Fast Monte-Carlo Algorithms for finding low-rank approximations

Alan Frieze[*]
Department of Mathematical Sciences,
Carnegie Mellon University,
Pittsburgh, PA 15213.
Email: af1p@andrew.cmu.edu.

Ravi Kannan
Computer Science Department,
Yale University,
New Haven, CT 06511.
Email: kannan@cs.yale.edu.

Santosh Vempala
Department of Mathematics and
Laboratory for Computer Science,
M.I.T., Cambridge, MA 02139.
Email: vempala@math.mit.edu

October 22, 1998

Abstract

In several applications, the data consists of an $m \times n$ matrix $\mathbf{A}$ and it is of interest to find an approximation $\mathbf{D}$ of a specified rank k to $\mathbf{A}$ where, k is much smaller than m and n. Traditional methods like the Singular Value Decomposition (SVD) help us find the "best" such approximation. However, these methods take time polynomial in m, n which is often too prohibitive.

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. Indeed, in the applications such sampling is possible.

Our **main result** *is that we can find* the description of *a matrix $\mathbf{D}^*$ of rank at most k so that*

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

holds with probability at least $1 - \delta$. (For any matrix $\mathbf{M}$, $||\mathbf{M}||_F^2$ denotes the sum of the squares of all the entries of $\mathbf{M}$.) The algorithm takes time polynomial in $k, 1/\varepsilon, \log(1/\delta)$ only, independent of m, n.

1 Introduction

In many applications, the data consists of an $m \times n$ matrix $\mathbf{A}$ and it is of interest to find an approximation $\mathbf{D}$ of a specified rank k to $\mathbf{A}$ where, k is much smaller than m and n. Traditional methods like the Singular Value Decomposition (SVD) help us find the "best" such approximation. However, these methods take time polynomial in m, n. In this paper, we essentially reduce the problem to a singular value problem in s dimensions where s depends only upon $k, 1/\varepsilon$.

[*] Supported in part by NSF grant CCR-9530974.

The traditional “random projection” method (where one projects the problem into a randomly chosen subspace of small dimension) would also accomplish a similar reduction in dimension; but carrying out the random projection amounts to premultiplying the given $m \times n$ matrix $\mathbf{A}$ by a $s \times m$ matrix which itself takes time dependent upon m, n (and in fact, it can be argued that this is not competitive with known Numerical Analysis techniques like the Lanczos method, in the case where the top few singular values dominate.)

In this paper, we describe an algorithm which is qualitatively faster provided we may sample the entries of the matrix according to a natural probability distribution which we describe presently.

For a matrix $\mathbf{M}$, $||\mathbf{M}||_F^2$ denotes $\sum_{i,j} \mathbf{M}_{i,j}^2$, where $\mathbf{M}_{i,j}$ denotes the i, jth entry of $\mathbf{M}$..

Our **main result** is expressed as 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*

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

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 $s = O(k^4 \varepsilon^{-3})$. The matrix $\mathbf{D}^*$ can be explicitly constructed from its description in $O(kmn)$ time.*

This depends on the following existence theorem. Let

$$\eta = \frac{1}{||\mathbf{A}||_F^2} \min_{\mathbf{D}, \text{rank}(\mathbf{D}) \leq k} ||\mathbf{A} - \mathbf{D}||_F^2. \tag{1}$$

Let c be as defined in Assumption 1 (below).

Theorem 2 *Let $\mathbf{A}$ be an $m \times n$ matrix. Let k, s be any positive integers. Suppose we independently choose a set S, $|S| = s$ of rows of $\mathbf{A}$ from a distribution satisfying (5) (below). Let V be the vector space spanned by the (at most) s rows of $\mathbf{A}$ chosen.*

With probability at least 9/10, there exist vectors $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots \mathbf{y}^{(k)}$ in V such that

$$||\mathbf{A} - \mathbf{A}(\sum_{j=1}^{k} \mathbf{y}^{(j)} \mathbf{y}^{(j)^T})||_F^2 \leq (\eta + \tfrac{10k}{cs}) ||\mathbf{A}||_F^2. \tag{2}$$

This theorem asserts the existence of “good” vectors $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots \mathbf{y}^{(k)}$ in the row space of $\mathbf{S}$.

It follows from Linear Algebra that we may take $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots \ldots \mathbf{y}^{(k)}$ to be the k largest generalized eigenvectors of $\mathbf{S}\mathbf{A}\mathbf{A}^T\mathbf{S}^T$ with respect to $\mathbf{S}\mathbf{S}^T$. Note that both of these matrices are $s \times s$, where we should take $s = O(k/\varepsilon)$. They can both be exactly computed (by direct multiplication) in time $O(smn)$; so it follows that in time $O(smn + \text{poly}\,(s))$, we can find the vectors of the Theorem. This time bound may be good enough for several applications.

The paper is organized as follows. After recalling the SVD and related definitions, we prove Theorem 2. As already remarked this immediately leads to an $O(mnk/\varepsilon + \text{poly}(k/\varepsilon))$ time algorithm.

We develop a theoretically better ("constant time") algorithm in Section 5, which relies heavily on sampling and in the last two sections we analyse its quality and efficiency.

Assumptions on sampling

We now state in detail the assumptions we make on the ability to sample. We discuss in the next section some prominent applications where these assumptions are naturally satisfied. Also, in the important "dense" case, uniformly sampling the entries satisfies the assumptions. (See Remark 1). In any case, after a one-pass preprocessing of the matrix, the assumptions can be satisfied. (See Remark 2).

For a matrix $\mathbf{M}$, $\mathbf{M}^{(i)}$ denotes the ith row, $\mathbf{M}_{(j)}$ denotes the jth column.

Assumption 1 We can choose row i of the matrix $\mathbf{A}$ with probability P_i satisfying $P_i \geq c|\mathbf{A}^{(i)}|^2/||\mathbf{A}||_F^2$ for some constant $c \leq 1$ independent of m, n. The P_i are known to us.

Assumption 2 For any given $i \in \{1, 2, \dots m\}$, we can pick a $j, j = 1, 2, \dots n$ with probabilities $Q_{j|i}$ satisfying $Q_{j|i} \geq cP_{i,j}/P_i$ where $P_{i,j} = \mathbf{A}_{i,j}^2/||\mathbf{A}||_F^2$. The $Q_{j|i}$ are known to us.

Remark 1: Note that if the matrix $\mathbf{A}$ is dense, i.e., $\mathbf{A}_{ij}^2 \leq c'||\mathbf{A}||_F^2/(mn)$ for some constant c', then we may take $P_{ij} = 1/(mn)$. Then of course we can take $P_i = 1/m$ and we may take $Q_{j|i} = 1/n$ for all i, j and $c = 1/c'$.

Remark 2: For any matrix at all, we claim that after making one pass through the entire matrix, we can set up data structures so that after that we can sample the entries fast - $O(1)$ time per sample, so as to satisfy Assumptions (1) and (2).

During the one pass, we do several things. Suppose M is such that for all i, j

$$\mathbf{A}_{ij}^2 \leq M$$

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

We create $O(\log M)$ bins; in the one-pass, we put into the lth bin all the entries (i, j) such that $\frac{1}{M}2^{l-1} \leq |\mathbf{A}_{ij}|^2 \leq \frac{1}{M}2^l$. We also keep track of the number of entries in each bin. After this, we pretend all entries in a bin are of equal absolute value and then it is easy to set up a sampler - the details of the data structures are elementary and left to the reader. In the pass, we also set up similar data structures for each row, so that the other assumptions can be staisfied.

Remark 3: If the matrix $\mathbf{A}$ has a known sparsity structure, i.e., if there is a simple set $S \subseteq \{(i, j) : i = 1, 2, \dots m; j = 1, 2, \dots n\}$ and $\mathbf{A}_{ij} = 0$ for all $(i, j) \notin S$, and in addition if $\mathbf{A}_{ij}^2 \leq c'||\mathbf{A}||_F^2/|S|$ for all i, j, then we may take $P_{ij} = 1/|S|$ for $(i, j) \in S$ and 0 otherwise. If the set S is simple enough, we can clearly find P_{ij}, P_i in unit time and we may take $Q_{j|i} = P_{ij}/P_i$.

2 Some applications

In this section we discuss our algorithm in the context of applications that rely on computing the (first k terms of the) SVD. 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.

2.1 Low-Rank Approximations and the Regularity Lemma

The fundamental Regularity Lemma of Szemerédi's in Graph Theory 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 [10]) 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 aand Yuster were finally able to give an algorithm to find such a partition in polynomial time [1]. In earlier papers [6, 7], we related low-rank approximations of the adjacency matrix of the graph to regular partitions and from that were able to derive both Szemerédi's Lemma and a "more user friendly version" and in fact showed that the partition could be constructed in constant time for any graph. While this connection is not directly relevant to this paper, we point this out here as one more case where low-rank approximations come in handy.

2.2 Latent Semantic Indexing

This is a general technique for analysing a collection of "documents" which are assumed to be related (for example, they are all documents dealing with a particular subject, or a portion of the web; see [2, 3, 4, 5] for details and empirical results). We give a very cursory description of this broad area here and discuss its relation to our main problem.

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 (because there are relationships among the documents), there are a small number k of main (unknown) "topics" which the documents are about. The first aim of the technique is to find a set of k topics which best describe the documents. (This is the only part which concerns us here.)

A topic is modelled as an $n-$vector of non-negative reals summing to 1, where the interpretation is that the jth component of a topic vector gives the frequency with which the jth term occurs in (a discussion of) the topic. With this model on hand, it is easy to argue (using Linear Algebra and a line of reasoning similar to the field of "Factor Analysis" in Statistics) that the k best topics are the top k singular vectors of the so-called "document-term" matrix, which 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, we can implement the assumptions of our algorithm. It is easy to see that if we are allowed one pass through each document, we can set up data structures for sampling (in a pragmatic situation one could have the creator of a document supply a vector of squared term frequencies). Otherwise, if no frequency is too large (this is typical 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"). In this case, assumption (1) is easily implemented — 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 $\frac{L_i}{\sum_j L_j}$). It can be then argued that the assumption with the squares is satisfied (because the frequencies are all in some small range). Assumption 2 is similarly implemented — given a document, we pick a word uniformly at random from it, i.e., $Q_{j|i} = \frac{\mathbf{A}_{ij}}{L_i}$.

2.3 Web Search model

Kleinberg [9] considered the ubiquitous problem of how to glean the most relevant documents from the (usually large) set of documents returned by a standard Web Search program for a key word. The intuition is to define a document to be an "authority" if a lot of other documents (returned by the search) point to (have a hypertext link to) it. He argues why it is not a good idea just to take documents which are pointed to by a lot of others. He defines a dual notion - a document is a "hub" if it **points to** a lot of other documents. More genarally, suppose n documents are returned by the search engine. Then, he defines an $n \times n$ matrix $\mathbf{A}$ where $\mathbf{A}_{ij}$ is 1 or 0 depending upon whether the ith document points to the j th. [He does not explicitly deal with this large matrix.]

He sets out to find two n-vectors - x, y where x_i is the "hub weight" of document i (the weight is higher if the document is a good hub) and y_j is the "authority weight" of document j. With the normalization $|x| = |y| = 1$, he argues (and we do not reproduce the argument here) that it is desirable to find $\max_{|x|=|y|=1} x^T \mathbf{A} y$, (since in the maximizing x, y we expect the hub weights and authority weights to be mutually consistent.)

This is of course the problem of finding the singular vectors of $\mathbf{A}$. Since $\mathbf{A}$ is large, he judiciously chooses a submatrix of $\mathbf{A}$ and computes only the singular vectors of it.

He also points out that especially in the case when the key word has multiple meanings, not only the top, but some of the other singular vectors (with large singular values) are interesting. For example, when the key word is "JAVA" the top few singular vectors put high authority weights on documents about the programming language JAVA whereas another set of singular vectors put high weights on documents about the Island, others on documents about the coffee etc. So, it is of interest to find the largest k singular vectors for some small k; this is indeed the problem we consider here.

We also find the singular vectors of a submatrix, but a randomly chosen one. 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.

3 Some Notation and Constants

For a matrix $\mathbf{M}$, $\mathbf{M}^{(i)}$ denotes the ith row, $\mathbf{M}_{(j)}$ denotes the jth column and $||\mathbf{M}||_F^2$ denotes $\sum_{i,j} \mathbf{M}_{i,j}^2$.

If $Z = (Z_1, Z_2, \ldots, Z_n)$ is a vector valued random variable, $\mathbf{E}(Z)$ denotes the expectation vector componentwise, $(\mathbf{E}(Z_1), \mathbf{E}(Z_2), \ldots, \mathbf{E}(Z_n))$. $\mathbf{Var}(Z)$ denotes $\sum_{i=1}^n \mathbf{Var}(Z_i)$ and the Chebychev inequality becomes

$$\mathbf{Pr}(|Z - \mathbf{E}(Z)|^2 \geq t\mathbf{Var}(Z)) \leq \frac{1}{t}$$

for any $t > 0$.

For a positive integer r, we let $[r] = \{1, 2, \ldots, r\}$.

For a matrix $\mathbf{M}$ and vectors $\mathbf{x}^{(i)}$, $i \in I$ we define

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

Using the fact that $||\mathbf{N}||_F^2 = \mathrm{Tr}(\mathbf{N}\mathbf{N}^T)$ for any matrix $\mathbf{N}$, we see that $\Delta(\mathbf{M}; \mathbf{x}^{(i)}, i \in I)$ equals

$$\begin{aligned}
&2\sum_{i \in I} \mathrm{Tr}(\mathbf{M}\mathbf{x}^{(i)}\mathbf{x}^{(i)^T}\mathbf{M}^T) \\
&- \sum_{i,i' \in I} (\mathbf{x}^{(i)^T}\mathbf{x}(i')^T)\mathrm{Tr}(\mathbf{M}\mathbf{x}^{(i)}\mathbf{x}^{(i')^T}\mathbf{M}^T \\
&= \sum_{i \in I} \mathbf{x}^{(i)^T}\mathbf{M}^T\mathbf{M}\mathbf{x}^{(i)} \\
&- \sum_{i \neq i' \in I} (\mathbf{x}^{(i)^T}\mathbf{x}(i')^T)\mathbf{x}^{(i')^T}\mathbf{M}^T\mathbf{M}\mathbf{x}^{(i)}.
\end{aligned} \tag{3}$$

4 Proof of Theorem 2

4.1 Singular Value Decomposition

Every real matrix 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 vectors and so 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 $\mathbf{A}$.

So if in (3) the vectors $\mathbf{x}^{(i)}$, $i \in I$ are singular vectors of **M** then

$$\Delta(\mathbf{M}; \mathbf{x}^{(i)}, i \in I) = \sum_{i \in I} \mathbf{x}^{(i)^T}\mathbf{M}^T\mathbf{M}\mathbf{x}^{(i)}. \tag{4}$$

From Linear Algebra, [8] we know that the matrix $\mathbf{D}_k$ producing the minimum of $||\mathbf{A} - \mathbf{D}||_F$ 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 (see (1) for the definition of η) that

$$\eta||\mathbf{A}||_F^2 = \sum_{t=k+1}^{r} \sigma_t^2.$$

We now show that we can find a good approximation to $\mathbf{D}_k$ by looking in a subspace generated by a small number of rows of $\mathbf{A}$. This will be done by independently choosing s rows of $\mathbf{A}$ (s sufficiently large) from a distribution $P_1, P_2, \ldots, P_m$ where the probability P_i that we choose row i satisfies

$$P_i \geq \frac{c|\mathbf{A}^{(i)}|^2}{||\mathbf{A}||_F^2} \tag{5}$$

for $1 \leq i \leq m$ and some absolute constant $0 < c \leq 1$.

4.2 The proof itself

Let S be the random set of rows described in the statement of the theorem. We define for $t = 1, 2, \ldots, r$ the vector random variable

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

Then

$$\mathbf{E}(\mathbf{w}^{(t)}) = \mathbf{A}^T \mathbf{u}^{(t)} = \sigma_t \mathbf{v}^{(t)},$$

and since $P_i \geq c|\mathbf{A}^{(i)}|^2/||\mathbf{A}||_F^2$,

$$\mathbf{E}(|\mathbf{w}^{(t)} - \sigma_t \mathbf{v}^{(t)}|^2) \leq \frac{1}{s} \sum_{i=1}^{m} |\mathbf{u}_i^{(t)}|^2 |\mathbf{A}^{(i)}|^2 / P_i \leq \frac{1}{sc} ||\mathbf{A}||_F^2. \tag{6}$$

Now let $\hat{\mathbf{y}}^{(t)} = \frac{1}{\sigma_t} \mathbf{w}^{(t)}$ for $t = 1, 2, \ldots, r$ and let $V_1 = \text{span}(\hat{\mathbf{y}}^{(1)}, \hat{\mathbf{y}}^{(2)}, \ldots, \hat{\mathbf{y}}^{(k)}) \subseteq V$.

Let $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(n)}$ be an orthonormal basis of $\boldsymbol{R}^n$ with $V_1 \subseteq \text{span}(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(k)})$. Let

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

Then

$$\begin{aligned} ||\mathbf{A} - \mathbf{F}||_F^2 = \sum_{i=1}^{n} |(\mathbf{A} - \mathbf{F}) \mathbf{y}^{(i)}|^2 = \sum_{i=k+1}^{n} |\mathbf{A} \mathbf{y}^{(i)}|^2 & \\ = \sum_{i=k+1}^{n} |(\mathbf{A} - \hat{\mathbf{F}}) \mathbf{y}^{(i)}|^2 \leq ||\mathbf{A} - \hat{\mathbf{F}}||_F^2. & \end{aligned} \tag{7}$$

Now

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

Taking expectations and using (6) we get

$$\mathbf{E}(||\mathbf{A}-\hat{\mathbf{F}}||_F^2 - \sum_{i=k+1}^{n} \sigma_i^2) \le \frac{k}{sc}||\mathbf{A}||_F^2. \tag{8}$$

Since $\hat{\mathbf{F}}$ is of rank $\le k$ we have

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

Thus $||\mathbf{A}-\hat{\mathbf{F}}||_F^2 - \eta||\mathbf{A}||_F^2$ is a non-negative random variable and (8) implies

$$\mathbf{Pr}(||\mathbf{A}-\hat{\mathbf{F}}||_F^2 - \eta||\mathbf{A}||_F^2 \ge \tfrac{10k}{sc}||\mathbf{A}||_F^2) \le \tfrac{1}{10}.$$

The result now follows from (7). □

We note next that a good low-rank approximation with respect to Frobenius norm, implies a good low-rank approximation with respect to the norm $||\mathbf{M}|| = \max_{|\mathbf{x}|=1}|\mathbf{Mx}|$.

Theorem 3 *If*

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

Then

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

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

$$||\mathbf{B}-\mathbf{Bxx}^T||_F^2 = ||\mathbf{B}||_F^2 - \lambda^2 < \sum_{i=k+1}^{n}\sigma_i^2 - \frac{1}{k+1}||\mathbf{A}||_F^2. \tag{9}$$

But the rank of $\mathbf{A}\sum_{t=1}^{k}\mathbf{y}^{(t)}\mathbf{y}^{(t)^T} + \mathbf{Bxx}^T$ is at most $k+1$, and we know that this cannot be better than the best rank $k+1$ approximation to $\mathbf{A}$, i.e.,

$$||\mathbf{A}-\mathbf{A}\sum_{t=1}^{k}\mathbf{y}^{(t)}\mathbf{y}^{(t)^T} - \mathbf{Bxx}^T||_F^2 \ge \sum_{t=k+2}^{r}\sigma_t^2$$

$$\ge \sum_{t=k+1}^{r}\sigma_t^2 - \frac{1}{k+1}||\mathbf{A}||_F^2,$$

which contradicts (9). □

5 Sampling Algorithm

The aim of this section is to develop a “constant time” algorithm to produce the approximation. What we do below is to first pick a set of p rows of $\mathbf{A}$. We form a matrix $\mathbf{S}$ from these rows after scaling them. We then pick again p columns of $\mathbf{S}$ from a probability distribution satisfying a condition of the type stated in Assumtion (1) and scale the columns to get a $p \times p$ matrix $\mathbf{W}$. We find the singular vectors of this matrix and argue that from those, we may get a good low-rank approxmation to $\mathbf{A}$. We first present the algorithm.

Algorithm

$$\varepsilon > 0 \text{ is given and } p = \frac{10^7 k^4}{c^3 \varepsilon^3}. \tag{10}$$

1. Independently choose $i_1, i_2, \ldots, i_p$ according to distribution $P = (P_1, P_2, \ldots, P_m)$ on $[m]$ which satisfies
$$P_i \geq c \frac{|\mathbf{A}^{(i)}|^2}{||\mathbf{A}||_F^2}.$$
Let $\mathbf{S}$ be the $p \times n$ matrix with rows $\mathbf{A}^{(i_t)}/\sqrt{pP_{i_t}}$ for $t = 1, 2, \ldots, p$.

2. Independently choose (columns) $j_1, j_2, \ldots, j_p$ (of $\mathbf{S}$) according to a distribution $P' = (P'_1, P'_2, \ldots, P'_n)$ on $[n]$ which satisfies
$$P'_j \geq \frac{c}{2} \frac{|\mathbf{S}_{(j)}|^2}{||\mathbf{S}||_F^2}.$$
(Later, we see that we can do such sampling.)

   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 the maximum of $\Delta(\mathbf{W}^T; \mathbf{u}^{(t)}, t \in [k])$ over all sets of k unit vectors in the column space of $\mathbf{W}$. (We may assume at this point that $\{\mathbf{u}^{(t)}\}_{t \in [k]}$ are the first k singular vectors of $\mathbf{W}$.)

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

5. Output $\mathbf{v}^{(t)}$ for $t \in T$. (I.e., output $\mathbf{A} \sum_{t \in T} \mathbf{v}^{(t)} \mathbf{v}^{(t)^T}$ as the approximation to $\mathbf{A}$).

Note that

$$||\mathbf{S}||_F^2 \leq \frac{||\mathbf{A}||_F^2}{c} \text{ and } ||\mathbf{W}||_F^2 \leq \frac{2||\mathbf{S}||_F^2}{c} \leq \frac{2}{c^2} ||\mathbf{A}||_F^2. \tag{11}$$

$$\mathbf{E}(||\mathbf{S}||_F^2) = ||\mathbf{A}||_F^2 \text{ and } \mathbf{E}(||\mathbf{W}||_F^2) = ||\mathbf{S}||_F^2. \tag{12}$$

5.1 Implementation Issues

Implementation Issues We explore some issues related to the implementation of the above algorithm.

First of all, 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 $j \in \{1, 2, \dots n\}$ with probabilities $Q_{j|i}$. This defines the probabilities P'_j. We then have (with $I = \{i : \mathbf{A}^{(i)}$ is a row of $\mathbf{S}\}$),

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

$$= \frac{c}{||\mathbf{A}||_F^2} \sum_{i \in I} \frac{\mathbf{A}_{i,j}^2}{pP_i} = c\frac{|\mathbf{S}_{(j)}|^2}{||\mathbf{A}||_F^2}.$$

Let

$$q = \mathbf{Pr}(||\mathbf{S}||_F^2 \leq ||\mathbf{A}||_F^2/2).$$

Then

$$||\mathbf{A}||_F^2 \leq \frac{q}{2}||\mathbf{A}||_F^2 + \frac{1-q}{c}||\mathbf{A}||_F^2$$

from which it follows that

$$q \leq \frac{2(1-c)}{2-c}.$$

It follows that with probablity

$$1 - q \geq \frac{c}{2-c}$$

we have,

$$P'_j \geq \frac{c}{2} \frac{|\mathbf{S}_{(j)}|^2}{||\mathbf{S}||_F^2}. \tag{13}$$

So let us assume from now on that

$$||\mathbf{S}||_F^2 \geq \tfrac{1}{2}||\mathbf{A}||_F^2 \text{ and also that } ||\mathbf{W}||_F^2 \geq \tfrac{1}{2}||\mathbf{S}||_F^2.$$

5.2 Basic Lemma

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

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

for some $0 < \alpha < 1$.

Let $\sigma = (i_1, i_2, \dots, i_p)$ 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

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

Then for all $\theta > 0$,

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

Proof

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

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

Thus

$$\mathbf{E}(||\mathbf{M}^T\mathbf{M} - \mathbf{N}^T\mathbf{N}||_F^2) =$$

$$\sum_{r,s=1}^{b} \mathbf{E}(|\mathbf{N}_{(r)}^T\mathbf{N}_{(s)} - \mathbf{M}_{(r)}^T\mathbf{M}_{(s)}|^2)$$

$$\leq \frac{||\mathbf{M}||_F^2}{\alpha p}\sum_{i=1}^{a}\frac{1}{|\mathbf{M}^{(i)}|^2}\sum_{r,s=1}^{b}(\mathbf{M}_{i,r}\mathbf{M}_{i,s})^2 = \frac{||\mathbf{M}||_F^4}{\alpha p}.$$

The result follows from the Markov inequality. □

It follows from the above lemma and the definition of p – (10) – that with probability at least 9/10 both of the following events hold:

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

and

$$\{||\mathbf{S}\mathbf{S}^T - \mathbf{W}\mathbf{W}^T||_F \leq \theta||\mathbf{S}||_F^2\} \quad (14)$$

where

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

Assume from now on that they do.

So if $\mathbf{z}, \mathbf{z}'$ are unit vectors in the row space of $\mathbf{A}$ then

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

and if $\mathbf{z}, \mathbf{z}'$ are unit vectors in the column space of $\mathbf{S}$

$$|\mathbf{z}^T\mathbf{S}^T\mathbf{S}\mathbf{z}' - \mathbf{z}^T\mathbf{W}^T\mathbf{W}\mathbf{z}'| \leq \theta||\mathbf{S}||_F^2 \tag{15}$$

It follows after a little calculation that if $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots, \mathbf{z}^{(\ell)}, \ell \leq k$ are unit vectors in the row space of $\mathbf{A}$ then

$$|\Delta(\mathbf{A}; \mathbf{z}^{(i)}, i \in [\ell]) - \Delta(\mathbf{S}; \mathbf{z}^{(i)}, i \in [\ell])| \leq k^2\theta||\mathbf{A}||_F^2. \tag{16}$$

Similarly, $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots, \mathbf{z}^{(k)}$ are unit vectors in the column space of $\mathbf{S}$ then

$$|\Delta(\mathbf{S}^T; \mathbf{z}^{(i)}, i \in [\ell]) - \Delta(\mathbf{W}^T; \mathbf{z}^{(i)}, i \in [\ell])| \leq k^2\theta||\mathbf{S}||_F^2. \tag{17}$$

5.3 Analysis of the Algorithm

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

$$\begin{aligned}\Delta(\mathbf{A}; \mathbf{x}^{(t)}, t \in [k]) &\geq (1 - \eta - \tfrac{10k}{cp}\varepsilon)||\mathbf{A}||_F^2 \\ &\geq (1 - \eta - \tfrac{1}{8}\varepsilon)||\mathbf{A}||_F^2.\end{aligned}$$

Applying (16) we see that

$$\begin{aligned}\Delta(\mathbf{S}; \mathbf{x}^{(t)}, t \in [k]) &\geq \Delta(\mathbf{A}; \mathbf{x}^{(t)}, t \in [k]) - k^2\theta||\mathbf{A}||_F^2 \\ &\geq (1 - \eta - \tfrac{1}{4}\varepsilon)||\mathbf{A}||_F^2.\end{aligned}$$

$\mathbf{S}$ 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

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

Applying Theorem 2 with $\mathbf{A}$ replaced by $\mathbf{S}^T$ and $\mathbf{S}$ replaced by $\mathbf{W}^T$ 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

$$\begin{aligned}\Delta(\mathbf{S}^T; \mathbf{z}^{(t)}, t \in [k]) &\geq \Delta(\mathbf{S}^T; \mathbf{y}^{(t)}, t \in [k]) - \tfrac{10k}{cp}||\mathbf{S}||_F^2 \\ &\geq (1 - \eta - \tfrac{3}{8}\varepsilon)||\mathbf{A}||_F^2.\end{aligned}$$

Applying (17) we see that

$$\begin{aligned}\Delta(\mathbf{W}^T; \mathbf{z}^{(t)}, t \in [k]) &\geq \Delta(\mathbf{S}^T; \mathbf{z}^{(t)}, t \in [k]) - k^2\theta||\mathbf{S}||_F^2 \\ &\geq (1 - \eta - \tfrac{1}{2}\varepsilon)||\mathbf{A}||_F^2.\end{aligned}$$

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

$$\Delta(\mathbf{W}^T; \mathbf{u}^{(t)}, t \in [k]) \geq (1 - \eta - \tfrac{1}{2}\varepsilon)||\mathbf{A}||_F^2.$$

Now because $\mathbf{u}^{(t)}, t \in [k]$ are singular vectors we see from (4) that

$$\begin{aligned}\Delta(\mathbf{W}^T; \mathbf{u}^{(t)}, t \in T) &\geq \Delta(\mathbf{W}^T; \mathbf{u}^{(t)}, t \in [k]) - k\gamma||\mathbf{W}||_F^2 \\ &\geq (1 - \eta - \tfrac{5}{8}\varepsilon)||\mathbf{A}||_F^2.\end{aligned}$$

It follows from (17) that

$$\begin{aligned}\Delta(\mathbf{S}^T; \mathbf{u}^{(t)}, t \in T) &\geq \Delta(\mathbf{W}^T; \mathbf{u}^{(t)}, t \in T) - k^2\theta||\mathbf{S}||_F^2 \\ &\geq (1 - \eta - \tfrac{3}{4}\varepsilon)||\mathbf{A}||_F^2.\end{aligned}$$

In Section 6 we prove that

$$\Delta(\mathbf{S}; \mathbf{v}^{(t)}, t \in T) \geq \Delta(\mathbf{S}^T; \mathbf{u}^{(t)}, t \in T) - \tfrac{1}{8}\varepsilon||\mathbf{A}||_F^2 \tag{18}$$

and that

$$|\mathbf{v}^{(t)}|^2 \leq 1 + \frac{\varepsilon}{16} \tag{19}$$

for $t \in T$.

It follows from (16) that

$$\begin{aligned}\Delta(\mathbf{A}; \mathbf{v}^{(t)}, t \in T) &\geq \Delta(\mathbf{S}; \mathbf{v}^{(t)}, t \in T) - 2k^2\theta||\mathbf{A}||_F^2 \\ &\geq (1 - \eta - \varepsilon)||\mathbf{A}||_F^2\end{aligned}$$

(assuming $\varepsilon \leq 16$) which completes the proof of Theorem 1. □

6 Proof of (18) and (19)

Observe first that

$$\begin{aligned}||\mathbf{S}\mathbf{S}^T\mathbf{S}\mathbf{S}^T - \mathbf{W}\mathbf{W}^T\mathbf{W}\mathbf{W}^T||_F &\leq \\ ||\mathbf{S}\mathbf{S}^T(\mathbf{S}\mathbf{S}^T - \mathbf{W}\mathbf{W}^T)||_F &+ ||(\mathbf{S}\mathbf{S}^T - \mathbf{W}\mathbf{W}^T)\mathbf{W}\mathbf{W}^T||_F \\ &\leq \theta||\mathbf{S}||_F^2(||\mathbf{S}||_F^2 + ||\mathbf{W}||_F^2),\end{aligned} \tag{20}$$

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

$$\begin{aligned}(\mathbf{v}^{(t)^T}\mathbf{v}^{(t')})(\mathbf{v}^{(t)^T}\mathbf{S}^T\mathbf{S}\mathbf{v}^{(t')}) &= \\ &\frac{(\mathbf{u}^{(t)^T}\mathbf{S}\mathbf{S}^T\mathbf{u}^{(t')})(\mathbf{u}^{(t)^T}\mathbf{S}\mathbf{S}^T\mathbf{S}\mathbf{S}^T\mathbf{u}^{(t')})}{|\mathbf{W}^T\mathbf{u}^{(t)}|^2|\mathbf{W}^T\mathbf{u}^{(t')}|^2}.\end{aligned}$$

Furthermore,

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

Similarly, using (20),

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

Hence,

$$\begin{aligned} |(\mathbf{v}^{(t)^T}\mathbf{v}^{(t')})(\mathbf{v}^{(t)^T}\mathbf{S}^T\mathbf{S}\mathbf{v}^{(t')})| &\leq \frac{\theta^2||\mathbf{S}||_F^2(||\mathbf{S}||_F^2 + ||\mathbf{W}||_F^2)}{\gamma^2||\mathbf{W}||_F^4} \\ &\leq \frac{12\theta^2}{\gamma^2c^2}||\mathbf{A}||_F^2 \qquad (21) \end{aligned}$$

For any vector $\mathbf{u}$ and any matrix $\mathbf{S}$

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

So for $t \in T$

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

Observe that (15) implies

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

So,

$$\left|\frac{|\mathbf{S}^T\mathbf{u}^{(t)}|^2}{|\mathbf{W}^T\mathbf{u}^{(t)}|^2} - 1\right| \leq \frac{2\theta}{\gamma} \leq \frac{\varepsilon}{16}. \qquad (22)$$

Equation (19) follows immediately.

We then have

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

So,

$$\Delta(\mathbf{S};\mathbf{v}^{(t)}, t \in T) \geq (1 - \tfrac{\varepsilon}{16})\Delta(\mathbf{S}^T;\mathbf{u}^{(t)}, t \in T) - \frac{12k^2\theta^2}{\gamma^2c^2}||\mathbf{A}||_F^2,$$

which completes the proof of (18).

References

[1] N. Alon, R. A. Duke, H Lefmann, V. Rödl and R. Yuster, “The algorithmic aspects of the Regularity Lemma,” Journal of Algorithms 16 (1994) 80-109.

[2] 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.

[3] 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.

[4] 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.

[5] S.T. Dumais, “Improving the retrieval of information from external sources”, Behavior Research Methods, Instruments and Computers, 23(2), 229-236, 1991.

[6] 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.

[7] A.M.Frieze and R. Kannan, “Quick approximations to matrices and applications,” to appear in Combinatorica. **http://www.math.cmu.edu/∼af1p/papers.html**.

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

[9] J. Kleinberg, “Authoritative sources in a hyperlinked environment,” Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.

[10] J. Komlós and M. Simonovits, “Szemerédi’s Regularity Lemma and its applications in graph theory”, to appear.