Low-Rank PSD Approximation in Input-Sparsity Time

Kenneth L. Clarkson
IBM Research – Almaden
klclarks@us.ibm.com

David P. Woodruff
IBM Research – Almaden
dpwoodru@us.ibm.com

Abstract

We give algorithms for approximation by low-rank positive semidefinite (PSD) matrices. For symmetric input matrix $A \in \mathbb{R}^{n\times n}$, target rank k, and error parameter $\varepsilon > 0$, one algorithm finds with constant probability a PSD matrix $\tilde{Y}$ of rank k such that $\|A - \tilde{Y}\|_F^2 \leq (1+\varepsilon)\|A - A_{k,+}\|_F^2$, where $A_{k,+}$ denotes the best rank-k PSD approximation to A, and the norm is Frobenius. The algorithm takes time $O(\mathtt{nnz}(A)\log n) + n\mathrm{poly}((\log n)k/\varepsilon) + \mathrm{poly}(k/\varepsilon)$, where $\mathtt{nnz}(A)$ denotes the number of nonzero entries of A, and $\mathrm{poly}(k/\varepsilon)$ denotes a polynomial in k/ε. (There are two different polynomials in the time bound.) Here the output matrix $\tilde{Y}$ has the form $CUC^\top$, where the $O(k/\varepsilon)$ columns of C are columns of A. In contrast to prior work, we do not require the input matrix A to be PSD, our output is rank k (not larger), and our running time is $O(\mathtt{nnz}(A)\log n)$ provided this is larger than $n\mathrm{poly}((\log n)k/\epsilon)$. We give a similar algorithm that is faster and simpler, but whose rank-k PSD output does not involve columns of A, and does not require A to be symmetric. We give similar algorithms for best rank-k approximation subject to the constraint of symmetry. We also show that there are asymmetric input matrices that cannot have good symmetric column-selected approximations.

1 Introduction

A number of matrices that arise in machine learning and data analysis are symmetric positive semidefinite (PSD), including covariance matrices, kernel matrices, Laplacian matrices, random dot product graph models [21], and others. A common task related to such matrices is to approximate them with a low-rank matrix, for efficiency or statistical inference; spectral clustering, kernel PCA, manifold learning, and Gaussian process regression can all involve this task.

These matrices can be very large, and so there has been an increasing emphasis on efficiency in the task of low-rank approximation, even at the cost of some reduction in approximation quality. In recent years, methods for low-rank approximation based on random projections, row and column sampling, and other such *sketching* techniques have been found, that are quite efficient, with running times that in some situations are dominated by the number of nonzero entries $\texttt{nnz}(A)$ for input matrix $A \in \mathbb{R}^{n \times n}$ [20]. However, many of these techniques do not readily yield low-rank approximations that satisfy the fundamental constraint of being PSD, or indeed, even symmetric. (We will consider only symmetric PSD matrices, that is, being PSD will imply being symmetric, as is the usual convention. We will generally assume that input matrix A is symmetric.)

There is, however, a substantial literature on the *Nyström* method and its descendants; for a PSD matrix A, integer k, and error parameter $\varepsilon > 0$, these methods return a PSD low-rank approximation $CUC^\top$, where the rank $\mathrm{rk}(U)$ of U is k, the columns of C are columns of A, and the number of columns of C depends on k and ε.

These methods can be quite fast: the classical Nyström algorithm obtains the rows of C by uniform sampling, and thereafter obtains U by operations on C; that is, it can be sublinear in $\texttt{nnz}(A)$. However, the approximation error bounds for this method are weak, and $\Omega(\sqrt{n})$ samples may be required [19]. Some sharper bounds have been found, for the restricted class of low coherence matrices [14]. More generally, Nyström-related algorithms select columns of A under a sampling distribution that is adaptive, that is, derived from A. The probability of sampling a column is commonly proportional to a leverage score, for example the squared Euclidean norm of the corresponding column of a matrix whose rows comprise the top k eigenvectors of A. Such adaptive sampling methods yield sharper bounds; for example, with $\tilde{O}(k/\varepsilon^2)$ column samples, the approximation error $\|A - CUC^\top\|_F$ can be bounded by $\|A - A_k\|_F + \varepsilon\|A - A_k\|_*$, where A_k is the best rank-k approximation to A (not necessarily PSD), $\|\|_F$ is the Frobenius norm, and $\|\|_*$ is the trace (nuclear) norm [12]. (Since $\|A\|_*$ is the ℓ_1 norm of the singular values of A, and $\|A\|_F$ their Euclidean norm, $\|A\|_* \geq \|A\|_F$ and can be much larger.) The best current bound with respect to approximation quality (but not run-time) seems to be $\|A - CUC^\top\|_F \leq (1+\varepsilon)\|A - A_k\|_F$, with $O(k/\varepsilon)$ columns sampled [18].

While populating C from the columns of A is very attractive, since such columns are "representative" and heuristically as sparse as A, it is also of interest to construct C using non-adaptive methods such as random projections, that are typically faster and require fewer passes over the data. Here such a sketching matrix S, perhaps comprising independent Gaussian entries, or a sparse embedding matrix [5, 15] or an OSNAP [17, 2, 8] would be applied to A to obtain the sketch AS, with fewer columns than A, to be used to construct a low-rank approximation to A.

A natural idea to apply such sketches for approximation by symmetric or PSD matrices would be to adapt schemes for the asymmetric case, such as [4], where the approximation matrix with guaranteed small Frobenius relative error has the form $AR(SAR)^+SA$. The sketching matrices S and R are sign (± 1) matrices, and $()^+$ is the Moore-Penrose pseudo-inverse. Here one might try analogously $AR(R^\top AR)^+R^\top A$ as a PSD approximation to PSD matrix A, but the analysis does

not seem to extend to this idea; a similar scheme $A \approx AQ(Q^\top AQ)^+Q^\top A$ was proposed but not analyzed [13], where Q is an orthonormal basis for the columspace of AR. The same authors do give an analysis for $QQ^\top AQQ^\top$; these constructions are more expensive to compute than our goal here, however.

We refer the reader to [12] for further discussion and references on low-rank PSD approximation.

1.1 Our results

We give algorithms for symmetric rank-k approximation, and PSD rank-k approximation, with relative error bounds with respect to the Frobenius norm. Our most notable result is the following.

Theorem 1 *[Quoting Theorem 20] For given integer $k \geq 1$ and $\varepsilon > 0$, and symmetric $A \in \mathbb{R}^{n\times n}$, there is $m_B = O(k/\varepsilon)$ such that there are matrices $B \in \mathbb{R}^{n\times m_B}$ with each column of B a column of A, and $U \in \mathbb{R}^{m_B\times m_B}$ with $\mathrm{rk}(U) = k$ and PSD, with $\|A - BUB^\top\|_F^2 \leq (1+\varepsilon)\|A - A_{k,+}\|_F^2$. These matrices can be found in $O(\mathtt{nnz}(A)\log n) + (n+d)\mathrm{poly}((\log n)k/\varepsilon) + \mathrm{poly}(k/\varepsilon)$ time, with constant probability.*

Here $A_{k,+}$ is the best rank-k PSD approximation to A; the form of $A_{k,+}$ is described in Lemma 19.

Note that in contrast to almost all previous results:

- the work is $O(\mathtt{nnz}(A)\log n)$ for at least some inputs (large enough ε, dense enough A);
- the returned matrix has the target rank k, and not larger (not bicriteria);
- the input matrix need not be PSD;
- the number of columns is $O(k/\varepsilon)$, which is optimal (see e.g. [3] and its references);
- the approximation quality is relative error in Frobenius norm.

Only [18] has the same number of columns and approximation quality, but it does not have the other features. Only classical Nyström has a similar or faster running time, but its quality bounds are poor, as noted, except for restricted inputs. Several previous algorithms could be easily modified to have outputs that are rank k and/or PSD, but there are few reported that do so and have quality guarantees. An algorithm with rank-k outputs [10] is slower than that here, and has weaker quality bounds. Algorithms with PSD output for non-PSD input, and quality guarantees, do not seem to be reported. Since numerical errors, input errors, and other issues can result in matrices that "should" be PSD and are not, this is significant.

A recent related paper featuring a fast algorithm with strong performance guarantees [16] is for a related, but different problem: for kernels of the form $BB^\top$, a good low-rank approximation to B is found. Such an approximation is useful, but simple examples show that in general a $(1+\epsilon)$-approximation to B can be an arbitrarily bad (as a function of the singular values of B) approximation to $BB^\top$, so this does not address the problem we study.

While our algorithm does not require input A to be PSD, we do require that it be symmetric. Note, though, in the non-column-selection case, the best symmetric approximation to A is the best symmetric approximation to $(A+A^\top)/2$, see Lemma 13, and $(A+A^\top)/2$ can be formed in $\mathtt{nnz}(A)$ time. Thus, we can also quickly provide a good rank-k symmetric or PSD approximation to A even

when A is not symmetric. There is a natural question here: for general asymmetric A, might it be possible to select columns of A to obtain a matrix C, and rows of A to obtain a matrix R, such that there is a matrix U with CUR symmetric, and CUR close to A? Theorem 26 states that this is not possible.

We have the following result on column selection, using adaptive sampling, for symmetric (not necessarily PSD) approximation. We use the notation $A_{-k} \equiv A - A_k$.

Theorem 2 *[Quoting Theorem 17] For given integer $k \geq 1$ and $\varepsilon > 0$, and symmetric $A \in \mathbb{R}^{n \times n}$, there is $m_B = O(k/\varepsilon)$ such that there are matrices $B \in \mathbb{R}^{n \times m_B}$ with each column of B a column of A, and $U \in \mathbb{R}^{m_B \times m_B}$ with $\mathrm{rk}(U) = k$, with $\|A - BUB^\top\|_F^2 \leq (1+\varepsilon)\|A_{-k}\|_F^2$. These matrices can be found in $O(\mathtt{nnz}(A)\log n) + (n+d)\mathrm{poly}((\log n)k/\varepsilon) + \mathrm{poly}(k/\varepsilon)$ time, with constant probability.*

Note that the quality guarantee is with respect to the best rank-k approximation to A.

We have an algorithm for symmetric approximation using oblivious sketching; this algorithm is faster than for PSD approximation, and since the sketching matrices are oblivious, only one pass is needed over the data to obtain the sketches.

Theorem 3 *[Quoting Theorem 18] A matrix $\tilde{X}D\tilde{X}^\top$, where $\tilde{X} \in \mathbb{R}^{n \times k}$ and D is diagonal, such that*

$$\|A - \tilde{X}D\tilde{X}^\top\|_F^2 \leq (1+\varepsilon)\|A_{-k}\|_F^2$$

can be found in $O(\mathtt{nnz}(A)) + O(n\varepsilon^{-2-\gamma}k^{3+\gamma}) + \mathrm{poly}(k/\varepsilon)$ time.

The quantity γ can be arbitrarily small, at the cost of an increase of a constant factor in the runtime.

We have a similar result for PSD approximation.

Theorem 4 *[Quoting Theorem 21] A matrix $\tilde{Y}\tilde{Y}^\top$, where $\tilde{Y} \in \mathbb{R}^{n \times k}$, such that*

$$\|A - \tilde{Y}\tilde{Y}^\top\|_F^2 \leq (1+\varepsilon)\|A - A_{k,+}\|_F^2$$

can be found in $O(\mathtt{nnz}(A)) + O(n\varepsilon^{-2-\gamma}k^{3+\gamma}) + \mathrm{poly}(k/\varepsilon)$ time.

Finally, we have a variant algorithm for rank-k PSD approximation.

Theorem 5 *[Quoting Theorem 25] Let $t \equiv 2k/\varepsilon$. A PSD rank-k matrix $\tilde{Y}$ such that*

$$\|A - \tilde{Y}\|_F^2 \leq \|A - A_{k,+}\|_F^2 + \|A_{t+k} - A_t\|_F^2$$

can be found in $O(\mathtt{nnz}(A)) + O(n+d)\mathrm{poly}(k/\varepsilon) + \mathrm{poly}(k/\varepsilon)$ time.

Note that $\|A_{t+k} - A_t\|_F^2 \leq \varepsilon\|A_{-k}\|_F^2$, and can be that large. That is, this result is no better for general A than Theorem 3. However, some input matrices might have rapidly enough decaying spectrum that $\|A_{t+k} - A_t\|_F^2$ is much smaller than $\varepsilon\|A_{-k}\|_F^2$. This is true trivially if $\mathrm{rk}(A) < t$, but would also be true for matrices comprising the sum of a low-rank matrix and small-enough random noise. An example matrix `Protein` with rapid spectral decay is discussed in [12].

1.2 Techniques, and Outline

The algorithmic technology we use is sketching, including leverage-score sampling [20] and sparse embeddings, or more generally OSNAP [5, 17, 15, 8], allowing fast reduction of input matrices to smaller matrices whose columnspaces and rowspaces contain good approximations to the rows and columns of the input.

The machinery of our use of these techniques is given in §3, stating that there are thin matrices Z based on sketching (Lemma 11) and sampling (Lemma 12) that are fast to compute, and such that minimizing $\|ZXZ^\top - A\|_F$ over rank-k symmetric X, or $\|ZYZ^\top - A\|_F$ over rank-k PSD Y, gives good rank-k symmetric and PSD approximations to A.

Lemma 8 of §3 shows quantitatively how a good columnspace translates to a good low-rank PSD approximation, and a good low-rank symmetric approximation. Its proof involves standard matrix machinery, such as properties of the trace, the matrix Pythagorean theorem, properties of projections, and von Neumann's trace inequality.

We also use fast sketching to accelerate our solution to the optimization problem of finding the promised low-rank approximations within columnspaces; Lemma 15 of §4 gives our use of this machinery for this purpose.

Having shown that good columnspaces can be found quickly (as columnspaces of sketches), and that they can be used quickly to find low-rank approximations, we put these tools to use, first to find low-rank symmetric approximations, in §5, and then to find low-rank PSD approximations, in §6.

Our variant algorithm is given in §7, and impossibility result (for symmetric column-selected approximation of asymmetric matrices) in §8.

2 Notation and Preliminaries

Let $[A]_k$ denote the best rank-k approximation to matrix A, also written A_k when convenient. Let $[A]_{k,+}$ and $A_{k,+}$ denote the best positive semidefinite (PSD) rank-k approximation to A. We will often write A_{-k} for $A - A_k$.

Let $\mathcal{P}$ denote the (symmetric) PSD matrices.

Recall that A^+, the Moore-Penrose pseudo-inverse of A is equal to $(A^\top A)^{-1}A^\top$ when $A^\top A$ is invertible. For square A, let $A^{-\top}$ denote $(A^{-1})^\top$.

The spectral norm $\|A\|_2$ is the maximum of the singular values of A, while $\|A\|_F$ is the Euclidean norm of those singular values. Throughout we freely use the fact that $\|QX\|_F = \|X\|_F$ when Q has orthonormal columns.

In the following, unless otherwise mentioned, A is symmetric and has eigendecomposition $A = UDU^\top$. We have $[A]_k = U[D]_kU^\top$, where $[D]_k = D_k$ has i'th diagonal entry D_{ii} when D_{ii} is among the top k entries of D in magnitude, and zero otherwise.

We use throughout properties of the matrix trace $\mathtt{tr}\, X$, such as its linearity, and $\mathtt{tr}\, XY = \mathtt{tr}\, YX$, and $\|X\|_F^2 = \mathtt{tr}\, X^\top X$. We use the matrix version of the Pythagorean theorem: if matrices X and Y have $\mathtt{tr}\, X^\top Y = 0$, then $\|X+Y\|_F^2 = \|X\|_F^2 + \|Y\|_F^2$.

We use throughout standard properties of projection matrices P, such as $P^2 = P$, and $\|AP\|_F \leq \|A\|_F$. For projection P, let $\bar{P}$ be the projection $I - P$, so $P\bar{P} = 0$.

3 Good Projections and Sketches

In this section, we first define a particular property of projections, then show that the property implies that the columnspaces (and rowspaces) of the projections are useful for us. We then show how such projections can be found quickly, via sketching and via sampling.

Definition 6 (SF(ε, k) projections) *For given $A \in \mathbb{R}^{n\times n}$, say that projection $P \in \mathbb{R}^{n\times n}$ is* **SF**(ε, k) *for A if*

$$\|A\bar{P}\|_2^2 \leq (\varepsilon/k)\|A_{-k}\|_F^2.$$

For symmetric A, $\|\bar{P}A\|_2^2 = \|A\bar{P}\|_2^2$, so the same bound and condition are equivalent for $\bar{P}A$.

Lemma 7 *For symmetric $A, B \in \mathbb{R}^{n\times n}$ with $(A-B)B = 0$, and projection $P \in \mathbb{R}^{n\times n}$,*

$$\begin{aligned}\|A - PBP\|_F^2 = \|A - B\|_F^2 + \|B - PBP\|_F^2 \\ + 2\,\mathtt{tr}(A-B)\bar{P}BP.\end{aligned}$$

Proof: Using standard properties of the trace and the Frobenius norm, we have

$$\begin{aligned}&\|A - PBP\|_F^2 \\ &= \|A - B + B - PBP\|_F^2 \\ &= \|A - B\|_F^2 + \|B - PBP\|_F^2 \\ &\quad + 2\,\mathtt{tr}(A-B)(B-PBP) \\ &= \|A - B\|_F^2 + \|B - PBP\|_F^2 \\ &\quad + 2\,\mathtt{tr}(A-B)\bar{P}BP\end{aligned}$$

Here the last equality uses $(A-B)B = 0$ and the linearity of the trace.

■

Lemma 8 *Suppose $P \in \mathbb{R}^{n\times n}$ is a projection that is* **SF**(ε, k) *for A, that is, $\|A\bar{P}\|_2^2 \leq (\varepsilon/k)\|A_{-k}\|_F^2$. Then*

$$\|A - PA_{k,+}P\|_F^2 \leq (1 + O(\varepsilon))\Delta_{k,+},$$

where $\Delta_{k,+} \equiv \|A - A_{k,+}\|_F^2$. Also

$$\|A - PA_kP\|_F^2 \leq (1 + O(\varepsilon))\|A_{-k}\|_F^2.$$

Proof: We apply Lemma 7 with $B = A_{k,+}$, using $(A - A_{k,+})A_{k,+} = 0$, so

$$\begin{aligned}&\|A - PA_{k,+}P\|_F^2 \\ &= \|A - A_{k,+}\|_F^2 + \|A_{k,+} - PA_{k,+}P\|_F^2 \\ &\quad + 2\,\mathtt{tr}(A - A_{k,+})\bar{P}A_{k,+}P.\end{aligned}$$

For any symmetric B and projection P, using $P\bar{P} = 0$ and matrix Pythagorus,

$$\begin{aligned}\|B - PBP\|_F^2 = \|(I-P)B\|_F^2 \\ + \|PB\bar{P}\|_F^2 \leq 2\|B\bar{P}\|_F^2.\end{aligned}$$

Therefore we have for the middle term, via $\operatorname{rk} A_{k,+} = k$ and $\|A_{k,+}x\| \le \|Ax\|$ for all x,

$$\begin{aligned}\|A_{k,+} - PA_{k,+}P\|_F^2 &\le 2\|A_{k,+}\bar{P}\|_F^2 \\ &\le 2k\|A_{k,+}\bar{P}\|_2^2 \\ &\le 2k\|A\bar{P}\|_2^2\end{aligned}$$

and for the last term, using the von Neumann trace inequality,

$$\begin{aligned}&2\,\mathtt{tr}(A - A_{k,+})\bar{P}A_{k,+}P \\ &= 2\,\mathtt{tr}(A - A_{k,+})\bar{P}\bar{P}A_{k,+}P \\ &\le 2\sum_i \sigma_i((A - A_{k,+})\bar{P})\sigma_i(\bar{P}A_{k,+}P) \\ &\le 2k\|(A - A_{k,+})\bar{P}\|_2\|\bar{P}A_{k,+}\|_2 \\ &\le 2k\|A\bar{P}\|_2^2,\end{aligned}$$

and so, using the $\mathbf{SF}(\varepsilon, k)$ condition,

$$\begin{aligned}\|A - PA_{k,+}P\|_F^2 &\le \Delta_{k,+} + 4k\|A\bar{P}\|_2^2 \\ &\le \Delta_{k,+} + 4k(\varepsilon/k)\|A_{-k}\|_F^2 \\ &\le \Delta_{k,+} + O(\varepsilon)\Delta_{k,+} \\ &= (1 + O(\varepsilon))\Delta_{k,+},\end{aligned}$$

as claimed. The proof for the claim for $\|A - PA_kP\|_F^2$ is entirely analogous, substituting $\|A_{-k}\|_F^2$ for $\Delta_{k,+} = \|A - A_{k,+}\|_F^2$, and using $(A_{-k})A_k = 0$. ■

We need two technical lemmas, rephrasing and extending Lemmas 11 and 18 of [7].

Lemma 9 *For a given integer k, there is a matrix M with $\|MM^\top\|_2 \le 1$, and for integer k',*

$$\|M\|_F^2/\|M\|_2^2 \le k'$$

such that the following holds. Suppose R is a matrix drawn from a distribution such that for any $\varepsilon' < 1$ and $\delta < 1/2$, the following holds with failure probability δ:

$$\|MRR^\top M^\top - MM^\top\|_2 \le \varepsilon' \tag{1}$$

$$\left|\|MR\|_F^2 - \|M\|_F^2\right| \le \varepsilon' k'. \tag{2}$$

Suppose for given $\varepsilon < 1$ and k, the above holds for $\varepsilon' = O(1)$ and $k' = O(k/\epsilon)$. Then P_{AR} is $\mathbf{SF}(\varepsilon, k)$ for symmetric matrix A, where P_{AR} is the orthogonal projection onto the column span of AR.

Specifically, M is the matrix $\frac{1}{2}[Z^T; \frac{\sqrt{k'}}{c} \cdot \bar{P}_Z A]$, where Z has orthonormal columns is such that $\|\bar{P}_Z A\|_F^2 \le 2\|A_{-k}\|_F^2$ and $\|\bar{P}_Z A\|_2^2 \le \frac{2}{k}\|A_{-k}\|_F^2$, and where $c = \Theta(\|A_{-k'}\|_F)$ is otherwise arbitrary. Here $\bar{P}_Z = I - ZZ^\top$.

The value $\|M\|_F^2/\|M\|_2^2$ is called the *stable rank* of M.

Proof: We follow the chain of arguments in [7]. The proof of Lemma 18 of [7] shows the following (taking transposes in that proof). Let $\bar{A} = AR$. Let Z be a matrix whose columns form an orthonormal basis for the column span of $\bar{A}$. Suppose

$$\|\bar{P}_{Z'}A\|_2^2 \leq O(1)\left(\|A_{-k'}\|_2^2 + \frac{1}{k'}\|A_{-k'}\|_F^2\right), \tag{3}$$

where $\bar{P}_{Z'} = I - Z'(Z')^\top$, and $Z' \in \mathbb{R}^{n\times k'}$ is such that its columns form an orthonormal basis for the column span of $\bar{A}_{k'}$, which is the best rank-k' approximation to $\bar{A}$, and where $k' = O(k/\epsilon)$. Then,

$$\|\bar{P}_Z A\|_2^2 \leq O(\epsilon/k)\|A_{-k}\|_F^2,$$

where $\bar{P}_Z = I - ZZ^\top$, that is, $P_Z = P_{AR}$ is $\mathbf{SF}(O(\varepsilon), k)$ for A.

In [7], it is shown that (3) holds for any $\bar{A}$ which is a rank-k' spectral norm projection-cost preserving sketch of A with error $\epsilon' = O(1)$ and rank $k' = O(k/\epsilon)$, that is, for which for all rank-k' orthogonal projection matrices P,

$$\begin{aligned}
&(1-\epsilon')\|A - PA\|_2^2 - \frac{\epsilon'}{k'}\|A - PA\|_F^2 \\
&\leq \|\bar{A} - P\bar{A}\|_2^2 \\
&\leq (1+\epsilon')\|A - PA\|_2^2 + \frac{\epsilon'}{k'}\|A - PA\|_F^2.
\end{aligned} \tag{4}$$

In Theorem 27 of [7], (4) is shown to hold, via Lemma 26 of [7], provided $\bar{A}$ satisfies the conditions of Lemma 10 of [7]. Further, Lemma 10 of [7] is shown to hold if the following holds. Let $M = \frac{1}{2}[Z^T; \frac{\sqrt{k'}}{c}\cdot\bar{P}_Z A]$, where Z has orthonormal columns is such that $\|\bar{P}_Z A\|_F^2 \leq 2\|A_{-k}\|_F^2$ and $\|\bar{P}_Z A\|_2^2 \leq \frac{2}{k}\|A_{-k}\|_F^2$, and where $c = \Theta(\|A_{-k'}\|_F)$ and $c \geq \|A_{-k'}\|_F$.

Then the stable rank of M is at most k', and in the paragraph before section 7.1 of [7], it is shown that provided that (1) and (2) hold, then Lemma 10 of [7] holds. This completes the proof. ■

Lemma 10 *Let R_1 and R_2 be sketching or sampling matrices drawn from distributions that each satisfy the conditions of Lemma 9. Then $P_{AR_1R_2}$ is $\mathbf{SF}(O(\varepsilon), k)$ for A.*

Proof: It is enough to show (1) and (2) for R_1R_2 using the corresponding properties of R_1 and R_2. For the second condition, using the triangle inequality we have

$$\begin{aligned}
&|\|MR_1R_2\|_F^2 - \|M\|_F^2| \\
&\quad\leq |\|MR_1R_2\|_F^2 - \|MR_1\|_F^2| + |\|MR_1\|_F^2 - \|M\|_F^2| \\
&\quad\leq \varepsilon\|MR_1\|_F^2 + \varepsilon k \\
&\quad\leq \varepsilon(1+\varepsilon)k + \varepsilon k \leq 3\varepsilon k
\end{aligned}$$

for small ε. Here we use $\|M\|_F^2 \leq k$.

To show that (1) holds for R_1R_2, we appeal to Appendix A.3 of [9], which proves this property, using also the norm preservation property (2) that we assume. ■

Lemma 11 *For symmetric $A \in \mathbb{R}^{n\times n}$, for fixed $\gamma > 0$, there is $m_1 = O((k/\varepsilon)^{1+\gamma})$ and $m_2 = O(k/\varepsilon)$ such that there are distributions of oblivious sketching matrices $R_1 \in \mathbb{R}^{n\times m_1}$ and $R_2 \in \mathbb{R}^{m_1\times m_2}$ such that:*

- *AR_1 can be computed in $O(\texttt{nnz}(A))$ time, and AR_1R_2 can be computed in an additional $O(nm_1m_2)$ time.*
- *Projection $P_{AR_1R_2}$ onto the columnspace of AR_1R_2 is $\mathbf{SF}(\varepsilon, k)$ for A.*
- *There is rank-k symmetric X^* such that*

$$\|AR_1R_2X^*R_2^\top R_1^\top A - A\|_F^2 \leq (1+O(\varepsilon))\|A_{-k}\|_F^2.$$

- *There is rank-k PSD Y^* such that*

$$\|AR_1R_2Y^*R_2^\top R_1^\top A - A\|_F^2 \leq (1+O(\varepsilon))\|A-A_{k,+}\|_F^2.$$

Proof: We apply Lemma 9 together with Lemma 11 of [7], which says that if R_1 is an OSNAP of the given dimensions, then it satisfies the conditions of Lemma 9, and also that if R_2 is a dense JL matrix of the given dimensions, then R_2 satisfies the conditions of Lemma 9. We now apply Lemma 10, to obtain that $P_{AR_1R_2}$ is $\mathbf{SF}(O(\varepsilon), k)$ for A. From Lemma 8, this implies that

$$\|A - PA_{k,+}P\|_2 \leq (1+O(\varepsilon))\|A-A_{k,+}\|_F^2.$$

Since $A_{k,+}$ is rank-k, and PSD, and $PA_{k,+}P$ is PSD and has columns in $\texttt{colspace}(AR_1R_2)$ and rows in $\texttt{colspace}(AR_1R_2)^\top$, it follows that there is some PSD rank-k matrix Y^* such that

$$\|AR_1R_2Y^*R_2^\top R_1^\top A - A\|_F^2 \leq (1+O(\varepsilon))\|A-A_{k,+}\|_F^2,$$

as claimed. A similar argument applies to show the existence of X^* with the claimed properties. ■

We say a matrix R is a *sampling and rescaling matrix* if R samples the columns of a symmetric $n \times n$ matrix A with replacement from a distribution $p = (p_1, \ldots, p_n)$ on the columns of A, and if column j is sampled in the i-th trial, then $R_{j,i} = 1/\sqrt{rp_j}$.

Lemma 12 *For symmetric $A \in \mathbb{R}^{n\times n}$, there is $m_1 = \tilde{O}(k/\varepsilon)$ and $m_2 = O(k/\varepsilon)$ such that there are distributions of sampling and rescaling matrices $R_1 \in \mathbb{R}^{n\times m_1}$ and $R_2 \in \mathbb{R}^{m_1\times m_2}$ such that:*

- *R_1 and R_2 can be found in $O(\texttt{nnz}(A)\log n) + n\text{poly}((\log n)k/\varepsilon)$ time.*
- *AR_1R_2 can be computed in $O(\texttt{nnz}(A)\log n) + n\text{poly}((\log n)k/\varepsilon)$ time.*
- *Projection $P_{AR_1R_2}$ onto the columnspace of AR_1R_2 is $\mathbf{SF}(\varepsilon, k)$ for A.*
- *There is rank-k PSD X^* such that*

$$\|AR_1R_2X^*R_2^\top R_1^\top A - A\|_F^2 \leq (1+O(\varepsilon))\|A-A_{k,+}\|_F^2.$$

- *There is rank-k symmetric Y^* such that*

$$\|AR_1R_2Y^*R_2^\top R_1^\top A - A\|_F^2 \leq (1+O(\varepsilon))\|A_{-k}\|_F^2.$$

Proof: Let $M = \frac{1}{2}[Z^T; \frac{\sqrt{k'}}{c} \cdot \bar{P}_Z A]$, where Z has orthonormal columns is such that $\|\bar{P}_Z A\|_F^2 \leq 2\|A_{-k}\|_F^2$ and $\|\bar{P}_Z A\|_2^2 \leq \frac{2}{k}\|A_{-k}\|_F^2$, and where $c = \Theta(\|A_{-k'}\|_F)$ and $c \geq \|A_{-k'}\|_F$. Here $\bar{P}_Z = I - ZZ^\top$.

We apply Lemma 9 together with Lemma 11 of [7], which says that if R_1 is a sampling and rescaling matrix of the stated dimensions, where the sampling probabilities are proportional to the squared column norms of M, then R_1 satisfies the conditions of Lemma 9. Also, if R_2 is a BSS sampling matrix formed by applying the BSS procedure [1] to the matrix $M' = \frac{1}{2}[W^T; \frac{\sqrt{k'}}{c'} \cdot \bar{P}_W AR_1; v]$, where W has orthonormal columns is such that $\|\bar{P}_W AR_1\|_F^2 \leq 2\|(AR_1)_{-k}\|_F^2$ and $\|\bar{P}_W AR_1\|_2^2 \leq \frac{2}{k}\|(AR_1)_{-k}\|_F^2$, where $c' = \Theta(\|(AR_1)_{-k}\|_F)$ and $c' \geq \|(AR_1)_{-k}\|_F$, where $\bar{P}_W = I - WW^\top$, and where v is a row vector in which the i-th entry is the norm of the i-th column of $\frac{\sqrt{k'}}{c'}\bar{P}_W AR_1$ (for a discussion on why this v is included, see the paragraph preceding section 7.1 of [7]), then if R_2 has $m_2 = O(k/\varepsilon)$ columns then R_2 satisfies the conditions of Lemma 9. We now apply Lemma 10, to obtain that $P_{AR_1R_2}$ is $\mathbf{SF}(O(\varepsilon), k)$ for A. From Lemma 8, this implies that

$$\|A - PA_{k,+}P\|_2 \leq (1 + O(\varepsilon))\|A - A_{k,+}\|_F^2.$$

Since $A_{k,+}$ is rank-k, and PSD, and $PA_{k,+}P$ is PSD and has columns in $\texttt{colspace}(AR_1R_2)$ and rows in $\texttt{colspace}(AR_1R_2)^\top$, it follows that there is some PSD rank-k matrix Y^* such that

$$\|AR_1R_2Y^*R_2^\top R_1^\top A - A\|_F^2 \leq (1 + O(\varepsilon))\|A - A_{k,+}\|_F^2,$$

as claimed. A similar argument applies to show the existence of X^* with the claimed properties.

It remains to bound the running times.

To find R_1, we first need to find the matrix Z. Any rank-$2k$ matrix Z for which $\|\bar{P}_Z A\|_F^2 \leq 2\|A_{-2k}\|_F^2$ can be used, since this condition implies $\|\bar{P}_Z A\|_F^2 \leq 2\|A_{-k}\|_F^2$ and $\|\bar{P}_Z A\|_2^2 \leq \frac{2}{k}\|A_{-k}\|_F^2$. Such a Z can be found in $\texttt{nnz}(A) + n\text{poly}(k)$ time [5]. One can also compute an estimate c to $\|A_{-k}\|_F$ in this amount of time [5]. Finally, one can compute the squared column norms of $\bar{P}_Z A$ by left-multiplying by a matrix of i.i.d. Gaussian random variables with $O(\log n)$ rows, taking $O(\texttt{nnz}(A)\log n) + n\text{poly}(k \log n)$ time. Consequently, one can compute the sampling probabilities and form R_1 in $O(\texttt{nnz}(A)\log n) + n\text{poly}(k \log n)$ time.

To find R_2, we first need to find the matrix W. As in the previous paragraph, this can be done in $\texttt{nnz}(A) + n\text{poly}(k)$ time, given AR_1, and one can also compute an estimate c' in this amount of time. Since M' only has $O((k/\epsilon)\log(k/\epsilon))$ columns, we can explicitly form it. To apply the BSS procedure to the matrix M', we first left-multiply M' by a sparse subspace embedding T with $O(k^2/\epsilon^2 \log(k/\epsilon))$ columns, and then run the BSS procedure on TM'. As the column space of M' has dimension $O(k/\epsilon \log(k/\epsilon))$, T is a subspace embedding of this space space [5, 15, 17] with arbitrarily large constant probability. Then TM' is a $\text{poly}(k/\epsilon) \times \text{poly}(k/\epsilon)$ matrix, and running the BSS procedure [1] on it takes $\text{poly}(k/\epsilon)$ time. Moreover, as T is a subspace embedding, it preserves spectral and Frobenius norms, and so the matrix R_2 found by the BSS procedure applied to TM' is a valid output for the BSS procedure applied to M', by readjusting ϵ by a constant factor.

Having found R_1 and R_2, they can be quickly applied since they are sampling and rescaling matrices. ■

4 Fast Rank-k Approximation

As shown in the last section, good columnspaces can be found, but it remains to use those columnspaces to find good rank-k approximations. In this section, we first show how to how

to reduce from an arbitrary square matrix to a symmetric one, then to minimizing $\|BXB^\top - A\|_F^2$ over X.

Lemma 13 *Given $B \in \mathbb{R}^{n\times n}$,*

$$\begin{aligned}&\operatorname{argmin}_{\substack{X=X^\top \\ \mathrm{rk}(X)=k}} \|X - B\|_F \\ &\quad = \operatorname{argmin}_{\substack{X=X^\top \\ \mathrm{rk}(X)=k}} \|X - (B+B^\top)/2\|_F.\end{aligned}$$

Proof: For symmetric X, $X - (B+B^\top)/2$ is symmetric; since $(B-B^\top)/2$ is anti-symmetric,

$$\mathtt{tr}[(X - (B+B^\top)/2)(B-B^\top)/2] = 0,$$

and so by matrix Pythagoras

$$\|X - B\|_F^2 = \|X - (B+B^\top)/2\|_F^2 + \|(B-B^\top)/2\|_F^2.$$

■

Thus when finding the best symmetric approximation to A, we can assume that A is symmetric.

Lemma 14 *For symmetric $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ with $m \le n$ and full column rank, the minimizer*

$$X^* \equiv \operatorname{argmin}_{\mathrm{rk}(X)=k} \|BXB^\top - A\|_F^2 \tag{5}$$

is $X^* = T^{-1}[Z^\top AZ]_k T^{-\top}$, a symmetric matrix, where $B = ZT$ with Z having orthonormal columns and T upper triangular. For Y^* the optimal rank-k solution under a PSD constraint, we have $Y^* = T^{-1}[Z^\top AZ]_{k,+}T^{-\top}$.

Proof: Since B has full column rank, the decomposition $B = ZT$ where Z has orthonormal columns and T is upper triangular, has the property that T is invertible, which we will now assume. To solve (5), we can obtain $X_0^* \equiv \operatorname{argmin}_{\mathrm{rk}(X)=k} \|ZXZ^\top - A\|_F^2$ and recover X^* as $T^{-1}X_0^*T^{-\top}$. Here we have $X_0^* = [Z^\top AZ]_k$. (This follows from a more general result [11], but here follows from the properties of the trace and Frobenius norm:

$$\begin{aligned}&\|ZXZ^\top - A\|_F^2 \\ &= \|ZXZ^\top\|_F^2 + \|A\|_F^2 - 2\,\mathtt{tr}\,AZXZ^\top \\ &= \|X\|_F^2 + \|Z^\top AZ\|_F^2 - 2\,\mathtt{tr}\,Z^\top AZX \\ &\quad + (\|A\|_F^2 - \|Z^\top AZ\|_F^2) \\ &= \|X - Z^\top AZ\|_F^2 + (\|A\|_F^2 - \|Z^\top AZ\|_F^2)\end{aligned}$$

so that

$$\begin{aligned}&\operatorname{argmin}_{\mathrm{rk}(X)=k} \|ZXZ^\top - A\|_F^2 \\ &\quad = \operatorname{argmin}_{\mathrm{rk}(X)=k} \|X - Z^\top AZ\|_F^2 = [Z^\top AZ]_k,\end{aligned}$$

as claimed.)

So the solution to (5) is, as claimed, $T^{-1}[Z^\top AZ]_k T^{-\top}$, a symmetric matrix. A similar argument holds for the PSD case. ■

Lemma 15 *Let k, A, B, X^*, and Y^* be as in Lemma 14. For given ε, a symmetric rank-k matrix $\tilde{X}$ can be found in*

$$O(\mathtt{nnz}(A) + \mathtt{nnz}(B)) + \mathrm{poly}(mk/\varepsilon)$$

time, such that

$$\|B\tilde{X}B^\top - A\|_F^2 \leq (1+\varepsilon)\|BX^*B^\top - A\|_F^2.$$

Moreover, a PSD rank-k matrix $\tilde{Y}$ can be found in the same time bound, such that

$$\|B\tilde{Y}B^\top - A\|_F^2 \leq (1+\varepsilon)\|BY^*B^\top - A\|_F^2.$$

Proof: A similar scheme to the following was used in [3] and elsewhere. We use oblivious sparse embeddings [5, 17, 15, 8], to quickly reduce to subproblems of size $\mathrm{poly}(k/\varepsilon)$.

First we use sparse embeddings to reduce to a more well-conditioned problem, as follows. We have that there is $m_S = O(m^2/\varepsilon^2)$ such that there is a sparse embedding matrix $S \in \mathbb{R}^{m_S \times n}$ such that with constant probability, S has the sparse embedding property that for all x, $\|SBx\|_2 = (1 \pm \varepsilon)\|Bx\|$, and SB can be computed in $O(\mathtt{nnz}(B))$ time. Suppose that indeed S is a sparse embedding. Given B, we compute the decomposition $SB = QT$, where Q has orthonormal columns and T is upper triangular. Moreover, since B has full rank, so does SB, and so T is invertible, and $Z \equiv BT^{-1}$ has singular values $1 \pm \varepsilon$. (Please note: we leave Z in factored form: we do not explicitly compute it.)

We will find rank-k symmetric X_0 with $\|ZX_0Z^\top - A\|_F^2 \leq (1+\varepsilon)\min_{\substack{\mathrm{rk}(X)=k \\ X=X^\top}} \|ZXZ^\top - A\|_F^2$, and then return $\tilde{X} = T^{-1}X_0T^{-\top}$. (Again, $T^{-\top} \equiv (T^{-1})^\top$.)

We now want to minimize $\|ZXZ^\top - A\|_F^2$, over rank-k symmetric X.

From [3], we have the following. There is $m_\ell, m_r = \mathrm{poly}(k/\varepsilon)$ so that there are sparse embedding distributions such that for $S_\ell \in \mathbb{R}^{m_\ell \times n}$ and $S_r \in \mathbb{R}^{n \times m_r}$ under those distributions,

$$\tilde{W}, \tilde{V} \equiv \mathrm{argmin}_{\substack{W \in \mathbb{R}^{m \times k} \\ V \in \mathbb{R}^{k \times m}}} \|S_\ell ZWVZ^\top S_r - S_\ell A S_r\|_F^2$$

satisfies

$$\|Z\tilde{W}\tilde{V}Z^\top - A\|_F^2 \leq (1+\varepsilon) \min_{\substack{W \in \mathbb{R}^{m \times k} \\ V \in \mathbb{R}^{k \times m}}} \|ZWVZ^\top - A\|_F^2.$$

Moreover $S_\ell A S_r$ can be computed in $O(\mathtt{nnz}(A))$ time, and $S_\ell Z = S_\ell BT^{-1}$ and $Z^\top S_r = T^{-\top}B^\top S_r$ can be computed in $O(\mathtt{nnz}(B) + m^2(m_\ell + m_r))$ time, by computing $(S_\ell B)T^{-1}$ and $T^{-\top}(B^\top S_r)$. The matrices $S_\ell Z$ and $Z^\top S_r$ are well-conditioned: they have singular values all $1 \pm \varepsilon$, and have full column (resp. row) rank.

Also: with constant probability, for S_ℓ and S_r under these distributions, the minimizer X^* to $\min_{\substack{\mathrm{rk}(X)=k \\ X=X^\top}} \|ZXZ^\top - A\|_F^2$ satisfies the condition that:

$$\|S_\ell(ZX^*Z^\top - A)S_r\|_F = (1 \pm \varepsilon)\|ZX^*Z^\top - A\|_F. \tag{6}$$

Therefore if we constrain $\tilde{W}, \tilde{V}$ to have $\tilde{W}\tilde{V}$ symmetric, the resulting solution X_0 will have cost $\|ZX_0Z^\top - A\|_F^2$ within $1+\varepsilon$ of the cost of using X^*, so it is enough to find

$$X_0 \equiv \mathrm{argmin}_{\substack{\mathrm{rk}(X)=k \\ X=X^\top}} \|S_\ell ZXZ^\top S_r - S_\ell A S_r\|_F^2.$$

Let $S_\ell Z$ and $Z^\top S_r$ have (economical) SVDs $S_\ell Z = U_\ell \Sigma_\ell V_\ell^\top$ and $Z^\top S_r = U_r \Sigma_r V_r^\top$, so that Σ_ℓ and Σ_r are invertible $m \times m$ matrices. Let $\hat{U}_\ell$ and $\hat{V}_r$ be such that $[\, U_\ell \ \hat{U}_\ell \,]$ and $[\, V_r \ \hat{V}_r \,]$ are orthogonal matrices. We have, using matrix Pythagorus,

$$\begin{aligned}
&\|S_\ell Z X Z^\top S_r - S_\ell A S_r\|_F^2 \\
&= \|U_\ell \Sigma_\ell V_\ell^\top X U_r \Sigma_r V_r^\top - S_\ell A S_r\|_F^2 \\
&= \|\Sigma_\ell V_\ell^\top X U_r \Sigma_r - U_\ell^\top S_\ell A S_r V_r\|_F^2 \\
&\quad + \|U_\ell^\top S_\ell A S_r \hat{V}_r\|_F^2 + \|\hat{U}_\ell^\top S_\ell A S_r V_r\|_F^2.
\end{aligned}$$

Since the last two terms are constant and nonnegative, it is enough to minimize

$$\begin{aligned}
&\|\Sigma_\ell V_\ell^\top (X - V_\ell \Sigma_\ell^{-1} U_\ell^\top S_\ell A S_r V_r \Sigma_r^{-1} U_r^\top) U_r \Sigma_r\|_F^2 \\
&= \|\Sigma_\ell V_\ell^\top (X - \hat{A}) U_r \Sigma_r\|_F^2,
\end{aligned}$$

where

$$\begin{aligned}
\hat{A} &\equiv V_\ell \Sigma_\ell^{-1} U_\ell^\top S_\ell A S_r V_r \Sigma_r^{-1} U_r^\top \\
&= (S_\ell Z)^+ S_\ell A S_r (Z^\top S_r)^+.
\end{aligned}$$

By the well-conditioned-ness properties, for any matrix D, $\|\Sigma_\ell V_\ell^\top D U_r \Sigma_r\|_F = (1 \pm \varepsilon)\|D\|_F$, and so it suffices to solve $\min_{\substack{\mathrm{rk}(X)=k \\ X=X^\top}} \|X - \hat{A}\|_F^2$. From Lemma 13 above, the minimizer is $[(\hat{A} + \hat{A}^\top)/2]_k$; this requires $\mathrm{poly}(k/\varepsilon)$ time to find, given $\hat{A}$. Note that $S_\ell Z \in \mathbb{R}^{m_\ell \times m}$, and is computable in $\mathtt{nnz}(B) + \mathrm{poly}(mk/\varepsilon)$ time, computing it by writing it as $(S_\ell B) T^{-1}$.

Recovering $\tilde{X}$ as $T^{-1} X_0 T^{-\top}$, as noted above, the lemma statement for symmetric approximation follows.

For the PSD case, the same argument applies, except that X^* becomes Y^*, and S_ℓ and S_r satisfy (6) for the corresponding expression with Y^*, and finally, we find the best rank-k PSD approximation to $\hat{A}$, which is $[(\hat{A} + \hat{A}^\top)/2]_{k,+}$. ■

5 Rank-k Symmetric Approximation

In this section, we put the machinery together to show that a sampling and sketching can be used to obtain good rank-k symmetric approximations. For the sampling case, we give a scheme using optimal rank-k *CUR decompositions*.

Lemma 16 ([3]) *For $A \in \mathbb{R}^{n \times d}$, given integer $k \geq 1$ and $\varepsilon > 0$, there are $m_C = O(k/\varepsilon)$ and $m_R = O(k/\varepsilon)$ such that there are are matrices*

- *$C \in \mathbb{R}^{n \times m_C}$ with each column of C a column of A;*
- *$U \in \mathbb{R}^{m_C \times m_R}$ with $\mathrm{rk}(U) = k$;*
- *$R \in \mathbb{R}^{m_R \times d}$, with each row of R a row of A, and with*
- *$\|A - CUR\|_F^2 \leq (1 + \varepsilon)\|A_{-k}\|_F^2$.*

These matrices can be found in $O(\mathtt{nnz}(A)\log n) + (n+d)\mathrm{poly}((\log n)k/\varepsilon) + \mathrm{poly}(k/\varepsilon)$ *time.*

Theorem 17 *For given integer* $k \geq 1$ *and* $\varepsilon > 0$*, and symmetric* $A \in \mathbb{R}^{n\times n}$*, there is* $m_B = O(k/\varepsilon)$ *such that there are matrices* $B \in \mathbb{R}^{n\times m_B}$ *with each column of* B *a column of* A*, and* $U \in \mathbb{R}^{m_B\times m_B}$ *with* $\mathrm{rk}(U) = k$*, with* $\|A - BUB^\top\|_F^2 \leq (1+\varepsilon)\|A_{-k}\|_F^2$*. These matrices can be found in* $O(\mathtt{nnz}(A)\log n) + (n+d)\mathrm{poly}((\log n)k/\varepsilon) + \mathrm{poly}(k/\varepsilon)$ *time, with constant probability.*

(Here poly() appears twice because the associated polynomials are different.)

Proof: As with [18], we use the optimal CUR decomposition of [3] as a black box.

Let C, U, and R be the matrices of Lemma 16 for the given k and ε. Let $B \equiv [\,C\ R^\top\,]$. Let U have the factorization WV for $W \in \mathbb{R}^{m_C\times k}$, $V \in \mathbb{R}^{k\times m_R}$. Then

$$\hat{U} \equiv \begin{bmatrix} W \\ 0_{m_R\times k} \end{bmatrix} [\,0_{k\times m_C}\ V\,]$$

has $B\hat{U}B^\top = CUR$, so $\hat{U}$ has

$$\|B\hat{U}B^\top - A\| \leq (1+\varepsilon)\|A_{-k}\|_F^2.$$

From Lemma 14 and Lemma 15, there are symmetric rank-k matrices X^* and $\tilde{X}$ with

$$\begin{aligned}\|B\tilde{X}B^\top - A\|_F^2 &\leq (1+\varepsilon)\|BX^*B^\top - A\|_F^2 \\ &\leq (1+\varepsilon)\|B\hat{U}B^\top - A\|_F^2 \\ &\leq (1+\varepsilon)^2\|A_{-k}\|_F^2,\end{aligned}$$

where $\tilde{X}$ can be found in $O(\mathtt{nnz}(B)) + \mathrm{poly}(mk/\varepsilon) = \mathrm{poly}(k/\varepsilon)$ time. Returning $\tilde{X}$ as U, and adjusting constants in the quality bounds, the theorem follows. ■

Theorem 18 *A matrix* $\tilde{X}D\tilde{X}^\top$*, where* $\tilde{X} \in \mathbb{R}^{n\times k}$ *and* D *is diagonal, such that*

$$\|A - \tilde{X}D\tilde{X}^\top\|_F^2 \leq (1+\varepsilon)\|A_{-k}\|_F^2$$

can be found in $O(\mathtt{nnz}(A)) + O(n\varepsilon^{-2-\gamma}k^{3+\gamma}) + \mathrm{poly}(k/\varepsilon)$ *time.*

Proof: Let R_1 and R_2 be as in Lemma 11. By that lemma, it suffices to solve

$$\min_{\substack{X=X^\top \\ \mathrm{rk}(X)=k}} \|AR_1R_2XR_2^\top R_1^\top A - A\|_F^2.$$

We apply Lemma 15, with B of that lemma AR_1R_2, and $m = \mathrm{poly}(k/\varepsilon)$; this yields a solution X_0 with $\tilde{A} \equiv AR_1R_2X_0R_2^\top R_1^\top A$ with distance to A within $1+\varepsilon$ of best possible, so that with the distance bound of Lemma 11 we have $\tilde{A}$ within $(1+O(\varepsilon))$ of best possible distance to A of a rank-k matrix.

We modify the procedure of Lemma 15 slightly, so that sketching by S, S_ℓ and S_r is done before sketching by R_1 and R_2; that is, the multiplication is $(SAR_1)R_2$, and so on. This implies that all such work takes $\mathtt{nnz}(A) + \mathrm{poly}(k/\varepsilon)$.

It remains to compute $\tilde{X} = AR_1R_2X_1$, where rank-k matrix X_0 has the eigenexpansion $X_0 = X_1DX_1^\top$. We compute in the order $AR_1(R_2X_1)$, taking $O(n\varepsilon^{-2-\gamma}k^{3+\gamma})$ as claimed. The result follows. ■

6 Rank-k PSD Approximation

The well-known general form of $A_{k,+}$ is given in the following lemma, with a proof included for completeness.

Lemma 19 *For $A = UDU^\top$ and $A_{k,+}$ as above, $A_{k,+} = UD_{k,+}U^\top$, and $D_{k,+}$ has i'th diagonal entry D_{ii} when D_{ii} is among the top k nonnegative entries of D, and zero otherwise.*

Proof: If B is PSD and rank k, then so is $U^\top BU$, and so (recalling the eigendecomposition $A = UDU^\top$)

$$\begin{aligned}\|B - A\|_F^2 &= \|U^\top BU - D\|_F^2 \\ &\geq \|D_{k,+} - D\|_F^2 \\ &= \|UD_{k,+}U^\top - A\|_F^2,\end{aligned}$$

and so $A_{k,+} = UD_{k,+}U^\top$.

Since nonzero off-diagonal entries for $D_{k,+}$ can only increase the distance to D, and similarly for positive diagonal entries for the i'th entry of $D_{k,+}$ when $D_{ii} < 0$, we have $D_{k,+} = [D_+]_k$, where D_+ has i'th entry equal to D_{ii} when that entry is positive, and all other entries zero. The lemma follows. ■

Theorem 20 *For given integer $k \geq 1$ and $\varepsilon > 0$, and symmetric $A \in \mathbb{R}^{n\times n}$, there is $m_B = O(k/\varepsilon)$ such that there are matrices $B \in \mathbb{R}^{n\times m_B}$ with each column of B a column of A, and $U \in \mathbb{R}^{m_B\times m_B}$ with $\mathrm{rk}(U) = k$ and PSD, with $\|A - BUB^\top\|_F^2 \leq (1+\varepsilon)\|A - A_{k,+}\|_F^2$. These matrices can be found in $O(\mathtt{nnz}(A)\log n) + n\mathrm{poly}((\log n)k/\varepsilon) + \mathrm{poly}(k/\varepsilon)$ time, with constant probability.*

Proof: We use the sampling matrices R_1 and R_2 of Lemma 12. It suffices to solve

$$\min_{\substack{Y\in\mathcal{P}\\ \mathrm{rk}(Y)=k}} \|AR_1R_2YR_2^\top R_1^\top A - A\|_F^2,$$

recalling that $\mathcal{P}$ is the set of PSD matrices. We apply Lemma 15 with $B = AR_1R_2$, and use the PSD case of the lemma. ■

Theorem 21 *A matrix $\tilde{Y}\tilde{Y}^\top$, where $\tilde{Y} \in \mathbb{R}^{n\times k}$, such that*

$$\|A - \tilde{Y}\tilde{Y}^\top\|_F^2 \leq (1+\varepsilon)\|A - A_{k,+}\|_F^2$$

can be found in $O(\mathtt{nnz}(A)) + O(n\varepsilon^{-2-\gamma}k^{3+\gamma}) + \mathrm{poly}(k/\varepsilon)$ time.

Proof: The proof is very close to that of Theorem 18. Let R_1 and R_2 be as in Lemma 11. By that lemma, it suffices to solve

$$\min_{\substack{Y\in\mathcal{P}\\ \mathrm{rk}(Y)=k}} \|AR_1R_2YR_2^\top R_1^\top A - A\|_F^2.$$

We apply Lemma 15, with B of that lemma AR_1R_2, and $m = \mathrm{poly}(k/\varepsilon)$, and using the PSD case; this yields a solution Y_0 with $\tilde{A} \equiv AR_1R_2Y_0R_2^\top R_1^\top A$ with distance to A within $1+\varepsilon$ of best possible,

so that with the distance bound of Lemma 11 we have $\tilde{A}$ within $(1+O(\varepsilon))$ of best possible distance to A of a rank-k PSD matrix.

We modify the procedure of Lemma 15 slightly, so that sketching by S, S_ℓ and S_r is done before sketching by R_1 and R_2; that is, the multiplication is $(S_\ell A R_1)R_2$, and so on. This implies that all such work takes $\mathtt{nnz}(A) + \mathrm{poly}(k/\varepsilon)$ time.

It remains to compute $\tilde{Y} = AR_1R_2Y_1D^{1/2}$, where rank-$k$ matrix Y_0 has the eigenexpansion $Y_0 = Y_1DY_1^\top$, and D has nonnegative entries. We compute in the order $AR_1(R_2Y_1D^{1/2})$, taking $O(n\varepsilon^{-2-\gamma}k^{3+\gamma})$ time as claimed. The result follows. ■

7 Sketched PSD: a Sharper Quality Bound

Here we give an alternative scheme for low-rank PSD approximation using sketching. As discussed in the introduction, while the quality bound for this scheme is no better in the worst case than that of Theorem 21, it can be better for input matrices A with rapidly decaying spectra.

Lemma 22 *If P is a rank-k projection, then $\|PA\|_F^2 \le \|A_k\|_F^2$.*

Proof: Omitted. ■

Lemma 23 *For symmetric $A, B \in \mathbb{R}^{n\times n}$ and projection $P \in \mathbb{R}^{n\times n}$,*

$$\|A - PBP\|_F^2 = \|A - PAP\|_F^2 + \|P(A-B)P\|_F^2. \tag{7}$$

Proof: We have

$$\begin{aligned}
&\mathtt{tr}(A-PAP)P(A-B)P \\
&\qquad = \mathtt{tr}\,P(A-PAP)P(A-B) \\
&\qquad = \mathtt{tr}(PAP - PAPP)(A-B) = 0,
\end{aligned}$$

and so by matrix Pythagoras,

$$\begin{aligned}
\|A - PBP\|_F^2 &= \|A - PAP + PAP - PBP\|_F^2 \\
&= \|A - PAP\|_F^2 + \|P(A-B)P\|_F^2,
\end{aligned} \tag{8}$$

as claimed. ■

Lemma 24 *Let symmetric $X \in \mathbb{R}^{n\times n}$ have rank $t \ge 2k/\varepsilon$, and $\|A - X\|_F^2 \le (1+\varepsilon/2)\Delta_t$, where $\Delta_t \equiv \|A - A_t\|_F^2$. Let P project onto the rowspace (or columnspace) of X. Then*

$$\|A - PA_{k,+}P\|_F^2 \le \|A - A_{k,+}\|_F^2 + \|A_{t+k} - A_t\|_F^2.$$

Proof: If the eigendecomposition of X is $X = ZLZ^\top$ and $P = ZZ^\top$, then

$$\begin{aligned}
\|A - PXP\|_F^2 &= \|A - ZZ^\top ZLZZ^\top ZX\|_F^2 \\
&= \|A - X\|_F^2 \le (1+\varepsilon/2)\|A - A_t\|_F^2.
\end{aligned}$$

That is, $\min_{\mathrm{rk}(W)=t} \|A - PWP\|_F^2 \le (1+\varepsilon/2)\Delta_t$.

Using (7),

$$\begin{aligned}(1+\varepsilon/2)\Delta_t &\geq \min_{\mathrm{rk}(W)=t} \|A - PWP\|_F^2 \\ &\geq \min_W \|A - PWP\|_F^2 \\ &= \min_W \|A - PAP\|_F^2 + \|P(A-W)P\|_F^2 \\ &= \|A - PAP\|_F^2. \end{aligned} \tag{9}$$

Using (7) again, and then (9) and Lemma 22,

$$\begin{aligned}\|A - PA_{k,+}P\|_F^2 &= \|A - PAP\|_F^2 + \|P(A - A_{k,+})P\|_F^2 \\ &\leq (1+\varepsilon/2)\|D - D_t\|_F^2 + \|[D - D_{k,+}]_t\|_F^2.\end{aligned}$$

Ordering the entries of D by magnitude, D_{ii} is counted in the sums for both of the last two norms just above only if $i \in (t, t+k]$, so up to a $(1+\varepsilon/2)$ factor,

$$\begin{aligned}\|A - PA_{k,+}P\|_F^2 &\leq \|A - A_{k,+}\|_F^2 + \sum_{i\in(t,t+k]} D_{ii}^2 \\ &\leq \|A - A_{k,+}\|_F^2 + \frac{k}{t}\|A_{-k}\|_F^2 \\ &\leq (1+\varepsilon)\|A - A_{k,+}\|_F^2\end{aligned}$$

for $t \geq 2k/\varepsilon$. The lemma follows by adjusting ε by a constant factor. ■

Theorem 25 *Let $t \equiv 2k/\varepsilon$. A PSD rank-k matrix $\tilde{Y}$ such that*

$$\|A - \tilde{Y}\|_F^2 \leq \|A - A_{k,+}\|_F^2 + \|A_{t+k} - A_t\|_F^2$$

can be found in $O(\mathtt{nnz}(A)) + O(n+d)\mathrm{poly}(k/\varepsilon) + \mathrm{poly}(k/\varepsilon)$ time.

Proof: We use the algorithm of Theorem 18 with k of that lemma equal to the given $t = 2k/\varepsilon$. The projection P onto the columnspace of $\tilde{X}$ satisfies Lemma 24 just above. Suppose $P = BB^\top$ for B with orthonormal columns. It is enough to solve

$$\min_{\substack{Y\in\mathcal{P}\\ \mathrm{rk}(Y)=k}} \|A - BYB^\top\|_F^2,$$

for which Lemma 15 gives a fast approximate solution. ■

8 Matrices With No Symmetric CUR Approximations

While we have found symmetric CUR decompositions of symmetric matrices in Theorem 17, one could ask if we can find symmetric CUR decompositions of asymmetric matrices. We show that this is not possible.

Theorem 26 *Let $k \geq 4$, and let A be an $n \times n$ asymmetric matrix. Suppose we want to find a subset C of columns of A and a subset R of rows of A for which*

1. There exists a matrix U for which

$$\|CUR - A\|_F^2 \leq \alpha \cdot \|A_{-k}\|_F^2$$

2. CUR is a symmetric matrix,

where $\alpha \geq 1$ is an approximation factor. Then there exist matrices A for which no such C and R exist, for arbitrarily large $\alpha \geq 1$.

Proof: We first prove the result for $k = 4$.

Suppose A has the following form: its first row equals $(\beta, 0, 0, \ldots, 0)^T$, its second row equals $(1, \beta, 1, 1, \ldots, 1, 0)^T$, and all remaining rows equal $(1, 0, 0, \ldots, 0)^T$, where β is an arbitrarily large real number.

The column span of A is

$$\texttt{span}\{(\beta, 1, 1, \ldots, 1), (0, 1, 0, \ldots, 0)\},$$

while the row span of A is equal to

$$\texttt{span}\{(1, 0, 0, \ldots, 0)^T, (1, \beta, 1, \ldots, 1, 0)^T\}.$$

Any non-zero vector in the column span of A has support in the set $\{\{1, 2, \ldots, n\}, \{1, 3, 4, \ldots, n\}, \{2\}, \emptyset\}$. Any non-zero vector in the row span of A has support in the set $\{\{1, 2, \ldots, n-1\}, \{1\}, \{2, 3, \ldots, n-1\}, \emptyset\}$. Thus, for $n \geq 4$, the supports of vectors in the column and row spaces of A only intersect in the empty set, implying that the column and row spaces only intersect in the 0 vector.

It follows that if CUR is a symmetric matrix then its column and row spans are equal, and so by the previous paragraph, $CUR = 0$. Hence $\|CUR - A\|_F^2 = \|A\|_F^2 = 2\beta^2 + (n-1) + (n-3) = 2\beta^2 + 2n - 4$.

Consider the anti-symmetric part $(A - A^\top)/2$ of A. This matrix has 0s on the diagonals and is entirely supported on the first two rows and columns, each entry being in $\{1/2, 0, -1/2\}$, and so $\|(A - A^\top)/2\|_F^2 \leq \frac{1}{4} \cdot 4n = n$. Consider the symmetric part $(A + A^\top)/2$ of A. This matrix is also entirely supported on the first two rows and columns, and has rank at most 4. Consequently,

$$\|A - A_4\|_F^2 \leq \|A - (A + A^\top)/2\|_F^2 = \|(A - A^\top)/2\|_F^2 \leq n.$$

Hence, there is no symmetric matrix CUR with C in the column span of A, and R in the row span of A, for which $\|CUR - A\|_F^2 \leq \frac{2\beta^2+2n-4}{n}\|A - A_4\|_F^2$, where the approximation factor $\frac{2\beta^2+2n-4}{n}$ can be made arbitrarily large by increasing β.

For larger k, we create a block matrix with two blocks, for which the first block is a $k-4 \times k-4$ identity matrix scaled by an arbitrarily large real number t, and the second block is the matrix A we have just constructed. Then in any symmetric CUR decomposition, the above analysis implies the support of any column in the span of C or row in the span of R is a (possibly empty) subset of $\{1, 2, \ldots, k-4\}$, and so we again have that $\|CUR - A\|_F^2 \leq 2\beta^2 + 2n - 4$. We also have $\|A - A_k\|_F^2 \leq n$, and so the same conclusion holds. ■

9 Concluding Remarks

We note that for column selection, the dependence on $\texttt{nnz}(A) \log n$ should be reducible to $\texttt{nnz}(A)$: the steps needing the log factor involve the computation of the norms of the columns to be sampled; this computation involves sketching by a matrix of independent Gaussian values. As in Theorem 41 of [6], sketching by a single column vector should be sufficient; the decrease in accuracy of estimation can be compensated for via a manageable increase in sample size.

Acknowledgments

We acknowledge the support of the XDATA program of the Defense Advanced Research Projects Agency (DARPA), administered through Air Force Research Laboratory contract FA8750-12-C-0323. We are grateful to Cameron Musco for helpful comments, in particular for clarifying parts of his work [7]. We also thank the SODA referees for valuable comments.

References

[1] Joshua D. Batson, Daniel A. Spielman, and Nikhil Srivastava. Twice-ramanujan sparsifiers. *SIAM Review*, 56(2):315–334, 2014.

[2] Jean Bourgain, Sjoerd Dirksen, and Jelani Nelson. Toward a unified theory of sparse dimensionality reduction in euclidean space. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 499–508, 2015.

[3] Christos Boutsidis and David P. Woodruff. Optimal CUR matrix decompositions. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, STOC '14, pages 353–362, New York, NY, USA, 2014. ACM.

[4] Kenneth L. Clarkson and David P. Woodruff. Numerical linear algebra in the streaming model. In *STOC*, pages 205–214, 2009.

[5] Kenneth L Clarkson and David P Woodruff. Low rank approximation and regression in input sparsity time. In *STOC*, 2013. Full version at `http://arxiv.org/abs/1207.6365`.

[6] Kenneth L Clarkson and David P Woodruff. Input sparsity and hardness for robust subspace approximation. In *FOCS*, 2015. Full version at `https://arxiv.org/abs/1510.06073`.

[7] M. B. Cohen, S. Elder, C. Musco, C. Musco, and M. Persu. Dimensionality Reduction for k-Means Clustering and Low Rank Approximation. *ArXiv e-prints*, October 2014.

[8] Michael B. Cohen. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '16, pages 278–287, Philadelphia, PA, USA, 2016. Society for Industrial and Applied Mathematics.

[9] Michael B. Cohen, Jelani Nelson, and David P. Woodruff. Optimal approximate matrix product in terms of stable rank. *CoRR*, abs/1507.02268, 2015.

[10] 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(Dec):2153–2175, 2005.

[11] Shmuel Friedland and Anatoli Torokhti. Generalized rank-constrained matrix approximations. *SIAM Journal on Matrix Analysis and Applications*, 29(2):656–659, 2007.

[12] Alex Gittens and Michael W. Mahoney. Revisiting the Nyström method for improved large-scale machine learning. *CoRR*, abs/1303.1849, 2013.

[13] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *ArXiv e-prints*, September 2009.

[14] Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. On sampling-based approximate spectral decomposition. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 553–560, New York, NY, USA, 2009. ACM.

[15] Xiangrui Meng and Michael W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *STOC*, pages 91–100, 2013.

[16] C. Musco and C. Musco. Provably Useful Kernel Matrix Approximation in Linear Time. *ArXiv e-prints*, May 2016.

[17] Jelani Nelson and Huy L. Nguyen. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *FOCS*, pages 117–126, 2013.

[18] S. Wang, L. Luo, and Z. Zhang. SPSD Matrix Approximation vis Column Selection: Theories, Algorithms, and Extensions. *ArXiv e-prints*, June 2014.

[19] S. Wang and Z. Zhang. Improving CUR Matrix Decomposition and the Nyström Approximation via Adaptive Sampling. *ArXiv e-prints*, March 2013.

[20] David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(12):1–157, 2014.

[21] 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.