Sublinear Time Low-Rank Approximation of Positive Semidefinite Matrices

Cameron Musco
MIT
cnmusco@mit.edu

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

Abstract

We show how to compute a *relative-error* low-rank approximation to any positive semidefinite (PSD) matrix in *sublinear time*, i.e., for any $n \times n$ PSD matrix A, in $\tilde{O}(n \cdot \text{poly}(k/\epsilon))$ time we output a rank-k matrix B, in factored form, for which $\|A-B\|_F^2 \leq (1+\epsilon)\|A-A_k\|_F^2$, where A_k is the best rank-k approximation to A. When k and $1/\epsilon$ are not too large compared to the sparsity of A, our algorithm does not need to read all entries of the matrix. Hence, we significantly improve upon previous $\text{nnz}(A)$ time algorithms based on oblivious subspace embeddings, and bypass an $\text{nnz}(A)$ time lower bound for general matrices (where $\text{nnz}(A)$ denotes the number of non-zero entries in the matrix). We prove time lower bounds for low-rank approximation of PSD matrices, showing that our algorithm is close to optimal. Finally, we extend our techniques to give sublinear time algorithms for low-rank approximation of A in the (often stronger) spectral norm metric $\|A - B\|_2^2$ and for ridge regression on PSD matrices.

1 Introduction

A fundamental task in numerical linear algebra is to compute a low-rank approximation of a matrix. Such an approximation can reveal underlying low-dimensional structure, can provide a compact way of storing a matrix in factored form, and can be quickly applied to a vector. Countless applications include clustering [DFK+04, FSS13, LBKW14, CEM+15], datamining [AFK+01], information retrieval [PRTV00], learning mixtures of distributions [AM05, KSV08], recommendation systems [DKR02], topic modeling [Hof03], and web search [AFKM01, Kle99].

One of the most well-studied versions of the problem is to compute a near optimal low-rank approximation with respect to the Frobenius norm. That is, given an $n \times n$ input matrix A and an accuracy parameter $\epsilon > 0$, output a rank-k matrix B for which:

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

where for a matrix C, $\|C\|_F^2 = \sum_{i,j} C_{i,j}^2$ is its squared Frobenius norm, and $A_k = \text{argmin}_{\text{rank-}k\ B}\|A - B\|_F$. A_k can be computed exactly using the singular value decomposition, but takes $O(n^3)$ time in practice and n^ω time in theory, where $\omega \approx 2.373$ is the exponent of matrix multiplication.

In seminal work, Frieze, Kannan, and Vempala [FKV04] and Achlioptas and McSherry [AM07] show that using randomization and approximation, much faster runtimes are possible. Specifically, [FKV04] gives an algorithm that, assuming access to the row norms of A, outputs rank-k B, in factored form, such that with good probability, $\|A - B\|_F^2 \leq \|A - A_k\|_F^2 + \epsilon\|A\|_F^2$. The algorithm runs in just $n \cdot \text{poly}(k/\epsilon)$ time. However $\text{nnz}(A)$ additional time is required to compute the row norms, where $\text{nnz}(A)$ denotes the number of non-zero entries of A. Further, the guarantee achieved can be significantly weaker than (1), since the error is of the form $\epsilon\|A\|_F^2$ rather than $\epsilon\|A - A_k\|_F^2$. Note that $\|A - A_k\|_F^2 \ll \|A\|_F^2$ precisely when A is well-approximated by a rank-k matrix. Related additive error algorithms with additional assumptions were given for tensors in [SWZ16].

Sarlós [Sar06] showed how to achieve (1) with constant probability in $\tilde{O}(\text{nnz}(A) \cdot k/\epsilon) + n \cdot \text{poly}(k/\epsilon)$ time. This was improved by Clarkson and Woodruff [CW13] who achieved $O(\text{nnz}(A)) + n \cdot \text{poly}(k/\epsilon)$ time. See also work by Bourgain, Dirksen, and Nelson [BDN15], Cohen [Coh16], Meng and Mahoney [MM13], and Nelson and Nguyen [NN13] which further improved the degree in the $\text{poly}(k/\epsilon)$ term. For a survey, see [Woo14].

In the special case that A is rank-k and so $\|A - A_k\|_F^2 = 0$, (1) is equivalent to the well studied low-rank matrix completion problem [CR09]. Much attention has focused on completing *incoherent* low-rank matrices, whose singular directions are represented uniformly throughout the rows and columns and hence can be identified via uniform sampling and without fully accessing the matrix. Under incoherence (and often condition number) assumptions, a number of methods are able to complete a rank-k matrix in $\tilde{O}(n \cdot \text{poly}(k))$ time [JNS13, Har14].

For general matrices, without incoherence, it is not hard to see that $\Omega(\text{nnz}(A))$ is a time lower bound: if one does not read a constant fraction of entries of A, with constant probability one can miss an entry much larger than all others, which needs to be included in the low-rank approximation.

1.1 Low-rank Approximation of Positive Semidefinite Matrices

An important class of matrices for which low-rank approximation is often applied is the set of positive semidefinite (PSD) matrices. These are real symmetric matrices with all non-negative eigenvalues. They arise for example as covariance matrices, graph Laplacians, Gram matrices (in particular, kernel matrices), and random dot product models [YS07]. In multidimensional scaling, low-rank approximation of PSD matrices in the Frobenius norm error metric (1) corresponds to the standard 'strain minimization' problem [CC00]. Completion of low-rank, or nearly low-rank

(i.e., when $\|A - A_k\|_F^2 \approx 0$), PSD matrices from few entries is important in applications such as quantum state tomography [GLF+10] and global positioning using local distances [SY05, YH38].

Due to its importance, a vast literature studies low-rank approximation of PSD matrices [DM05, ZTK08, KMT09, BW09, LKL10, GM13, WZ13, DLWZ14, WLZ16, TYUC16, LJS16, MM16, CW17]. However, known algorithms either run in at least $\text{nnz}(A)$ time, do not achieve the relative-error guarantee of (1), or require strong incoherence assumptions[1] (see Table 1).

At the same time, the simple $\Omega(\text{nnz}(A))$ time lower bound for general matrices *does not hold in the PSD case*. Positive semidefiniteness ensures that for all i, j, $|A_{i,j}| \leq \max(A_{i,i}, A_{j,j})$. So 'hiding' a large entry in A requires creating a corresponding large diagonal entry. By reading the n diagonal elements, an algorithm can avoid being tricked by this approach. While far from an algorithm, this argument raises the possibility that improved runtimes could be possible for PSD matrices.

1.2 Our Results

We give the first sublinear time relative-error low-rank approximation algorithm for PSD matrices. Our algorithm reads just $nk \cdot \text{poly}(\log n/\epsilon)$ entries of A and runs in $nk^{\omega-1} \cdot \text{poly}(\log n/\epsilon)$ time (Theorem 9). With probability 99/100 it outputs a matrix B in factored form which satisfies (1). We critically exploit the intuition that large entries cannot 'hide' in PSD matrices, but surprisingly require *no additional assumptions* on A, such as incoherence or bounded condition number.

We complement our algorithm with an $\Omega(nk/\epsilon)$ time lower bound. The lower bound is information-theoretic, showing that any algorithm which reads fewer than this number of entries in the input PSD matrix cannot achieve the relative-error guarantee of (1) with constant probability. As our algorithm only reads $nk \cdot \text{poly}(\log n/\epsilon)$ entries of A, this is nearly optimal for constant ϵ. We note that the actual time complexity of our algorithm is slower by a factor of $k^{\omega-2}$.

Finally, we show that our techniques can be extended to compute B satisfying the spectral norm guarantee: $\|A-B\|_2^2 \leq (1+\epsilon)\|A-A_k\|_2^2 + \frac{\epsilon}{k}\|A-A_k\|_F^2$ using just $nk^2 \cdot \text{poly}(\log n/\epsilon)$ accesses to A and $nk^{\omega} \cdot \text{poly}(\log n/\epsilon)$ time (Theorem 25). This guarantee is often stronger than (1) when $\|A-A_k\|_F^2$ is large, and is important in many applications. For example, we use this result to solve the ridge regression problem $\min_{x\in\mathbb{R}^n} \|Ax - y\|_2^2 + \lambda\|x\|_2^2$ up to $(1+\epsilon)$ relative error in $\tilde{O}\left(\frac{ns_\lambda^\omega}{\epsilon^{2\omega}}\right)$ time, where $s_\lambda = \text{tr}((A^2+\lambda I)^{-1}A^2)$ is the *statistical dimension* of the problem (see Theorem 27). Typically $s_\lambda \ll n$, so our runtime is sublinear and improves significantly on existing input-sparsity time results [ACW16]. For a summary of our results and comparison to prior work see, Table 1.

1.3 Algorithm Overview

The starting point for our approach is the fundamental fact that *any* matrix A contains a subset of $O(k/\epsilon)$ columns, call them C, that span a relative-error rank-k approximation to A [DRVW06, DV06, DMM06]. Computing the best low-rank approximation to A using an SVD requires access to all $\Theta(n^2)$ dot products between the columns of the matrix. However, given C, just $n \cdot O(k/\epsilon)$ dot products are needed – to project the remaining columns of the matrix to the span of the subset.

Additionally, a subset of size $\text{poly}(k/\epsilon)$ can be identified using an intuitive approach known as *adaptive sampling* [DV06]: columns are iteratively added to the subset, with each new column being sampled with probability proportional to its norm *outside the column span* of the current subset. Formally, column a_i is selected with probability $\frac{\|a_i - P_C a_i\|_2^2}{\|A-P_C A\|_F^2}$ where P_C is the projection onto the current subset C. Computing these sampling probabilities requires knowing the norm of each

[1] Many of these algorithms satisfy the additional constraint that the low-rank approximation B is PSD. This is also now known to be possible in $O(\text{nnz}(A))$ time using sketching-based algorithms for general matrices [CW17].

[2] Note that this bound is stated incorrectly as $\|A - B\|_F \leq \|A - A_k\|_F + \epsilon\sum_{i=1}^n (A_{ii})^2$ in [DM05].

Source	Runtime	Approximation Bound
[DM05]	$nk^{\omega-1} \cdot \text{poly}(1/\epsilon)$	$\|A-B\|_F \leq \|A-A_k\|_F + \epsilon\|A\|_*$[2]
[KMT09]	$nk^{\omega-1} \cdot \text{poly}(1/\epsilon)$	$\|A-B\|_F \leq \|A-A_k\|_F + \epsilon n \cdot \max_i A_{ii}$
[GM13]	$\tilde{O}(n^2) + nk^{\omega-1}\,\text{poly}(\log n/\epsilon)$	$\|A-B\|_F \leq \|A-A_k\|_F + \epsilon\|A-A_k\|_*$
[AGR16] + [BW09]	$n \log n \cdot \text{poly}(k)$	$\|A-B\|_F \leq (k+1)\|A-A_k\|_*$
[MM16]	$nk^{w-1} \cdot \text{poly}(\log k/\epsilon)$	$\|A^{1/2}-B\|_F^2 \leq (1+\epsilon)\|A^{1/2}-A_k^{1/2}\|_F^2$
[CW17]	$O(\text{nnz}(A)) + n\,\text{poly}(k/\epsilon)$	$\|A-B\|_F^2 \leq (1+\epsilon)\|A-A_k\|_F^2$
Our Results		
Theorem 9	$nk^{\omega-1} \cdot \text{poly}(\log n/\epsilon)$	$\|A-B\|_F^2 \leq (1+\epsilon)\|A-A_k\|_F^2$
Theorem 25	$nk^{\omega} \cdot \text{poly}(\log n/\epsilon)$	$\|A-B\|_2^2 \leq (1+\epsilon)\|A-A_k\|_2^2 + \frac{\epsilon}{k}\|A-A_k\|_F^2$

Table 1: Comparison of our results to prior work on low-rank approximation of PSD matrices. $\|M\|_* = \sum_{i=1}^n \sigma_i(M)$ denotes the nuclear norm of matrix M. The cited results all output B (in factored form), which is itself PSD. In Theorem 12 we show how to modify our algorithm to satisfy this condition, and run in $nk^\omega \cdot \text{poly}(\log n/\epsilon)$ time. The table shows results that do not require incoherence assumptions on A. For general PSD matrices, all known incoherence based results (see e.g., [Git11, GM13]) degrade to $\Theta(n^\omega)$ runtime. Additionally, as discussed, any general low-rank approximation algorithm can be applied to PSD matrices, with state-of-the-art approaches running in input-sparsity time [CW13, MM13, NN13]. [CW17] extends these results to the case where the output B is restricted to be PSD. [TYUC16] does the same but outputs B with rank $2k$. For a more in depth discussion of bounds obtained in prior work, see Section 1.4.

a_i along with its dot product with each column currently in C. So, overall this approach gives a relative-error low-rank approximation using just $n \cdot \text{poly}(k/\epsilon)$ dot products between columns of A.

The above observation is surprising – not only does every matrix contain a small column subset witnessing a near optimal low-rank approximation, but also, such a witness can be found using significantly less information about the column span of the matrix than is required by a full SVD.

This fact is not immediately algorithmically useful, as computing the required dot products takes $\text{nnz}(A) \cdot \text{poly}(k/\epsilon)$ time. However, given PSD A, we can write the eigendecomposition $A = U\Lambda U^T$ where Λ is a non-negative diagonal matrix of eigenvalues, and let $A^{1/2} = U\Lambda^{1/2}U^T$ be the matrix square root of A. Since $A^{1/2}A^{1/2} = A$, the entry $A_{i,j}$ is just the dot product between the i^{th} and j^{th} columns of $A^{1/2}$. So with A in hand, the dot products have been 'precomputed' and the above approach yields a low-rank approximation algorithm for $A^{1/2}$ running in just $n \cdot \text{poly}(k/\epsilon)$ time. Note that, aligning with our initial intuition that reading the diagonal entries of A is necessary to avoid the $\text{nnz}(A)$ time lower bound for general matrices, the diagonal entries of A are the column norms of $A^{1/2}$, and hence their values are critical to computing the adaptive sampling probabilities.

By the above argument, given PSD A, we can compute in $n \cdot \text{poly}(k/\epsilon)$ time a rank-k orthogonal projection matrix $P \in \mathbb{R}^{n \times n}$ (in factored form) for which $\|A^{1/2} - A^{1/2}P\|_F^2 \leq (1+\epsilon)\|A^{1/2} - A_k^{1/2}\|_F^2$. This approach can be implemented using adaptive sampling [DV06], sublinear time volume sampling [AGR16], or as shown in [MM16], recursive *ridge leverage score sampling*. The ridge leverage scores are a natural interpolation between adaptive sampling and the widely studied leverage scores, which, as we will see, have a number of additional algorithmically useful properties. As discussed in [MM16], the guarantee for $A^{1/2}$ is useful for a number of kernel learning methods such as kernel ridge regression. However, it is very different from our final goal. In fact, one can show that projecting to P can yield an *arbitrarily bad* low-rank approximation to A itself (see Appendix A).

We note that, since P is constructed via column selection methods, it is possible to efficiently

compute a factorization of $A^{1/2}PA^{1/2}$ (see Appendix A). Further, this matrix gives a near optimal low-rank approximation of A if we use error parameter $\epsilon' = \epsilon/\sqrt{n}$. This approach gives a first sublinear time algorithm, but it is significantly suboptimal. Namely, it requires reading $\tilde{O}(nk/\epsilon') = \tilde{O}(n^{3/2}k/\epsilon)$ entries of A and takes $n^{1.69} \cdot \text{poly}(k/\epsilon)$ time using fast matrix multiplication.

To improve the dependence on n, we need a better understanding of how to perform ridge leverage score sampling on A itself. We start by showing that the ridge leverage scores of $A^{1/2}$ are within a factor of $O(\sqrt{n/k})$ of the ridge leverage scores of A. By this bound, if we over-sample columns of A by a factor of $O(\sqrt{n/k})$ using the ridge leverage scores of $A^{1/2}$ (computable via [MM16]), obtaining a sample of $\tilde{O}(\sqrt{n/k} \cdot k/\epsilon^2)$ columns, the sample will be a so-called *projection-cost preserving sketch* (PCP) of A. The notion of a PCP was introduced in [CEM+15], where a matrix C is defined to be an (ϵ, k)-column PCP of A if for all rank-k projection matrices P:

$$(1-\epsilon)\|A - PA\|_F^2 \leq \|C - PC\|_F^2 \leq (1+\epsilon)\|A - PA\|_F^2. \tag{2}$$

One important property of a PCP is that good low-rank approximations to C translate to good low-rank approximations of A. More precisely, if U is an $n \times k$ matrix with orthonormal columns for which $\|C - UU^TC\|_F^2 \leq (1+\epsilon)\|C - C_k\|_F^2$, then $\|A - UU^TA\|_F^2 \leq \frac{(1+\epsilon)^2}{(1-\epsilon)}\|A - A_k\|_F^2$.

Letting C be the $n \times \tilde{O}(\sqrt{nk}/\epsilon^2)$ submatrix which we sample via ridge leverage scores, we can apply an $\text{nnz}(C)$ time algorithm to compute a subspace $U \in \mathbb{R}^{n \times k}$ whose columns span a near-optimal low-rank approximation of C, and hence of A by the PCP property. Using standard sampling techniques, we can approximately project the columns of A to U, producing our final solution. This gives time complexity $n^{3/2} \cdot \text{poly}(k/\epsilon)$, improving slightly upon our first approach.

To reduce the time to linear in n, we must further reduce the size of C by sampling a small subset of its rows, which themselves form a PCP. To find these rows, we cannot afford to use oblivious sketching techniques, which would take at least $\text{nnz}(C)$ time, nor can we use our previous method for providing $O(\sqrt{n/k})$ overestimates to the ridge leverage scores, since C is no longer PSD. In fact, the row ridge leverage scores of C can be arbitrarily large compared to those of $A^{1/2}$.

The key idea to getting around this issue is that, since C is a column PCP of A, projecting its columns onto A's top eigenvectors gives a near optimal low-rank approximation. Further, we can show that the ridge leverage scores of $A^{1/2}$ (appropriately scaled) upper bound the *standard leverage scores* of this low-rank approximation. Sampling by these leverage scores is not enough to give a guarantee like (2) – they ignore the entire component of C not falling in the span of A's top eigenvectors and so may significantly distort projection costs over the matrix. Further, at this point, we have no idea how to estimate the row norms of C, or even its Frobenius norm, with $n \,\text{poly}(k/\epsilon)$ samples, which are necessary to implement any kind of adaptive sampling approach.

Fortunately, using that sampling at least preserves the matrix in expectation, along with a few other properties of the ridge leverage scores of $A^{1/2}$, we show that, with good probability, sampling $\tilde{O}(\sqrt{nk}/\,\text{poly}(\epsilon))$ rows of C by these scores yields R satisfying for all rank-k projection matrices P:

$$(1-\epsilon)\|C - CP\|_F^2 \leq \|R - RP\|_F^2 + \Delta \leq (1+\epsilon)\|C - CP\|_F^2$$

where Δ is a fixed value, independent of P, with $|\Delta| \leq c\|C - C_k\|_F^2$ for some constant c. Since the same Δ distortion applies to all P, and since it is at most a constant times the true optimum, a near optimal low-rank approximation for R still translates to a near optimal approximation for C.

At this point R is a small matrix, and we can run any $O(\text{nnz}(R))$ time algorithm to find a good low-rank factorization EF^T to it, where F^T is $k \times \tilde{O}(\sqrt{nk}/\,\text{poly}(\epsilon))$. Since R is a row PCP for C, by regressing the rows of C to the span of F, we can obtain a near optimal low-rank approximation to C. We can solve this multi-response regression approximately in sublinear time via standard

sampling techniques. Approximately regressing A to the span of this approximation using similar techniques yields our final result. The total runtime is dominated by the input-sparsity low-rank approximation of R requiring $O(\text{nnz}(R)) = \tilde{O}(nk/\operatorname{poly}(\epsilon))$ time.

To improve ϵ dependencies in our final runtime, achieving sample complexity $\tilde{O}\left(\frac{nk}{\epsilon^{2.5}}\right)$, we modify this approach somewhat, showing that R actually satisfies a stronger *spectral norm PCP* property for C. This property lets us find a low-rank span Z with $\|C - CZZ^T\|_2^2 \leq \frac{\epsilon}{k}\|A - A_k\|_F^2$, from which, through a series of approximate regression steps, we can extract a low-rank approximation to A satisfying (1). This stronger spectral guarantee also lies at the core of our extensions to near optimal spectral norm low-rank approximation (Theorem 25), ridge regression (Theorem 27), and low-rank approximation where B is restricted to be PSD (Theorem 12).

1.4 Some Further Intuition on Error Guarantees

Observe that in computing a low-rank approximation of A, we read just $\tilde{O}(n \cdot \operatorname{poly}(k/\epsilon))$ entries of the matrix, which is, up to lower order terms, the same number of entries (corresponding to column dot products of $A^{1/2}$) that we accessed to compute a low-rank approximation of $A^{1/2}$ in our description above. However, these sets of entries are very different. While low-rank approximation of $A^{1/2}$ looks at an $n \times \operatorname{poly}(k/\epsilon)$ sized submatrix of A together with the diagonal entries, our algorithm considers a carefully chosen $\sqrt{nk}\operatorname{poly}(\log n/\epsilon) \times \sqrt{nk}\operatorname{poly}(\log n/\epsilon)$ submatrix together with the diagonal entries, which gives significantly more information about the spectrum of A.

As a simple example, consider A with top eigenvalue $\lambda_1 = \sqrt{n}$, and $\lambda_i = 1$ for $i = 2, ...n$. $\|A^{1/2}\|_F^2 = \sum_{i=1}^n \lambda_i = \sqrt{n} + n - 1$ while $\|A^{1/2} - A_1^{1/2}\|_F^2 = \sum_{i=2}^n \lambda_i = n - 1$. So, $A^{1/2}$ has no good rank-1 approximation. Unless we set $\epsilon = O(1/\sqrt{n})$, a low-rank approximation algorithm for $A^{1/2}$ can learn nothing about λ_1 and still be near optimal. In contrast, $\|A\|_F^2 = \sum_{i=1}^n \lambda_i^2 = 2n - 1$ and $\|A - A_1\|_F^2 = \sum_{i=2}^n \lambda_i^2 = n - 1$. So, even with $\epsilon = 1/2$, any rank-1 approximation algorithm for A must identify the presence of λ_1 and project this direction off the matrix. In this sense, our algorithm is able to obtain a much more accurate picture of A's spectrum.

With incoherence assumptions, prior work on PSD low-rank approximation [GM13] obtains the bound $\|A - B\|_* \leq (1 + \epsilon)\|A - A_k\|_*$ in sublinear time, where $\|M\|_* = \sum_{i=1}^n \sigma_i(M)$ is the nuclear norm of M. Recent work ([AGR16] in combination with [BW09]) gives $\|A - B\|_F \leq (k + 1)\|A - A_k\|_*$ without the incoherence assumption. These nuclear norm bounds are closely related to approximation bounds for $A^{1/2}$ and it is not hard to see that neither require λ_1 to be detected in the example above, and so in this sense are weaker than our Frobenius norm bound.

A natural question if even stronger bounds are possible: e.g., can we compute B with $\|A - B\|_2^2 \leq (1 + \epsilon)\|A - A_k\|_2^2$ in sublinear time? We partially answer this question in Theorem 25. In $\tilde{O}(nk^\omega \operatorname{poly}(\log n/\epsilon))$ time, we can find B satisfying $\|A - B\|_2^2 \leq (1 + \epsilon)\|A - A_k\|_2^2 + \frac{\epsilon}{k}\|A - A_k\|_F^2$.

Significantly improving the above bound seems difficult: it is easy to see that a relative error spectral norm guarantee requires $\Omega(n^2)$ time. Consider A which is the identity except with $A_{i,j} = A_{j,i} = 1$ for some uniform random pair (i, j). Finding (i, j) requires $\Omega(n^2)$ queries to A. However, it is necessary to achieve a relative error spectral norm guarantee with $\epsilon < 3$ since $\|A\|_2^2 = 4$ while $\|A - A_1\|_2^2 = 1$ where A_1 is all zeros with ones at its (i, i), (j, j), (i, j), and (j, i) entries.

A similar argument shows that relative error low-rank approximation in higher Schatten-p norms, i.e., $\|A - B\|_p^p$ for $p > 2$ requires superlinear dependence on n (where $\|M\|_p^p = \sum_{i=1}^n \sigma_i^p(M)$.) We can set A to be the identity but with an all ones block on a uniform random subset of $n^{1/p}$ indices. This block has associated eigenvalue $\lambda_1 = n^{1/p}$ and so, since all other $(n - n^{1/p})$ eigenvalues of A are 1, $\|A\|_p^p = \Theta(n)$, and the block must be recovered to give a relative error approximation to $\|A - A_1\|_p^p$. However, as the block is placed uniformly at random and contains just $n^{2/p}$ entries, finding even a single entry requires $n^{2-2/p}$ queries to A – superlinear for $p > 2$.

1.5 Open Questions

While it is apparent that obtaining stronger error guarantees than (1) may require increased runtime, understanding exactly what can be achieved in sublinear time is an interesting direction for future work. We also note that it is still unknown how to compute a number of basic properties of PSD matrices in sublinear time. For example, while we can output B satisfying $\|A - B\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2$, surprisingly it is not clear how to actually estimate the value $\|A - A_k\|_F^2$ to within a $(1 \pm \epsilon)$ factor. This can be achieved in $n^{3/2}\,\mathrm{poly}(k/\epsilon)$ time using our PCP techniques. However, obtaining linear runtime in n is open. Estimating $\|A - A_k\|_F^2$ seems strongly connected to estimating other important quantities such as the statistical dimension of A for ridge regression (see Theorem 27) which we do not know how to do in $o(n^{3/2})$ time.

Finally, an open question is if these techniques can be generalized to a broader class of matrices. As discussed, in the matrix completion literature, much attention has focused on incoherent low-rank matrices [CR09] which can be approximated with uniform sampling. PSD matrices are not incoherent in general, which is highlighted by the fact that our sampling schemes are far from uniform and very adaptive to previously seen matrix entries. However, perhaps there is some other parameter (maybe relating to a measure of diagonal dominance) which characterizes when low-rank approximation can be performed with just a small number of adaptive accesses to A.

1.6 Paper Outline

Section 2: Ridge Leverage Score Sampling. We show that the ridge leverage scores of A are within an $O(\sqrt{n/k})$ factor of those of $A^{1/2}$, letting us use the fast ridge leverage score sampling algorithm of [MM16] to sample $\tilde{O}(\sqrt{nk}/\epsilon^2)$ columns of A that form a column PCP of the matrix.

Section 3: Row Sampling. We discuss how to further accelerate our algorithm by obtaining a row PCP for our column sample, allowing us to achieve runtime linear in n.

Section 4: Full Algorithm. We use the primitives in the previous sections along with approximate regression techniques to give our full sublinear time low-rank approximation algorithm.

Section 5: Lower Bounds. We show that our algorithm is nearly optimal – any relative error low-rank approximation algorithm must read $\Omega(nk/\epsilon)$ entries of A.

Section 6: Spectral Norm Bounds. We modify the algorithm of Section 4 to give a tighter approximation in the spectral norm and discuss applications to sublinear time ridge regression.

2 Ridge Leverage Score Sampling

Our main algorithmic tool will be ridge leverage score sampling, which is used to identify a small subset of columns of A that span a good low-rank approximation of the matrix. Following the definition of [CMM17], the rank-k ridge leverage scores of any matrix A are given by:

Definition 1 (Ridge Leverage Scores)**.** *For any $A \in \mathbb{R}^{n\times d}$, letting $a_i \in \mathbb{R}^n$ be the i^{th} column of A, the i^{th} rank-k column ridge leverage score of A is:*

$$\tau_i^k(A) = a_i^T \left(AA^T + \frac{\|A - A_k\|_F^2}{k} I \right)^+ a_i.$$

Above I is the appropriately sized identity matrix and M^+ denotes the matrix pseudoinverse, equivalent to the inverse unless $\|A - A_k\|_F^2 = 0$ and A is singular. Analogous scores can be defined

for the rows of A by simply transposing the matrix. It is not hard to see that $0 < \tau_i^k(A) < 1$ for all i. Since we use these scores as sampling probabilities, it is critical that the sum of scores, and hence the size of the subsets we sample, is not too large. We have the following (see Appendix B):

Lemma 2 (Sum of Ridge Leverage Scores)**.** *For any $A \in \mathbb{R}^{n\times d}$, $\sum_{i=1}^{d} \tau_i^k(A) \leq 2k$.*

Intuitively, the ridge leverage scores are similar to the standard leverage scores of A, which are given by $a_i^T(AA^T)^+a_i$. By writing $A = U\Sigma V^T$ in its SVD, one sees that standard leverage scores are just the squared column norms of V^T. Sampling columns by ridge leverage scores yields a spectral approximation to the matrix. The addition of the weighted identity (or 'ridge') $\frac{\|A-A_k\|_F^2}{k}I$ 'dampens' contributions from smaller singular directions of A, decreasing the sum of the scores and allowing us to sample fewer columns. At the same time, it introduces error dependent on the size of the tail $\|A - A_k\|_F^2$, ultimately giving an approximation from which it is possible to output a near optimal low-rank approximation to the original matrix. Specifically, sampling by ridge leverage scores yields a *projection-cost preserving* sketch (PCP) of A:

Lemma 3 (Theorem 6 of [CMM17])**.** *For any $A \in \mathbb{R}^{n\times d}$, for $i \in \{1, \ldots, d\}$, let $\tilde{\tau}_i^k \geq \tau_i^k(A)$ be an overestimate for the i^{th} rank-k ridge leverage score. Let $p_i = \frac{\tilde{\tau}_i^k}{\sum_i \tilde{\tau}_i^k}$ and $t = \frac{c\log(k/\delta)}{\epsilon^2}\sum_i \tilde{\tau}_i^k$ for any $\epsilon < 1$ and sufficiently large constant c. Construct C by sampling t columns of A, each set to $\frac{1}{\sqrt{tp_i}}a_i$ with probability p_i. With probability $1-\delta$, for any rank-k orthogonal projection $P \in \mathbb{R}^{n\times n}$,*

$$(1-\epsilon)\|A - PA\|_F^2 \leq \|C - PC\|_F^2 \leq (1+\epsilon)\|A - PA\|_F^2.$$

We refer to C as an (ϵ, k)-column PCP of A.

Since the 'cost' $\|A - PA\|_F^2$ of any rank-k projection of A is preserved by C, any near-optimal low-rank approximation of C yields a near optimal low-rank approximation of A. Further, C is much smaller than A, so such a low-rank approximation can be computed quickly. The difficulty is in computing the approximate leverage scores. To do this, we use the main result from [MM16]:

Lemma 4 (Corollary of Theorem 20 of [MM16])**.** *There is an algorithm that given any PSD matrix $A \in \mathbb{R}^{n\times n}$, runs in $O(n(k\log(k/\delta))^{\omega-1})$ time, accesses $O(nk\log(k/\delta))$ entries of A, and returns for each $i \in [1, .., n]$, $\tilde{\tau}_i^k(A^{1/2})$ such that with probability $1-\delta$, for all i:*

$$\tau_i^k(A^{1/2}) \leq \tilde{\tau}_i^k(A^{1/2}) \leq 3\tau_i^k(A^{1/2}).$$

Proof. Theorem 20 of [MM16] shows that by using a recursive ridge leverage score sampling algorithm, it is possible to return (with probability $1-\delta$) a sampling matrix $S \in \mathbb{R}^{n\times s}$ with $s = O\left(k\log(k/\delta)\right)$ such that, letting $\lambda = \frac{1}{k}\|A^{1/2} - A_k^{1/2}\|_F^2$:

$$\frac{1}{2}\left(A + \lambda I\right) \preceq \left(A^{1/2}SS^TA^{1/2} + \lambda I\right) \preceq \frac{3}{2}\left(A + \lambda I\right)$$

where $M \preceq N$ indicates $x^TMx \leq x^TNx$ for all x. If we set $\tilde{\tau}_i^k(A^{1/2}) = 2{\cdot}x_i^T\left(A^{1/2}SS^TA^{1/2} + \lambda I\right)^+ x_i$, where x_i is the i^{th} column of $A^{1/2}$ we have the desired bound. Of course, we cannot directly compute this value without factoring A to form $A^{1/2}$. However, as shown in Lemma 6 of [MM16]:

$$x_i^T\left(A^{1/2}SS^TA^{1/2} + \lambda I\right)^{-1}x_i = \frac{1}{\lambda}\left(A - AS(S^TAS + \lambda I)^{-1}S^TA\right)_{i,i}.$$

Computing $(S^TAS + \lambda I)^{-1}$ requires $O(s^2) = O((k\log(k/\delta))^2)$ accesses to A and $O(s^\omega) = O((k\log(k/\delta))^\omega)$ time. Computing all n diagonal entries of $AS(S^TAS + \lambda I)^{-1}S^TA$ then requires

$O(nk\log(k/\delta))$ accesses to A and $O(n(k\log(k/\delta))^{\omega-1})$ time. With these entries in hand we can simply subtract from the diagonal entries of A and rescale to give the final leverage score approximation. Critically, this calculation always reads *all diagonal entries* of A, allowing it to identify rows containing large off diagonal entries and skirt the $\text{nnz}(A)$ time lower bound for general matrices.

Note that the stated runtime in [MM16] for outputting S is $\tilde{O}(nk)$ accesses to A (*kernel evaluations* in the language of [MM16]) and $\tilde{O}(nk^2)$ runtime. However this runtime is improved to $\tilde{O}(nk^{\omega-1})$ using fast matrix multiplication. □

In order to apply Lemmas 3 and 4 to low-rank approximation of A, we now show that the ridge leverage scores of $A^{1/2}$ coarsely approximate those of A:

Lemma 5 (Ridge Leverage Score Bound)**.** *For any PSD matrix $A \in \mathbb{R}^{n\times n}$:*

$$\tau_i^k(A) \leq 2\sqrt{\frac{n}{k}} \cdot \tau_i^k(A^{1/2}).$$

Proof. We write $A^{1/2}$ in its eigendecomposition $A^{1/2} = U\Lambda^{1/2}U^T$, where $\Lambda_{i,i} = \lambda_i$ is the i^{th} eigenvalue of A. Letting x_i denote the i^{th} column of $A^{1/2}$ we have:

$$\tau_i^k(A^{1/2}) = x_i^T\left(A + \frac{\|A^{1/2} - A_k^{1/2}\|_F^2}{k}I\right)^{-1} x_i = x_i^T U\bar{\Lambda}U^T x_i$$

where $\bar{\Lambda}_{i,i} \stackrel{\text{def}}{=} \frac{1}{\lambda_i + \frac{1}{k}\sum_{j=k+1}^n \lambda_j}$. We can similarly write:

$$\begin{aligned}
\tau_i^k(A) &= a_i^T\left(A^2 + \frac{\|A - A_k\|_F^2}{k}I\right)^{-1} a_i \\
&= x_i^T A^{1/2}\left(A^2 + \frac{\|A - A_k\|_F^2}{k}I\right)^{-1} A^{1/2}x_i \\
&= x_i^T U\hat{\Lambda}U^T x_i
\end{aligned}$$

where $\hat{\Lambda}_{i,i} \stackrel{\text{def}}{=} \frac{\lambda_i}{\lambda_i^2 + \frac{1}{k}\sum_{j=k+1}^n \lambda_j^2}$. Showing $\hat{\Lambda} \preceq 2\sqrt{\frac{n}{k}} \cdot \bar{\Lambda}$ is enough to give the lemma. Specifically we must show, for all i, $\hat{\Lambda}_{i,i} \leq 2\sqrt{\frac{n}{k}} \cdot \bar{\Lambda}_{i,i}$ which after cross-multiplying is equivalent to:

$$\lambda_i^2 + \frac{1}{k}\lambda_i \sum_{j=k+1}^n \lambda_j \leq 2\sqrt{\frac{n}{k}}\left(\lambda_i^2 + \frac{1}{k}\sum_{j=k+1}^n \lambda_j^2\right). \tag{3}$$

First consider the relatively large eigenvalues. Say we have $\frac{1}{k}\sum_{j=k+1}^n \lambda_j \leq \sqrt{\frac{n}{k}}\lambda_i$. Then:

$$\lambda_i^2 + \frac{1}{k}\lambda_i \sum_{j=k+1}^n \lambda_j \leq \left(1 + \sqrt{n/k}\right)\lambda_i^2$$

which gives (3). Next consider small eigenvalues with $\frac{1}{k}\sum_{j=k+1}^{n}\lambda_j \geq \sqrt{\frac{n}{k}}\lambda_i$. In this case:

$$\begin{aligned}\lambda_i^2 + \frac{1}{k}\lambda_i \sum_{j=k+1}^{n}\lambda_j &\leq \lambda_i^2 + \frac{1}{\sqrt{n}\cdot k^{3/2}}\left(\sum_{j=k+1}^{n}\lambda_j\right)^2 \\ &\leq \lambda_i^2 + \frac{1}{\sqrt{n}\cdot k^{3/2}}\cdot n\sum_{j=k+1}^{n}\lambda_j^2 \qquad \text{(Norm bound: } \|\cdot\|_1^2 \leq n\|\cdot\|_2^2\text{)} \\ &\leq \sqrt{\frac{n}{k}}\left(\lambda_i^2 + \frac{1}{k}\sum_{j=k+1}^{n}\lambda_j^2\right)\end{aligned}$$

which gives (3), completing the proof. □

Combining Lemmas 2, 3, 4, 5 we have:

Corollary 6 (Fast PSD Ridge Leverage Score Sampling)**.** *There is an algorithm that given any PSD matrix $A \in \mathbb{R}^{n\times n}$ runs in $\tilde{O}(nk^{\omega-1})$ time, accesses $\tilde{O}(nk)$ entries of A, and with prob. $1-\delta$ outputs a weighted sampling matrix $S_1 \in \mathbb{R}^{n\times \tilde{O}\left(\frac{\sqrt{nk}}{\epsilon^2}\right)}$ such that AS_1 is an (ϵ,k)-column PCP of A.*

Proof. By Lemma 4 we can compute constant factor approximations to the ridge leverage scores of $A^{1/2}$ in time $\tilde{O}(nk^{\omega-1})$. Applying Lemma 5, if we scale these scores up by $2\sqrt{n/k}$ they will be overestimates of the ridge leverage scores of A. If we set $t = O\left(\frac{\log(k/\delta)}{\epsilon^2}\cdot\sum\tilde{\tau}_i^k\right)$, and generate S_1 by sampling t columns of A with probabilities proportional to these estimated scores, by Lemma 3, AS_1 will be an (ϵ,k)-column PCP of A with probability $1-\delta$. By Lemma 2, $\sum_{i=1}^{n}\tau_i^k(A^{1/2}) \leq 2k$. So we have $t = \tilde{O}(\sum\tilde{\tau}_i^k/\epsilon^2) = \tilde{O}(\sqrt{nk}/\epsilon^2)$. □

Forming AS_1 requires reading just $\tilde{O}(n^{3/2}\sqrt{k}/\epsilon^2)$ entries of A. At this point, we could employ any input sparsity time algorithm to find a near optimal rank-k projection P for approximating AS_1 in $O(\text{nnz}(AS_1)) + n\,\text{poly}(k/\epsilon) = n^{3/2}\cdot\text{poly}(k/\epsilon)$ time. This would in turn yield a near optimal low-rank approximation of A. However, as we will see in the next section, by further sampling the rows of AS_1, we can significantly improve this runtime.

3 Row Sampling

To achieve near linear dependence on n, we sample roughly $\sqrt{nk}$ rows from AS_1, producing an even smaller matrix $S_2^TAS_1$, which we can afford to fully read and from which we can form a near optimal low-rank approximation to AS_1 and consequently to A. However, sampling AS_1 is challenging: we cannot employ input sparsity time methods as we cannot afford to read the full matrix, and since it is no longer PSD, we cannot apply the same approach we used for A, approximating the ridge leverage scores with those of $A^{1/2}$.

Rewriting Definition 1 using the SVD $AS_1 = U\Sigma V^T$ (and transposing AS_1 to give row instead of column scores) we see that the row ridge leverage scores of AS_1 are the diagonal entries of:

$$AS_1\left(S_1^TA^TAS_1 + \frac{\|AS_1 - (AS_1)_k\|_F^2}{k}I\right)^+ S_1^TA^T = U\bar{\Sigma}U^T$$

where $\bar{\Sigma}_{i,i} = \frac{\Sigma_{i,i}^2}{\Sigma_{i,i}^2 + \frac{\|AS_1-(AS_1)_k\|_F^2}{k}}$. That is, the row ridge leverage scores depend only on the column span U of AS_1 and its spectrum. Since AS_1 is a column PCP of A this gives hope that the two matrices have similar row ridge leverage scores.

Unfortunately, this is *not the case*. It is possible to have rows in AS_1 with ridge leverage scores significantly higher than in A. Thus, even if we knew the ridge leverage scores of A, we would have to scale them up significantly to sample from AS_1. As an example, consider A with relatively uniform ridge leverage scores: $\tau_i(A) \approx k/n$ for all i. When a column is selected to be included in AS_1 it will be reweighted by roughly a factor of $\sqrt{n/k}$. Now, append a number of rows to A each with very small norm and just a containing single non-zero entry. These rows will have little effect on the ridge leverage scores if their norms are small enough. However, if the column corresponding to the nonzero in a row is selected, the row will appear in AS_1 with $\sqrt{n/k}$ times the weight that it appears in A, and its ridge leverage score will be roughly a factor n/k times higher.

Fortunately, we are still able to show that sampling the rows of AS_1 by the rank $k' = O(k/\epsilon)$ leverage scores of $A^{1/2}$ scaled up by a $\sqrt{n/k'}$ factor yields a row PCP for this matrix. Our proof works not with the ridge scores of AS_1 but with the *standard leverage scores* of a near optimal low-rank approximation to this matrix – specifically the approximation given by projecting onto the top eigenvectors of A. We have:

Lemma 7 (Row PCP)**.** *For any PSD $A \in \mathbb{R}^{n\times n}$ and $\epsilon \leq 1$ let $k' = \lceil ck/\epsilon \rceil$ and let $\tilde{\tau}_i^{k'}(A^{1/2}) \geq \tau_i^{k'}(A^{1/2})$ be an overestimate for the i^{th} rank-k' ridge leverage score of $A^{1/2}$. Let $\tilde{\ell}_i = \sqrt{\frac{16n\epsilon}{k}} \cdot \tilde{\tau}_i^{k'}(A^{1/2})$, $p_i = \frac{\tilde{\ell}_i}{\sum_i \tilde{\ell}_i}$, and $t = \frac{c' \log n}{\epsilon^2} \sum_i \tilde{\ell}_i$. Construct weighted sampling matrices $S_1, S_2 \in \mathbb{R}^{n\times t}$ each whose j^{th} column is set to $\frac{1}{\sqrt{tp_i}} e_i$ with probability p_i.*

For sufficiently large constants c, c', with probability $\frac{99}{100}$, letting $\tilde{A} = S_2^T A S_1$, for any rank-k orthogonal projection $P \in \mathbb{R}^{t\times t}$:

$$(1-\epsilon)\|AS_1(I-P)\|_F^2 \leq \|\tilde{A}(I-P)\|_F^2 + \Delta \leq (1+\epsilon)\|AS_1(I-P)\|_F^2$$

for some fixed Δ (independent of P) with $|\Delta| \leq 600\|A - A_k\|_F^2$. We refer to $\tilde{A}$ as an (ϵ, k)-row PCP of AS_1.

Note that by Lemma 2, $\sum_i \tau_i^{k'}(A^{1/2}) = O(k/\epsilon)$. So if $\tilde{\tau}_i^{k'}(A^{1/2})$ is a constant factor approximation to $\tau_i^{k'}(A^{1/2})$, $t = O\left(\frac{\sqrt{nk}\log n}{\epsilon^{2.5}}\right)$. Also note that the Lemma requires both S_1 and S_2 to be sampled using the rank k' ridge scores. If we sample S_1 using a sum of the rank-k and rank-k' ridge scores (appropriately scaled) Lemma 7 and Lemma 3 will hold simultaneously.

By applying an input sparsity time low-rank approximation algorithm to $\tilde{A}$ (which has just $\tilde{O}\left(\frac{nk}{\epsilon^5}\right)$ entries) we can find a near optimal low-rank approximation of AS_1, and thus for A. However, in our final algorithm, we will take a somewhat different approach. We are able to show that using appropriate sampling probabilities, we can in fact sample $\tilde{A}$ which is a projection-cost preserving sketch of AS_1 for *spectral norm* error. As we will see, recovering a near optimal spectral norm low-rank approximation to AS_1 suffices to recover a near optimal Frobenius norm approximation to A, and allows us to improve ϵ dependencies in our final runtime.

Lemma 8 (Spectral Norm Row PCP)**.** *For any PSD $A \in \mathbb{R}^{n\times n}$, and $\epsilon < 1$ let $k' = \lceil ck/\epsilon^2 \rceil$ and $\tilde{\tau}_i^{k'}(A^{1/2}) \geq \tau_i^{k'}(A^{1/2})$ be an overestimate for the i^{th} rank-k' ridge leverage score of $A^{1/2}$. Let $\tilde{\ell}_i = 4\epsilon\sqrt{\frac{n}{k}}\tau_i^{k'}(A^{1/2})$, $p_i = \frac{\tilde{\ell}_i}{\sum_i \tilde{\ell}_i}$, and $t = \frac{c' \log n}{\epsilon^2} \cdot \sum_i \tilde{\ell}_i$. Construct weighted sampling matrices $S_1, S_2 \in \mathbb{R}^{n\times t}$, each whose j^{th} column is set to $\frac{1}{\sqrt{tp_i}} e_i$ with probability p_i.*

For sufficiently large constants c, c', with high probability (i.e. probability $\geq 1 - 1/n^d$ for some large constant d), letting $\tilde{A} = S_2^T A S_1$, for any orthogonal projection $P \in \mathbb{R}^{t\times t}$:

$$(1-\epsilon)\|AS_1(I-P)\|_2^2 - \frac{\epsilon}{k}\|A - A_k\|_F^2 \leq \|\tilde{A}(I-P)\|_2^2 \leq (1+\epsilon)\|AS_1(I-P)\|_2^2 + \frac{\epsilon}{k}\|A - A_k\|_F^2.$$

We refer to $\tilde{A}$ as an (ϵ, k)-spectral PCP of AS_1.

Note that if $\tilde{\tau}_i^{k'}(A^{1/2})$ is a constant factor approximation to $\tau_i^{k'}(A^{1/2})$, $t = O\left(\frac{\sqrt{nk}\log n}{\epsilon^3}\right)$. We defer the proofs of Lemmas 7 and 8 to Appendix B.

4 Full Low-Rank Approximation Algorithm

We are finally ready to give our main algorithm for relative error low-rank approximation of PSD matrices in $\tilde{O}(n\,\mathrm{poly}(k/\epsilon))$ time, Algorithm 1. We set $k_1 \stackrel{\text{def}}{=} \lceil ck/\epsilon \rceil$ and estimate the both the rank-k and rank-$c'k_1$ ridge leverage scores of $A^{1/2}$ using the algorithm of [MM16] (Step 1). If c, c' are sufficiently large, sampling by the sum of these scores (Steps 2-3) ensures that AS_1 is an (ϵ, k)-column PCP for A and simultaneously, by applying Lemmas 7 and 8 that $\tilde{A} = S_2^T AS_1$ is a row PCP in both spectral and Frobenius norm with rank k_1 and error $\epsilon = 1/2$ for AS_1

In conjunction, these guarantees ensure that we can apply an input sparsity time algorithm to $\tilde{A}$ (Step 4) to find a rank-k_1 Z satisfying $\|AS_1 - AZZ^T\|_2^2 = O(\|AS_1 - (AS_1)_{k_1}\|_2^2 + \frac{1}{k_1}\|A - A_k\|_F^2) = O\left(\frac{\epsilon}{k}\|A - A_k\|_F^2\right)$, where the final bound holds since $k_1 = \Theta(k/\epsilon)$. It is not hard to show that, due to this strong spectral norm bound, projecting AS_1 to Z and taking the best rank-k approximation in the span gives a near optimal Frobenius norm low-rank approximation to AS_1 and hence A.

We can still not afford to read AS_1 in its entirety, so we employ a number of standard leverage score sampling techniques to perform this projection approximately. In Step 5, we sample $\tilde{O}(k/\epsilon^2)$ columns of AS_1 using the leverage scores of Z (its row norms since it is an orthonormal matrix) to form AS_1S_3. We argue that there is a good rank-k approximation to AS_1 lying in both the column span of AS_1S_3 and the row span of Z^T. In Step 6 we find a near optimal such approximation by further sampling $\tilde{O}(k/\epsilon^4)$ rows AS_1 by the leverage scores of AS_1S_3 (the row norms of V, an orthonormal basis for its span), and computing the best rank-k approximation to the sampled matrix falling in the column span of AS_1S_3 and the row span of Z^T.

Finally, in Step 7 we approximately project A itself to the span of this rank-k approximation by first sampling by the leverage scores of the approximation (the row norms of Q) and projecting.

Algorithm 1 PSD Low-Rank Approximation

1. Let $k_1 = \lceil ck/\epsilon \rceil$. For all $i \in [1, .., n]$ compute $\tilde{\tau}_i^k(A^{1/2})$ and $\tilde{\tau}_i^{c'k_1}(A^{1/2})$ which are constant factor approximations to the ridge leverage scores $\tau_i^k(A^{1/2})$ and $\tau_i^{c'k_1}(A^{1/2})$ respectively.
2. Set $\ell_i^{(1)} = \sqrt{\frac{n}{k}}\tilde{\tau}_i^k(A^{1/2}) + \sqrt{\frac{n\epsilon^4}{k_1}}\tilde{\tau}_i^{c'k_1}(A^{1/2})$ and $\ell_i^{(2)} = \sqrt{\frac{n}{k_1}}\tilde{\tau}_i^{c'k_1}(A^{1/2})$. Set $p_i^{(1)} = \frac{\ell_i^{(1)}}{\sum_i \ell_i^{(1)}}$ and $p_i^{(2)} = \frac{\ell_i^{(2)}}{\sum_i \ell_i^{(2)}}$.
3. Set $t_1 = \frac{c_1 \log n}{\epsilon^2}\sum_i \ell_i^{(1)}$ and $t_2 = c_2 \log n \sum_i \ell_i^{(2)}$. Sample $S_1 \in \mathbb{R}^{n \times t_1}$ whose j^{th} column is set to $\frac{1}{\sqrt{tp_i^{(1)}}}e_i$ with probability $p_i^{(1)}$. Sample $S_2 \in \mathbb{R}^{n \times t_2}$ analogously using $p_i^{(2)}$.
4. Let $\tilde{A} = S_2^T AS_1$, and use an input sparsity time algorithm to compute orthonormal $Z \in \mathbb{R}^{t_1 \times k_1}$ satisfying the spectral guarantee $\|\tilde{A} - \tilde{A}ZZ^T\|_2^2 \le 2\|\tilde{A} - \tilde{A}_{k_1}\|_2^2 + \frac{2}{k_1}\|\tilde{A} - \tilde{A}_{k_1}\|_F^2$.
5. Let $t_3 = c_3\left(\frac{k\log(k/\epsilon)}{\epsilon} + \frac{k}{\epsilon^2}\right)$, set $p_i^{(3)} = \frac{\|z_i\|_2^2}{\|Z\|_F^2}$, and sample $S_3 \in \mathbb{R}^{t_1 \times t_3}$ where the j^{th} column is set to $\frac{1}{\sqrt{t_3 p_i^{(3)}}}e_i$ with probability $p_i^{(3)}$. Compute $V \in \mathbb{R}^{n \times t_3}$ which is an orthogonal basis for the column span of AS_1S_3.

6. Let $p_i^{(4)} = \frac{\|v_i\|_2^2}{\|V\|_F^2}$ and $t_4 = c_4\left(\frac{t_3 \log t_3}{\epsilon^2}\right)$. Sample $S_4 \in \mathbb{R}^{n \times t_4}$ where the j^{th} column is set to $\frac{1}{\sqrt{t_4 p_i^{(4)}}} e_i$ with probability $p_i^{(4)}$. Compute $W \in \mathbb{R}^{t_3 \times t_1}$ satisfying:

$$W = \underset{W|\,\mathrm{rank}(W)=k}{\arg\min} \|S_4^T A S_1 S_3 W Z^T - S_4^T A S_1\|_F^2.$$

7. Compute an orthogonal basis $Q \in \mathbb{R}^{n \times k}$ for the column span of AS_1S_3W. Let $t_5 = c_5\left(k \log k + \frac{k}{\epsilon}\right)$, set $p_i^{(5)} = \frac{\|q_i\|_2^2}{\|Q\|_F^2}$, where q_i is the i^{th} row of Q. Sample $S_5 \in \mathbb{R}^{n \times t_5}$ where the j^{th} column is set of $\frac{1}{\sqrt{t_5 p_i^{(5)}}} e_i$ with probability $p_i^{(5)}$. Solve:

$$N = \underset{N \in \mathbb{R}^{n \times k}}{\arg\min} \|S_5^T Q N^T - S_5^T A\|_F^2.$$

8. Return $Q, N \in \mathbb{R}^{n \times k}$.

Theorem 9 (Sublinear Time Low-Rank Approximation)**.** *Given any PSD $A \in \mathbb{R}^{n \times n}$, for sufficiently large constants $c, c', c_1, c_2, c_3, c_4, c_5$, for any $\epsilon < 1$, Algorithm 1 accesses $O(\frac{n \cdot k \log^2 n}{\epsilon^{2.5}} + \sqrt{n}k^{1.5} \cdot \log^2 n \cdot \mathrm{poly}(1/\epsilon))$ entries of A, runs in $\tilde{O}\left(\frac{nk^{\omega-1}}{\epsilon^{2(\omega-1)}} + \sqrt{n}k^{\omega-.5} \cdot \mathrm{poly}(1/\epsilon)\right)$ time, and with probability at least $9/10$ outputs $M, N \in \mathbb{R}^{n \times k}$ with:*

$$\|A - MN^T\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2.$$

For simplicity, we will show $\|A - MN^T\|_F^2 \leq (1 + O(\epsilon))\|A - A_k\|_F^2$. By scaling up the constants $c, c', c_1, c_2, c_3, c_4, c_5$ we can then scale down ϵ, achieving the claimed result. We first show:

Lemma 10. *For sufficiently large constants c, c', c_1, c_2, with probability $98/100$, Steps 1-4 of Algorithm 1 produce $Z \in \mathbb{R}^{t_1 \times k_1}$ satisfying:*

$$\|AS_1 - AS_1 Z Z^T\|_2^2 \leq \frac{\epsilon}{k}\|A - A_k\|_F^2.$$

Proof. In Step 3, S_1 is sampled using probabilities proportional to the scores $\ell_i^{(1)} = \sqrt{\frac{n}{k}}\tilde{\tau}_i^k(A^{1/2}) + \sqrt{\frac{n\epsilon^4}{k_1}}\tilde{\tau}_i^{c'k_1}(A^{1/2})$. The first part of this score ensures that by Lemma 3, with high probability AS_1 is an (ϵ, k)-column PCP of A. The second half, in combination with the construction of S_2 ensures via Lemmas 7 and 8 that if c' is large enough, $\tilde{A} = S_2^T A S_1$ is both a $(1/2, k_1)$-row PCP of AS_1 for Frobenius norm error with probability $99/100$ and a $(1/2, k_1)$-spectral PCP of AS_1 with high probability. Note that in $\ell_i^{(1)}$ the scores corresponding to $\tilde{\tau}_i^{c'k_1}(A^{1/2})$ are scaled down by an ϵ^2 factor from what is required to apply Lemmas 7 and 8 with rank k_1 and error $1/2$. However, t_1 is oversampled by a $1/\epsilon^2$ factor, balancing this scaling. By a union bound, all three PCP properties hold simultaneously with probability $98/100$.

By the Frobenius error PCP property of Lemma 7, letting U_{k_1} contain the top k_1 row singular vectors of AS_1 we have:

$$\begin{aligned}\|\tilde{A} - \tilde{A}_{k_1}\|_F^2 &\leq \|\tilde{A} - \tilde{A}U_{k_1}U_{k_1}^T\|_F^2 \\ &\leq \frac{3}{2}\|AS_1 - (AS_1)_{k_1}\|_F^2 - \Delta \leq 603\|A - A_k\|_F^2 \end{aligned} \tag{4}$$

by the fact that $|\Delta| \leq 600\|A - A_k\|_F^2$ and $\|AS_1 - (AS_1)_{k_1}\|_F^2 \leq \|AS_1 - (AS_1)_k\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2 \leq 2\|A - A_k\|_F^2$ since AS_1 is an (ϵ, k)-column PCP of A. Further, via Lemma 8, for any rank-k_1 projection P:

$$\frac{1}{2}\|AS_1(I-P)\|_2^2 - \frac{1}{2k_1}\|A - A_{k_1}\|_F^2 \leq \|\tilde{A}(I-P)\|_2^2 \leq \frac{3}{2}\|AS_1(I-P)\|_2^2 + \frac{1}{2k_1}\|A - A_{k_1}\|_F^2. \quad (5)$$

By this bound, we have

$$\|\tilde{A} - \tilde{A}_{k_1}\|_2^2 \leq \|\tilde{A} - \tilde{A}U_{k_1}U_{k_1}^T\|_2^2 \leq \frac{3}{2}\|AS_1 - (AS_1)_{k_1}\|_2^2 + \frac{1}{2k_1}\|A - A_{k_1}\| \leq \frac{4\epsilon}{k}\|A - A_k\|_F^2 \quad (6)$$

where the final bound follows because if we set $c \geq 2$, $k_1 \geq \lceil 2k/\epsilon \rceil$ so $\|AS_1 - (AS_1)_{k_1}\|_2^2 \leq \frac{\epsilon}{k}\|AS_1 - (AS_1)_k\|_F^2 \leq \frac{2\epsilon}{k}\|A - A_k\|_F^2$. With (4) and (6) in place, applying (5) again, if we compute Z satisfying the guarantee in Step 4 we have:

$$\begin{aligned}
\|AS_1 - AS_1ZZ^T\|_2^2 &\leq 2\|\tilde{A} - \tilde{A}ZZ^T\|_2^2 + \frac{1}{k_1}\|A - A_{k_1}\|_F^2 \\
&\leq 4\|\tilde{A} - \tilde{A}_{k_1}\|_2^2 + \frac{4}{k_1}\|\tilde{A} - \tilde{A}_{k_1}\|_F^2 + \frac{1}{k_1}\|A - A_{k_1}\|_F^2 \\
&\qquad \text{(By the guarantee on } Z \text{ in Step 4)} \\
&\leq \frac{(4 + 4 \cdot 603 + 1)\epsilon}{k} \cdot \|A - A_k\|_F^2
\end{aligned}$$

which gives the lemma after adjusting constants on ϵ by making c, c', c_1 and c_2 sufficiently large. □

Lemma 10 ensures that the rank-k matrix W computed in Step 6 gives a near optimal low-rank approximation of AS_1. Specifically we have:

Lemma 11. *With probability* 95/100*, for sufficiently large* $c, c', c_1, c_2, c_3, c_4$*, Step 5 of Algorithm 1 produces* W *such that, letting* $Q \in \mathbb{R}^{n \times k}$ *be an orthonormal basis for the column span of* AS_1S_3W*:*

$$\|AS_1 - QQ^TAS_1\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2.$$

Proof. We first consider the optimization problem:

$$M^* = \underset{M|\,\mathrm{rank}(M)=k}{\arg\min} \|MZ^T - AS_1\|_F^2.$$

We have:

$$\begin{aligned}
\|M^*Z^T - AS_1\|_F^2 &\leq \|(AS_1)_kZZ^T - AS_1\|_F^2 \\
&= \|[AS_1 - (AS_1)_k] + (AS_1)_k(I - ZZ^T)\|_F^2 \\
&= \|AS_1 - (AS_1)_k\|_F^2 + \|(AS_1)_k(I - ZZ^T)\|_F^2 \\
&\leq \|AS_1 - (AS_1)_k\|_F^2 + k \cdot \|AS_1(I - ZZ^T)\|_2^2 \\
&\leq \|AS_1 - (AS_1)_k\|_F^2 + \epsilon\|A - A_k\|_F^2 \\
&\leq (1+2\epsilon)\|A - A_k\|_F^2 \quad (7)
\end{aligned}$$

where the second to last step uses that $\|AS_1(I - ZZ^T)\|_2^2 \leq \frac{\epsilon}{k}\|A - A_k\|_F^2$ by Lemma 10 and the last step uses that $\|AS_1 - (AS_1)_k\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2$ since AS_1 is an (ϵ, k)-column PCP of A.

Note that since it is rank-k we can write $M^* = Y^* N^*$ where $Y^* \in \mathbb{R}^{n\times k}$ and $N^* \in \mathbb{R}^{k\times k_1}$ has orthonormal rows. Y^* is the solution to the unconstrained low-rank approximation problem:

$$Y^* = \arg\min_{Y\in\mathbb{R}^{n\times k}} \|YN^*Z^T - AS_1\|_F^2.$$

N^*Z^T has orthonormal rows, so its column norms are its leverage scores. S_3 is sampled using the column norms of Z^T, which upper bound those of N^*Z^T. Since $\|Z\|_F^2 = \lceil ck/\epsilon \rceil$, $t_3 = \Theta\left(\|Z\|_F^2 \log(\|Z\|_F^2) + \|Z\|_F^2/\epsilon\right)$. It is thus well known (see e.g. Theorem 38 of [CW13], and [DMM08] for earlier work) that if we set

$$\tilde{Y} = \arg\min_{Y\in\mathbb{R}^{n\times k}} \|YN^*Z^TS_3 - AS_1S_3\|_F^2$$

we have with probability 99/100:

$$\|\tilde{Y}N^*Z^T - AS_1\|_F^2 \le (1+\epsilon)\|Y^*N^*Z^T - AS_1\|_F^2. \tag{8}$$

$\tilde{Y}$ can be computed in closed form as $\tilde{Y} = AS_1S_3(N^*Z^TS_3)^+$, which is in the column span of AS_1S_3. Thus (8) demonstrates that there is some rank-k M in this span satisfying $\|MZ^T - AS_1\|_F^2 \le (1+\epsilon)\|M^*Z^T - AS_1\|_F^2 \le (1+4\epsilon)\|A - A_k\|_F^2$ by (7). Thus if we compute

$$W^* = \arg\min_{W\,|\,\mathrm{rank}(W)=k} \|AS_1S_3WZ^T - AS_1\|_F^2.$$

we will have

$$\|AS_1S_3W^*Z^T - AS_1\|_F^2 \le (1+4\epsilon)\|A - A_k\|_F^2. \tag{9}$$

We can compute W^* approximately by sampling the rows of AS_1S_3 by their leverage scores. These are given by the row norms of the orthogonal basis V as computed in Step 5 and sampled with to obtain S_4 in Step 6. For any W, by the Pythagorean theorem, since V spans AS_1S_3,

$$\|AS_1S_3WZ^T - AS_1\|_F^2 = \|AS_1S_3WZ^T - VV^TAS_1\|_F^2 + \|(I-VV^T)AS_1\|_F^2. \tag{10}$$

Similarly

$$\begin{aligned}\|S_4^TAS_1S_3WZ^T - S_4^TAS_1\|_F^2 = \|S_4^TAS_1S_3WZ^T - S_4^TVV^TAS_1\|_F^2 + \|S_4^T(I-VV^T)AS_1\|_F^2 \\ + 2\,\mathrm{tr}\left((ZW^TS_3^TS_1^TA^T - S_1^TA^TVV^T)S_4S_4^T(I-VV^T)AS_1\right).\end{aligned} \tag{11}$$

For the first term, since S_4 is sampled via the leverage scores of AS_1S_3 and $t_4 = c_4\left(\frac{t_3\log t_3}{\epsilon^2}\right)$, S_4 gives a subspace embedding for the column span of AS_1S_3 with high probability and:

$$\|S_4^TAS_1S_3WZ^T - S_4^TVV^TAS_1\|_F^2 \in (1\pm\epsilon)\|AS_1S_3WZ^T - VV^TAS_1\|_F^2 \tag{12}$$

For the cross term, again since V spans the columns of AS_1S_3 we have:

$$\begin{aligned}&\mathrm{tr}\left((ZW^TS_3^TS_1^TA^T - S_1^TA^TVV^T)S_4S_4^T(I-VV^T)AS_1\right)\\ &\qquad= \mathrm{tr}\left((ZW^TS_3^TS_1^TA^T - S_1^TA^T)VV^TS_4S_4^T(I-VV^T)AS_1\right)\\ &\qquad\le \|AS_1S_3WZ^T - AS_1\|_F\cdot\|VV^TS_4S_4^T(I-VV^T)AS_1\|_F\\ &\qquad\le \|AS_1S_3WZ^T - AS_1\|_F\cdot\epsilon\|(I-VV^T)AS_1\|_F\end{aligned}$$

where the last bound follows with probability 99/100 from a standard approximate matrix multiplication result [DKM06] since S_4 is sampled using the row norms of V. $\|(I - VV^T)AS_1\|_F \leq \|AS_1S_3W^*Z^T - AS_1\|_F \leq \|AS_1S_3WZ^T - AS_1\|_F$ and so overall:

$$\operatorname{tr}\left((ZW^TS_3^TS_1^TA^T - S_1^TA^TVV^T)S_4S_4^T(I - VV^T)AS_1\right) \leq \epsilon\|AS_1S_3WZ^T - AS_1\|_F^2. \tag{13}$$

Combining (10), (11), (12), and (13) for *any* W,

$$\|S_4^TAS_1S_3WZ^T - S_4^TAS_1\|_F^2 \in (1 \pm 3\epsilon)\|AS_1S_3WZ^T - AS_1\|_F^2 + \Delta$$

where $\Delta = \|S_4^T(I - VV^T)AS_1\|_F^2 - \|(I - VV^T)AS_1\|_F^2$ is fixed independent of W. Thus, if we set:

$$\tilde{W} = \underset{W|\operatorname{rank}(W)=k}{\arg\min} \|S_4^TAS_1S_3WZ^T - S_4^TAS_1\|_F^2.$$

with probability 95/100 union bounding over all the probabilistic events above, and applying (9):

$$\|AS_1S_3\tilde{W}Z^T - AS_1\|_F^2 \leq \frac{(1+3\epsilon)}{(1-3\epsilon)}\|AS_1S_3W^*Z^T - AS_1\|_F^2 \leq \frac{(1+3\epsilon)(1+4\epsilon)}{(1-3\epsilon)}\|A - A_k\|_F^2$$

which gives the lemma after adjusting constants on ϵ by making $c, c', c_1, c_2, c_3, c_4$ large enough. □

Proof of Theorem 9. By Lemma 11 and since AS_1 is an (ϵ, k)-column PCP of A, we have $\|A - QQ^TA\|_F^2 \leq \frac{1+\epsilon}{1-\epsilon}\|A - A_k\|_F^2$. In Step 7 we sample $c_5(k\log k + k/\epsilon)$ rows of Q by their standard leverage scores (their row norms) and so have with probability 99/100:

$$\|QN^T - A\|_F^2 \leq (1+\epsilon)\min_{Y\in\mathbb{R}^{n\times k}}\|QY^T - A\|_F^2 = (1+\epsilon)\|A - QQ^TA\|_F^2 \leq \frac{(1+\epsilon)^2}{1-\epsilon}\|A - A_k\|_F^2.$$

Union bounding over failure probabilities and adjusting constants on ϵ yields the final claim.

It just remains to discuss Algorithm 1's runtime and query complexity. In Step 3, forming $\tilde{A} = S_2^TAS_1$ requires reading $t_1 \cdot t_2 = O\left(\frac{nk\log^2 n}{\epsilon^{2.5}}\right)$ entries of A. This dominates the access cost of Step 1, which requires $O(nk_1\log k_1) = O\left(\frac{nk\log(k/\epsilon)}{\epsilon}\right)$ accesses by Lemma 4, forming AS_1S_3 in Step 5 which requires $O\left(\frac{nk\log(k/\epsilon)}{\epsilon} + \frac{nk}{\epsilon^2}\right)$ accesses, and forming S_5^TA in Step 7 which requires $O\left(nk\log k + \frac{nk}{\epsilon}\right)$ accesses. Forming $S_4^TAS_1$ in Step 6 requires $t_1 \cdot t_4 = O\left(\frac{\sqrt{n}k^{1.5}\log(k/\epsilon)\log n}{\epsilon^6}\right)$ accesses, which when n is large compared to k and $1/\epsilon$ will be dominated by our linear in n terms.

For time complexity, Step 1 requires $O(n(k_1\log k_1)^{\omega-1}) = \tilde{O}\left(\frac{nk^{\omega-1}}{\epsilon^{\omega-1}}\right)$ time. Computing Z in Step 4 can be done using an input sparsity time algorithm for spectral norm error with rank k_1 and error parameter $\epsilon' = \Theta(1)$. By Theorem 27 of [CEM+15] or using input sparsity time ridge leverage score sampling [CMM17] in conjunction with the spectral norm PCP result of Lemma 24, the total runtime required is $O(\operatorname{nnz}(\tilde{A})) + \tilde{O}(\frac{\sqrt{n}k^{\omega-1}}{\epsilon^{\omega-1}})$.

We can compute V in Step 5 in $O(nt_3^{\omega-1}) = \tilde{O}\left(\frac{nk^{\omega-1}}{\epsilon^{2(\omega-1)}}\right)$. In Step 6, we can compute W by first multiplying $S_4^TAS_1$ by Z and then multiplying by $(S_4^TAS_1S_3)^+$ and taking the best rank-k approximation of the result. The total runtime is $\tilde{O}\left(t_1 \cdot k^{\omega-1} + k^\omega\right) \cdot \operatorname{poly}(1/\epsilon) = \tilde{O}(\sqrt{n}k^{\omega-.5}) \cdot \operatorname{poly}(1/\epsilon)$. Finally, we can compute Q in Step 7 by first computing a $t_3 \times k$ span of the column space of W, multiplying this by AS_1S_3, and then taking an orthogonal basis for the result, requiring total time $\tilde{O}(k^\omega \cdot \operatorname{poly}(1/\epsilon)) + O\left(\frac{nk^{\omega-1}\log(k/\epsilon)}{\epsilon^2} + nk^{\omega-1}\right)$. The final regression problem can be solved by forming the pseudoinverse of S_5^TQ in $O(k^\omega(\log k + 1/\epsilon))$ time and applying it to S_5^TA in $O(nk^{\omega-1}(\log k + 1/\epsilon))$ time. So overall our runtime cost with linear dependence on n is dominated by the cost of Step 5. We additionally must add the $\tilde{O}(\sqrt{n}k^{\omega-.5}) \cdot \operatorname{poly}(1/\epsilon)$ term from Step 6, which only dominates if n is relatively small. □

In many applications it is desirable that the low-rank approximation to A is also symmetric and positive semidefinite. We show in Appendix C that a modification to Algorithm 1 can satisfy this constraint also in $\tilde{O}(n \operatorname{poly}(k/\epsilon))$ time. The upshot is:

Theorem 12 (Sublinear Time Low-Rank Approximation – PSD Output)**.** *There is an algorithm that given any PSD $A \in \mathbb{R}^{n\times n}$, accesses $\tilde{O}\left(\frac{nk^2}{\epsilon^2} + \frac{nk}{\epsilon^3}\right)$ entries of A, runs in $\tilde{O}\left(\frac{nk^\omega}{\epsilon^\omega} + \frac{nk^{\omega-1}}{\epsilon^{3(\omega-1)}}\right)$ time and with probability at least $9/10$ outputs $M \in \mathbb{R}^{n\times k}$ with:*

$$\|A - MM^T\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2.$$

5 Query Lower Bound

We now present our lower bound on the number of accesses to A required to compute a near-optimal low-rank approximation, matching the query complexity of Algorithm 4 up to a $\tilde{O}(1/\epsilon^{1.5})$ factor.

Theorem 13. *Assume that k, ϵ are such that $nk/\epsilon = o(n^2)$. Consider any (possibly randomized) algorithm $\mathcal{A}$ that, given any PSD $A \in \mathbb{R}^{n\times n}$, outputs a $(1+\epsilon)$-approximate rank-k approximation to A (in the Frobenius norm) with probability at least $2/3$. Then there must be some input A of which $\mathcal{A}$ reads at least $\Omega(nk/\epsilon)$ positions, possibly adaptively, in expectation.*

5.1 Lower Bound Approach

We prove Theorem 13 via Yao's minimax principle [Yao77], proving a lower bound for randomized algorithms on worst-case inputs via a lower bound for deterministic algorithms on a hard input distribution. Specifically, we will draw the input A from a distribution over binary matrices. A has all 1's on its diagonal, along with k randomly positioned (non-contiguous) blocks of all 1's, each of size $\sqrt{2\epsilon n/k} \times \sqrt{2\epsilon n/k}$. In other words, A is the adjacency matrix (plus identity) of a graph with k cliques of size $\sqrt{2\epsilon n/k}$, placed on random subsets of the vertices, with all other vertices isolated.

It is easy to see that every A chosen from the above distribution is PSD since applying a permutation yields a block diagonal matrix, each of whose blocks is a PSD matrix (either a single 1 entry or a rank-1 all 1's block). Additionally, for every A chosen from the distribution, the optimal rank-k approximation to A projects off each of the k blocks, achieving Frobenius norm error $\|A - A_k\|_F^2 = n - k\sqrt{2\epsilon n/k} \approx n$. To match this up to a $1+\epsilon$ factor, any near-optimal rank-k approximation must at least capture a constant fraction of the Frobenius norm mass in the blocks since this mass is $k \cdot 2\epsilon n/k = 2\epsilon n$.

Doing so requires identifying at least a constant fraction of the blocks. However, since block positions are chosen uniformly at random, and since the diagonal entries of A are identical and so convey no information about the positions, to identify a single block, any algorithm essentially must read arbitrary off-diagonal entries until it finds a 1. There are $\approx n^2$ off-diagonal entries with just $2\epsilon n$ of them 1's, so identifying a first block requires $\Omega(n/\epsilon)$ queries to A in expectation (over the random choice of A). Since the vast majority of vertices are isolated and not contained within a block, finding this first block does little to make finding future blocks easier. So overall, the algorithm must make $\Omega(nk/\epsilon)$ queries in expectation to find a constant fraction of the k blocks and output a near-optimal low-rank approximation with good probability.

While the above intuition is the key idea behind the lower bound, a rigorous proof requires a number of additional steps and modifications, detailed in the remainder of this section. In Section 5.2 we introduce the notion of a *primitive approximation* to a matrix, which we will employ in our lower bound for low-rank approximation. In Section 5.3 we show that any deterministic

low-rank approximation algorithm that succeeds with good probability on the input distribution described above can be used to give an algorithm that computes a primitive approximation with good probability on a matrix drawn from a related input distribution (Lemma 20). This reduction yields a lower bound for deterministic low-rank approximation algorithms (Theorem 23), which gives Theorem 13 after an application of Yao's principle.

5.2 Primitive Approximation

We first define the notion of an *ϵ-primitive approximation* to a matrix and establish some basic properties of these approximations.

Definition 14 (ϵ-primitive Approximation)**.** *A matrix $A' \in \mathbb{R}^{m\times m}$ is said to be ϵ-primitive for $A \in \mathbb{R}^{m\times m}$ if the squared Frobenius norm of $A - A'$, restricted to its off-diagonal entries is $\leq \epsilon m$. Note that A' is allowed to have any rank.*

We also define a distribution on matrices with a randomly placed block of all one entries that will will later use in our hard input distribution construction.

Definition 15 (Random Block Matrix)**.** *For any m, ϵ with $m/\epsilon \leq m^2$, let $\mu(m,\epsilon)$ be the distribution on $A \in \mathbb{R}^{m\times m}$ defined as follows. We choose a uniformly random subset S of $[m]$ where $|S| = \sqrt{16\epsilon m}$, where we assume for simplicity that $|S|$ is an integer. We generate a random matrix A by setting for each $i \neq j \in S$, $A_{i,j} = 1$. We then set $A_{i,i} = 1$ for all i and set all remaining entries of A to equal 0.*

Note that we associate a random subset S with the sampling of a matrix A according to $\mu(m,\epsilon)$. It is clear that any A in the support of $\mu(m,\epsilon)$ is PSD. This is since any A in the support of $\mu(m,\epsilon)$, after a permutation, is composed of an $|S| \times |S|$ all ones block and an $(m-|S|)\times(m-|S|)$ identity.

If A' is *ϵ-primitive* for A in the support of $\mu(m,\epsilon)$, it approximates A up to error ϵm on its off-diagonal entries. Let R denote the set of off-diagonal entries of A restricted to the intersection of the rows and columns indexed by S. The error on the entries of R is at most ϵm. Further, restricted to these entries, A is an all ones matrix. Thus, on the entries of R, both A and A' look far from an identity matrix. Formally, we show:

Lemma 16. *If $A' \in \mathbb{R}^{m\times m}$ is ϵ-primitive for an $A \in \mathbb{R}^{m\times m}$ in the support of $\mu(m,\epsilon)$, then A' is not ϵ-primitive for I.*

Proof. By definition, any ϵ-primitive matrix A' for A has squared Frobenius norm restricted to A's off-diagonal entries of $\leq \epsilon m$. Let R denote the set of off-diagonal entries in the intersection of the rows and columns of A indexed by S. Assuming $|S| \geq 4$, which follows from the assumption in Definition 15 that $m/\epsilon \leq m^2$, $|R| = |S|^2 - S \geq 12\epsilon m$.

Restricted to the entries in R, A is an all ones matrix. Thus, on at least $8\epsilon m$ of these entries A' must have value at least $1/2$. Otherwise it would have greater than $|R| - 4\epsilon m \geq 4\epsilon m$ entries with value $\leq 1/2$ and so squared Frobenius norm error on A's off-diagonal entries greater than $\frac{1}{2^2}\cdot 4\epsilon m \geq \epsilon m$, contradicting the assumption that it is ϵ-primitive for A.

Restricted to the entries in R, the identity matrix is all zero. Thus, A' has Frobenius norm error on these entries at least $\frac{1}{2^2}\cdot 8\epsilon m \geq 2\epsilon m$. Thus A' is not ϵ-primitive for I, giving the lemma. ∎

Consider A drawn from $\mu(m,\epsilon)$ with probability $1/2$ and set to I with probability $1/2$. By Lemma 16, any (possibly randomized) algorithm $\mathcal{A}$ that returns an ϵ-primitive approximation to A with good probability can be used to distinguish with good probability whether A is in the support of $\mu(m,\epsilon)$ or $A = I$. This is because, by Lemma 16, the output of such an algorithm can

only be correct either for some A in the support of $\mu(m,\epsilon)$ or for the identity. We first define the distribution over matrices to which this result applies:

Definition 17. *For any m, ϵ with $m/\epsilon \leq m^2$, let $\gamma(m,\epsilon)$ be the distribution on $A \in \mathbb{R}^{m\times m}$ defined as follows. With probability $1/2$ draw A from $\mu(m,\epsilon)$ (Definition 15). Otherwise, draw A from $\nu(m)$, which is the distribution whose support is only the $m \times m$ identity matrix.*

Note that, as with $\mu(m,\epsilon)$, we associate a random subset S with $|S| = \sqrt{16\epsilon m}$ with the sampling of a matrix A according to $\gamma(m,\epsilon)$. S is not used in the case that A is drawn from $\nu(m)$.

Formally, since $\mathcal{A}$ can distinguish whether A is in the support of $\mu(m,\epsilon)$ or $\nu(m)$ (i.e., $A = I$) with good probability we can prove that the distribution of $\mathcal{A}$'s access pattern (over randomness in the input and possible randomization in the algorithm) is significantly different when it is given input A drawn from $\mu(m,\epsilon)$ than when it is given A drawn from $\nu(m)$. Recall for distributions α and β supported on elements s of a finite set S, that the total variation distance $D_{TV}(\alpha,\beta) = \sum_{s\in S} |\alpha(s) - \beta(s)|$, where $\alpha(s)$ is the probability of s in distribution α.

Corollary 18. *Suppose that a (possibly randomized) algorithm $\mathcal{A}$, with probability at least $7/12$ over its random coin flips and random input $A \in \mathbb{R}^{n\times n}$ drawn from $\gamma(m,\epsilon)$, outputs an ϵ-primitive matrix for A. Further, suppose that $\mathcal{A}$ reads at most r positions of A, possibly adaptively.*[3] *Let S be a random variable indicating the list of positions read and their corresponding values.*[4] *Let $L(\mu)$ denote the distribution of S conditioned on $A \sim \mu(m,\epsilon)$, and let $L(\nu)$ denote the distribution of S conditioned on $A \sim \nu(m)$.*[5] *Then*

$$D_{TV}(L(\mu), L(\nu)) \geq 1/6.$$

Proof. By Lemma 16, if algorithm $\mathcal{A}$ succeeds then its output can be used to decide if $A \sim \mu(m,\epsilon)$ or if $A \sim \nu(m)$. The success probability of any such algorithm is well-known (see, e.g., Proposition 2.58 of [BY02]) to be at most $1/2 + D_{TV}(L(\mu), L(\nu))/2$. Making this quantity at least $7/12$ and solving for $D_{TV}(L(\mu), L(\nu))$ proves the corollary. □

5.3 Lower Bound for Low-Rank Approximation

We now give a reduction, showing that any deterministic relative error k-rank approximation algorithm that succeeds with good probability on the hard input distribution described in Section 5.1 can be used to compute an ϵ-primitive approximation to A that is drawn from $\gamma(n/(2k),\epsilon)$ with good probability. We first formally define this input distribution.

Definition 19 (Hard Input Distribution – Low-Rank Approximation)**.** *Suppose $2nk/\epsilon \leq n^2$. Let γ_b be the distribution on $A \in \mathbb{R}^{n\times n}$ determined as follows. We draw a uniformly random subset S of $[n]$ where $|S| = n/2$, where we assume for simplicity that $|S|$ is an integer. We further partition S into k subsets $S^1, S^2, ..., S^k$ chosen uniformly at random. For all $\ell \in [k]$, $|S^\ell| = n/(2k)$, which we also assume to be an integer.*

Letting A^ℓ denote the entries of A restricted to the intersection of the rows and columns indexed by S^ℓ, we independently draw each A^ℓ from $\gamma(n/(2k),\epsilon)$ (Definition 17).[6] *We then set $A_{i,i} = 1$ for all i and set all remaining entries of A to equal 0.*

[3] That is, for any input A, in any random execution, $\mathcal{A}$ reads at most r entries of A.

[4] S is determined by the random input A and the random choices of $\mathcal{A}$. Since $\mathcal{A}$ reads at most r positions of any input, we always have $|S| \leq r$.

[5] Here and throughout, we let $A \sim \mu(m,\epsilon)$ denote the event that, when A is drawn from the distribution $\gamma(m,\epsilon)$ (Definition 17), which is a mixture of the distributions $\mu(m,\epsilon)$ and $\nu(m)$, that A is drawn from $\mu(m,\epsilon)$. $A \sim \nu(m)$ denotes the analogous event for $\nu(m)$.

[6] By the assumption that $2nk/\epsilon \leq n^2$ we have $\frac{n}{2k\epsilon} \leq \left(\frac{n}{2k}\right)^2$ and so this is a valid setting of the parameters for $\gamma(m,\epsilon)$.

Our reduction from ϵ-primitive approximation to low-rank approximation is as follows:

Lemma 20 (Reduction from Primitive Approximation to Low-Rank Approximation)**.** *Suppose that $nk/\epsilon = o(n^2)$ and that $\mathcal{A}$ is a deterministic algorithm that, with probability $\geq 2/3$ on random input $A \in \mathbb{R}^{n\times n}$ drawn from γ_b, outputs a $(1+\epsilon/26)$-approximate rank-k approximation to A. Further suppose $\mathcal{A}$ reads at most r positions of A, possibly adaptively.*

Then there is a randomized algorithm $\mathcal{B}$ that, with probability $\geq 7/12$ over its random coin flips and random input B drawn from $\gamma(n/(2k),\epsilon)$, outputs an ϵ-primitive matrix for B. Further, letting $L(\mu), L(\nu)$ be as defined in Corollary 18 for $\mathcal{B}$,

$$D_{TV}(L(\mu), L(\nu)) \leq \frac{2\epsilon r}{nk}.$$

Proof. Consider a randomized algorithm $\mathcal{B}$ that, given B drawn from $\gamma(n/(2k),\epsilon)$ generates a random matrix $A^{n\times n}$ drawn from γ_b as follows. Choose a uniformly random subset S of $[n]$ with $|S| = n/2$. Partition S into k subsets $S^1, S^2, ..., S^k$ chosen uniformly at random. Note that for $\ell \in [k]$ $|S^\ell| = n/(2k)$. Letting A^ℓ denote the entries of A restricted to the intersection of the rows and columns indexed by S^ℓ, set $A^1 = B$, and for $\ell = 2, ..., k$ independently draw A^ℓ from $\gamma(n/(2k),\epsilon)$. Set $A_{i,i} = 1$ for all i and set all remaining entries of A equal to 0.

After generating A, $\mathcal{B}$ then applies $\mathcal{A}$ to A to compute a rank-k approximation A'. $\mathcal{B}$ then outputs $(A')^1$, the $n/(2k) \times n/(2k)$ submatrix of A' corresponding to the intersection of the rows and columns indexed by S^1. We have the following:

Claim 21. *With probability $\geq 7/12$ over the random choices of $\mathcal{B}$ and over the random input B drawn from $\gamma(n/(2k),\epsilon)$, $(A')^1$ is ϵ-primitive for B.*

Proof. It is clear that A generated by $\mathcal{B}$ is distributed according to γ_b (Definition 19). By construction, for any $A \sim \gamma_b$ there is a rank-k approximation of cost at most n; indeed this follows by choosing the best rank-1 solution for each A^ℓ. Consequently if $\mathcal{A}$ succeeds on A, then its output A' satisfies $\|A - A'\|_F^2 \leq n + \epsilon n/26$.

Note that A has $A_{i,i} = 1$ for all $i \in [n]$. So the squared Frobenius norm cost of any rank-k approximation restricted to the diagonal is at least $n - k$. Let c_ℓ be the squared Frobenius norm cost of A' restricted to the off diagonal entries in A^ℓ. Note that c_ℓ is a random variable. Then,

$$n - k + \sum_{\ell=1}^{k} c_\ell \leq \|A' - A\|_F^2 \leq n + \epsilon n/26.$$

By averaging for at least a 11/12 fraction of the blocks i,

$$c_i \leq 12 + 12\epsilon n/(26k) \leq 13\epsilon n/(26k) < \epsilon n/(2k),$$

assuming $\epsilon n/(26k) \geq 8$, which holds if $\epsilon n/k = \omega(1)$, which follows from our assumption that $nk/\epsilon = o(n^2)$.

It follows by symmetry of γ_b with respect to the k blocks that with probability at least 11/12, if $\mathcal{A}$ succeeds on A, $c_1 \leq \epsilon n/(2k)$. This gives that $(A')^1$ is ϵ-primitive (Definition 14) for $A^1 = B$. This yields the claim after applying a union bound, since $\mathcal{A}$ succeeds with probability at least 2/3. ☐

It remains to bound the total variation distance between $\mathcal{B}$'s access pattern when $B \sim \mu(n/(2k),\epsilon)$ and when $B \sim \nu(n/(2k))$. We have:

Claim 22. *For the algorithm $\mathcal{B}$ defined above*

$$D_{TV}(L(\mu), L(\nu)) \leq \frac{2\epsilon r}{nk}.$$

Proof. Let $\mathcal{W}$ denote the random variable that encompasses $\mathcal{B}$'s random choices in choosing the indices in $S^2, ..., S^k$ and in setting the entries in $A^2, ..., A^k$. Let Ω denote the set of all possible values of $\mathcal{W}$. Let S denote the random subset with $|S| = \sqrt{n/(2k) \cdot \epsilon}$ associated with the drawing of B from $\gamma(n/(2k), \epsilon)$ (see Definition 17). Let $\bar{S}$ denote the set of off-diagonal entries of A in the intersection of the rows and columns indexed by S.

If $B \sim \mu(n/(2k), \epsilon)$ then all entries of $\bar{S}$ are 1. If $B \sim \nu(n/(2k))$ then $B = I$ and so all entries of $\bar{S}$ are 0. Outside of the entries in $\bar{S}$, the entries of B are identical in the two cases that $B \sim \mu(n/(2k), \epsilon)$ and $B \sim \nu(n/(2k))$ (in particular, they are all 0 off the diagonal and 1 on the diagonal).

So, for any $w \in \Omega$, conditioned on $\mathcal{W} = w$, all entries outside $\bar{S}$ are fixed. Further conditioned on $B \sim \nu(n/(2k))$, all entries in $\bar{S}$ are 0. So all entries of A are fixed and $\mathcal{A}$ always reads the same sequence of entries $(i_{w,1}, j_{w,1}), ..., (i_{w,r}, j_{w,r})$. Further, for any $w \in \Omega$, conditioned on $\mathcal{W} = w$, S is a uniform random subset of $R = [n] \setminus (S^2 \cup ... \cup S^k)$. $|R| = n - (k-1) \cdot n/(2k) \geq n/2$. So, for any $\ell \in [r]$ we have:

$$\Pr[(i_{w,\ell}, j_{w,\ell}) \in \bar{S} | \mathcal{W} = w] \leq \frac{|\bar{S}|}{|R|^2} \leq \frac{|S|^2}{n^2/4} \leq \frac{2\epsilon}{nk}.$$

By a union bound, letting $\mathcal{E}$ be the event that $\mathcal{A}$ reads $(i_{w,1}, j_{w,1}), ..., (i_{w,r}, j_{w,r})$ and does not read any entry of $\bar{S}$ in its r accesses to A we have for any $w \in \Omega$:

$$\Pr[\mathcal{E} | \mathcal{W} = w] \geq 1 - \frac{2\epsilon r}{nk}.$$

Thus, for any $w \in \Omega$, regardless of whether $B \sim \nu(n/(2k))$ or $B \sim \mu(n/(2k), \epsilon)$, $\mathcal{A}$ has access pattern $(i_{w,1}, j_{w,1}), ..., (i_{w,r}, j_{w,r})$ to A with probability $\geq 1 - \frac{2\epsilon r}{nk}$. Correspondingly, $\mathcal{B}$ has a fixed access pattern to B with probability $\geq 1 - \frac{2\epsilon r}{nk}$. Thus,

$$D_{TV}(L(\mu), L(\nu)) \leq \frac{2\epsilon r}{nk}$$

yielding the claim. □

In combination Claims 21 and 22 give Lemma 20. □

We can now use Lemma 20 to argue that if a deterministic low-rank approximation algorithm succeeding with good probability on a random input drawn from γ_b accesses too few entries, then it can be used to give a primitive approximation algorithm violating Corollary 18.

Theorem 23 (Lower Bound for Deterministic Algorithms)**.** *Assume that n, k, ϵ are such that $nk/\epsilon = o(n^2)$. Consider any deterministic algorithm $\mathcal{A}$ that, given random input A drawn from γ_b, outputs a $(1+\epsilon)$-approximate rank-k approximation to A (in the Frobenius norm) with probability at least $2/3$. Further, suppose $\mathcal{A}$ reads at most r positions of A, possibly adaptively. Then $r = \Omega(nk/\epsilon)$.*

Proof. Assume towards a contradiction that $r = o(nk/\epsilon)$. Then applying Lemma 20, $\mathcal{A}$ can be used to give a randomized algorithm $\mathcal{B}$ that with probability $\geq 7/12$ over its random coin flips and random input $B \in n/(2k) \times n/(2k)$ drawn from $\gamma(n/(2k), \epsilon)$, outputs an ϵ-primitive matrix for B. Further, letting $L(\mu), L(\nu)$ be as defined in Corollary 18 for $\mathcal{B}$, by Lemma 20, $D_{TV}(L(\mu), L(\nu)) \leq \frac{2\epsilon r}{nk}$. For $r = o(nk/\epsilon)$, $\frac{2\epsilon r}{nk} = o(1)$, contradicting Corollary 18, and giving the theorem. □

Theorem 13 follows directly from applying Yao's minimax principle ([Yao77], Theorem 3) to Theorem 23.

6 Spectral Norm Error Bounds

We conclude by showing how to modify Algorithm 1 to output a low-rank approximation B achieving the spectral norm guarantee:

$$\|A - B\|_2^2 \leq (1+\epsilon)\|A - A_k\|_2^2 + \frac{\epsilon}{k}\|A - A_k\|_F^2. \tag{14}$$

This can be significantly stronger than the Frobenius guarantee (1) when $\|A - A_k\|_F^2$ is large, and, for example is critical in our application to ridge regression, discussed in Section 6.2.

It is not hard to see that since additive error in the Frobenius norm upper bounds additive error in the spectral norm (see e.g. Theorem 3.4 of [Gu14]) that for B satisfying the Frobenius norm guarantee $\|A - B\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2$, we immediately have the spectral bound $\|A - B\|_2^2 \leq \|A - A_k\|_2^2 + \epsilon\|A - A_k\|_F^2$. Thus, we can achieve 14 simply by running Algorithm 1 with error parameter ϵ/k. However, this approach is suboptimal. Applying Theorem 9, our query complexity would be $\Theta\left(\frac{nk^{3.5}\log^2 n}{\epsilon^{2.5}}\right)$. We improve this k dependence significantly in Algorithm 2. Since (14) is often applied (see for example Section 6.2) with $k' = k/\epsilon$ and $\epsilon = \Theta(1)$ to give $\|A - B\|_2^2 \leq O\left(\frac{\epsilon}{k}\|A - A_k\|_F^2\right)$, optimizing k dependence is especially important.

We first give an extension of Lemma 3 to the spectral norm case. This lemma provides the column sampling analog to Lemma 8.

Lemma 24 (Spectral Norm PCP)**.** *For any $A \in \mathbb{R}^{n\times d}$, for $i \in \{1,\ldots,d\}$, let $\tilde{\tau}_i^k \geq \tau_i^k(A)$ be an overestimate for the i^{th} rank-k ridge leverage score. Let $p_i = \frac{\tilde{\tau}_i^k}{\sum_i \tilde{\tau}_i^k}$ and $t = \frac{c\log(k/\delta)}{\epsilon^2}\sum_i \tilde{\tau}_i^k$ for any $\epsilon < 1$ and sufficiently large constant c. Construct C by sampling t columns of A, each set to $\frac{1}{\sqrt{tp_i}}a_i$ with probability p_i. With probability $1-\delta$, for any orthogonal projection $P \in \mathbb{R}^{n\times n}$,*

$$(1-\epsilon)\|A - PA\|_2^2 - \frac{\epsilon}{k}\|A - A_k\|_F^2 \leq \|C - PC\|_2^2 \leq (1+\epsilon)\|A - PA\|_2^2 + \frac{\epsilon}{k}\|A - A_k\|_F^2.$$

We refer to C as an (ϵ, k)-spectral PCP of A.

Proof. This follows from Corollary 34. With probability $\geq 1-\delta$, sampling by the rank-k ridge leverage scores gives C satisfying:

$$(1-\epsilon)CC^T - \frac{\epsilon}{k}\|A - A_k\|_F^2 I \preceq AA^T \preceq (1+\epsilon)CC^T + \frac{\epsilon}{k}\|A - A_k\|_F^2 I. \tag{15}$$

We can write for any M, $\|M\|_2^2 = \max_{x:\|x\|_2^2=1} x^T M x$. When $\|x\|_2^2 = 1$, $\|(I-P)x\|_2^2 \leq 1$ so by (15) we have for any unit norm x:

$$\begin{aligned} x^T(I-P)CC^T(I-P)x &\leq x^T(I-P)AA^T(I-P)x + \frac{\epsilon}{k}\|A - A_k\|_F^2 \\ &\leq \|A - PA\|_2^2 + \frac{\epsilon}{k}\|A - A_k\|_F^2 \end{aligned}$$

which gives $\|C - PC\|_2^2 \leq \frac{1}{1-\epsilon}\|A - PA\|_2^2 + \frac{\epsilon}{(1-\epsilon)k}\|A - A_k\|_F^2$. Similarly we have:

$$\begin{aligned} x^T(I-P)AA^T(I-P)x &\leq (1+\epsilon)x^T(I-P)CC^T(I-P)x + \frac{\epsilon}{k}\|A - A_k\|_F^2 \\ &\leq (1+\epsilon)\|C - PC\|_2^2 + \frac{\epsilon}{k}\|A - A_k\|_F^2 \end{aligned}$$

which gives $\frac{1}{1+\epsilon}\|A - PA\|_2^2 - \frac{\epsilon}{(1+\epsilon)k}\|A - A_k\|_F^2 \leq \|C - PC\|_2^2$. The lemma follows from combining these upper and lower bounds after adjusting constant factors on ϵ (by making c large enough). ∎

6.1 Spectral Norm Low-Rank Approximation Algorithm

We now use Lemma 24, along with its row sampling counterpart, Lemma 8, to give an algorithm (Algorithm 2) for computing a near optimal spectral norm low-rank approximation to A.

In Steps 1-3 we sample both rows and columns of A via the rank $\Theta(k/\epsilon)$ ridge leverage scores of $A^{1/2}$, ensuring with high probability that AS_1 is an (ϵ, k)-spectral PCP of A and $\tilde{A}$ is in turn an (ϵ, k)-spectral row PCP of AS_1. Thus, if we compute (using an input sparsity time algorithm) a span Z which gives a near optimal spectral norm low-rank approximation to $\tilde{A}$ (Step 3), this span will also be nearly optimal for AS_1. Since we cannot afford to fully read AS_1, we approximately project it to Z by further sampling its columns using Z's leverage scores (Step 4). We use leverage score sampling again in Step 5 to approximately project A to the span of the result. This yields our final approximation, using the fact that AS_1 is a spectral PCP for A.

Algorithm 2 PSD Low-Rank Approximation – Spectral Error

1. Let $k_1 = \lceil ck/\epsilon^2 \rceil$. For all $i \in [1, .., n]$ compute $\tilde{\tau}_i^{k_1}(A^{1/2})$ which is a constant factor approximation to the ridge leverage score $\tau_i^{k_1}(A^{1/2})$.

2. Set $\ell_i^{(1)} = 4\epsilon\sqrt{\frac{n}{k}}\tilde{\tau}_i^{k_1}(A^{1/2})$. Set $p_i^{(1)} = \frac{\ell_i^{(1)}}{\sum_i \ell_i^{(1)}}$ and $t_1 = \frac{c_1 \log n}{\epsilon^2}\sum_i \ell_i^{(1)}$. Sample $S_1, S_2 \in \mathbb{R}^{n \times t_1}$ each whose j^{th} column is set to $\frac{1}{\sqrt{t p_i^{(1)}}} e_i$ with probability $p_i^{(1)}$.

3. Let $\tilde{A} = S_2^T A S_1$, and use an input sparsity time algorithm to compute orthonormal $Z \in \mathbb{R}^{t_1 \times k}$ satisfying the Frobenius guarantee $\|\tilde{A} - \tilde{A}ZZ^T\|_F^2 \leq 2\|\tilde{A} - \tilde{A}_k\|_F^2$ along with the spectral guarantee $\|\tilde{A} - \tilde{A}ZZ^T\|_2^2 \leq (1+\epsilon)\|\tilde{A} - \tilde{A}_k\|_2^2 + \frac{\epsilon}{k}\|\tilde{A} - \tilde{A}_k\|_F^2$.

4. Let $t_3 = c_3\left(k \log k + \frac{k^2}{\epsilon}\right)$, set $p_i^{(3)} = \frac{\|z_i\|_2^2}{\|Z\|_F^2}$, and sample $S_3 \in \mathbb{R}^{t_1 \times t_3}$ where the j^{th} column is set to $\frac{1}{\sqrt{t_3 p_i^{(3)}}} e_i$ with probability $p_i^{(3)}$. Solve:

$$M = \arg\min_{M \in \mathbb{R}^{n \times k}} \|AS_1S_3 - MZ^TS_3\|_F^2$$

5. Compute an orthogonal basis $Q \in \mathbb{R}^{n \times k}$ for the column span of M. Let $t_4 = c_4\left(k \log k + \frac{k^2}{\epsilon}\right)$, set $p_i^{(4)} = \frac{\|q_i\|_2^2}{\|Q\|_F^2}$, and sample $S_4 \in \mathbb{R}^{n \times t_4}$ where the j^{th} column is set to $\frac{1}{\sqrt{t_4 p_i^{(4)}}} e_i$ with probability $p_i^{(4)}$. Solve:

$$N = \arg\min_{N \in \mathbb{R}^{n \times k}} \|S_4^T Q N^T - S_4^T A\|_F^2.$$

6. Return $Q, N \in \mathbb{R}^{n \times k}$.

Theorem 25 (Sublinear Time Low-Rank Approximation –Spectral Norm Error)**.** *Given any PSD $A \in \mathbb{R}^{n \times n}$, for sufficiently large constants c, c_1, c_2, c_3, c_4, Algorithm 2 accesses $O(\frac{n \cdot k \log^2 n}{\epsilon^6} + \frac{nk^2}{\epsilon})$ entries of A, runs in $\tilde{O}\left(\frac{nk^\omega}{\epsilon} + \frac{nk}{\epsilon^6} + (\sqrt{n}k^{\omega-1} + k^{\omega+1}) \cdot \text{poly}(1/\epsilon)\right)$ time and with probability at least 9/10 outputs $M, N \in \mathbb{R}^{n \times k}$ with:*

$$\|A - MN^T\|_2^2 \leq (1+\epsilon)\|A - A_k\|_2^2 + \frac{\epsilon}{k}\|A - A_k\|_F^2.$$

Proof. $\ell_i^{(1)} = 4\epsilon\sqrt{\frac{n}{k}}\tilde{\tau}_i^{k_1}(A^{1/2})$ which by Lemma 5 is within a constant factor of upper bounding the rank $k_1 = \lceil ck/\epsilon^2 \rceil$ ridge leverage scores of A. As long as $c \geq 1$, these scores upper bound the rank-k ridge leverage scores. So by Lemma 24, with high probability AS_1 an (ϵ, k)-spectral PCP of A as long as c_1 is set large enough. Additionally, by Lemma 8, with high probability $\tilde{A}$ is an (ϵ, k)-spectral row PCP of AS_1.

Frobenius norm PCP bounds also hold. By Lemma 3, AS_1 is an (ϵ, k_1)-column PCP of A with high probability and $\tilde{A}$ is an (ϵ, k_1)-row PCP for AS_1 with probability 99/100. These bounds trivially give that AS_1 and $\tilde{A}$ are (ϵ, k) PCPs for A and AS_1 respectively, which gives:

$$\|\tilde{A} - \tilde{A}_k\|_F^2 \leq c_5\|A - A_k\|_F^2 \tag{16}$$

for some constant c_5. By (16) and the fact that $\tilde{A}$ is an (ϵ, k)-spectral PCP of AS_1, for Z computed in Step 3 of Algorithm 2 we have:

$$\begin{aligned}
\|AS_1 - AS_1ZZ^T\|_2^2 &\leq \frac{(1+\epsilon)}{(1-\epsilon)}\left(\|\tilde{A} - \tilde{A}_k\|_2^2 + \frac{\epsilon}{k}\|\tilde{A} - \tilde{A}_k\|_F^2\right) + \frac{\epsilon}{k(1-\epsilon)}\|A - A_k\|_F^2 \\
&\leq (1+3\epsilon)\|\tilde{A} - \tilde{A}_k\|_2^2 + \frac{(3c_5+2)\epsilon}{k}\|A - A_k\|_F^2 \\
&\leq (1+3\epsilon)(1+\epsilon)\|AS_1 - (AS_1)_k\|_2^2 + \frac{(3c_5+2+1)\epsilon}{k}\|A - A_k\|_F^2
\end{aligned} \tag{17}$$

where we assume without loss of generality that $\epsilon < 1/2$ as this can be achieved by scaling our constants sufficiently.

Now, by standard approximate regression, since S_3 is sampled by the leverage scores of Z, and since $t_3 = \Theta\left(k\log k + k/\epsilon'\right)$ for $\epsilon' = \epsilon/k$, for M computed in Step 4, with probability 99/100:

$$\|AS_1 - MZ^T\|_F^2 \leq \left(1 + \frac{\epsilon}{k}\right)\|AS_1 - AS_1ZZ^T\|_F^2. \tag{18}$$

This Frobenius norm bound also implies a spectral norm bound. Specifically, we can write:

$$\|AS_1 - MZ^T\|_F^2 = \|AS_1ZZ^T - MZ^T\|_F^2 + \|AS_1(I - ZZ^T)\|_F^2$$

so by (18) we must have $\|AS_1ZZ^T - MZ^T\|_F^2 \leq \frac{\epsilon}{k}\|AS_1(I - ZZ^T)\|_F^2$. This gives:

$$\begin{aligned}
\|AS_1 - MZ^T\|_2^2 &\leq \|AS_1ZZ^T - MZ^T\|_2^2 + \|AS_1(I - ZZ^T)\|_2^2 \\
&\leq \|AS_1(I - ZZ^T)\|_2^2 + \|AS_1ZZ^T - MZ^T\|_F^2 \\
&\leq \|AS_1(I - ZZ^T)\|_2^2 + \frac{\epsilon}{k}\|AS_1(I - ZZ^T)\|_F^2.
\end{aligned}$$

Note that $\|AS_1(I - ZZ^T)\|_F^2 = O(\|A - A_k\|_F^2)$ by the Frobenius norm guarantee required in Step 3, and the fact that $\tilde{A}$ and AS_1 are (ϵ, k) PCPs for AS_1 and A respectively. So combining with (16) the above implies that for Q spanning the columns of M,

$$\|AS_1 - QQ^TAS_1\|_2^2 \leq \|AS_1 - MZ^T\|_2^2 \leq (1 + O(\epsilon))\|AS_1 - (AS_1)_k\|_2^2 + O\left(\frac{\epsilon}{k}\right)\|A - A_k\|_F^2.$$

Since AS_1 is an (ϵ, k)-spectral PCP for A we also have $\|A - QQ^TA\|_2^2 \leq (1+O(\epsilon))\|AS_1 - (AS_1)_k\|_2^2 + O\left(\frac{\epsilon}{k}\right)\|A - A_k\|_F^2$. The theorem follows by applying an identical approximate regression argument for N computed in Step 5 to show that

$$\|A - QN^T\|_2^2 \leq (1+\epsilon)\|A - QQ^TA\|_2^2 + \frac{\epsilon}{k}\|A - QQ^TA\|_F^2 = (1+O(\epsilon))\|A - A_k\|_2^2 + O\left(\frac{\epsilon}{k}\right)\|A - A_k\|_F^2$$

with probability 99/100. Adjusting constants on ϵ and union bounding over failure probabilities yields the final bound.

It just remains to discuss runtime and sample complexity. Constructing $S_2^T A S_1$ in Step 3 requires reading $t_1^2 = O\left(\frac{nk\log^2 n}{\epsilon^6}\right)$ entries of A. Constructing AS_1S_3 in Step 4 and AS_4 in Step 5 both require reading $O\left(nk\log k + \frac{nk^2}{\epsilon}\right)$ entries.

For runtime, computing Z in Step 3 requires $O(\text{nnz}(\tilde{A})) + \tilde{O}(\sqrt{n}k^{\omega-1}\cdot\text{poly}(1/\epsilon))$ time using an input sparsity time algorithm (e.g. by Theorem 27 of [CEM+15] or using input sparsity time ridge leverage score sampling [CMM17] in conjunction with the spectral norm PCP result of Lemma 24). Computing M in Step 4 and N in Step 5 both require computing the pseudoinverse of a $k \times O\left(k\log k + \frac{k^2}{\epsilon}\right)$ matrix in $\tilde{O}(k^{\omega+1}\,\text{poly}(1/\epsilon))$ time, and then applying this to an $n \times O\left(k\log k + \frac{k^2}{\epsilon}\right)$ matrix requiring $\tilde{O}\left(\frac{nk^\omega}{\epsilon}\right)$ time. □

6.2 Sublinear Time Ridge Regression

We now demonstrate how Theorem 25 can be leveraged to give a sublinear time, relative error algorithm for approximately solving the ridge regression problem:

$$\min_{x\in\mathbb{R}^n} \|Ax - y\|_2^2 + \lambda\|x\|_2^2 \tag{19}$$

for PSD A. We begin with a lemma showing that any approximation to A with small spectral norm error can be used to approximately solve (19) up to relative error.

Lemma 26 (Ridge Regression via Spectral Norm Low-Rank Approximation)**.** *For any PSD $A \in \mathbb{R}^{n\times n}$, $y \in \mathbb{R}^n$, regularization parameter $\lambda \geq 0$ and B with $\|A - B\|_2^2 \leq \epsilon^2\lambda$, let $\tilde{x} \in \mathbb{R}^n$ be any vector satisfying:*

$$\|B\tilde{x} - y\|_2^2 + \lambda\|\tilde{x}\|_2^2 \leq (1+\alpha)\cdot\min_{x\in\mathbb{R}^n} \|Bx - y\|_2^2 + \lambda\|x\|_2^2$$

we have:

$$\|A\tilde{x} - y\|_2^2 + \lambda\|\tilde{x}\|_2^2 \leq (1+\alpha)(1+5\epsilon)\min_{x\in\mathbb{R}^n} \|Ax - y\|_2^2 + \lambda\|x\|_2^2.$$

Proof. For any $x \in \mathbb{R}^n$ we have:

$$\|Bx - y\|_2^2 + \lambda\|x\|_2^2 = \|Ax - y\|_2^2 + \|(B-A)x\|_2^2 + 2x^T(B-A)^T(Ax - y) + \lambda\|x\|_2^2$$

Now $\|(B-A)x\|_2^2 \leq \epsilon^2\lambda\|x\|_2^2$ and further

$$\begin{aligned}
|2x^T(B-A)^T(Ax-y)| &\leq 2\|(A-B)x\|_2\|Ax-y\|_2 \\
&\leq 2\epsilon\sqrt{\lambda}\|x\|_2\|Ax-y\|_2 \\
&\leq \epsilon\left(\|Ax-y\|_2^2 + \lambda\|x\|_2^2\right).
\end{aligned}$$

So for any x, $\|Bx - y\|_2^2 + \lambda\|x\|_2^2 \in (1\pm 2\epsilon)\left(\|Ax-y\|_2^2 + \lambda\|x\|_2^2\right)$ which gives the lemma since any nearly optimal $\tilde{x}$ for the ridge regression problem on B will also be nearly optimal for A. □

Combining Lemma 26 with Theorem 25 gives the following:

Theorem 27 (Sublinear Time Ridge Regression)**.** *Given any PSD $A \in \mathbb{R}^{n\times n}$, regularization parameter $\lambda \geq 0$, $y \in \mathbb{R}^n$, and upper bound $\tilde{s}_\lambda$ on the statistical dimension $s_\lambda \stackrel{\text{def}}{=} \text{tr}((A^2+\lambda I)^{-1}A^2)$, there is an algorithm accessing $\tilde{O}\left(\frac{n\tilde{s}_\lambda^2}{\epsilon^4}\right)$ entries of A and running in $\tilde{O}\left(\frac{n\tilde{s}_\lambda^\omega}{\epsilon^{2\omega}}\right)$ time, which outputs $\tilde{x}$ satisfying:*

$$\|A\tilde{x}-y\|_2^2+\lambda\|\tilde{x}\|_2^2 \leq (1+\epsilon)\cdot\min_{x\in\mathbb{R}^n}\|Ax-y\|_2^2+\lambda\|x\|_2^2.$$

When $\tilde{s}_\lambda \ll n$ as is often the case, the above significantly improves upon state-of-the-art input sparsity time runtimes for general matrices [ACW16].

Proof. Let $k = \frac{c\tilde{s}_\lambda}{\epsilon^2}$ for sufficiently large constant c. We have:

$$s_\lambda = \sum_{i=1}^n \frac{\lambda_i^2(A)}{\lambda_i^2(A)+\lambda} \geq \sum_{i:\lambda_i^2(A)\geq\epsilon^2\lambda} \frac{\lambda_i^2(A)}{(1+1/\epsilon^2)\lambda_i^2(A)} \geq \frac{\epsilon^2}{2}\cdot|\{i:\lambda_i^2(A)\geq\epsilon^2\lambda\}|.$$

So $|\{i:\lambda_i^2(A)\geq\epsilon^2\lambda\}| \leq \frac{2s_\lambda}{\epsilon^2}$ and for large enough c and $k=\frac{c\tilde{s}_\lambda}{\epsilon^2} \geq \frac{cs_\lambda}{\epsilon^2}$ we have $\|A-A_k\|_2^2 \leq \frac{\epsilon^2\lambda}{2}$ and can run Algorithm 2 with error parameter $\epsilon=\Theta(1)$ to find $M,N\in\mathbb{R}^{n\times k}$ with $\|A-MN^T\|_2^2\leq\epsilon^2\lambda$. We can then apply Lemma 26 – solving $\tilde{x}=\min_{x\in\mathbb{R}^n}\|MN^Tx-y\|_2^2+\lambda\|x\|_2^2$ exactly using an SVD in $O(nk^{\omega-1})$ time. $\tilde{x}$ will be a $(1+O(\epsilon))$ approximate solution for A, which, after adjusting constants on ϵ, gives the lemma. The runtime follows from Theorem 25 with $k=c\tilde{s}_\lambda/\epsilon^2$ and $\epsilon'=\Theta(1)$. The $O(nk^{\omega-1})$ regression cost is dominated by the cost of computing the low-rank approximation. □

Note that Theorem 27 ensures that if the $k\geq\frac{cs_\lambda}{\epsilon^2}$ for some constant c, $\tilde{x}$ is a good approximation to the ridge regression problem. Setting k properly requires some knowledge of an upper bound $\tilde{s}_\lambda$ on s_λ. A constant factor approximation to s_λ can be computed in $\tilde{O}(n^{3/2}\cdot\text{poly}(s_\lambda))$ time using for example a column PCP as given by Lemma 3 and binary searching for an appropriate k value.

An interesting open question is if s_λ be be approximated more quickly – specifically with linear dependence on n. This question is closely related to if it is possible to estimate the cost $\|A-A_k\|_F^2$ in $\tilde{O}(n\cdot\text{poly}(k))$, which surprisingly is also open.

Acknowledgements

The authors thank IBM Almaden where part of this work was done. David Woodruff also thanks the Simons Institute program on Machine Learning and the XDATA program of DARPA for support.

References

[ACW16] Haim Avron, Kenneth L. Clarkson, and David P. Woodruff. Sharper bounds for regression and low-rank approximation with regularization, 2016.

[AFK+01] Yossi Azar, Amos Fiat, Anna R. Karlin, Frank McSherry, and Jared Saia. Spectral analysis of data. In *Proceedings of the 33rd Annual ACM Symposium on Theory of Computing (STOC)*, pages 619–626, 2001.

[AFKM01] Dimitris Achlioptas, Amos Fiat, Anna R. Karlin, and Frank McSherry. Web search via hub synthesis. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 500–509, 2001.

[AGR16] Nima Anari, Shayan Oveis Gharan, and Alireza Rezaei. Monte Carlo Markov chain algorithms for sampling strongly Rayleigh distributions and determinantal point processes. In *Proceedings of the 29th Annual Conference on Computational Learning Theory (COLT)*, pages 103–115, 2016.

[AM05] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *Proceedings of the 18th Annual Conference on Computational Learning Theory (COLT)*, pages 458–469, 2005.

[AM07] Dimitris Achlioptas and Frank McSherry. Fast computation of low-rank matrix approximations. *Journal of the ACM*, 54(2), 2007.

[BDN15] Jean Bourgain, Sjoerd Dirksen, and Jelani Nelson. Toward a unified theory of sparse dimensionality reduction in Euclidean space. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC)*, pages 499–508, 2015.

[BW09] Mohamed-Ali Belabbas and Patrick J Wolfe. Spectral methods in machine learning and new strategies for very large datasets. *Proceedings of the National Academy of Sciences*, 106(2):369–374, 2009.

[BY02] Ziv Bar-Yossef. *The Complexity of Massive Data Set Computations*. PhD thesis, University of California at Berkeley, 2002.

[CC00] Trevor F Cox and Michael AA Cox. *Multidimensional scaling*. CRC press, 2000.

[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 47th Annual ACM Symposium on Theory of Computing (STOC)*, pages 163–172, 2015.

[CMM17] Michael B Cohen, Cameron Musco, and Christopher Musco. Input sparsity time low-rank approximation via ridge leverage score sampling. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2017.

[Coh16] Michael B. Cohen. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the 27th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 278–287, 2016.

[CR09] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[CW13] Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing (STOC)*, pages 81–90, 2013.

[CW17] Ken Clarkson and David P. Woodruff. Low-rank PSD approximation in input-sparsity time. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2017.

[DFK+04] Petros Drineas, Alan M. Frieze, Ravi Kannan, Santosh Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Machine Learning*, 56(1-3):9–33, 2004.

[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.

[DKR02] Petros Drineas, Iordanis Kerenidis, and Prabhakar Raghavan. Competitive recommendation systems. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, pages 82–90, 2002.

[DLWZ14] Xuefeng Duan, Jiaofen Li, Qingwen Wang, and Xinjun Zhang. Low rank approximation of the symmetric positive semidefinite matrix. *Journal of Computational and Applied Mathematics*, 260:236–243, 2014.

[DM05] Petros Drineas and Michael W Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2153–2175, 2005.

[DMM06] Petros Drineas, Michael W Mahoney, and S Muthukrishnan. Subspace sampling and relative-error matrix approximation: Column-row-based methods. In *European Symposium on Algorithms*, pages 304–314. Springer, 2006.

[DMM08] Petros Drineas, Michael W Mahoney, and S Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2), 2008.

[DRVW06] Amit Deshpande, Luis Rademacher, Santosh Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2006.

[DV06] Amit Deshpande and Santosh Vempala. Adaptive sampling and fast low-rank matrix approximation. In *Approximation, Randomization, and Combinatorial Optimization: Algorithms and Techniques*, pages 292–303. Springer, 2006.

[FKV04] Alan M. Frieze, Ravi Kannan, and Santosh Vempala. Fast Monte-Carlo algorithms for finding low-rank approximations. *Journal of the ACM*, 51(6):1025–1041, 2004.

[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 24th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1434–1453, 2013.

[Git11] Alex Gittens. The spectral norm error of the naive Nyström extension. *arXiv:1110.5305*, 2011.

[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.

[GM13] Alex Gittens and Michael W. Mahoney. Revisiting the Nyström method for improved large-scale machine learning. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013. Preliminary version at arXiv:1303.1849.

[Gu14] Ming Gu. Subspace iteration randomization and singular value problems. *arXiv:1408.2208*, 2014.

[Har14] Moritz Hardt. Understanding alternating minimization for matrix completion. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 651–660, 2014.

[Hof03] Thomas Hofmann. Collaborative filtering via Gaussian probabilistic latent semantic analysis. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 259–266, 2003.

[JNS13] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing (STOC)*, pages 665–674, 2013.

[Kle99] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[KMT09] Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling techniques for the Nyström method. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 304–311, 2009.

[KSV08] Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. *SIAM Journal on Computing*, 38(3):1141–1156, 2008.

[LBKW14] Yingyu Liang, Maria-Florina Balcan, Vandana Kanchanapally, and David P. Woodruff. Improved distributed principal component analysis. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 3113–3121, 2014.

[LJS16] Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Fast DPP sampling for Nyström with application to kernel methods. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 2061–2070, 2016.

[LKL10] Mu Li, James Tin-Yau Kwok, and Baoliang Lu. Making large-scale Nyström approximation possible. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, page 631, 2010.

[MM13] Xiangrui Meng and Michael W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing (STOC)*, pages 91–100, 2013.

[MM15] Cameron Musco and Christopher Musco. Randomized block Krylov methods for stronger and faster approximate singular value decomposition. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 1396–1404, 2015.

[MM16] Cameron Musco and Christopher Musco. Recursive sampling for the Nyström method. *arXiv:1605.07583*, 2016.

[NN13] Jelani Nelson and Huy L. Nguyen. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 117–126, 2013.

[PRTV00] Christos H. Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, 61(2):217–235, 2000.

[Sar06] Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 143–152, 2006.

[SWZ16] Zhao Song, David P. Woodruff, and Huan Zhang. Sublinear time orthogonal tensor decomposition. In *Advances in Neural Information Processing Systems 29 (NIPS)*, 2016.

[SY05] Anthony Man-Cho So and Yinyu Ye. Theory of semidefinite programming for sensor network localization. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 405–414, 2005.

[Tro15] Joel A Tropp. An introduction to matrix concentration inequalities. *arXiv:1501.01571*, 2015.

[TYUC16] Joel A Tropp, Alp Yurtsever, Madeleine Udell, and Volkan Cevher. Randomized single-view algorithms for low-rank matrix approximation. *arXiv:1609.00048*, 2016.

[WLZ16] Shusen Wang, Luo Luo, and Zhihua Zhang. SPSD matrix approximation vis column selection: theories, algorithms, and extensions. *Journal of Machine Learning Research*, 17(49):1–49, 2016.

[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.

[WZ13] Shusen Wang and Zhihua Zhang. Improving CUR matrix decomposition and the Nyström approximation via adaptive sampling. *Journal of Machine Learning Research*, 14(1):2729–2769, 2013.

[Yao77] Andrew Chi-Chin Yao. Probabilistic computations: Toward a unified measure of complexity. In *Proceedings of the 18th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 222–227, 1977.

[YH38] Gale Young and Alston S Householder. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3(1):19–22, 1938.

[YS07] Stephen J Young and Edward R Scheinerman. Random dot product graph models for social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 138–149. Springer, 2007.

[ZTK08] Kai Zhang, Ivor W Tsang, and James T Kwok. Improved Nyström low-rank approximation and error analysis. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 1232–1239, 2008.

A Low-Rank Approximation of A via Approximation of $A^{1/2}$

We first observe that a low-rank approximation for $A^{1/2}$ does not imply a good low-rank approximation for A. Intuitively, if A has a large top singular value, the low-rank approximation for A must capture the corresponding singular direction with significantly more accuracy than a good low-rank approximation for $A^{1/2}$, in which the singular value is relatively much smaller.

Theorem 28. *For any k, ϵ there exists a PSD matrix A and a rank k matrix B such that $\|A^{1/2} - B\|_F^2 \leq (1+\epsilon)\|A^{1/2} - A_k^{1/2}\|_F^2$ but for every matrix C in the rowspan of B,*

$$\|A - C\|_F^2 \geq \left(1 + \epsilon \cdot \frac{(n-k-1)\lambda_1(A)}{\lambda_{k+1}(A)}\right)\|A - A_k\|_F^2.$$

Notably, if we set $B = A^{1/2}P$ for some rank k orthogonal projection P, AP can be an arbitrarily bad low-rank approximation of A.

Proof. Let $A \in \mathbb{R}^{n\times n}$ be a diagonal matrix with $A_{i,i} = \alpha^2$ for $i = 1, ..., k$, $A_{k+1,k+1} = 0$ and all other diagonal entries equal to β^2, where $\alpha > \beta > 0$. Let B be a rank k matrix which has its last $n-k$ rows all zero. For $i = 1, ..., k$, let $B_{i,i} = A_{i,i}^{1/2}$ and $B_{1,k+2} = \sqrt{\epsilon(n-k-1)} \cdot \beta$. We have: $\|A^{1/2} - A_k^{1/2}\|_F^2 = (n-k-1)\beta^2$ and $\|A^{1/2} - B\|_F^2 = (1+\epsilon)(n-k-1)\beta^2$. Note that the first row b_1 aligns somewhat well with a_1, but as we will see, not well enough to give a good low-rank approximation for A itself.

Let C be the projection of A onto the rowspan of B, which gives the optimal low-rank approximation to A within this span. For $i = 2, ..., k$, $c_i = a_i$, since A and B match exactly on these rows up to a scaling. For $i > k$, $c_i = \vec{0}$. Finally, $c_1 = \frac{b_1}{\|b_1\|_2^2} \cdot \langle b_1, a_1\rangle = b_1 \cdot \left(\frac{\alpha^3}{\alpha^2 + \epsilon(n-k-1)\beta^2}\right)$. Overall:

$$\begin{aligned}
\|A - C\|_F^2 &= (n-k-1)\beta^4 + (A_{1,1} - C_{1,1})^2 + (A_{1,k+2} - C_{1,k+2})^2 \\
&\geq (n-k-1)\beta^4 + \left(\frac{\sqrt{\epsilon(n-k-1)} \cdot \beta\alpha^3}{\alpha^2 + \epsilon(n-k-1)\beta^2}\right)^2 \\
&\geq (n-k-1)\beta^4 \cdot (1 + \epsilon(n-k-1)\alpha^2/4\beta^2) \\
&= (1 + \epsilon(n-k-1)\alpha^2/\beta^2) \cdot \|A - A_k\|_F^2.
\end{aligned}$$

By setting $\alpha \gg \beta$ we can make this approximation arbitrarily bad. Note that $\alpha^2/\beta^2 = \lambda_1(A)/\lambda_{k+1}(A)$. This ratio will be large whenever A is well approximated by a low-rank matrix. □

Despite the above example, we can show that for a projection P, if $A^{1/2}P$ is a very near optimal low-rank approximation of $A^{1/2}$ then $A^{1/2}PA^{1/2}$ is a relative error low-rank approximation of A:

Theorem 29. *Let $P \in \mathbb{R}^{n\times n}$ be an orthogonal projection matrix such that $\|A^{1/2} - A^{1/2}P\|_F^2 \leq (1 + \epsilon/\sqrt{n})\|A^{1/2} - A_k^{1/2}\|_F^2$. Then:*

$$\|A - A^{1/2}PA^{1/2}\|_F^2 \leq (1+3\epsilon)\|A - A_k\|_F^2.$$

Proof. We can rewrite using the fact that $(I-P) = (I-P)^2$ since it is a projection:

$$\|A - A^{1/2}PA^{1/2}\|_F^2 = \|A^{1/2}(I-P)^2A^{1/2}\|_F^2 = \|A^{1/2}(I-P)\|_4^4.$$

Let $\delta_i = \sigma_i(A^{1/2}(I-P))$ denote the i^{th} singular value of $A^{1/2}(I-P)$. Let λ_i be the i^{th} eigenvalue of A. By the assumption that P gives a near optimal low-rank approximation of $A^{1/2}$:

$$\sum_{i=1}^{n-k} \delta_i^2 \leq \sum_{i=k+1}^{n} \lambda_i + \epsilon/\sqrt{n}\|A^{1/2} - A_k^{1/2}\|_F^2.$$

Additionally, by Weyl's monotonicity theorem (see e.g. Theorem 3.2 in [Gu14] and proof of Lemma 15 in [MM15]), for all i, $\delta_i \geq \lambda_{i+k}^{1/2}$. We thus have:

$$\|A - A^{1/2}PA^{1/2}\|_F^2 = \sum_{i=1}^{n-k} \delta_i^4 \leq \sum_{i=k+2}^{n} \lambda_i^2 + \left(\lambda_{k+1} + \epsilon/\sqrt{n}\|A^{1/2} - A_k^{1/2}\|_F^2\right)^2.$$

If $\lambda_{k+1} \geq 1/\sqrt{n} \cdot \|A^{1/2} - A_k^{1/2}\|_F^2$ then $\left(\lambda_{k+1} + \epsilon/\sqrt{n}\|A^{1/2} - A_k^{1/2}\|_F^2\right)^2 \leq (1+\epsilon)^2\lambda_{k+1}^2 \leq (1+3\epsilon)\lambda_{k+1}^2$ and hence:

$$\|A - A^{1/2}PA^{1/2}\|_F^2 \leq (1+3\epsilon)\sum_{i=k+1}^{n} \lambda_i^2 = (1+3\epsilon)\|A - A_k\|_F^2.$$

Alternatively if $\lambda_{k+1} \leq 1/\sqrt{n} \cdot \|A^{1/2} - A_k^{1/2}\|_F^2$ then

$$\left(\lambda_{k+1} + \epsilon/\sqrt{n}\|A^{1/2} - A_k^{1/2}\|_F^2\right)^2 \leq \left((1+\epsilon)/\sqrt{n}\|A^{1/2} - A_k^{1/2}\|_F^2\right)^2 \leq (1+\epsilon)^2\|A - A_k\|_F^2$$

which also gives the theorem. □

A.1 PSD Low-Rank Approximation in $n^{1.69} \cdot \text{poly}(k/\epsilon)$ Time

We now combine Theorem 29 with the ridge leverage score sampling algorithm of [MM16] to give a sublinear time algorithm for low-rank approximation of A reading $n^{3/2} \cdot \text{poly}(k/\epsilon)$ entries of the matrix and running in $n^{1.69} \cdot \text{poly}(k/\epsilon)$ time. We note that this approach could also be used with adaptive sampling [DV06] or volume sampling techniques [AGR16], as outlined in the introduction.

Theorem 30. *There is an algorithm based off ridge leverage score sampling which, given PSD $A \in \mathbb{R}^{n\times n}$ with high probability outputs $M \in \mathbb{R}^{n\times k}$ with $\|A - MM^T\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2$. The algorithm reads $\tilde{O}(n^{3/2}k/\epsilon)$ entries of A and runs in $\tilde{O}\left(n^{(\omega+1)/2} \cdot (k/\epsilon)^{\omega-1}\right)$ time where $\omega < 2.38$ is the exponent of matrix multiplication.*

This follows from Lemma 4, adapted from Theorem 8 of [MM16], which shows that it is possible to estimate the ridge leverage scores of $A^{1/2}$ with $\tilde{O}(nk)$ accesses to A and $O(nk^{\omega-1})$ time. We can use these scores to sample a set of rows from $A^{1/2}$ whose span contains a near optimal low-rank approximation. Specifically we have:

Lemma 31 (Theorem 7 of [CMM17])**.** *For any $B \in \mathbb{R}^{n\times n}$, for $i \in \{1, \ldots, n\}$, let $\tilde{\tau}_i^k \geq \tau_i^k(B)$ be an overestimate for the i^{th} rank-k ridge leverage score of B. Let $p_i = \frac{\tilde{\tau}_i^k}{\sum_i \tilde{\tau}_i^k}$ and $t = c\left(\log k + \frac{\log(1/\delta)}{\epsilon}\right)\sum_i^k \tilde{\tau}_i^k$ for $\epsilon < 1$ and some sufficiently large constant c. Construct R by sampling t rows of B, each set to row b_i with probability p_i. With probability $1-\delta$, letting P_R be the projection onto the rows of R,*

$$\|B - (BP_R)_k\|_F^2 \leq (1+\epsilon)\|B - B_k\|_F^2.$$

Note that $(BP_R)_k$ can be written as a row projection of B – onto the top k singular vectors of BP_R. So, if we compute for each i, $\tilde{\tau}_i^k \geq \tau_i^k(A^{1/2})$ using Lemma 4, set $\epsilon' = \epsilon/3\sqrt{n}$, and let S be a sampling matrix selecting $\tilde{O}(\sum \tilde{\tau}_i^k/\epsilon') = \tilde{O}(k\sqrt{n}/\epsilon)$ rows of $A^{1/2}$, then by Theorem 29, letting P be the projection onto the rows of $SA^{1/2}$, we have $\|A - (A^{1/2}P)_k(A^{1/2}P)_k^T\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2$.

Finally, $(A^{1/2}P)_k(A^{1/2}P)_k^T = (A^{1/2}PA^{1/2})_k = (AS(S^TAS)^+S^TA)_k$. We can compute a factorization of this matrix by computing $(S^TAS)^{+/2}$, then computing $AS(S^TAS)^{+/2}$ and taking the SVD of this matrix. Since S has $\tilde{O}(k\sqrt{n}/\epsilon)$ columns, using fast matrix multiplication this requires time $\tilde{O}(n \cdot (k\sqrt{n}/\epsilon)^{\omega-1}) = \tilde{O}(n^{(\omega+1)/2} \cdot (k/\epsilon)^{\omega-1})$ and $\tilde{O}(n^{3/2}k/\epsilon)$ accesses to A (to read the entries of AS), giving Theorem 30.

B Additional Proofs for Main Algorithm

Lemma 2 (Sum of Ridge Leverage Scores)**.** *For any $A \in \mathbb{R}^{n\times d}$, $\sum_{i=1}^d \tau_i(A) \leq 2k$.*

Proof. We rewrite Definition 1 using A's singular value decomposition $A = U\Sigma V^T$.

$$\begin{aligned}\tau_i(A) &= a_i^T\left(U\Sigma^2U^T + \frac{\|A-A_k\|_F^2}{k}UU^T\right)^{-1}a_i \\ &= a_i^T\left(U\bar{\Sigma}U^T\right)a_i,\end{aligned}$$

where $\bar{\Sigma}_{i,i} = \frac{1}{\sigma_i^2(A)+\frac{\|A-A_k\|_F^2}{k}}$. We then have:

$$\sum_{i=1}^n \tau_i(A) = \text{tr}\left(A^TU\bar{\Sigma}U^TA\right) = \text{tr}\left(V\Sigma\bar{\Sigma}\Sigma V^T\right) = \text{tr}(\Sigma^2\bar{\Sigma})$$

$(\Sigma^2\bar{\Sigma})_{i,i} = \frac{\sigma_i^2(A)}{\sigma_i^2(A)+\frac{\|A-A_k\|_F^2}{k}}$. For $i \leq k$ we simply upper bound this by 1. So:

$$\text{tr}(\Sigma^2\bar{\Sigma}) = k + \sum_{i=k+1}^n \frac{\sigma_i^2(A)}{\sigma_i^2(A)+\frac{\|A-A_k\|_F^2}{k}} \leq k + k\sum_{i=k+1}^n \frac{\sigma_i^2(A)}{\|A-A_k\|_F^2} = 2k.$$

□

We first prove our row sampling PCP result for spectral norm error. We follow with the closely related proof Lemma 7 which gives Frobenius norm error.

Lemma 8 (Spectral Norm Row PCP)**.** *For any PSD $A \in \mathbb{R}^{n\times n}$, and $\epsilon < 1$ let $k' = \lceil ck/\epsilon^2 \rceil$ and $\tilde{\tau}_i^{k'}(A^{1/2}) \geq \tau_i^{k'}(A^{1/2})$ be an overestimate for the i^{th} rank-k' ridge leverage score of $A^{1/2}$. Let $\tilde{\ell}_i = 4\epsilon\sqrt{\frac{n}{k}}\tau_i^{k'}(A^{1/2})$, $p_i = \frac{\tilde{\ell}_i}{\sum_i \tilde{\ell}_i}$, and $t = \frac{c'\log n}{\epsilon^2} \cdot \sum_i \tilde{\ell}_i$. Construct weighted sampling matrices $S_1, S_2 \in \mathbb{R}^{n\times t}$, where the j^{th} column is set to $\frac{1}{\sqrt{tp_i}}e_i$ with probability p_i.*

For sufficiently large constants c, c', with high probability, letting $\tilde{A} = S_2^TAS_1$, for any orthogonal projection $P \in \mathbb{R}^{t\times t}$:

$$(1-\epsilon)\|AS_1(I-P)\|_2^2 - \frac{\epsilon}{k}\|A-A_k\|_F^2 \leq \|\tilde{A}(I-P)\|_2^2 \leq (1+\epsilon)\|AS_1(I-P)\|_2^2 + \frac{\epsilon}{k}\|A-A_k\|_F^2.$$

We refer to $\tilde{A}$ as an (ϵ, k)-spectral PCP of AS_1.

Note that if $\tilde{\tau}_i^{k'}(A^{1/2})$ is a constant factor approximation to $\tau_i^{k'}(A^{1/2})$, $t = O\left(\frac{\sqrt{nk}\log n}{\epsilon^3}\right)$. We use 'with high probability' to mean with probability $\geq 1 - 1/n^d$ for some large constant d.

Proof. For conciseness write $C \stackrel{\text{def}}{=} AS_1$ and write the eigendecomposition $A = U\Lambda U^T$ with $\lambda_i = \Lambda_{i,i}$. Applying the triangle inequality we have:

$$\begin{aligned}\|\tilde{A}(I-P)\|_2^2 = \|(I-P)\tilde{A}^T\tilde{A}(I-P)\|_2 &= \|(I-P)[C^TC + (\tilde{A}^T\tilde{A} - C^TC)](I-P)\|_2 \\ &\in \|C(I-P)\|_2^2 \pm \|(I-P)(\tilde{A}^T\tilde{A} - C^TC)(I-P)\|_2\end{aligned}$$

Thus to show the Lemma it suffices to show:

$$\|(I-P)(\tilde{A}^T\tilde{A} - C^TC)(I-P)\|_2 \leq \epsilon\|C(I-P)\|_2^2 + \frac{\epsilon}{k}\|A - A_k\|_F^2. \tag{20}$$

Let m be the largest index with $\lambda_m^2 \geq \frac{\epsilon^2}{k}\|A - A_k\|_F^2$. Let $U_H \stackrel{\text{def}}{=} U_m$ contain the top m 'head' eigenvectors of A and let U_T contain the remaining 'tail' eigenvectors. Let $C_H = U_H U_H^T C$ and $C_T = U_T U_T^T C$. $C_H + C_T = C$ and so by the triangle inequality:

$$\begin{aligned}\|(I-P)(\tilde{A}^T\tilde{A} - C^TC)(I-P)\|_2 \leq &\|(I-P)(C_H^TS_2S_2^TC_H - C_H^TC_H)(I-P)\|_2 \\ &+\|(I-P)(C_T^TS_2S_2^TC_T - C_T^TC_T)(I-P)\|_2 \\ &+2\|(I-P)C_H^TS_2S_2^TC_T(I-P)\|_2.\end{aligned} \tag{21}$$

We bound each of the terms in the above sum separately. Specifically we show:

- Head Term: $\|(I-P)(C_H^TS_2S_2^TC_H - C_H^TC_H)(I-P)\|_2 \leq \epsilon\|C(I-P)\|_2^2$
- Tail Term: $\|(I-P)(C_T^TS_2S_2^TC_T - C_T^TC_T)(I-P)\|_2 \leq \frac{13\epsilon^2}{k}\|A - A_k\|_F^2$.
- Cross Term: $\|(I-P)C_H^TS_2S_2^TC_T(I-P)\|_2 \leq \frac{10\epsilon}{k}\|A - A_k\|_F^2 + 4\epsilon\|C(I-P)\|_F^2$.

Combining these three bounds, after adjusting constant factors on ϵ by making the constants c and c' in the rank parameter k' and sample size t large enough, gives (20) and thus the lemma. For the remainder of the proof we thus fix $c = 1$ so $k' = \lceil k/\epsilon^2 \rceil$.

Head Term:

We first show that the ridge scores of $A^{1/2}$ upper bound the standard leverage scores of U_m.

Lemma 32. *For any p with $\lambda_p^2(A) \geq \frac{1}{k}\|A - A_k\|_F^2$ we have:*

$$\sqrt{\frac{16n}{k}} \cdot \tau_i^k(A^{1/2}) \geq \|(U_p)_i\|_2^2.$$

where $\|(U_p)_i\|_2^2$ is the i^{th} row norm of U_p, whose columns are the top p eigenvectors of A.

Proof. If U_p contains no eigenvectors, this is true vacuously as all row norms are 0. Otherwise

$$\begin{aligned}\tau_i^k(A) &= a_i^T\left(A^2 + \frac{\|A - A_k\|_F^2}{k}I\right)^{-1}a_i \\ &= e_i^TU\hat{\Lambda}U^Te_i\end{aligned}$$

where $\hat{\Lambda}_{j,j} = \frac{\lambda_j^2}{\lambda_j^2 + \frac{\|A-A_k\|_F^2}{k}}$. We can then write:

$$\begin{aligned}
\tau_i^k(A) = \sum_{j=1}^{n} U_{i,j}^2 \cdot \hat{\Lambda}_{j,j} &\geq \sum_{j=1}^{p} U_{i,j}^2 \cdot \hat{\Lambda}_{j,j} && \text{(truncate sum)} \\
&\geq \sum_{j=1}^{p} \left(U_{i,j}^2 \cdot \frac{\lambda_j^2}{2\lambda_j^2} \right) && \text{(By assumption, } \lambda_j^2(A) \geq \lambda_p^2(A) \geq \tfrac{1}{k}\|A-A_k\|_F^2\text{)} \\
&\geq \frac{1}{2} \sum_{j=1}^{m} U_{i,j}^2 = \frac{1}{2}\|(U_p)_i\|_2^2.
\end{aligned}$$

This gives the lemma combined with the fact that $\tau_i^k(A) \leq 2\sqrt{\frac{n}{k}}\tau_i^k(A^{1/2})$ by Lemma 5. □

Applying Lemma 32 with $k' = \lceil k/\epsilon^2 \rceil$, since $\lambda_m^2(A) \geq \frac{1}{k'}\|A - A_{k'}\|_F^2$, we have:

$$\tilde{\ell}_i = \sqrt{\frac{16n\epsilon^2}{k}} \cdot \tilde{\tau}_i^{k'}(A^{1/2}) \geq \|(U_m)_i\|_2^2 = \|(U_H)_i\|_2^2$$

Further, $t = \frac{c' \log n}{\epsilon^2} \cdot \sum_i \tilde{\ell}_i$ so if we set c' large enough, we have by a standard matrix Chernoff bound (see Lemma 33) since U_H is an orthogonal span for the columns of C_H and hence and its row norms are the leverage scores of C_H, with high probability:

$$(1-\epsilon)C_H^T C_H \preceq C_H^T S_2 S_2^T C_H \prec (1+\epsilon)C_H^T C_H. \tag{22}$$

This in turn gives:

$$\begin{aligned}
\|(I-P)(C_H^T S_2 S_2^T C_H - C_H^T C_H)(I-P)\|_2 &\leq \epsilon \|(I-P)C_H^T C_H (I-P)\|_2 \\
&= \epsilon \|C_H(I-P)\|_2^2 \leq \epsilon \|C(I-P)\|_2^2.
\end{aligned} \tag{23}$$

Tail Term:

We can loosely bound via the triangle inequality and the fact that $\|I-P\|_2 \leq 1$:

$$\|(I-P)(C_T^T S_2 S_2^T C_T - C_T^T C_T)(I-P)\|_2 \leq \|S_2 C_T\|_2^2 + \|C_T\|_2^2. \tag{24}$$

Since $k' = \lceil k/\epsilon^2 \rceil$:

$$\tilde{\ell}_i = 4\epsilon\sqrt{\frac{n}{k}} \cdot \tilde{\tau}_i^{k'}(A^{1/2}) \geq x_i^T \left(A + \frac{\epsilon \|A^{1/2} - A_k^{1/2}\|_F^2}{\sqrt{nk}} \right)^+ x_i.$$

Thus for $S = S_1$ or $S = S_2$, for sufficiently large c' in our sample size $t = \frac{c' \log n}{\epsilon^2}$, by a matrix Chernoff bound (Corollary 34), with high probability

$$(1-\epsilon)A^{1/2}SS^TA^{1/2} - \frac{\epsilon\|A^{1/2} - A_k^{1/2}\|_F^2}{\sqrt{nk}} \preceq A \preceq (1+\epsilon)A^{1/2}SS^TA^{1/2} + \frac{\epsilon\|A^{1/2} - A_k^{1/2}\|_F^2}{\sqrt{nk}}.$$

Which in turn implies

$$\sigma_1^2(S^TA^{1/2}(I - U_T U_T^T)) \leq (1+\epsilon)\sigma_1^2(A^{1/2}(I - U_T U_T^T)) + \frac{\epsilon\|A^{1/2} - A_k^{1/2}\|_F^2}{\sqrt{nk}}.$$

and so, applying the AMGM inequality,

$$\begin{aligned}
\|S_2^T C_T\|_2^2 &= \|S_2^T(I - U_T U_T^T)AS_1\|_2^2 \\
&\leq \sigma_1^2(S_2^T A^{1/2}(I - U_T U_T^T)) \cdot \sigma_1^2((I - U_T U_T^T)A^{1/2}S_1) \\
&\leq 2(1+\epsilon)^2\sigma_1^4(A^{1/2}(I - U_T U_T)^T) + \frac{2\epsilon^2\|A^{1/2} - A_k^{1/2}\|_F^4}{nk} \\
&\leq 8\|A(I - U_T U_T^T)\|_2^2 + \frac{2\epsilon^2\|A - A_k\|_F^2}{k} && \text{(by } \ell_1/\ell_2 \text{ bound.)} \\
&\leq \frac{10\epsilon^2\|A - A_k\|_F^2}{k}. && (\|A(I - U_T U_T^T)\|_2^2 \leq \tfrac{\epsilon^2\|A-A_k\|_F^2}{k} \text{ by definition.})
\end{aligned}$$

Similarly we have:

$$\begin{aligned}
\|C_T\|_2^2 &= \sigma_1^2((I - U_T U_T^T)AS_1) \\
&\leq \sigma_1^2(S_1^T A^{1/2}(I - U_T U_T^T)) \cdot \sigma_1^2(A^{1/2}(I - U_T U_T^T)) \\
&\leq (1+\epsilon)\sigma_1^4(A^{1/2}(I - U_T U_T^T)) + \frac{\epsilon\|A^{1/2} - A_k^{1/2}\|_F^2}{\sqrt{nk}} \cdot \sigma_1^2(A^{1/2}(I - U_T U_T^T)) \\
&\leq \frac{3\epsilon^2\|A - A_k\|_F^2}{k}.
\end{aligned}$$

So overall, plugging back into (24) gives:

$$\|(I-P)(C_T^T S_2 S_2^T C_T - C_T^T C_T)(I-P)\|_2 \leq \frac{13\epsilon^2}{k}\|A - A_k\|_F^2. \tag{25}$$

Cross Term:

By submultiplicativity of the spectral norm:

$$\|(I-P)C_T^T S_2 S_2^T C_H(I-P)\|_2 \leq \|C_T^T S_2\|_2 \cdot \|S_2^T C_H(I-P)\|_2.$$

As shown above for our tail bound, $\|C_T^T S_2\|_2 \leq \sqrt{\frac{10\epsilon^2}{k}}\|A - A_k\|_F$. Further, as shown in (22), S_2 gives a subspace embedding for C_H so we have:

$$\begin{aligned}
\|(I-P)C_T^T S_2 S_2^T C_H(I-P)\|_2 &\leq \sqrt{\frac{10\epsilon^2}{k}}\|A - A_k\|_F \cdot (1+\epsilon)\|C_H(I-P)\|_2 \\
&\leq \frac{10\epsilon}{k}\|A - A_k\|_F^2 + 4\epsilon\|C(I-P)\|_2^2.
\end{aligned} \tag{26}$$

where the last bound follows form the AMGM inequality.

Plugging our head (23), tail (25) and cross term (26) bounds in (21), and adjusting constants on ϵ by making c and c' sufficiently large gives the lemma.

□

Lemma 7 (Frobenius Norm Row PCP)**.** *For any PSD $A \in \mathbb{R}^{n\times n}$ and $\epsilon \leq 1$ let $k' = \lceil ck/\epsilon \rceil$ and let $\tilde{\tau}_i^{k'}(A^{1/2}) \geq \tau_i^{k'}(A^{1/2})$ be an overestimate for the i^{th} rank-k' ridge leverage score of $A^{1/2}$. Let $\tilde{\ell}_i = \sqrt{\frac{16n\epsilon}{k}} \cdot \tilde{\tau}_i^{k'}(A^{1/2})$, $p_i = \frac{\tilde{\ell}_i}{\sum_i \tilde{\ell}_i}$, and $t = \frac{c'\log n}{\epsilon^2}\sum_i \tilde{\ell}_i$. Construct weighted sampling matrices $S_1, S_2 \in \mathbb{R}^{n\times t}$ each whose j^{th} column is set to $\frac{1}{\sqrt{tp_i}}e_i$ with probability p_i.*

For sufficiently large constants c, c', with probability $\frac{99}{100}$, letting $\tilde{A} = S_2^T A S_1$, for any rank-k orthogonal projection $P \in \mathbb{R}^{t\times t}$:

$$(1-\epsilon)\|AS_1(I-P)\|_F^2 \leq \|\tilde{A}(I-P)\|_F^2 + \Delta \leq (1+\epsilon)\|AS_1(I-P)\|_F^2$$

for some fixed Δ (independent of P) with $|\Delta| \leq 600\|A-A_k\|_F^2$.

Note that if $\tilde{\tau}_i^{k'}(A^{1/2})$ is a constant factor approximation to $\tau_i^{k'}(A^{1/2})$, $t = O\left(\frac{\sqrt{nk}\log n}{\epsilon^{2.5}}\right)$.

Proof. The proof is similar to that of Lemma 8. Denote $C \stackrel{\text{def}}{=} AS_1$ and write the eigendecomposition $A = U\Lambda U^T$ with $\lambda_i = \Lambda_{i,i}$. Let m be the largest index with $\lambda_m^2 \geq \frac{\epsilon}{k}\|A-A_{k'}\|_F^2$. Let $U_H \stackrel{\text{def}}{=} U_m$ contain the top m 'head' eigenvectors of A and let U_T contain the remaining 'tail' eigenvectors. Let $C_H = U_H U_H^T C$ and $C_T = U_T U_T^T C$ and note that $C_H + C_T = C$.

By the Pythagorean theorem we have $\|C(I-P)\|_F^2 = \|C_H(I-P)\|_F^2 + \|C_T(I-P)\|_F^2$. Expanding using the identity $\|M\|_F^2 = \operatorname{tr}(M^T M)$,

$$\|\tilde{A}(I-P)\|_F^2 = \|S_2^T C_H(I-P)\|_F^2 + \|S_2^T C_T(I-P)\|_F^2 + 2\operatorname{tr}\left((I-P)C_H^T S_2 S_2^T C_T(I-P)\right) \quad (27)$$

We bound each of the terms in the above sum separately. Specifically we show:

- Head Term: $\|S_2^T C_H(I-P)\|_F^2 \in (1\pm\epsilon)\|C_H(I-P)\|_F^2$
- Tail Term: $\|S_2^T C_T(I-P)\|_F^2 + \Delta \in \|C_T(I-P)\|_F^2 \pm \epsilon\|A-A_k\|_F^2$ where $|\Delta| \leq 600\|A-A_k\|_F^2$.
- Cross Term: $|\operatorname{tr}\left((I-P)C_H^T S_2 S_2^T C_T(I-P)\right)| \leq \epsilon\|C(I-P)\|_F^2$.

It is not hard to see that combining these bounds gives the Lemma. $\|A-A_k\|_F^2 \leq (1+3\epsilon)\|C(I-P)\|_F^2$ since C is an (ϵ,k)-column PCP of A by Lemmas 3, 5, and the fact that rank-k' ridge scores upper bound the rank-k scores (as long as $c>1$). Additionally, $\|A-A_k\|_F^2 \leq \|A(I-P)\|_F^2$ for any rank-k P. So plugging into (27) we have:

$$\begin{aligned}\|\tilde{A}(I-P)\| + \Delta &\in (1\pm\epsilon)\|C(I-P)\| \pm (2+3\epsilon)\epsilon\|C(I-P)\|_F^2 \\ &\in (1\pm O(\epsilon))\|C(I-P)\|.\end{aligned}$$

This gives the lemma by adjusting constants on ϵ by making c and c' large enough. For the remainder of the proof we thus fix $c=1$ and so $k' = \lceil k/\epsilon\rceil$.

Head Term:

By Lemma 32 applied to $k' = \lceil k/\epsilon\rceil$, since $\lambda_m^2 \geq \frac{1}{k'}\|A-A_{k'}\|_F^2$:

$$\tilde{\ell}_i \geq \sqrt{\frac{16n\epsilon}{k}}\cdot\tilde{\tau}_i^{k'}(A^{1/2}) \geq \|(U_H)_i\|_2^2.$$

Further, $t = \frac{c'\log n}{\epsilon^2}\cdot\sum_i \tilde{\ell}_i$ so if we set c' large enough, by a standard matrix Chernoff bound (see Lemma 33) since U_H is an orthogonal span for the columns of C_H and hence and its row norms are the leverage scores of C_H, with high probability:

$$(1-\epsilon)C_H^T C_H \preceq C_H^T S_2 S_2^T C_H \prec (1+\epsilon)C_H^T C_H.$$

This gives the bound $\|S_2^T C_H(I-P)\|_F^2 \in (1\pm\epsilon)\|C_H(I-P)\|_F^2$.

Tail Term: We want to show:

$$\|S_2^T C_T(I-P)\|_F^2 + \Delta \in \|C_T(I-P)\|_F^2 \pm \epsilon\|A-A_k\|_F^2 \tag{28}$$

where $|\Delta| \leq 600\|A-A_k\|_F^2$.

We again split using Pythagorean theorem, $\|S_2^T C_T(I-P)\|_F^2 = \|S_2^T C_T\|_F^2 - \|S_2^T C_T P\|_F^2$. We set $\Delta = \|C_T\|_F^2 - \|S_2^T C_T\|_F^2$. We have $\mathbb{E}\,\|S_2 C_T\|_F^2 = \|C_T\|_F^2 \leq (1+\epsilon)\|A-A_k\|_F^2$, where the last bound follows since C is an (ϵ,k)-column PCP for A. Thus by a Markov bound, with probability 299/300, $|\Delta| \leq (1+\epsilon)300\|A-A_k\|_F^2 \leq 600\|A-A_k\|_F^2$.

Additionally, we have $\|C_T\|_2^2 \leq \frac{10\epsilon}{k}\|A-A_k\|_F^2$ and $\|S_2^T C_T\|_2^2 \leq \frac{10\epsilon}{k}\|A-A_k\|_F^2$ with high probability by an identical argument to that used for Lemma 8. This gives:

$$\left|\|S_2^T C_T P\|_F^2 - \|C_T P\|_F^2\right| \leq k(\|S_2^T C_T\|_2^2 + \|C_T\|_2^2) \leq 20\epsilon\|A-A_k\|_F^2$$

which gives the main bound (28) after adjusting constants on ϵ.

Cross Term: We want to show:

$$|\operatorname{tr}\left((I-P)C_T^T S_2 S_2^T C_H (I-P)\right)| \leq \epsilon\|C(I-P)\|_F^2. \tag{29}$$

We can write:

$$\begin{aligned}\left|\operatorname{tr}\left((I-P)C_T^T S_2 S_2^T C_H (I-P)\right)\right| &= \left|\operatorname{tr}\left(C_T^T S_2 S_2^T C_H (I-P)\right)\right| \\ &\qquad \text{(Cyclic property of trace and } (I-P)=(I-P)^2\text{)} \\ &= \left|\operatorname{tr}\left(C_T^T S_2 S_2^T C_H (C^TC)^+(C^TC)(I-P)\right)\right|\end{aligned}$$

where in the last step, inserting $(C^TC)^+(C^TC)$, which is the projection onto the row span of C has no effect as the rows of $C_H = U_H U_H^T C$ already lie in this span. $\langle M,N\rangle = \operatorname{tr}(M(C^TC)^+N^T)$ is a semi-inner product since C^TC is positive semidefinite, so by Cauchy-Schwarz:

$$\left|\operatorname{tr}\left((I-P)C_T^T S_2 S_2^T C_H (I-P)\right)\right| \leq \|C_T^T S_2 S_2^T C_H (C^TC)^{+/2}\|_F \cdot \|C(I-P)\|_F.$$

Using the singular value decomposition $C = XSY^T$:

$$\begin{aligned}\left|\operatorname{tr}\left(C_T^T S_2 S_2^T C_H (I-P)\right)\right| &\leq \|C_T^T S_2 S_2^T U_H U_H^T X\|_F \cdot \|C(I-P)\|_F \\ &\leq \|C_T^T S_2 S_2^T U_H\|_F \cdot \|C(I-P)\|_F \end{aligned} \tag{30}$$

As argued, by Lemma 32 we have $\tilde{\ell}_i \geq \|(U_H)_i\|_2^2$, and so by a standard approximate matrix multiplication result [DKM06], with probability 299/300 if we set the constant c' in our sample size $> c''$ for some fixed c'':

$$\begin{aligned}\|C_T^T S_2 S_2^T U_H\|_F &\leq \frac{\|U_H\|_F\|C_T\|_F}{\sqrt{\frac{t\cdot c''\|U_H\|_F^2}{\sum_i \tilde{\ell}_i}}}. \\ &\leq \epsilon\|C_T\|_F = O(\epsilon\|A-A_k\|_F)\end{aligned}$$

where the last step follows from the fact that $\|C_T\|_F^2 = O(\|A-A_k\|_F^2)$ since C is an (ϵ,k)-column PCP for A. The final bound then follows from combining with (30) with the fact that $\|A-A_k\|_F^2 \leq (1+\epsilon)\|C(I-P)\|_F^2$ and adjusting constants on ϵ by increasing the constant c' in the sample size t and c in the rank parameter $k' = \lceil ck/\epsilon\rceil$.

The full lemma follows simply from noting that we can union bound over our failure probability for each term so all hold simultaneously with probability 99/100. ∎

Lemma 33 (Leverage Score Sampling Matrix Chernoff)**.** *For any $A \in \mathbb{R}^{n\times d}$ for $i \in \{1,\ldots,d\}$, let $\ell_i(A) \stackrel{\text{def}}{=} a_i^T(AA^T)^+a_i$ be the i^{th} column leverage score of A and let $\tilde{\ell}_i \geq \ell_i(A)$ be an overestimate for this score. Let $p_i = \frac{\tilde{\ell}_i}{\sum_i \tilde{\ell}_i}$ and $t = \frac{c\log(d/\delta)}{\epsilon^2}\sum_i \tilde{\ell}_i$ for sufficiently large c. Construct C by sampling t columns of A, each set to $\frac{1}{\sqrt{tp_i}}a_i$ with probability p_i. With probability $1-\delta$:*

$$(1-\epsilon)CC^T \preceq AA^T \preceq (1+\epsilon)CC^T. \tag{31}$$

Proof. Write the singular value decomposition $A = U\Sigma V^T$. Note that:

$$\ell_i = e_i^T V\Sigma U^T\left(U\Sigma^2U^T\right)^+U\Sigma V^Te_i = \|v_i\|_2^2.$$

Let $Y = \Sigma^{-1}U^T\left(CC^T - AA^T\right)U\Sigma^{-1}$. We can write:

$$Y = \sum_{j=1}^t\left[\Sigma^{-1}U^T\left(c_jc_j^T - \frac{1}{t}AA^T\right)U\Sigma^{-1}\right] \stackrel{\text{def}}{=} \sum_{j=1}^t[X_j].$$

For each $j \in 1,\ldots,t$, X_j is given by:

$$X_j = \frac{1}{t}\cdot\Sigma^{-1}U^T\left(\frac{1}{p_i}a_ia_i^T - AA^T\right)U\Sigma^{-1} \text{ with probability } p_i.$$

$\mathbb{E}\,Y = 0$ since $\mathbb{E}\,X_j = \sum_{i=1}^d p_i\left[\frac{1}{p_i}a_ia_i^T - AA^T\right] = 0$. Furthermore, $CC^T = U\Sigma Y\Sigma U + AA^T$. Showing $\|Y\|_2 \leq \epsilon$ gives $-\epsilon I \preceq Y \preceq \epsilon I$, which gives the lemma. We prove this bound using a matrix Bernstein inequality from [Tro15]. This inequality requires upper bounds on the spectral norm of each X_j and on variance of Y. We first note that for any i, $\frac{1}{\ell_i(A)}a_ia_i^T \preceq AA^T$. This follows from writing any x in the column span of A as $(AA^T)^{+/2}y$ and then noting:

$$x^T\left(a_ia_i^T\right)x = y^T(AA^T)^{+/2}a_ia_i^T(AA^T)^{+/2}y \leq \ell_i(A)\|y\|_2^2$$

since $(AA^T)^{+/2}a_ia_i^T(AA^T)^{+/2}$ is rank-1 and so has maximum eigenvalue $\text{tr}\left((AA^T)^{+/2}a_ia_i^T(AA^T)^{+/2}\right) = \ell_i(A)$ by the cyclic property of trace. Further $x^TAA^Tx = y^T(AA^T)^{+/2}AA^T(AA^T)^{+/2}y = \|y\|_2^2$, giving us the bound. We then have:

$$\frac{1}{\ell_i(A)}\cdot\Sigma^{-1}U^Ta_ia_i^TU\Sigma^{-1} \preceq \Sigma^{-1}U^TAA^TU\Sigma^{-1} = I.$$

And so $\frac{1}{tp_i}\Sigma^{-1}U^Ta_ia_i^TU\Sigma^{-1} \preceq \frac{\epsilon^2}{c\log(d/\delta)\tilde{\ell}_i}\Sigma^{-1}U^Ta_ia_i^TU\Sigma^{-1} \preceq \frac{\epsilon^2}{c\log(d/\delta)}I$. Additionally,

$$\frac{1}{tp_i}\Sigma^{-1}U^TAA^TU\Sigma^{-1} \preceq \frac{\epsilon^2}{c\log(d/\delta)}I$$

as long as $\tilde{\ell}_i \leq 1$ which we may as well enforce since $\ell_i(A) \leq 1$. Overall this gives $\|X_j\|_2 \leq \frac{\epsilon^2}{c\log(d/\delta)}$. Next we bound the variance of Y.

$$\begin{aligned}
\mathbb{E}(Y^2) = t\cdot\mathbb{E}(X_j^2) &= \frac{1}{t}\sum p_i\cdot\left(\frac{1}{p_i^2}\Sigma^{-1}U^Ta_ia_i^TU\Sigma^{-2}U^Ta_ia_i^TU\Sigma^{-1}\right.\\
&\left.-2\frac{1}{p_i}\Sigma^{-1}U^Ta_ia_i^TU\Sigma^{-2}U^TAA^TU\Sigma^{-1} + \Sigma^{-1}U^TAA^TU\Sigma^{-2}U^TAA^TU\Sigma^{-1}\right)\\
&\preceq \frac{1}{t}\sum\left[\frac{\sum\tilde{\ell}_i}{\tilde{\ell}_i}\cdot\ell_i(A)\cdot\Sigma^{-1}Ua_ia_i^TU\Sigma^{-1}\right] - \frac{1}{t}\Sigma^{-1}U^TAA^TU\Sigma^{-1}\\
&\preceq \frac{\epsilon^2}{c\log(d/\delta)}\Sigma^{-1}U^TAA^TU\Sigma^{-1} \preceq \frac{\epsilon^2}{c\log(d/\delta)}I.
\end{aligned}$$

By Theorem 7.3.1 of [Tro15], for $\epsilon < 1$,

$$\Pr\left[\|Y\|_2 \geq \epsilon\right] \leq 4d \cdot e^{\frac{-\epsilon^2/2}{\left(\frac{\epsilon^2}{c\log(d/\delta)}(1+\epsilon/3)\right)}}.$$

Which gives $\Pr\left[\|Y\|_2 \geq \epsilon\right] \leq 4de^{-\frac{c\log(d/\delta)}{4}} \leq \delta$ if we choose c large enough. ☐

The above Lemma also gives an easy Corollary for the approximation obtained when sampling with ridge leverage scores:

Corollary 34 (Ridge Leverage Scores Sampling Matrix Chernoff)**.** *For any $A \in \mathbb{R}^{n\times d}$ for $i \in \{1,\ldots,d\}$, let $\tau_i^\lambda(A) \stackrel{\text{def}}{=} a_i^T(AA^T+\lambda I)^+ a_i$ be the i^{th} λ-ridge leverage score of A and let $\tilde{\tau}_i^\lambda \geq \tau_i^\lambda(A)$ be an overestimate for this score. Let $p_i = \frac{\tilde{\tau}_i^\lambda}{\sum_i \tilde{\tau}_i^\lambda}$ and $t = \frac{c\log(d/\delta)}{\epsilon^2}\sum_i \tilde{\tau}_i^\lambda$ for sufficiently large c. Construct C by sampling t columns of A, each set to $\frac{1}{\sqrt{tp_i}}a_i$ with probability p_i. With probability $1-\delta$:*

$$(1-\epsilon)CC^T - \epsilon\lambda I \preceq AA^T \preceq (1+\epsilon)CC^T + \epsilon\lambda I. \tag{32}$$

Proof. We can instantiate Lemma 33 with $[A, \sqrt{\lambda I}]$ setting $\tilde{\ell}_i = \tilde{\tau}_i^\lambda$. We simply fix the columns of the identity to appear in our sample. This only decreases variance, all calculations go through, and with probability δ:

$$(1-\epsilon)[C,\sqrt{\lambda}I][C,\sqrt{\lambda}I]^T \preceq [A,\sqrt{\lambda}I][A,\sqrt{\lambda}I]^T \preceq (1+\epsilon)[C,\sqrt{\lambda}I][C,\sqrt{\lambda}I]^T$$

which gives the desired bound if we subtract λI from all sides. ☐

C Outputting a PSD Low-Rank Approximation

In this section, we prove Theorem 12, showing to to efficiently output a low-rank approximation to A under the restriction that the approximation is also PSD. We start with the following lemma, which shows that as long as we have a good low rank subspace for approximating A in the spectral norm (computable using Theorem 25), we can quickly find a near optimal PSD low-rank approximation:

Lemma 35. *Given PSD $A \in \mathbb{R}^{n\times n}$ and orthonormal basis $Z \in R^{n\times m}$ with $\|A - AZZ^T\|_2^2 \leq \frac{\epsilon}{k}\|A-A_k\|_F^2$, there is an algorithm which accesses $O\left(\frac{nm\log m}{\epsilon^2}\right)$ entries of A, runs in $\tilde{O}\left(\frac{nm^{\omega-1}}{\epsilon^{2(\omega-1)}}\right)$ time, and with probability 99/100 outputs $M \in \mathbb{R}^{n\times k}$ satisfying $\|A-MM^T\|_F^2 \leq (1+\epsilon)\|A-A_k\|_F^2$.*

Proof. By Lemma 10 of [CW17] for any basis $Z \in \mathbb{R}^{n\times m}$ with $\|A-AZZ^T\|_2^2 \leq \frac{\epsilon}{k}\|A-A_k\|_F^2$,

$$\min_{X:rank(X)=k, X\succeq 0}\|A-ZXZ^T\|_F^2 \leq (1+O(\epsilon))\|A-A_k\|_F^2.$$

As in Algorithm 1 we can find a near optimal X by further sampling Z using its leverage scores. If we sample $t_1 = O\left(\frac{m\log m}{\epsilon^2}\right)$ rows of Z by their leverage scores (their norms since Z is orthonormal) to form $S_1 \in R^{n\times t_1}$, by Theorem 39 of [CW13], we will have an *affine embedding* of Z. Specifically, letting $B^* = \arg\min_B \|A-ZB\|_F^2$ and $E^* = A - ZB^*$, for any B we have:

$$\|S_1^TA - S_1ZB\|_F^2 + \left(\|E^*\|_F^2 - \|S_1^TE^*\|_F^2\right) \in \left[(1-\epsilon)\|A-ZB\|_F^2, (1+\epsilon)\|A-ZB\|_F^2\right].$$

Note that this is similar to the embedding property used in the proof of Theorem 9 to show that W computed in Step 6 of Algorithm 1 gave a near optimal low-rank approximation to AS_1.

By a Markov bound, since $\mathbb{E}\,\|S_1^T E^*\|_F^2 = \|E^*\|_F^2$, with probability 99/100, $\left|\|E^*\|_F^2 - \|S_1^T E^*\|_F^2\right| \leq 100\|E^*\|_F^2 = O(1)\|A - A_k\|_F^2$. This guarantees that a $(1+\epsilon)$ approximation to the sketched problem gives a $(1 + O(\epsilon))$ approximation to the original. That is, for any PSD $\tilde{X}$ with $\text{rank}(\tilde{X}) = k$, and

$$\|S_1^T A - S_1 Z \tilde{X} Z^T\|_F^2 \leq (1+\epsilon) \min_{X:\text{rank}(X)=k, X\succeq 0} \|S_1^T A - S_1 Z X Z^T\|_F^2$$

we have $\|A - Z\tilde{X}Z^T\|_F^2 \leq (1 + O(\epsilon))\|A - A_k\|_F^2$. Following [CW17], we write $S_1^T Z$ in its SVD $S_1^T Z = U_z \Sigma_z V_z^T$. Since $S_1 Z X Z^T$ falls in the column span of $S_1 Z$ and the row span of Z^T, we write $\delta \stackrel{\text{def}}{=} \|(I - U_z U_z^T) S_1^T A\|_F^2 + \|U_z U_z^T S_1^T A (I - ZZ^T)\|_F^2$ and by Pythagorean theorem have:

$$\begin{aligned}
\|S_1^T A - S_1 Z X Z^T\|_F^2 &= \|U_z U_z^T S_1 A Z Z^T - U_z \Sigma_z V_z^T X Z^T\|_F^2 + \delta \\
&= \|U_z^T S_1^T A Z - \Sigma_z V_z^T X\|_F^2 + \delta \\
&= \|\Sigma_z V_z^T \left(V_z \Sigma_z^{-1} U_z^T S_1^T A Z - X\right)\|_F^2 + \delta
\end{aligned}$$

Since S_1 is sampled via Z's leverage scores, $(1-\epsilon)I \preceq S_1^T Z \preceq (1+\epsilon)I$ and so:

$$\|S_1^T A Z - S_1 Z X Z^T\|_F^2 = (1 \pm \epsilon)\|V_z^T \Sigma_z^{-1} U_z^T S_1^T A Z - X\|_F^2 + \delta.$$

Finally, following [CW17], letting $B = V_z^T \Sigma_z^{-1} U_z^T S_1^T A Z$, we have

$$\tilde{X} = \arg\min_{X | X\succeq 0, \text{rank}(X)=k} \|B - X\|_F^2 = \left(B/2 + B^T/2\right)_{k,+}$$

where $N_{k,+}$ has all but the top k *positive* eigenvalues of N set to 0. We output $M = Z\tilde{X}^{1/2}$. Overall, the above algorithm requires accessing $O\left(\frac{nm \log m}{\epsilon^2}\right)$ entries of A (the entries of AS_1) and has runtime $\tilde{O}\left(\frac{nm^{\omega-1}}{\epsilon^{2(\omega-1)}}\right)$ giving the lemma. □

We can obtain Z with rank $m = \Theta(k/\epsilon)$ by applying Theorem 25 with rank $k' = \Theta(k/\epsilon)$ and error parameter $\epsilon' = \Theta(1)$. Combined with Lemma 35 this yields:

Theorem 12 (Sublinear Time Low-Rank Approximation – PSD Output)**.** *There is an algorithm that given any PSD $A \in \mathbb{R}^{n\times n}$, accesses $\tilde{O}\left(\frac{nk^2}{\epsilon^2} + \frac{nk}{\epsilon^3}\right)$ entries of A, runs in $\tilde{O}\left(\frac{nk^\omega}{\epsilon^\omega} + \frac{nk^{\omega-1}}{\epsilon^{3(\omega-1)}}\right)$ time and with probability at least 9/10 outputs $M \in \mathbb{R}^{n\times k}$ with:*

$$\|A - MM^T\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2.$$