{\max}^2 nk/\epsilon)$ 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 I.13. *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 I.14. *Our techniques extend to low-rank approximation of correlation matrices, and we obtain a sample complexity of $\widetilde{O}(nk/\epsilon)$, which is optimal. In fact, the hard instance in [10] implies an $\widetilde{\Omega}(nk/\epsilon)$ 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 $\sqrt{\eta}\|\mathbf{A}\|_F^2$.*

II. 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 $\Omega(\mathsf{nnz}(\mathbf{A}))$ 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.

A. 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 [10] 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 $\widetilde{O}(nk/\epsilon)$ entries in $\mathbf{A}$ (see Theorem I.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* 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 [31] 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 $\mathbf{P}$:

$$\|\mathbf{C}-\mathbf{PC}\|_F^2 = (1\pm\epsilon_0)\|\mathbf{A}-\mathbf{PA}\|_F^2 \tag{3}$$

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 3. 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 $\mathbf{P}$,

$$\|\mathbf{R}-\mathbf{RP}\|_F^2 + O(\|\mathbf{A}-\mathbf{A}_k\|_F^2) = (1\pm\epsilon_0)\|\mathbf{C}-\mathbf{CP}\|_F^2 \tag{4}$$

Since $\mathbf{R}$ is a much smaller matrix, they run an input-sparsity time algorithm to compute a low-rank approximation for it [32]. Using standard regression techniques (described in [10], [11], [31]) along with equations 3 and 4, 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 3 can be strengthen to a mixed *Spectral-Frobenius PCP* guarantee (also introduced by [10]): for all rank-(k/ϵ) projection matrices $\mathbf{P}$, the column PCP $\mathbf{C}$ satisfies :

$$\begin{aligned}&(1-0.1)\|\mathbf{A}-\mathbf{PA}\|_2^2-\frac{\epsilon}{10k}\|\mathbf{A}-\mathbf{A}_{k/\epsilon}\|_F^2\\&\leq\|\mathbf{C}-\mathbf{PC}\|_2^2\\&\leq(1+0.1)\|\mathbf{A}-\mathbf{PA}\|_2^2+\frac{\epsilon}{10k}\|\mathbf{A}-\mathbf{A}_{k/\epsilon}\|_F^2\end{aligned}\tag{5}$$

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 $\mathbf{P}$,

$$\begin{aligned}&(1-0.1)\|\mathbf{C}-\mathbf{CP}\|_2^2-\frac{\epsilon}{10k}\|\mathbf{A}-\mathbf{A}_{k/\epsilon}\|_F^2\\&\leq\|\mathbf{R}-\mathbf{RP}\|_2^2\\&\leq(1+0.1)\|\mathbf{C}-\mathbf{CP}\|_2^2+\frac{\epsilon}{10k}\|\mathbf{A}-\mathbf{A}_{k/\epsilon}\|_F^2\end{aligned}\tag{6}$$

We then use an *input-sparsity* spectral-low-rank approximation algorithm by [33], to obtain a low-dimensional subspace, represented by a $\sqrt{nk/\epsilon}\times k/\epsilon$ matrix $\mathbf{Z}$ with orthonormal columns such that

$$\|\mathbf{R}-\mathbf{RZZ}^\top\|_2^2\leq\frac{\epsilon}{k}\|\mathbf{R}-\mathbf{R}_{k/\epsilon}\|_F^2\tag{7}$$

Following the notation of Clarkson and Woodruff [34], we refer to the projection matrix $\mathbf{ZZ}^\top$ as a Spectral-Frobenius (SF) projection. A key property of an SF projection is that it spans a $(1+\epsilon)$-relative-error low-rank approximation to $\mathbf{R}$, i.e.

$$\begin{aligned}&\|\mathbf{R}-\mathbf{ZZ}^\top\mathbf{R}_{k/\epsilon}\mathbf{ZZ}^\top\|_F^2\\&\leq(1+\epsilon)\|\mathbf{R}-\mathbf{R}_{k/\epsilon}\|_F^2\end{aligned}\tag{8}$$

Now, using the fact that $\mathbf{R}$ is a Spectral-Frobenius PCP, plugging in $\mathbf{P}=\mathbf{ZZ}^\top$ in Equation 6, we can bound $\|\mathbf{C}-\mathbf{CZZ}^\top\|_2^2$ as follows :

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

where the second inequality follows from Equation 7 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 $\mathbf{A}$. 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:

$$\min_{\text{rank}(\mathbf{X})\leq k}\|\mathbf{A}-\mathbf{QXQ}^\top\|_F^2\tag{10}$$

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:

$$\min_{\text{rank}(X)\leq k}\|\mathbf{SAT}-\mathbf{SQXQ}^\top\mathbf{T}\|_F^2\tag{11}$$

Here, we are faced with an intriguing phenomenon: our sketched optimization problem does not have the property that the minimum cost for Equation 11 is a $(1+\epsilon)$-approximation to the minimum cost for Equation 10. 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 [10], it is open to estimate this cost in $o(n^{3/2})$ time.

However, we can apply the structural result above twice, to show that the optimal solution to Equation 11, when plugged in to Equation 10 obtains a $(1+\epsilon)$-approximation to the minimum cost. Formally, $\mathbf{S}$ and $\mathbf{T}$ have the property that if $\widehat{\mathbf{X}}=\arg\min_{\mathbf{X}}\|\mathbf{SAT}-\mathbf{SQXQ}^\top\mathbf{T}\|_F^2$, then

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

The optimization problem in Equation 11 is called *Generalized Low-Rank Approximation* and admits a closed form solution [35]. 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 poly(k/ϵ) 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 $\mathbf{Q}\widehat{\mathbf{X}}$. In order to compute the second factor, we set up and approximately solve a regression problem Efficiently solving such a regression problem is now standard in low-rank approximation literature [10], [11], [31]. Therefore, we can output a low-rank approximation to $\mathbf{A}$ by querying only $\widetilde{O}(nk/\epsilon)$ entries.

We have now reduced our problem to computing an SF projection for $\mathbf{A}$, while reading only nk/ϵ entries. Recall, in Equation 9, 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 3, it would also work for $\mathbf{A}$ and we would be done. To this end, we consider the following optimization problem:

$$\min_{\mathbf{W}\in\mathbb{R}^{n\times k'}} \|\mathbf{C}-\mathbf{W}\mathbf{Z}^\top\|_2^2 \tag{12}$$

We show that an orthonormal basis $\mathbf{Q}$ for an approximate minimizer to Equation 12 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 12 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 [36]. They construct an oblivious sketch, consisting of random entries in $\{-1,1\}$, for Equation 12 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 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 12 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

$$\|\mathbf{C}-\widehat{\mathbf{W}}\mathbf{Z}^\top\|_2^2 \le O(1)\left(\min_{\mathbf{W}\in\mathbb{R}^{n\times k/\epsilon}}\|\mathbf{C}-\mathbf{W}\mathbf{Z}^\top\|_2^2 + \frac{\epsilon}{k}\|\mathbf{C}-\mathbf{C}_{k/\epsilon}\|_F^2\right) \tag{13}$$

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 13 by $O(\epsilon/k)\cdot\|\mathbf{C}-\mathbf{C}_{k/\epsilon}\|_F^2$ and the Spectral-Frobenius PCP from Equation 3 incurs this term anyway. We then show that we can construct such a sketch $\mathbf{S}$ satisfying Equation 13 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 we prove here is a *weak approximate matrix product* for $\mathbf{C}^*$ and $\mathbf{Z}^\top$ where $\mathbf{C}^*=\mathbf{C}(\mathbb{I}-\mathbf{P}_{\mathbf{Z}^\top})$ is the projection onto the orthogonal complement of $\mathbf{Z}^\top$. While *approximate matrix product* has been extensively studied [1], [32], [37], even for spectral norm [38], 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 [39]. They show that sampling q rows of a matrix $\mathbf{M}$ uniformly at random, indexed by the set $\mathcal{Q}$, results in a matrix $\mathbf{M}_{|\mathcal{Q}}$ such that

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

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 $\mathbf{Z}^T$ 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

$$\mathbb{E}\left[\left\|\mathbf{M}_{|\mathcal{Q}}\right\|_2\right] = O\left(\sqrt{\frac{q}{n}}\|\mathbf{M}\|_2 + \sqrt{\log(q)/b}\|\mathbf{M}\|_{(n/q)}\right) \tag{14}$$

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 14, up to log factors. 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 [11]) 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 $\widetilde{O}(nk/\epsilon^{2.5})$ 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 $\|\mathbf{A}-\mathbf{B}\|_F^2 \leq (1+\epsilon)\|\mathbf{A}-\mathbf{A}_k\|_F^2$.

The key observation they make is that negative-type metrics can be realized as the distances corresponding to a point set $\mathcal{P} = \{x_1, x_2, \ldots x_n\}$ 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 [40] 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, $\|\mathbf{A} - \mathbf{\Omega}\mathbf{A}_k\mathbf{\Omega}\|_F^2 = \|\mathbf{A} - \mathbf{A}_k\|_F^2 + \|\mathbf{A}_k - \mathbf{\Omega}\mathbf{A}_k\mathbf{\Omega}\|_F^2 + 2\mathrm{Tr}\left((\mathbf{A} - \mathbf{A}_k)(\mathbb{I} - \mathbf{\Omega})\mathbf{A}_k\mathbf{\Omega}\right)$. A simple calculation using Von-Neumann's trace inequality bounds the deviation by $O(k\|\mathbf{A}(\mathbb{I} - \mathbf{\Omega})\|_2^2)$. 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 $\|\mathbf{A}-\mathbf{\Omega}\mathbf{A}_k\mathbf{\Omega}\|_F^2 \leq (1+O(\epsilon))\|\mathbf{A}-\mathbf{A}_k\|_F^2$. 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 coreset 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 coreset 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 coreset simultaneously preserves the ridge regression cost for all x, y, it suffices to query the coreset on tuples of (scaled) eigenvectors and learn the positions of the blocks. However, a priori, we do not know what the eingenvectors of $\mathbf{A}$ are. Instead, we query the coreset 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.

B. 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 $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{N}$. 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 $\tilde{\mathbf{A}}$, we can determine the row norms exactly. Therefore, we can run the algorithm of Frieze-Kannan-Vempala [29] to obtain a rank-k matrix $\mathbf{X}\mathbf{Y}^\top$ (in factored form) such that with probability at least $99/100$,

$$\begin{aligned}\|\tilde{\mathbf{A}} - \mathbf{X}\mathbf{Y}^\top\|_F^2 &\leq \|\tilde{\mathbf{A}} - (\tilde{\mathbf{A}})_k\|_F^2 + \epsilon\|\tilde{\mathbf{A}}\|_F^2 \\ &\leq \|\tilde{\mathbf{A}} - \mathbf{A}_k\|_F^2 + \epsilon\|\tilde{\mathbf{A}}\|_F^2 \\ &\leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + \|\mathbf{N}\|_F^2 + 2\langle \mathbf{A} - \mathbf{A}_k, \mathbf{N}\rangle \\ &\quad + (3+\eta)\epsilon\|\mathbf{A}\|_F^2 \\ &\leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + O(\epsilon + \sqrt{\eta})\|\mathbf{A}\|_F^2 \end{aligned} \tag{15}$$

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 inequaliity uses Cauchy-Schwarz on $2|\langle \mathbf{A} - \mathbf{A}_k, \mathbf{N}\rangle| \leq 2\|\mathbf{A}\|_F \cdot \|\mathbf{N}\|_F \leq 2\sqrt{\eta}\|\mathbf{A}\|_F^2$, which follows from the assumption on $\mathbf{N}$. Additionally

$$\begin{aligned}\|\tilde{\mathbf{A}} - \mathbf{X}\mathbf{Y}^\top\|_F^2 &= \|\mathbf{A} - \mathbf{X}\mathbf{Y}^\top\|_F^2 + \|\mathbf{N}\|_F^2 \\ &\quad + 2\langle \mathbf{A} - \mathbf{X}\mathbf{Y}^\top, \mathbf{N}\rangle \\ &\geq \|\mathbf{A} - \mathbf{X}\mathbf{Y}^\top\|_F^2 - 2\sqrt{\eta}\|\mathbf{A}\|_F^2 \end{aligned} \tag{16}$$

Combining Equations 15 and 16, we have $\|\mathbf{A} - \mathbf{X}\mathbf{Y}^\top\|_F^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + O(\epsilon + \sqrt{\eta})\|\mathbf{A}\|_F^2$. 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 I.9. *(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 $\Omega\left(\phi_{\max}^2 nk/\epsilon\right)$ 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 $\mathbb{I}$ 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 $\phi_{\max}$. 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 $\widetilde{O}\left(\phi_{\max}^2\sqrt{n}k^2/\epsilon^2\right)$ 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 II.1 (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 $q = \{q_1, q_2 \ldots q_n\}$ be a probability distribution over the columns of $\mathbf{A}$ such that $q_j = (\mathbf{A}+\mathbf{N})_{j,j}/\mathrm{Tr}(\mathbf{A}+\mathbf{N})$. Construct $\mathbf{C}$ by sampling $O(\phi_{\max}^2\sqrt{n}k^2/\epsilon^2)$ 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. [31] works for arbitrary $\mathbf{A}$, but requires $nnz(\mathbf{A})$ time and queries to compute the approximate ridge-leverage scores of $\mathbf{A}$. Musco and Musco [41] 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 [10] use this method to compute the ridge leverage scores of $\mathbf{A}^{1/2}$ with $\Theta(nk)$ queries and show that the this provides a $(\sqrt{n/k})$-approximation to the ridge leverage scores of $\mathbf{A}$. We note that the guarantees obtained by [10], [31] are relative error, as opposed to the additive error guarantee in the theorem above. Finally, Bakshi and Woodruff [11] 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 II.2 (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 $p_j = (\mathbf{A}+\mathbf{N})_{j,j}/\mathrm{Tr}\,(\mathbf{A}+\mathbf{N})$. Construct $\mathbf{R}$ by sampling $\widetilde{O}(\phi_{\max}\sqrt{n}k^2/\epsilon^2)$ 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 $t = \widetilde{O}(\phi_{\max}\sqrt{n}k^2/\epsilon^2)$. 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 $\epsilon^3 t^2/k^3 = \widetilde{O}(\phi_{\max}^2 nk/\epsilon)$ 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 $\mathbf{R}$ [29]. 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 $\mathcal{Y}_i$ [29]. As shown in [16] the following two conditions are a relaxation of those required in [29], 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 [29]. 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 [9], [12]–[15] that uses Frieze-Kannan-Vempala ℓ_2^2 sampling as a key algorithmic primitive, our analysis may be of independent interest.

Lemma II.1 (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 $\mathbf{X}_i = \sum_{j\in[\epsilon^3 t/k^3]} \mathbf{X}_{i,j}$ such that $\mathbf{X}_{i,j} = k^3\mathbf{R}_{i,j'}^2/\epsilon^3$ with probability $1/t$, for all $j' \in [t]$. Then, for all $i \in [t]$, $\mathbf{X}_i = (1 \pm 0.1)\|\mathbf{R}_{i,*}\|_2^2$ 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 [29]. 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 II.1. 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 $\mathcal{H} = \left\{i \mid \|\mathbf{R}_{i,*}\|_2^2 \geq \phi_{\max}^2 n/t^2 \|\mathbf{R}\|_F^2\right\}$ be the set of heavy rows and $[t] \setminus \mathcal{H}$ be the remaining rows. Note, $|\mathcal{H}| = O\big(t^2/\phi_{\max}^2 n\big) = O\big(k^4 \log^4(n)/\epsilon^4\big)$. 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-Vemapala 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 [10], [11], [31], 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 I.12.

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 $\widetilde{O}(nk/\epsilon)$ 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.

ACKNOWLEDGMENT

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

[1] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay, "Clustering large graphs via the singular value decomposition," *Machine learning*, vol. 56, no. 1-3, pp. 9–33, 2004.

[2] F. McSherry, "Spectral partitioning of random graphs," in *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*. IEEE, 2001, pp. 529–537.

[3] P. Drineas, I. Kerenidis, and P. Raghavan, "Competitive recommendation systems," in *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*. ACM, 2002, pp. 82–90.

[4] D. Achlioptas, A. Fiat, A. R. Karlin, and F. McSherry, "Web search via hub synthesis," in *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*. IEEE, 2001, pp. 500–509.

[5] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[6] D. Achlioptas and F. McSherry, "On spectral learning of mixtures of distributions," in *International Conference on Computational Learning Theory*. Springer, 2005, pp. 458–469.

[7] R. Kannan, H. Salmasian, and S. Vempala, "The spectral method for general mixture models," in *International Conference on Computational Learning Theory*. Springer, 2005, pp. 444–457.

[8] D. P. Woodruff, "Sketching as a tool for numerical linear algebra," *Foundations and Trends in Theoretical Computer Science*, vol. 10, no. 1-2, pp. 1–157, 2014.

[9] I. Kerenidis and A. Prakash, "Quantum recommendation systems," *arXiv preprint arXiv:1603.08675*, 2016.

[10] C. Musco and D. 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*, 2017, pp. 672–683.

[11] A. Bakshi and D. Woodruff, "Sublinear time low-rank approximation of distance matrices," in *Advances in Neural Information Processing Systems*, 2018, pp. 3782–3792.

[12] N.-H. Chia, H.-H. Lin, and C. Wang, "Quantum-inspired sublinear classical algorithms for solving low-rank linear systems," *arXiv preprint arXiv:1811.04852*, 2018.

[13] E. Tang, "A quantum-inspired classical algorithm for recommendation systems," in *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 2019, pp. 217–228.

[14] P. Rebentrost, A. Steffens, I. Marvian, and S. Lloyd, "Quantum singular-value decomposition of nonsparse low-rank matrices," *Physical review A*, vol. 97, no. 1, p. 012327, 2018.

[15] A. Gilyén, S. Lloyd, and E. Tang, "Quantum-inspired low-rank stochastic regression with logarithmic dependence on the dimension," *arXiv preprint arXiv:1811.04909*, 2018.

[16] P. Indyk, A. Vakilian, T. Wagner, and D. Woodruff, "Sample-optimal low-rank approximation of distance matrices," *arXiv preprint arXiv:1906.00339*, 2019.

[17] X. Shi and D. 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.*, 2019, pp. 4918–4925.

[18] A. Gilyén, Y. Su, G. H. Low, and N. 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*. ACM, 2019, pp. 193–204.

[19] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, "Quantum state tomography via compressed sensing," *Physical review letters*, vol. 105, no. 15, p. 150401, 2010.

[20] M. M. Deza and M. Laurent, *Geometry of cuts and metrics*. Springer, 2009, vol. 15.

[21] P. Terwilliger and M. Deza, "The classification of finite connected hypermetric spaces," *Graphs and Combinatorics*, vol. 3, no. 1, pp. 293–298, 1987.

[22] A. K. Chandra, P. Raghavan, W. L. Ruzzo, R. Smolensky, and P. Tiwari, "The electrical resistance of a graph captures its commute and cover times," *Computational Complexity*, vol. 6, no. 4, pp. 312–340, 1996.

[23] P. Christiano, J. A. Kelner, A. Madry, D. A. Spielman, and S.-H. 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*. ACM, 2011, pp. 273–282.

[24] S. Arora, J. Lee, and A. Naor, "Euclidean distortion and the sparsest cut," *Journal of the American Mathematical Society*, vol. 21, no. 1, pp. 1–21, 2008.

[25] D. A. Spielman and N. Srivastava, "Graph sparsification by effective resistances," *SIAM Journal on Computing*, vol. 40, no. 6, pp. 1913–1926, 2011.

[26] I. Koutis, G. L. Miller, and R. Peng, "Approaching optimality for solving sdd linear systems," *SIAM Journal on Computing*, vol. 43, no. 1, pp. 337–354, 2014.

[27] A. Madry, D. Straszak, and J. 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*. Society for Industrial and Applied Mathematics, 2015, pp. 2019–2036.

[28] N. J. Higham, "Computing the nearest correlation matrixa problem from finance," *IMA journal of Numerical Analysis*, vol. 22, no. 3, pp. 329–343, 2002.

[29] A. M. Frieze, R. Kannan, and S. Vempala, "Fast monte-carlo algorithms for finding low-rank approximations," *J. ACM*, vol. 51, no. 6, pp. 1025–1041, 2004.

[30] M. Gruber, *Improving Efficiency by Shrinkage: The James–Stein and Ridge Regression Estimators*. Routledge, 2017.

[31] M. B. Cohen, C. Musco, and C. 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*, 2017, pp. 1758–1777.

[32] K. L. Clarkson and D. P. Woodruff, "Low rank approximation and regression in input sparsity time," in *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. ACM, 2013, pp. 81–90.

[33] M. B. Cohen, S. Elder, C. Musco, C. Musco, and M. Persu, "Dimensionality reduction for k-means clustering and low rank approximation," in *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. ACM, 2015, pp. 163–172.

[34] K. L. Clarkson and D. P. Woodruff, "Low-rank psd approximation in input-sparsity time," in *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 2017, pp. 2061–2072.

[35] S. Friedland and A. Torokhti, "Generalized rank-constrained matrix approximations," *SIAM Journal on Matrix Analysis and Applications*, vol. 29, no. 2, pp. 656–659, 2007.

[36] K. L. Clarkson and D. P. Woodruff, "Numerical linear algebra in the streaming model," in *Proceedings of the forty-first annual ACM symposium on Theory of computing*. ACM, 2009, pp. 205–214.

[37] T. Sarlos, "Improved approximation algorithms for large matrices via random projections," in *FOCS*, 2006, pp. 143–152.

[38] M. B. Cohen, J. Nelson, and D. P. Woodruff, "Optimal approximate matrix product in terms of stable rank," *arXiv preprint arXiv:1507.02268*, 2015.

[39] M. Rudelson and R. Vershynin, "Sampling from large matrices: An approach through geometric functional analysis," *Journal of the ACM (JACM)*, vol. 54, no. 4, p. 21, 2007.

[40] S. Fisk, "A very short proof of cauchy's interlace theorem for eigenvalues of hermitian matrices," *arXiv preprint math/0502408*, 2005.

[41] C. Musco and C. Musco, "Recursive sampling for the nystrom method," in *Advances in Neural Information Processing Systems*, 2017, pp. 3833–3845.