Fast Monte-Carlo Algorithms for Finding Low-Rank Approximations

ALAN FRIEZE
Carnegie Mellon University
RAVI KANNAN
Yale University
and
SANTOSH VEMPALA
M.I.T.

We consider the problem of approximating a given $m \times n$ matrix $\mathbf{A}$ by another matrix of specified rank k, which is smaller than m and n. The Singular Value Decomposition (SVD) can be used to find the "best" such approximation. However, it takes time polynomial in m, n which is prohibitive for some modern applications. In this paper, we develop an algorithm which is qualitatively faster, provided we may sample the entries of the matrix according to a natural probability distribution. In many applications such sampling can be done efficiently. Our **main result** is a randomized algorithm to find *the description of* a matrix $\mathbf{D}^*$ of rank at most k so that

$$||\mathbf{A} - \mathbf{D}^*||_F^2 \leq \min_{\mathbf{D}, \mathrm{rank}(\mathbf{D}) \leq k} ||\mathbf{A} - \mathbf{D}||_F^2 + \varepsilon ||\mathbf{A}||_F^2$$

holds with probability at least $1 - \delta$ (where $||.||_F$ is the Frobenius norm). The algorithm takes time polynomial in $k, 1/\varepsilon, \log(1/\delta)$ only and is *independent of m and n*. In particular, this implies that in constant time, it can be determined if a given matrix of arbitrary size has a good low-rank approximation.

Categories and Subject Descriptors: F2 [**Theory of Computation**]: Analysis of Algorithms and Problem Complexity; H3 [**Information Systems**]: Information Storage and Retrieval; G1.3 [**Mathematics of Computing**]: Numerical Linear Algebra

General Terms: Algorithms, Theory

Additional Key Words and Phrases: Matrix algorithms, sampling, low-rank approximation

---

Author's addresses: A. Frieze, Dept. of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213. `af1p@andrew.cmu.edu`
R. Kannan, Computer Science Dept., Yale University, New Haven, CT 06511. `kannan@cs.yale.edu`
Santosh Vempala, Mathematics Dept., M.I.T., Cambridge, MA 02139. `vempala@math.mit.edu`
Alan Frieze was supported in part by NSF grant CCR-9530974. Ravi Kannan was supported in part by NSF grant CCR-9820850. Santosh Vempala was supported by NSF grant CCR-0307536 and a Sloan foundation Fellowship.

1. INTRODUCTION

Real-world data often has a large a number of attributes (features/dimensions). A natural question is whether in fact it is generated by a small model, i.e., with a much smaller number of parameters than the number of attributes. One way to formalize this question is the problem of finding a low-rank approximation: given an $m \times n$ matrix $\mathbf{A}$, find a matrix $\mathbf{D}$ of rank at most k so that $||\mathbf{A} - \mathbf{D}||_F$ is as small as possible (for any matrix $\mathbf{M}$, the Frobenius norm, $||.||_F$, is defined as $||\mathbf{M}||_F^2 = \sum_{i,j} M_{ij}^2$). Alternatively, if we view the rows of $\mathbf{A}$ as points in $\boldsymbol{R}^n$, then it is the problem of finding a k-dimensional linear subspace that minimizes the sum of squared distances to the points. This problem arises in many contexts, partly because some matrix algorithms are more efficient for low-rank matrices. It is also interesting to consider other norms, e.g., the 2-norm (see Section 6), but we will mostly focus on the Frobenius norm.

The traditional Singular Value Decomposition (SVD) can be used to solve the problem in time $O(\min\{mn^2, nm^2\})$. For many applications motivated by information retrieval and the web, this is too slow and one needs a linear or sublinear algorithm. To speed up SVD-based low-rank approximation, [Papadimitriou et al. 2000] suggested random projection as a pre-processing step, i.e., project the rows of $\mathbf{A}$ to an $O(\log n)$-dimensional subspace and then find the SVD in that subspace. This reduces the worst-case complexity to $O(mn \log n)$ for a small loss in approximation quality. This is still too high.

How fast can the problem be solved? At first sight, it seems that $\Omega(mn)$ is a lower bound — if $\mathbf{A}$ has only a single non-zero entry, one has to examine all its entries to find a good approximation. But suppose we can sample the entries with probability proportional to their magnitudes. Then a constant-sized sample would suffice in this case!

In this paper, we show that a rank k approximation can be found in time polynomial in k and $1/\varepsilon$ where ε is an error parameter, provided we can sample the entries of $\mathbf{A}$ from a natural probability distribution. The sampling assumptions will be made explicit shortly and also discussed in the context of some applications. Our main result is the following.

THEOREM 1. *Given an $m \times n$ matrix $\mathbf{A}$, and k, ε, δ, there is a randomized algorithm which finds* the description of *a matrix $\mathbf{D}^*$ of rank at most k so that*

$$||\mathbf{A} - \mathbf{D}^*||_F^2 \leq \min_{\mathbf{D}, rank(\mathbf{D}) \leq k} ||\mathbf{A} - \mathbf{D}||_F^2 + \varepsilon ||\mathbf{A}||_F^2$$

holds with probability at least $1 - \delta$. The algorithm takes time polynomial in $k, 1/\varepsilon, \log(1/\delta)$ only, independent of m, n. *The most complex computational task is to find the first k singular values of a randomly chosen $s \times s$ submatrix where $s = O(\max\{k^4\varepsilon^{-3}, k^2\varepsilon^{-4}\})$. The matrix $\mathbf{D}^*$ can be explicitly constructed from its description in $O(kmn)$ time.*

As a consequence, in $poly(k, \frac{1}{\varepsilon})$ time, we can determine (with high probability) if $\mathbf{A}$ has a rank k approximation with error at most $\varepsilon ||\mathbf{A}||_F$. The error probability δ can be boosted by standard techniques and we will prove the theorem for a fixed error probability.

The central idea of our approach is described as follows : We pick p rows of A

independently at random, each according to a probability distribution satisfying Assumption A1 (see Section 1.1). Suppose these rows form a $p \times m$ matrix S'. The rows will be scaled to form a matrix $\mathbf{S}$ (step 1 of the Algorithm in section 4). It will be relatively easy (Lemma 2) to show that $\mathbf{S}^T\mathbf{S}$ approximately equals $\mathbf{A}^T\mathbf{A}$. The intuition for this is that the (i,j)th entry of $\mathbf{A}^T\mathbf{A}$ is the dot product of the ith and jth columns of $\mathbf{A}$ and indeed, since $\mathbf{S}$ has a random sample of rows of $\mathbf{A}$, the entry $(\mathbf{S}^T\mathbf{S})_{i,j}$ estimates this; the scaling is done to make this estimate unbiased. Now from standard Linear Algebra, we can get the SVD of $\mathbf{A}$ from the spectral decomposition (SD) of $\mathbf{A}^T\mathbf{A}$ [1], and therefore approximately from the SD of $\mathbf{S}^T\mathbf{S}$. Repeating this, the SD of $\mathbf{S}^T\mathbf{S}$ can be read off from the SVD of $\mathbf{S}$ which in turn can be obtained from the SD of $\mathbf{S}\mathbf{S}^T$. Since $\mathbf{S}\mathbf{S}^T$ is just a $p \times p$ matrix, the problem is reduced to computing the SVD of a constant sized matrix! This still leaves the computation of $\mathbf{S}\mathbf{S}^T$. For this, we apply the sampling trick a second time — we pick a sample of p columns of S, to form a $p \times p$ matrix $\mathbf{W}$ (step 2 of the algorithm), then $\mathbf{W}\mathbf{W}^T$ approximates $\mathbf{S}\mathbf{S}^T$. Now the SD of $\mathbf{W}\mathbf{W}^T$ is all that is needed for which the SVD of $\mathbf{W}$ suffices. This then is the central computational task of the algorithm. We present the algorithm in Section 4. Besides Lemma 2, the key step in the analysis is showing that we can go from approximate left singular vectors of $\mathbf{S}$ to approximate right singular vectors with only a small loss.

A key insight of the paper, and the basis of the algorithm, is the existence of a good low-rank approximation to $\mathbf{A}$ in the subspace spanned by a small sample of its rows. We state this below formally. The constant c is defined in Assumption A1.

THEOREM 2. *Let $\mathbf{A}$ be an $m \times n$ matrix and S be a sample of s rows of $\mathbf{A}$ from a distribution satisfying Assumption 1. Let V be the vector space spanned by S. Then with probability at least 9/10, there exist an orthonormal set of vectors $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \dots \mathbf{y}^{(k)}$ in V such that*

$$||\mathbf{A} - \mathbf{A}(\sum_{j=1}^{k} \mathbf{y}^{(j)}\mathbf{y}^{(j)^T})||_F^2 \le \min_{\mathbf{D}, rank(\mathbf{D}) \le k} ||\mathbf{A} - \mathbf{D}||_F^2 + \tfrac{10k}{cs}||\mathbf{A}||_F^2. \qquad (1)$$

It is useful to note that that $\mathbf{A}\sum_{j=1}^{k} \mathbf{y}^{(j)}\mathbf{y}^{(j)^T}$ is the "restriction" of the linear transformation $\mathbf{A}$ to the subspace spanned by the $\mathbf{y}^{(j)}$, namely for any x in that subspace,

$$\mathbf{A}x = \mathbf{A}\sum_{j=1}^{k} \mathbf{y}^{(j)}\mathbf{y}^{(j)^T} x$$

and further, for x orthogonal to this subspace,

$$\mathbf{A}\sum_{j=1}^{k} \mathbf{y}^{(j)}\mathbf{y}^{(j)^T} x = 0.$$

By elementary linear algebra, the matrix $\mathbf{A}\sum_{j=1}^{k} \mathbf{y}^{(j)}\mathbf{y}^{(j)^T}$ has its rows in the span of the vectors $\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(k)}$ and therefore its rank is at most k. To describe this

[1]The right singular vectors of $\mathbf{A}$ are precisely the eigenvectors of $\mathbf{A}^T\mathbf{A}$

approximation, $\mathbf{D}^*$, it suffices to give the vectors $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(k)}$. In the algorithm, these vectors are themselves computed by multiplying a submatrix of $\mathbf{A}$ with a set of vectors in $\boldsymbol{R}^p$. Thus the matrix D^* can be recovered from a set of k p-dimensional vectors and a set of p numbers indicating a submatrix of $\mathbf{A}$ spanned by the corresponding rows. It also follows that the first k singular values of $\mathbf{A}$ can be computed to within a cumulative additive error of $\varepsilon||\mathbf{A}||_F$.

In Section 3, we give the proof of this existence theorem. The theorem directly gives an $O(mnk/\varepsilon+\text{poly}(k/\varepsilon))$ algorithm and suggests an algorithm whose running time is linear in m and a small polynomial in k and $1/\varepsilon$. Such an algorithm was developed following this paper in [Drineas et al. 2004a]. This and other subsequent developments are discussed in Section 6.

Throughout the paper, $\mathbf{M}^{(i)}$ denotes the ith row of the matrix $\mathbf{M}$ (as a row vector), $\mathbf{M}_{(j)}$ denotes the jth column (as a column vector) and $M_{i,j}$ is the entry in the ith row and jth column. Also, for any positive integer r, $[r]$ denotes the set $\{1, 2, \ldots, r\}$.

1.1 Sampling assumptions

We now describe our sampling assumptions in detail.

Assumption 1.. We can sample the rows of $\mathbf{A}$ so that row i is chosen with probability P_i satisfying

$$P_i \geq c\frac{|\mathbf{A}^{(i)}|^2}{||\mathbf{A}||_F^2}$$

for some constant $c \leq 1$ (independent of m, n). Here $|.|$ denotes Euclidean length. The P_i's are known to us (if $c = 1$, then we we don't need to know the P_i).

Assumption 2.. From any row i we can sample entry j with probability $Q_{j|i}$ satisfying

$$Q_{j|i} \geq c\frac{P_{i,j}}{P_i}, \quad \text{where} \quad P_{i,j} = \frac{\mathbf{A}_{i,j}^2}{||\mathbf{A}||_F^2}.$$

The $Q_{j|i}$ are known to us (again if $c = 1$, we don't need to know the values).

If the matrix $\mathbf{A}$ has a known sparsity structure, then we might be able to set up the sampling with very little preprocessing. In particular, if the matrix $\mathbf{A}$ is dense, i.e., if for all i, j,

$$\mathbf{A}_{ij}^2 \leq \frac{c'}{mn}||\mathbf{A}||_F^2$$

for some constant c', then we we can take $P_i = 1/m$ and $Q_{j|i} = 1/n$ for all i, j and $c = 1/c'$.

In general, for any matrix, by making one pass through the entire matrix, we can set up data structures that let us sample the entries fast from then onwards — $O(1)$ time per sample — so as to satisfy Assumptions 1 and 2. During the pass, we do several things. Suppose M is such that for all i, j

$$\mathbf{A}_{ij}^2 = 0 \qquad \text{OR} \qquad \frac{1}{M} \leq |\mathbf{A}_{ij}|^2 \leq M.$$

We create $O(\log M)$ bins; during the pass, we put into the lth bin all the entries (i, j) such that $\frac{2^{l-1}}{M} \leq |\mathbf{A}_{ij}|^2 \leq \frac{2^l}{M}$. We also keep track of the number of entries in each bin. After this, we treat all entries in a bin as being of the same value. To sample, we pick a bin with probability proportional to the total sum of squares in that bin. Then we pick an entry uniformly from the set of entries in the bin. In the pass, we also set up similar data structures for each row.

1.2 Applications

In this section we discuss our algorithm in the context of applications that rely on low-rank approximation. We show that in several situations we can satisfy the sampling assumptions of our algorithm and thus obtain the SVD approximation more efficiently. Applications that we do not discuss include face recognition and picture compression.

1.2.1 *Latent Semantic Indexing.* This is a general technique for processing a collection of "documents". We give a very cursory description of this broad area here and discuss its relation to our main problem (see [Berry et al. 1995; Deerwester et al. 1990; Dumais et al. 1988; Dumais 1991] for details and empirical results).

Suppose there are m documents and n "terms" which occur in the documents (terms may be all the words that occur in the documents or key words that occur in them). The model hypothesizes that there are a small number (k) of unknown "topics" which the documents are about. A topic is modelled as a probability distribution on the n terms, i.e., an $n-$vector of non-negative reals summing to 1. With this model on hand, it is shown (with additional assumptions) that the subspace spanned by the k best topics is close to the span of the top k singular vectors of the so-called "document-term" matrix [Papadimitriou et al. 2000]. The latter is an $m \times n$ matrix $\mathbf{A}$ with $\mathbf{A}_{ij}$ being the frequency of the jth term in the ith document. Alternatively, one can define $\mathbf{A}_{ij}$ as 0 or 1 depending upon whether the jth term occurs in the ith document.

Here we argue that, in practice, the assumptions of our algorithm are satisfied and it can be used in place of the full SVD algorithm. It is easy to see that if we are allowed one pass through each document, we can set up data structures for sampling (ideally, the creator of a document could supply a vector of squared term frequencies). Otherwise, if no frequency is too large (this is not unreasonable since words that occur too often, so-called "buzz words", are removed from the analysis), all we need to precompute is the length ($L_i = \sum_j \mathbf{A}_{ij}$), of each document. This is typically available (as say "file size") and we pick a document with probability proportional to its length. This is easily seen to satisfy Assumption 1, but without the squares (i.e. we sample the ith entry with probability $L_i / \sum_j L_j$). The assumption with the squares is satisfied if all the frequencies are small. Assumption 2 is similarly implemented — given a document, we pick a word uniformly at random from it, i.e., $Q_{j|i} = \mathbf{A}_{ij}/L_i$.

1.2.2 *Web Search model.* Kleinberg [Kleinberg 1999] proposed an algorithm for the problem of finding the most "important" documents from the set of documents returned by a web search that works by analyzing the $m \times m$ hyperlink matrix. This matrix $\mathbf{A}$ has entries $\mathbf{A}_{ij}$ equal to 1 or 0 depending upon whether the i'th document points to the j'th. The algorithm sets out to find two unit-length m-vectors $\mathbf{x}, \mathbf{y}$ such that $\mathbf{x}^T\mathbf{A}\mathbf{y}$ is maximized. This is of course the problem of finding the singular vectors of $\mathbf{A}$. When the keyword has multiple meanings, not only the top, but some of the other singular vectors (with large singular values) are interesting. So, it is of interest to find the largest k singular vectors for some small k.

It is worthwhile to consider our assumptions in this case. For Assumption 1, it is sufficient to sample the documents (roughly) according to the number of hypertext links from them. For Assumption 2, it is sufficient to be able to follow a random link from a document.

1.2.3 *Low-Rank Approximations and the Regularity Lemma.* The fundamental Regularity Lemma of Szemerédi's in graph theory [Szemeredi 1978] gives a partition of the vertex set of any graph so that "most" pairs of parts are "nearly regular". (We do not give details here.) This lemma has a host of applications (see [Komlós and Simonovits 1996]) in graph theory. The lemma was non-constructive in that it only asserted the existence of the partition (but did not give an algorithm to find it.) Alon, Duke, Lefmann, Rödl and Yuster were finally able to give an algorithm to find such a partition in polynomial time [Alon et al. 1994]. In [Frieze and Kannan 1996; Frieze and Kannan 1999a], low-rank approximations of the adjacency matrix of the graph were related to regular partitions. Szemerédi's Lemma and an algorithm for constructing the partition were derived from this connection. While this is not directly relevant to our results, we point it out here as one more case where low-rank approximations are very useful. A more direct application of eigenvector computation and Szemerédi's partition is given in [Frieze and Kannan 1999b].

2. THE SINGULAR VALUE DECOMPOSITION

The matrix $\mathbf{A}$ can be expressed

$$\mathbf{A} = \sum_{t=1}^{r} \sigma_t \mathbf{u}^{(t)} \mathbf{v}^{(t)^T}$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r \geq 0$ and the $\mathbf{u}^{(t)}$ form an orthonormal set of column vectors as do the $\mathbf{v}^{(t)}$. Also $\mathbf{u}^{(t)^T}\mathbf{A} = \sigma_t \mathbf{v}^{(t)^T}$ and $\mathbf{A}\mathbf{v}^{(t)} = \sigma_t \mathbf{u}^{(t)}$ for $1 \leq t \leq r$. This is called the *singular value decomposition* of $\mathbf{A}$. Here r is the rank of A.

By a theorem of Eckart and Young [Golub and Van Loan 1989], the matrix $\mathbf{D}_k$ that minimizes $||\mathbf{A} - \mathbf{D}||_F$ among all matrices $\mathbf{D}$ of rank k or less is given by

$$\mathbf{D}_k = \sum_{t=1}^{k} \mathbf{A}\mathbf{v}^{(t)} \mathbf{v}^{(t)^T}.$$

This implies that

$$||\mathbf{D}_k||_F^2 = \sum_{t=1}^{k} \sigma_t^2 \quad \text{and} \quad ||\mathbf{A} - \mathbf{D}_k||_F^2 = \sum_{t=k+1}^{r} \sigma_t^2.$$

We use this notation throughout the paper.

3. A SMALL SAMPLE INDUCES A GOOD APPROXIMATION

The goal of this section is to prove Theorem 2, namely the subspace spanned by a sample of rows chosen according to Assumption 1 contains an approximation to $\mathbf{A}$ that is nearly the best possible. If, by chance, $\mathbf{v}^{(1)}, \ldots \mathbf{v}^{(k)}$ belong to this subspace, we would be done, since then, $\sum_{t=1}^{k} \mathbf{A}\mathbf{v}^{(t)}\mathbf{v}^{(t)^T}$ would provide the required approximation to $\mathbf{A}$. What we show is that the vectors $\mathbf{w}^{(t)}$ (to be defined shortly) in the subspace do approximate scaled versions of the respective $\mathbf{v}^{(t)}$.

Let S be a random sample of s rows chosen from a distribution that satisfies Assumption 1. For $t = 1, 2, \ldots, r$, we define the vector-valued random variable

$$\mathbf{w}^{(t)} = \frac{1}{s} \sum_{i \in S} \frac{\mathbf{u}_i^{(t)}}{P_i} \mathbf{A}^{(i)}.$$

Note that S is, in general, a multiset, i.e., rows of $\mathbf{A}$ might be picked multiple times. The vectors $\mathbf{w}^{(t)}$ are clearly in the subspace generated by S. We first compute the expectation of $\mathbf{w}^t$. For this, we can view it as the average of s i.i.d. random variables $X_1, \ldots, X_s$ where each X_j has the following distribution:

$$X_j = \frac{\mathbf{u}_i^{(t)}}{P_i} \mathbf{A}^{(i)} \text{ with probability } P_i, \qquad \text{for } i = 1, 2, \ldots m.$$

Then, taking expectations, [2]

$$\mathbf{E}(X_j) = \sum_{i=1}^{m} \frac{\mathbf{u}_i^{(t)}}{P_i} \mathbf{A}^{(i)} P_i = \mathbf{u}^{(t)^T} \mathbf{A} = \sigma_t \mathbf{v}^{(t)^T},$$

and so,

$$\mathbf{E}(\mathbf{w}^{(t)}) = \sigma_t \mathbf{v}^{(t)^T}.$$

Further, since $P_i \geq c \frac{|\mathbf{A}^{(i)}|^2}{||\mathbf{A}||_F^2}$, we have (since $|\mathbf{u}^{(t)}| = 1$),

$$\begin{aligned} \mathbf{E}(|\mathbf{w}^{(t)} - \sigma_t \mathbf{v}^{(t)^T}|^2) &= \frac{1}{s} \left(\sum_{i=1}^{m} \frac{|\mathbf{u}_i^{(t)}|^2 |\mathbf{A}^{(i)}|^2}{P_i} \right) - \frac{\sigma_t^2}{s} \\ &\leq \frac{1}{sc} ||\mathbf{A}||_F^2. \end{aligned} \tag{2}$$

If we had $\mathbf{w}^{(t)}$ exactly equal to $\sigma_t \mathbf{v}^{(t)^T}$ (instead of just in expectation), then

$$\mathbf{A} \sum_{t=1}^{k} \mathbf{v}^{(t)} \mathbf{v}^{(t)^T} = \mathbf{A} \sum_{t=1}^{k} \frac{1}{\sigma_t^2} \mathbf{w}^{(t)^T} \mathbf{w}^{(t)}$$

and this would be sufficient to prove the theorem. We wish to carry this out approximately.

To this end, define $\hat{\mathbf{y}}^{(t)} = \frac{1}{\sigma_t} \mathbf{w}^{(t)^T}$ for $t = 1, 2, \ldots, r$ and let

$$V_1 = \operatorname{span}(\hat{\mathbf{y}}^{(1)}, \hat{\mathbf{y}}^{(2)}, \ldots, \hat{\mathbf{y}}^{(k)}) \subseteq V.$$

[2] The expectation and variance of a vector valued random variable are taken separately for each component.

Let $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(n)}$ be an orthonormal basis of $\boldsymbol{R}^n$ so that

$$V_1 = \text{span}(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(l)}),$$

where l is the dimension of V_1. Let

$$\mathbf{F} = \sum_{t=1}^{l} \mathbf{A}\mathbf{y}^{(t)}\mathbf{y}^{(t)^T} \quad \text{and} \quad \hat{\mathbf{F}} = \sum_{t=1}^{k} \mathbf{A}\mathbf{v}^{(t)}\hat{\mathbf{y}}^{(t)^T}$$

The matrix $\mathbf{F}$ will be our candidate approximation to $\mathbf{A}$ in the span of S. We will bound its error using $\hat{\mathbf{F}}$. Note that for any $i \leq k$ and $j > l$, we have $\hat{\mathbf{y}}^{((i)^T)}\mathbf{y}^{(j)} = 0$. Thus,

$$\begin{aligned}
||\mathbf{A} - \mathbf{F}||_F^2 &= \sum_{i=1}^{n} |(\mathbf{A} - \mathbf{F})\mathbf{y}^{(i)}|^2 \\
&= \sum_{i=l+1}^{n} |\mathbf{A}\mathbf{y}^{(i)}|^2 \\
&= \sum_{i=l+1}^{n} |(\mathbf{A} - \hat{\mathbf{F}})\mathbf{y}^{(i)}|^2 \\
&\leq ||\mathbf{A} - \hat{\mathbf{F}}||_F^2. \qquad (3)
\end{aligned}$$

Also,

$$\begin{aligned}
||\mathbf{A} - \hat{\mathbf{F}}||_F^2 &= \sum_{i=1}^{n} |\mathbf{u}^{(i)^T}(\mathbf{A} - \hat{\mathbf{F}})|^2 \\
&= \sum_{i=1}^{k} |\sigma_i \mathbf{v}^{(i)^T} - \mathbf{w}^{(i)}|^2 + \sum_{i=k+1}^{n} \sigma_i^2.
\end{aligned}$$

Taking expectations and using (2) we get

$$\mathbf{E}(||\mathbf{A} - \hat{\mathbf{F}}||_F^2) \leq \sum_{i=k+1}^{n} \sigma_i^2 + \frac{k}{sc}||\mathbf{A}||_F^2. \qquad (4)$$

$\hat{\mathbf{F}}$ is of rank at most k and $\mathbf{D}_k$ is the best rank k approximation to $\mathbf{A}$. So, we have

$$||\mathbf{A} - \hat{\mathbf{F}}||_F^2 \geq ||\mathbf{A} - \mathbf{D}_k||_F^2 = \sum_{i=k+1}^{n} \sigma_i^2.$$

Thus $||\mathbf{A} - \hat{\mathbf{F}}||_F^2 - ||\mathbf{A} - \mathbf{D}_k||_F^2$ is a non-negative random variable and (4) implies

$$\mathbf{Pr}(||\mathbf{A} - \hat{\mathbf{F}}||_F^2 - ||\mathbf{A} - \mathbf{D}_k||_F^2 \geq \tfrac{10k}{sc}||\mathbf{A}||_F^2) \leq \tfrac{1}{10}.$$

From (3) it follows that

$$\mathbf{Pr}(||\mathbf{A} - \mathbf{F}||_F^2 - ||\mathbf{A} - \mathbf{D}_k||_F^2 \geq \tfrac{10k}{sc}||\mathbf{A}||_F^2) \leq \tfrac{1}{10}$$

as required. □

We next observe that a good low-rank approximation with respect to Frobenius norm implies a good low-rank approximation with respect to the 2-norm, $||\mathbf{M}|| =$

$\max_{|\mathbf{x}|=1} |\mathbf{Mx}|$. Better approximations have since been obtained for the 2-norm (see Section 6).

THEOREM 3. *If*

$$||\mathbf{A} - \mathbf{A}\sum_{t=1}^{k} \mathbf{y}^{(t)}\mathbf{y}^{(t)^T}||_F^2 \leq ||\mathbf{A} - \mathbf{D}_k||_F^2 + \varepsilon||\mathbf{A}||_F^2.$$

Then

$$||\mathbf{A} - \mathbf{A}\sum_{t=1}^{k} \mathbf{y}^{(t)}\mathbf{y}^{(t)^T}||^2 \leq (\tfrac{1}{k+1} + \varepsilon)||\mathbf{A}||_F^2.$$

Proof. Let $\mathbf{B} = \mathbf{A} - \mathbf{A}\sum_{t=1}^{k} \mathbf{y}^{(t)}\mathbf{y}^{(t)^T}$. Suppose that $\mathbf{B}$ has a *unit eigenvector* $\mathbf{x}$ with eigenvalue λ such that

$$\lambda^2 > (\tfrac{1}{k+1} + \varepsilon)||\mathbf{A}||_F^2.$$

Then we see that

$$\begin{aligned} ||\mathbf{B} - \mathbf{B}\mathbf{x}\mathbf{x}^T||_F^2 &= ||\mathbf{B}||_F^2 - \lambda^2 \\ &< \sum_{i=k+1}^{n} \sigma_i^2 - \frac{1}{k+1}||\mathbf{A}||_F^2. \end{aligned} \tag{5}$$

The rank of the matrix $\mathbf{A}\sum_{t=1}^{k} \mathbf{y}^{(t)}\mathbf{y}^{(t)^T} + \mathbf{B}\mathbf{x}\mathbf{x}^T$ is at most $k+1$, and so

$$\begin{aligned} ||\mathbf{A} - \mathbf{A}\sum_{t=1}^{k} \mathbf{y}^{(t)}\mathbf{y}^{(t)^T} - \mathbf{B}\mathbf{x}\mathbf{x}^T||_F^2 &\geq ||\mathbf{A} - \mathbf{D}_{k+1}||_F^2 \\ &= \sum_{t=k+2}^{n} \sigma_t^2 \\ &\geq ||\mathbf{A} - \mathbf{D}_k||_F^2 - \frac{1}{k+1}||\mathbf{A}||_F^2, \end{aligned}$$

since $\sigma_{k+1}^2 \leq \frac{1}{k+1}||\mathbf{A}||_F^2$. This contradicts (5). □

4. SAMPLING ALGORITHM

In this section we present the main constant-time algorithm to produce the approximation of Theorem 1. What we do below is to first pick a set of p rows of $\mathbf{A}$ from a distribution satisfying Assumption 1. We form a matrix $\mathbf{S}$ from these rows after scaling them. We then pick p columns of $\mathbf{S}$ from a probability distribution satisfying Assumption 2 and scale the columns to get a $p \times p$ matrix $\mathbf{W}$. We find the singular vectors of this matrix and from those, show how to get a good low-rank approximation to $\mathbf{A}$. The reader might want to consult the discussion between the statements of Theorems 1 and 2 in the introduction for an intuitive idea of how the algorithm works. In the description below, the parameter c is the one from the assumptions.

Algorithm

Input: Matrix $\mathbf{A}$, integer $k > 0$ and error parameter $\varepsilon > 0$. Set

$$p = 10^7 \max\{\frac{k^4}{c^3\varepsilon^3}, \frac{k^2}{c^3\varepsilon^4}\}.$$

(1) (**Sample rows**) Independently choose (rows) $i_1, i_2, \ldots, i_p$ according to distribution $P = (P_1, P_2, \ldots, P_m)$ which satisfies Assumption 1, i.e.,

$$P_i \geq c\frac{|\mathbf{A}^{(i)}|^2}{||\mathbf{A}||_F^2}.$$

Let $\mathbf{S}$ be the $p \times n$ matrix with rows $\mathbf{A}^{(i_t)}/\sqrt{pP_{i_t}}$ for $t = 1, 2, \ldots, p$. Note that if $c = 1$, the this scaling amounts to normalizing all rows to be of the same length.

(2) (**Sample columns**) Independently choose (columns) $j_1, j_2, \ldots, j_p$ (of $\mathbf{S}$) according to a distribution $P' = (P'_1, P'_2, \ldots, P'_n)$ which satisfies

$$P'_j \geq \frac{c}{2}\frac{|\mathbf{S}_{(j)}|^2}{||\mathbf{S}||_F^2}.$$

(we show below how to do this using Assumption 2.)
Let $\mathbf{W}$ be the $p \times p$ matrix with columns $\mathbf{S}^{(j_t)}/\sqrt{pP'_{j_t}}$ for $t = 1, 2, \ldots, p$.

(3) (**Compute SVD**) Compute the top k singular vectors $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(k)}$ in the column space of $\mathbf{W}$.

(4) (**Filter**) Let

$$T = \{t : \ |\mathbf{W}^T\mathbf{u}^{(t)}|^2 \geq \gamma||\mathbf{W}||_F^2\},$$

where

$$\gamma = \frac{c\varepsilon}{8k}.$$

For $t \in T$ let

$$\mathbf{v}^{(t)} = \frac{\mathbf{S}^T\mathbf{u}^{(t)}}{|\mathbf{W}^T\mathbf{u}^{(t)}|}$$

Output $\mathbf{v}^{(t)}$ for $t \in T$ (the low rank approximation to $\mathbf{A}$ is $\mathbf{A}\sum_{t\in T}\mathbf{v}^{(t)}\mathbf{v}^{(t)^T}$).

We next discuss the implementation of the above algorithm. In particular, how do we carry out Step 2? We first pick a row of $\mathbf{S}$, each row with probability $1/p$; suppose the chosen row is the ith row of $\mathbf{A}$. Then pick $j \in \{1, 2, \ldots n\}$ with probabilities $Q_{j|i}$ as in Assumption 2. This defines the probabilities P'_j. We then have (with $I = \{i : \mathbf{A}^{(i)}$ is a row of $\mathbf{S}\}$),

$$P'_j = \sum_{i\in I}\frac{Q_{j|i}}{p} \geq \sum_{i\in I}\frac{cP_{i,j}}{pP_i} = \sum_{i\in I}\frac{c\mathbf{A}_{i,j}^2}{pP_i||\mathbf{A}||_F^2} = \frac{c}{||\mathbf{A}||_F^2}\sum_{i\in I}\frac{\mathbf{A}_{i,j}^2}{pP_i} = c\frac{|\mathbf{S}_{(j)}|^2}{||\mathbf{A}||_F^2} \geq \frac{c}{2}\frac{|\mathbf{S}_{(j)}|^2}{||\mathbf{S}||_F^2}$$

where the last step is implied by the next lemma.

LEMMA 1. *For* $\mathbf{S}$ *and* $\mathbf{W}$ *chosen as in the algorithm, with probability at least* $1 - \frac{16}{c^2 p}$,

$$\tfrac{1}{2}||\mathbf{A}||_F^2 \le ||\mathbf{S}||_F^2 \le \tfrac{3}{2}||\mathbf{A}||_F^2 \text{ and } \tfrac{1}{2}||\mathbf{S}||_F^2 \le ||\mathbf{W}||_F^2 \le \tfrac{3}{2}||\mathbf{S}||_F^2.$$

Proof. By a routine calculation,

$$\mathbf{E}(||\mathbf{S}||_F^2) = ||\mathbf{A}||_F^2$$

Next, observe that for any row i of S,

$$||\mathbf{S}^{(i)}||_F^2 \le \frac{||\mathbf{A}||_F^2}{cp}$$

The random variable $||\mathbf{S}||_F^2$ is a sum of p independent random variables. Therefore,

$$\mathbf{Var}(||\mathbf{S}||_F^2||) = p\mathbf{Var}(||\mathbf{S}^{(i)}||_F^2) \le p\mathbf{E}(||\mathbf{S}^{(i)}||_F^4) \le \frac{1}{c^2 p}||\mathbf{A}||_F^4.$$

The first part of the lemma now follows using Chebychev's inequality. The proof of the second part is similar. □

5. ANALYSIS

The next lemma asserts that a sample $\mathbf{N}$ of rows of a matrix $\mathbf{M}$ provides a good approximation to $\mathbf{M}$ in the sense that $\mathbf{N}^T\mathbf{N}$ is close to $\mathbf{M}^T\mathbf{M}$. This will be a key tool in the analysis.

LEMMA 2. *Let* $\mathbf{M}$ *be an* $a \times b$ *matrix and let* $Q = Q_1, Q_2, \ldots, Q_a$ *be a probability distribution on* $\{1, 2, \ldots, a\}$ *such that*

$$Q_i \ge \alpha \frac{|\mathbf{M}^{(i)}|^2}{||\mathbf{M}||_F^2}, \qquad i = 1, 2, \ldots, a$$

for some $0 < \alpha < 1$. *Let* $\sigma = (i_1, i_2, \ldots, i_p)$ *be a sequence of* p *independent samples from* $[a]$, *each chosen according to distribution* Q. *Let* $\mathbf{N}$ *be the* $p \times b$ *matrix with*

$$\mathbf{N}^{(t)} = \frac{\mathbf{M}^{(i_t)}}{\sqrt{pQ_{i_t}}} \qquad t = 1, 2, \ldots, p.$$

Then for all $\theta > 0$,

$$\mathbf{Pr}(||\mathbf{M}^T\mathbf{M} - \mathbf{N}^T\mathbf{N}||_F \ge \theta||\mathbf{M}||_F^2) \le \frac{1}{\theta^2 \alpha p}.$$

Proof.

$$\begin{aligned}
||\mathbf{M}^T\mathbf{M} - \mathbf{N}^T\mathbf{N}||_F^2 &= \sum_{r,s=1}^{b} |\mathbf{M}_{(r)}^T\mathbf{M}_{(s)} - \mathbf{N}_{(r)}^T\mathbf{N}_{(s)}|^2 \\
\mathbf{E}(\mathbf{N}_{(r)}^T\mathbf{N}_{(s)}) &= \sum_{t=1}^{p} \mathbf{E}(\mathbf{N}_{i_t,r}\mathbf{N}_{i_t,s}) \\
&= \sum_{t=1}^{p}\sum_{i=1}^{a} Q_i \frac{\mathbf{M}_{i,r}\mathbf{M}_{i,s}}{pQ_i} \\
&= \mathbf{M}_{(r)}^T\mathbf{M}_{(s)}
\end{aligned}$$

$$\begin{aligned}
\mathbf{E}(|\mathbf{N}_{(r)}^T\mathbf{N}_{(s)} - \mathbf{M}_{(r)}^T\mathbf{M}_{(s)}|^2) &\le \sum_{t=1}^{p} \mathbf{E}((\mathbf{N}_{i_t,r}\mathbf{N}_{i_t,s})^2) \\
&= \sum_{t=1}^{p}\sum_{i=1}^{a} Q_i \frac{\mathbf{M}_{i,r}^2\mathbf{M}_{i,s}^2}{p^2Q_i^2} \\
&\le \frac{||\mathbf{M}||_F^2}{\alpha p^2} \sum_{t=1}^{p}\sum_{i=1}^{a} \frac{\mathbf{M}_{i,r}^2\mathbf{M}_{i,s}^2}{|\mathbf{M}^{(i)}|^2} \\
&= \frac{||\mathbf{M}||_F^2}{\alpha p} \sum_{i=1}^{a} \frac{\mathbf{M}_{i,r}^2\mathbf{M}_{i,s}^2}{|\mathbf{M}^{(i)}|^2}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbf{E}(||\mathbf{M}^T\mathbf{M} - \mathbf{N}^T\mathbf{N}||_F^2) &= \sum_{r,s=1}^{b} \mathbf{E}(|\mathbf{N}_{(r)}^T\mathbf{N}_{(s)} - \mathbf{M}_{(r)}^T\mathbf{M}_{(s)}|^2) \\
&\le \frac{||\mathbf{M}||_F^2}{\alpha p} \sum_{i=1}^{a} \frac{1}{|\mathbf{M}^{(i)}|^2} \sum_{r,s=1}^{b} (\mathbf{M}_{i,r}\mathbf{M}_{i,s})^2 \\
&\le \frac{||\mathbf{M}||_F^2}{\alpha p} \sum_{i=1}^{a} \frac{1}{|\mathbf{M}^{(i)}|^2} \left(\sum_{r=1}^{b} \mathbf{M}_{i,r}^2\right)^2 \\
&= \frac{||\mathbf{M}||_F^4}{\alpha p}.
\end{aligned}$$

The result follows from Markov's inequality. □

We introduce some notation for the rest of the analysis. For a matrix $\mathbf{M}$ and vectors $\mathbf{x}^{(i)}, i \in I$ we define

$$\Delta(\mathbf{M}; \mathbf{x}^{(i)}, i \in I) = ||\mathbf{M}||_F^2 - ||\mathbf{M} - \mathbf{M}\sum_{i\in I} \mathbf{x}^{(i)}\mathbf{x}^{(i)^T}||_F^2$$

When the $\mathbf{x}^{(i)}$ are orthogonal unit vectors, this represents the norm of the projection of $\mathbf{M}$ onto the subspace spanned by the $\mathbf{x}^{(i)}$. In this case,

$$\Delta(\mathbf{M}; \mathbf{x}^{(i)}, i \in I) = \sum_{i\in I} \mathbf{x}^{(i)^T}\mathbf{M}^T\mathbf{M}\mathbf{x}^{(i)}.$$

Thus, if $\mathbf{x}^{(t)}, t \in [k]$ are the top k singular vectors of $\mathbf{A}$, then

$$\Delta(\mathbf{A}; \mathbf{x}^{(t)}, t \in [k]) = \sum_{t=1}^{k} \sigma_t^2.$$

LEMMA 3. *Let $\mathbf{A}, \mathbf{S}$ be matrices with the same number of columns, and*

$$||\mathbf{A}^T\mathbf{A} - \mathbf{S}^T\mathbf{S}|| \leq \theta ||\mathbf{A}||_F^2.$$

(1) For any pair of unit vectors $\mathbf{z}, \mathbf{z}'$ in the row space of $\mathbf{A}$,

$$|\mathbf{z}^T\mathbf{A}^T\mathbf{A}\mathbf{z}' - \mathbf{z}^T\mathbf{S}^T\mathbf{S}\mathbf{z}'| \leq \theta ||\mathbf{A}||_F^2.$$

(2) For any set of unit vectors $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots, \mathbf{z}^{(\ell)}, \ell \leq k$ in the row space of $\mathbf{A}$,

$$|\Delta(\mathbf{A}; \mathbf{z}^{(i)}, i \in [\ell]) - \Delta(\mathbf{S}; \mathbf{z}^{(i)}, i \in [\ell])| \leq k^2\theta ||\mathbf{A}||_F^2.$$

Proof. The first part of the lemma is easy. For the second, using the fact that $||\mathbf{N}||_F^2 = \mathrm{Tr}(\mathbf{N}\mathbf{N}^T)$ for any matrix $\mathbf{N}$, we see that $\Delta(\mathbf{A}; \mathbf{x}^{(i)}, i \in I)$ equals

$$\begin{aligned}
&2\sum_{i\in I} \mathrm{Tr}(\mathbf{A}\mathbf{z}^{(i)}\mathbf{z}^{(i)^T}\mathbf{A}^T) - \sum_{i,i'\in I} (\mathbf{z}^{(i)^T}\mathbf{z}(i')^T)\mathrm{Tr}(\mathbf{A}\mathbf{z}^{(i)}\mathbf{z}^{(i')^T}\mathbf{A}^T) \\
&= 2\sum_{i\in I} \mathbf{z}^{(i)^T}\mathbf{A}^T\mathbf{A}\mathbf{z}^{(i)} - \sum_{i\in I} (|\mathbf{z}^{(i)}|^2)\mathbf{z}^{(i)^T}\mathbf{A}^T\mathbf{A}\mathbf{z}^{(i)} - \sum_{i\neq i'\in I} (\mathbf{z}^{(i)^T}\mathbf{z}^{(i')^T})\mathbf{z}^{(i')^T}\mathbf{A}^T\mathbf{A}\mathbf{z}^{(i)} \\
&= \sum_{i\in I} \mathbf{z}^{(i)^T}\mathbf{A}^T\mathbf{A}\mathbf{z}^{(i)} - \sum_{i\neq i'\in I} (\mathbf{z}^{(i)^T}\mathbf{z}^{(i')^T})\mathbf{z}^{(i')^T}\mathbf{A}^T\mathbf{A}\mathbf{z}^{(i)}.
\end{aligned}$$

By exactly similar reasoning, we have

$$\Delta(\mathbf{S}; \mathbf{z}^{(i)}, i \in I) = \sum_{i\in I} \mathbf{z}^{(i)^T}\mathbf{S}^T\mathbf{S}\mathbf{z}^{(i)} - \sum_{i\neq i'\in I} (\mathbf{z}^{(i)^T}\mathbf{z}^{(i')^T})\mathbf{z}^{(i')^T}\mathbf{S}^T\mathbf{S}\mathbf{z}^{(i)}.$$

Now using the first part of the lemma, the second part follows. □

We are now ready to prove the main theorem.

Proof of Theorem 1. We will prove that the conclusion of the theorem holds with probability at least 3/4. We will apply Lemma 2 *twice*, once to the row sample and once to the induced column sample. It follows from the Lemma 2 that with probability at least 9/10 both of the following events hold:

$$||\mathbf{A}^T\mathbf{A} - \mathbf{S}^T\mathbf{S}||_F \leq \theta ||\mathbf{A}||_F^2 \quad \text{and } ||\mathbf{S}\mathbf{S}^T - \mathbf{W}\mathbf{W}^T||_F \leq \theta ||\mathbf{S}||_F^2 \tag{6}$$

where

$$\theta = \sqrt{\frac{40}{cp}} = \min\{\frac{c\varepsilon^2}{500k}, \frac{c\varepsilon^{3/2}}{500k^2}\}.$$

Assume from now on that these events occur. Throughout the proof, we will approximate constants that arise somewhat crudely (by convenient rationals).

It follows from Theorem 2 that with probability at least 9/10 there are unit vectors $\mathbf{x}^{(t)}, t \in [k]$ in the row space of $\mathbf{S}$ such that

$$\Delta(\mathbf{A}; \mathbf{x}^{(t)}, t \in [k]) \geq ||\mathbf{A}||_F^2 - ||\mathbf{A} - \mathbf{D}_k||_F^2 - \tfrac{10k}{cp}||\mathbf{A}||_F^2 \geq ||\mathbf{D}_k||_F^2 - \tfrac{\varepsilon}{8}||\mathbf{A}||_F^2. \tag{7}$$

Applying the second part of Lemma 3 to $\mathbf{A}, \mathbf{S}$ and the vectors $\mathbf{x}^{(i)}$, we see that

$$\begin{aligned}\Delta(\mathbf{S}; \mathbf{x}^{(t)}, t \in [k]) &\geq \Delta(\mathbf{A}; \mathbf{x}^{(t)}, t \in [k]) - k^2\theta ||\mathbf{A}||_F^2 \\ &\geq ||\mathbf{D}_k||_F^2 - \tfrac{\varepsilon}{4}||\mathbf{A}||_F^2\end{aligned}$$

(using (7). Now, $\mathbf{S}$ and $\mathbf{S}^T$ have the same singular values and so there exist unit vectors $\mathbf{y}^{(t)}, t \in [k]$ in the column space of $\mathbf{S}$ such that

$$\Delta(\mathbf{S}^T; \mathbf{y}^{(t)}, t \in [k]) \geq ||\mathbf{D}_k||_F^2 - \tfrac{1}{4}\varepsilon ||\mathbf{A}||_F^2.$$

Applying Theorem 2 to $\mathbf{S}^T$ and $\mathbf{W}^T$, we see that with probability at least 9/10 there are unit vectors $\mathbf{z}^{(t)}, t \in [k]$ in the column space of $\mathbf{W}$ such that

$$\Delta(\mathbf{S}^T; \mathbf{z}^{(t)}, t \in [k]) \geq \Delta(\mathbf{S}^T; \mathbf{y}^{(t)}, t \in [k]) - \tfrac{20k}{cp}||\mathbf{S}||_F^2 \geq ||\mathbf{D}_k||_F^2 - \tfrac{3\varepsilon}{8}||\mathbf{A}||_F^2.$$

Applying the second part of Lemma 3 to $\mathbf{S}^T, \mathbf{W}^T$ and the vectors $\mathbf{z}^{(t)}$, we see that

$$\Delta(\mathbf{W}^T; \mathbf{z}^{(t)}, t \in [k]) \geq \Delta(\mathbf{S}^T; \mathbf{z}^{(t)}, t \in [k]) - k^2\theta ||\mathbf{S}||_F^2 \geq ||\mathbf{D}_k||_F^2 - \tfrac{\varepsilon}{2}||\mathbf{A}||_F^2.$$

Therefore, the vectors $\mathbf{u}^{(t)}, t \in [k]$ computed by the algorithm satisfy

$$\Delta(\mathbf{W}^T; \mathbf{u}^{(t)}, t \in [k]) \geq ||\mathbf{D}_k||_F^2 - \tfrac{\varepsilon}{2}||\mathbf{A}||_F^2.$$

Note that the highest possible value of $\Delta(\mathbf{A}; \mathbf{x}^{(t)}, t \in [k])$ is $||\mathbf{D}_k||_F^2$. All that remains to show is that in fact $\Delta(\mathbf{W}^T, .)$ being large implies that $\Delta(\mathbf{A}, .)$ is large. For this, we construct a suitable set of vectors (as in the algorithm).

Since $\mathbf{u}^{(t)}, t \in [k]$ are orthonormal singular vectors,

$$\Delta(\mathbf{W}^T; \mathbf{u}^{(t)}, t \in T) \geq \Delta(\mathbf{W}^T; \mathbf{u}^{(t)}, t \in [k]) - k\gamma ||\mathbf{W}||_F^2 \geq ||\mathbf{D}_k||_F^2 - \tfrac{5}{8}\varepsilon ||\mathbf{A}||_F^2.$$

Applying Lemma 3 again, this time to $\mathbf{S}^T, \mathbf{W}^T$ and the vectors $\mathbf{u}^{(t)}, t \in T$, it follows that

$$\Delta(\mathbf{S}^T; \mathbf{u}^{(t)}, t \in T) \geq \Delta(\mathbf{W}^T; \mathbf{u}^{(t)}, t \in T) - k^2\theta ||\mathbf{S}||_F^2 \geq ||\mathbf{D}_k||_F^2 - \tfrac{3}{4}\varepsilon ||\mathbf{A}||_F^2.$$

The next and crucial step, is to switch from $\mathbf{u}^{(t)}$ in the column space of $\mathbf{S}$ to $\mathbf{v}^{(t)}$ in the row space of $\mathbf{S}$. This is achieved by the following claims whose proof we defer to Section 5.1. For $t \in T$,

Claim 1. $\Delta(\mathbf{S}; \mathbf{v}^{(t)}, t \in T) \geq \Delta(\mathbf{S}^T; \mathbf{u}^{(t)}, t \in T) - \frac{1}{8}\varepsilon ||\mathbf{A}||_F^2$

Claim 2. $|\mathbf{v}^{(t)}|^2 \leq 1 + \frac{\varepsilon}{16}$.

It follows from Lemma 3 that

$$\Delta(\mathbf{A}; \mathbf{v}^{(t)}, t \in T) \geq \Delta(\mathbf{S}; \mathbf{v}^{(t)}, t \in T) - (1 + \frac{\varepsilon}{16})k^2\theta ||\mathbf{A}||_F^2 \geq ||\mathbf{D}_k||_F^2 - \varepsilon ||\mathbf{A}||_F^2$$

(assuming $\varepsilon \leq 16$). Thus,

$$||\mathbf{A}||_F^2 - ||\mathbf{A} - \mathbf{A}\sum_{t \in T} \mathbf{v}^{(t)}\mathbf{v}^{(t)^T}||_F^2 \geq ||\mathbf{D}_k||_F^2 - \varepsilon ||\mathbf{A}||_F^2.$$

Rearranging terms, we get the conclusion of the theorem:

$$||\mathbf{A} - \mathbf{A}\sum_{t \in T} \mathbf{v}^{(t)}\mathbf{v}^{(t)^T}||_F^2 \leq ||\mathbf{A} - \mathbf{D}_k||_F^2 + \varepsilon ||\mathbf{A}||_F^2.$$

□

5.1 Proof of Claims 1 and 2

Observe first that

$$\begin{aligned} & ||\mathbf{S}\mathbf{S}^T\mathbf{S}\mathbf{S}^T - \mathbf{W}\mathbf{W}^T\mathbf{W}\mathbf{W}^T||_F \\ \le\ & ||\mathbf{S}\mathbf{S}^T(\mathbf{S}\mathbf{S}^T - \mathbf{W}\mathbf{W}^T)||_F + ||(\mathbf{S}\mathbf{S}^T - \mathbf{W}\mathbf{W}^T)\mathbf{W}\mathbf{W}^T||_F \\ \le\ & \theta||\mathbf{S}||_F^2(||\mathbf{S}||_F^2 + ||\mathbf{W}||_F^2), \end{aligned} \tag{8}$$

and that for $t \neq t' \in T$,

$$\mathbf{u}^{(t)^T}\mathbf{W}\mathbf{W}^T\mathbf{u}^{(t')} = \mathbf{u}^{(t)^T}\mathbf{W}\mathbf{W}^T\mathbf{W}\mathbf{W}^T\mathbf{u}^{(t')} = 0.$$

Now consider $t \neq t' \in T$. Then

$$(\mathbf{v}^{(t)^T}\mathbf{v}^{(t')})(\mathbf{v}^{(t)^T}\mathbf{S}^T\mathbf{S}\mathbf{v}^{(t')}) = \frac{(\mathbf{u}^{(t)^T}\mathbf{S}\mathbf{S}^T\mathbf{u}^{(t')})(\mathbf{u}^{(t)^T}\mathbf{S}\mathbf{S}^T\mathbf{S}\mathbf{S}^T\mathbf{u}^{(t')})}{|\mathbf{W}^T\mathbf{u}^{(t)}|^2|\mathbf{W}^T\mathbf{u}^{(t')}|^2}. \tag{9}$$

Furthermore,

$$|\mathbf{u}^{(t)^T}\mathbf{S}\mathbf{S}^T\mathbf{u}^{(t')}| = |\mathbf{u}^{(t)^T}(\mathbf{S}\mathbf{S}^T - \mathbf{W}\mathbf{W}^T)\mathbf{u}^{(t')}| \le \theta||\mathbf{S}||_F^2. \tag{10}$$

Similarly, using (8),

$$|\mathbf{u}^{(t)^T}\mathbf{S}\mathbf{S}^T\mathbf{S}\mathbf{S}^T\mathbf{u}^{(t')}| \le \theta||\mathbf{S}||_F^2(||\mathbf{S}||_F^2 + ||\mathbf{W}||_F^2). \tag{11}$$

Using the bounds (10) and (11) in (9),

$$\begin{aligned} |(\mathbf{v}^{(t)^T}\mathbf{v}^{(t')})(\mathbf{v}^{(t)^T}\mathbf{S}^T\mathbf{S}\mathbf{v}^{(t')})| &\le \frac{\theta^2||\mathbf{S}||_F^4(||\mathbf{S}||_F^2 + ||\mathbf{W}||_F^2)}{\gamma^2||\mathbf{W}||_F^4} \\ &\le \frac{9\theta^2}{\gamma^2}||\mathbf{A}||_F^2 \end{aligned}$$

using the bounds from Lemma 1. Next, for any vector $\mathbf{u}$ and any matrix $\mathbf{S}$

$$\frac{|\mathbf{S}\mathbf{S}^T\mathbf{u}|}{|\mathbf{S}^T\mathbf{u}|} \ge \frac{|\mathbf{S}^T\mathbf{u}|}{|\mathbf{u}|}.$$

So for $t \in T$

$$\mathbf{v}^{(t)^T}\mathbf{S}^T\mathbf{S}\mathbf{v}^{(t)} = \frac{\mathbf{u}^{(t)^T}\mathbf{S}\mathbf{S}^T\mathbf{S}\mathbf{S}^T\mathbf{u}^{(t)}}{|\mathbf{W}^T\mathbf{u}^{(t)}|^2} \ge \frac{|\mathbf{S}^T\mathbf{u}^{(t)}|^4}{|\mathbf{W}^T\mathbf{u}^{(t)}|^2}.$$

Observe that the first part of Lemma 3 implies

$$|\mathbf{S}^T\mathbf{u}^{(t)}|^2 - |\mathbf{W}^T\mathbf{u}^{(t)}|^2 \le \theta||\mathbf{S}||_F^2.$$

So,

$$\left|\frac{|\mathbf{S}^T\mathbf{u}^{(t)}|^2}{|\mathbf{W}^T\mathbf{u}^{(t)}|^2} - 1\right| \le \frac{2\theta}{\gamma} \le \frac{\varepsilon}{16}. \tag{12}$$

Claim 2 follows immediately. We then have

$$\sum_{t\in T}\mathbf{v}^{(t)^T}\mathbf{S}^T\mathbf{S}\mathbf{v}^{(t)} \ge (1 - \tfrac{\varepsilon}{16})\sum_{t\in T}\mathbf{u}^{(t)^T}\mathbf{S}\mathbf{S}^T\mathbf{u}^{(t)} = (1 - \tfrac{\varepsilon}{16})\Delta(\mathbf{S}^T; \mathbf{u}^{(t)}, t \in T).$$

So,

$$\Delta(\mathbf{S}; \mathbf{v}^{(t)}, t \in T) \geq (1 - \tfrac{\varepsilon}{16})\Delta(\mathbf{S}^T; \mathbf{u}^{(t)}, t \in T) - \frac{9k^2\theta^2}{\gamma^2}||\mathbf{A}||_F^2,$$

which completes the proof of Claim 1, since $9k^2\theta^2/\gamma^2 < \varepsilon/16$. The value of p was chosen to satisfy this and (12).

6. RECENT WORK

There have been several developments on the problem of low-rank approximation since a preliminary version of this paper [Frieze et al. 1998] appeared. Drineas et al. [Drineas et al. 2004a] give an algorithm whose running time is $O(mr^2 + r^3)$ where $r = O(k/\varepsilon^2)$. Although this is theoretically much slower (due to the dependence on m), in practice, the better dependence on k and $1/\varepsilon$ might make it more practical. An alternative sampling-based algorithm was given in [Achlioptas and McSherry 2001] with comparable bounds for the Frobenius norm and significantly better bounds for the 2-norm. There, the main idea is to sparsify the matrix by randomly sampling its entries, one by one, and then compute the SVD of a sparse matrix (which is faster). In [Bar-Yossef 2003], a lower bound for low-rank approximation is given, which essentially matches the bound of [Drineas et al. 2004a]. It is also shown there that an algorithm with this complexity is not possible using just uniform sampling. In [Drineas et al. 2004a], (and the current paper), we only get an implicitly defined low-rank approximation to $\mathbf{A}$. A more explicit approximation is given in [Drineas and Kannan 2003], which shows that if $\mathbf{C}$ is a $m \times s$ matrix of s columns of $\mathbf{A}$ picked by sampling and $\mathbf{R}$ is a $s \times n$ matrix of s rows of $\mathbf{A}$ picked at random, then we have $\mathbf{A} \approx \mathbf{CUR}$, where $\mathbf{U}$ is a $s \times s$ matrix computed from $\mathbf{C}$. This thus is a more explicit approximation preserving the sparsity structure of $\mathbf{A}$. The paper [Drineas and Kannan 2001] applied the sampling idea here to the basic problem of multiplying two matrices $\mathbf{A}$ and $\mathbf{B}$ and showed that if we sample a few columns of $\mathbf{A}$ (according to probability distributions similar to this paper) and take the corresponding rows of $\mathbf{B}$, their product approximates the whole product $\mathbf{AB}$. Finally, an improvement of our analysis, in terms of the number of rows that need to be sampled has been obtained in [Drineas et al. 2004b].

REFERENCES

D. Achlioptas and F. McSherry, "Fast Computation of Low Rank Approximations" Proceedings of the 33rd Annual Symposium on Theory of Computing, 611-618, 2001.

N. Alon, R. A. Duke, H Lefmann, V. Rödl and R. Yuster, "The algorithmic aspects of the Regularity Lemma," *Journal of Algorithms* 16, 80-109, 1994.

Z. Bar-Yossef, "Sampling Lower Bounds via Information Theory", Proceedings of the 35th Annual Symposium on Theory of Computing, 335-344, 2003.

M. W. Berry, S. T. Dumais, and G. W. O'Brien. "Using linear algebra for intelligent information retrieval", *SIAM Review* 37(4), 1995, 573-595, 1995.

S. Deerwester, S. T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman. "Indexing by latent semantic analysis," *Journal of the Society for Information Science* 41(6), 391-407, 1990.

P. Drineas and R. Kannan, "Fast Monte-Carlo Algorithms for approximate Matrix Multiplication", Proceedings of the 42nd IEEE Annual Symposium on the Foundations of Computer Science, 452-459, 2001.

P. Drineas and R. Kannan, "Pass Efficient Algorithms for approximating large matrices" Proceedings of the Symposium on Discrete Algortihms, 223-232, 2003.

P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay, "Clustering Large Graphs via the Singular Value Decomposition," *Machine Learning* 56, 9-33, 2004.

P. Drineas, M. W. Mahoney, and R. Kannan, "Fast Monte Carlo Algorithms for Matrices II: Computing Low-Rank Approximations to a Matrix," Technical Report, Yale University, YALEU/DCS/TR-1270, 2004.

S.T. Dumais, G.W. Furnas, T.K. Landauer, and S. Deerwester, "Using latent semantic analysis to improve information retrieval," In Proceedings of CHI'88: Conference on Human Factors in Computing, New York: ACM, 281-285, 1988.

S.T. Dumais, "Improving the retrieval of information from external sources", *Behavior Research Methods, Instruments and Computers* 23(2), 229-236, 1991.

A.M.Frieze and R. Kannan, "The Regularity Lemma and approximation schemes for dense problems",Proceedings of the 37th Annual IEEE Symposium on Foundations of Computing, (1996) 12-20.

A.M.Frieze and R. Kannan, "Quick approximations to matrices and applications," *Combinatorica* 19 (1999) 175-220.

A.M.Frieze and R. Kannan, "A simple algorithm for constructing Szemeredi's Regularity Partition", *Electronic Journal of Combinatorics* 6(1) (1999) R17.

A. Frieze, R. Kannan and S. Vempala, "Fast Monte-Carlo algorithms for finding low-rank approximations", *Proceedings of 39th Symposium on Foundations of Computer Science*, 370-378, 1998.

G. H. Golub and C. F. Van Loan, Matrix Computations, Johns Hopkins University Press, London, 1989.

J. Kleinberg, "Authoritative sources in a hyperlinked environment," *JACM* 46(5), 604-632, 1999.

J. Komlós and M. Simonovits, "Szemerédi's Regularity Lemma and its applications in graph theory", Combinatorics, Paul Erdos is Eighty, (D. Miklos et. al, eds.), Bolyai Society Mathematical Studies, 2, 295-352, 1996.

C. Papadimitriou, P. Raghavan, H. Tamaki and S. Vempala, *Latent Semantic Indexing: A Probabilistic Analysis*, *JCSS* 61, 217–235, 2000.

E. Szemeredi, "Regular partitions of graphs," Proceedings, Colloque Inter. CNRS (J.-C. Bermond, J.-C.Fournier, M.Las Vergnas and D.Sotteau, Eds.), 399-401, 1978.

Received October 2003; May 2004; accepted June 2004.