Electronic Colloquium on Computational Complexity, Report No. 42 (2006)

Adaptive Sampling and Fast Low-Rank Matrix Approximation

Amit Deshpande* Santosh Vempala*

Abstract

We prove that any real matrix A contains a subset of at most $4k/\epsilon + 2k\log(k+1)$ rows whose span "contains" a matrix of rank at most k with error only $(1+\epsilon)$ times the error of the best rank-k approximation of A. This leads to an algorithm to find such an approximation with complexity essentially $O(Mk/\epsilon)$, where M is the number of nonzero entries of A. The algorithm maintains sparsity, and in the streaming model, it can be implemented using only $2(k+1)(\log(k+1)+1)$ passes over the input matrix. Previous algorithms for low-rank approximation use only one or two passes but obtain an additive approximation.

1 Introduction

Given an $m \times n$ matrix A of reals and an integer k, the problem of finding a matrix B of rank at most k that minimizes $\|A - B\|_F^2 = \sum_{i,j}(A_{ij} - B_{ij})^2$ has received much attention in the past decade. The classical optimal solution to this problem is the matrix A_k consisting of the first k terms in the Singular Value Decomposition (SVD) of A:

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^T$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$ are the singular values and $\{u_i\}_1^n, \{v_i\}_1^n$ are orthonormal sets of vectors called left and right singular vectors, respectively. Computing the SVD and hence the best low-rank approximation takes $O(\min\{mn^2, m^2n\})$ time.

Recent work on this problem has focussed on reducing the complexity while allowing an approximation to A_k. Frieze et al. [10] introduced the following sampling approach.

Theorem 1 ([10]). *Let S be an i.i.d. sample of s rows of an $m \times n$ matrix A, from the following distribution: row i is picked with probability*

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

Then there is a matrix $\tilde{A}_k$ whose rows lie in $\mathrm{span}(S)$ *such that*

$$\mathsf{E}\left[\|A - \tilde{A}_k\|_F^2\right] \leq \|A - A_k\|_F^2 + \frac{k}{cs}\|A\|_F^2.$$

*Mathematics Department and CSAIL, MIT. Email:{`amitd, vempala`}`@mit.edu`.

ISSN 1433-8092

Setting $s = k/c\epsilon$ in the theorem, we get

$$\mathsf{E}\left[\|A - \tilde{A}_k\|_F^2\right] \le \|A - A_k\|_F^2 + \epsilon\|A\|_F^2.$$

The theorem suggests a randomized algorithm (analyzed in [10], [5] and later in [7]) that makes two passes through the matrix A and finds such an approximation using $O(\min\{m,n\}k^2/\epsilon^4)$ additional time. So overall, it takes $O(M + \min\{m,n\}k^2/\epsilon^4)$ time. A different sampling approach that uses only one pass and has comparable guarantees (in particular, additive error) was given in [1].

The additive error $\epsilon\|A\|_F^2$ could be arbitrarily large compared to the true error, $\|A - A_k\|_F^2$. Is it possible to get a $(1+\epsilon)$-relative approximation efficiently, i.e., in linear or sublinear time? Related to this, is there a small witness, i.e., Is there a $(1+\epsilon)$-approximation of rank k whose rows lie in a small subset of the rows of A? Addressing these questions, it was shown in [8] that any matrix A contains a subset S of $O(k^2/\epsilon)$ rows such that there is a matrix $\tilde{A}_k$ of rank at most k whose rows lie in span(S) and

$$\|A - \tilde{A}_k\|_F^2 \le (1+\epsilon)\|A - A_k\|_F^2.$$

This existence result was applied to derive an approximation algorithm for a projective clustering problem: find j linear subspaces, each of dimension at most k, that minimize the sum of squared distances of each point to its nearest subspace. However, the question of efficiently finding such a $(1+\epsilon)$-relative approximation to A_k was left open.

1.1 Our Results

Our first result is the following improved existence theorem.

Theorem 2. *Any $m \times n$ matrix A contains a subset S of $4k/\epsilon + 2k\log(k+1)$ rows such that there is a matrix $\tilde{A}_k$ of rank at most k whose rows lie in* span(S) *and*

$$\|A - \tilde{A}_k\|_F^2 \le (1+\epsilon)\|A - A_k\|_F^2.$$

Based on this, we give an effcient algorithm in Section 3.2 that exploits any sparsity of the input matrix. For a matrix with M nonzero entries, a rank-k approximation is computed in

$$O\left(M\left(\frac{k}{\epsilon} + k^2\log k\right) + (m+n)\left(\frac{k^2}{\epsilon^2} + \frac{k^3\log k}{\epsilon} + k^4\log^2 k\right)\right)$$

time using $O(n(\frac{k}{\epsilon} + k^2\log k))$ space (Theorem 9). In the streaming model, the algorithm requires $2(k+1)(\log(k+1)+1)$ passes over the input matrix. The running time is $O\left(M(k/\epsilon + k^2\log k)\right)$ for M sufficiently larger than m, n; when k is a constant it is $O(M/\epsilon + 1/\epsilon^2)$. We note that while the analysis is new, most of the algorithmic ideas were proposed in [8].

We complement the existence result with following lower bound (Prop. 11): there exist matrices for which the span of any subset of $k/2\epsilon$ rows does not contain a $(1+\epsilon)$-relative approximation.

Finally, the improved existence bound also leads to better PTAS for the projective clustering problem. The complexity becomes $d(n/\epsilon)^{O(jk^2/\epsilon + jk^2\log k)}$ reducing the dependence on k in the exponent from k^3 and resolving an open question of [8].

Notation. Henceforth, we will use $\pi_V(A)$ to denote the matrix obtained by projecting each row of A onto a linear subspace V. If V is spanned by a subset S of rows, we denote the projection of A onto V by $\pi_{\text{span}(S)}(A)$. We use $\pi_{\text{span}(S),k}(A)$ for the best rank-k approximation to A whose rows lie in span(S). Thus, the approximation $\tilde{A}_k$ in Theorem 2 is $\tilde{A}_k = \pi_{\text{span}(S),k}(A)$ for a suitable S.

2 Sampling Techniques

We now describe the two sampling techniques that will be used.

2.1 Adaptive Sampling

One way to generalize the sampling procedure of Frieze et al. [10] is to do the sampling in multiple rounds, and in an adaptive fashion. Here is the t-round adaptive sampling algorithm, introduced in [8].

1. Start with an linear subspace V. Let $E_0 = A - \pi_V(A)$, and $S = \emptyset$.
2. For $j = 1$ to t, do:
   (a) Pick a sample S_j of s_j rows of A independently from the following distribution: row i is picked with probability $P_i^{(j-1)} \geq c\frac{\|E_{j-1}^{(i)}\|^2}{\|E_{j-1}\|_F^2}$.
   (b) $S = S \cup S_j$.
   (c) $E_j = A - \pi_{\text{span}(V \cup S)}(A)$.

The next theorem, from [8] is a generalization of Theorem 1.

Theorem 3 ([8])**.** *After one round of the adaptive sampling procedure described above,*

$$\mathsf{E}_{S_1}\left[\|A - \pi_{\text{span}(V\cup S_1),k}(A)\|_F^2\right] \leq \|A - A_k\|_F^2 + \frac{k}{cs_1}\|E_0\|_F^2.$$

We can now prove the following corollary of Theorem 3, for t-round adaptive sampling, using induction on the number of rounds.

Corollary 4. *After t rounds of the adaptive sampling procedure described above,*

$$\mathsf{E}_{S_1,\dots,S_t}\left[\|A - \pi_{\text{span}(V\cup S),k}(A)\|_F^2\right] \leq \left(1 + \frac{k}{cs_t} + \frac{k^2}{c^2 s_t s_{t-1}} + \dots + \frac{k^{t-1}}{c^{t-1} s_t s_{t-1}\dots s_2}\right)\|A - A_k\|_F^2 + \frac{k^t}{c^t s_t s_{t-1}\dots s_1}\|E_0\|_F^2.$$

Proof. We prove the theorem by induction on t. The case $t = 1$ is precisely Theorem 3. For the inductive step, using Theorem 3 with $\text{span}(V \cup S_1 \cup \dots \cup S_{t-1})$ as our initial subspace, we have

$$\mathsf{E}_{S_t}\left[\|A - \pi_{\text{span}(V\cup S),k}(A)\|_F^2\right] \leq \|A - A_k\|_F^2 + \frac{k}{cs_t}\|E_{t-1}\|_F^2.$$

Combining this inequality with the fact that

$$\|E_{t-1}\|_F^2 = \|A - \pi_{\text{span}(V\cup S_1\cup\dots\cup S_{t-1})}(A)\|_F^2 \leq \|A - \pi_{\text{span}(V\cup S_1\cup\dots\cup S_{t-1}),k}(A)\|_F^2$$

we get

$$\mathsf{E}_{S_t}\left[\|A - \pi_{\text{span}(S'),k}(A)\|_F^2\right] \leq \|A - A_k\|_F^2 + \frac{k}{cs_t}\|A - \pi_{\text{span}(V\cup S_1\cup\dots\cup S_{t-1}),k}(A)\|_F^2.$$

Finally, taking the expectation over $S_1, \ldots, S_{t-1}$:

$$\mathsf{E}_{S_1,\ldots,S_t}\left[\|A - \pi_{\mathrm{span}(V\cup S),k}(A)\|_F^2\right] \leq \|A - A_k\|_F^2 + \frac{k}{cs_t}\mathsf{E}_{S_1,\ldots,S_{t-1}}\left[\|A - \pi_{\mathrm{span}(V\cup S_1\cup\cdots\cup S_{t-1}),k}(A)\|_F^2\right]$$

and the result follows from the induction hypothesis for $t-1$. $\square$

From Corollary 4, it is clear that if we can get a good initial subspace V such that $dim(V) = k$ and the error given by V is within some multiplicative factor of $\|A - A_k\|_F^2$, then we can hope to prove something about relative rank-k approximation. This motivates a different generalization of the sampling method of [10].

2.2 Volume Sampling

Another way to generalize the sampling scheme of Frieze et al. [10] is by sampling subsets of rows instead of individual rows. Let S be a subset of k rows of A, and $\Delta(S)$ be the simplex formed by these rows and the origin. Volume sampling corresponds to the following distribution: we pick subset S with probability equal to

$$P_S = \frac{\mathrm{vol}(\Delta(S))^2}{\sum_{T:|T|=k}\mathrm{vol}(\Delta(T))^2}.$$

This was also introduced in [8] to prove the next theorem.

Theorem 5 ([8])**.** *Let S be a random subset of k rows of a given matrix A chosen with probability P_S defined as above. Then.*

$$\mathsf{E}_S\left[\|A - \pi_{\mathrm{span}(S)}(A)\|_F^2\right] \leq (k+1)\|A - A_k\|_F^2.$$

The next lemma was used crucially in the analysis of volume sampling.

Lemma 6 ([8])**.**

$$\sum_{S,|S|=k}\mathrm{vol}(\Delta_S)^2 = \frac{1}{(k!)^2}\sum_{1\leq t_1<t_2<\ldots<t_k\leq n}\sigma_{t_1}^2\sigma_{t_2}^2\ldots\sigma_{t_k}^2,$$

where $\sigma_1, \sigma_2, \ldots, \sigma_r > 0 = \sigma_{r+1} = \ldots = \sigma_n$ are the singular values of A.

2.3 Approximate Volume Sampling via Adaptive Sampling

Here we give an algorithm for approximate volume sampling. In brief, we run a k-round adaptive sampling procedure, picking one row in each round.

1. $S = \emptyset$, $E_0 = A$.
2. For $j = 1$ to k, do:
   (a) Pick row i with probability proportional to $P_i^{(j-1)} \geq c\frac{\|E_{j-1}^{(i)}\|^2}{\|E_{j-1}\|_F^2}$.
   (b) Add this new row to subset S.
   (c) $E_j = A - \pi_{\mathrm{span}(S)}(A)$.

Next we show that the above procedure gives an approximate implementation of volume sampling.

Proposition 7. *Suppose the k-round adaptive procedure mentioned above picks a subset S with probability $\tilde{P}_S$. Then,*

$$\tilde{P}_S \leq k! \; P_S$$

Proof. Let $S = \{A^{i_1}, A^{i_2}, \ldots, A^{i_k}\}$ be a subset of k rows, and let $\tau \in \Pi_k$, the set of all permutations of $\{i_1, i_2, \ldots, i_k\}$. By $H_{\tau,t}$ we denote the linear subspace $\text{span}(A^{\tau(i_1)}, A^{\tau(i_2)}, \ldots, A^{\tau(i_t)})$, and by $d(A^i, H_{\tau,t})$ we denote the orthogonal distance of A^i from this subspace. Our adaptive procedure picks a subset S with probability equal to

$$\begin{aligned}
\tilde{P}_S &= \sum_{\tau \in \Pi_k} \frac{\|A^{\tau(i_1)}\|^2}{\|A\|_F^2} \frac{d(A^{\tau(i_2)}, H_{\tau,1})^2}{\sum_{i=1}^m d(A^i, H_{\tau,1})^2} \cdots \frac{d(A^{\tau(i_k)}, H_{\tau,k-1})^2}{\sum_{i=1}^m d(A^i, H_{\tau,k-1})^2} \\
&\leq \frac{\sum_{\tau \in \Pi_k} \|A^{\tau(i_1)}\|^2 \; d(A^{\tau(i_2)}, H_{\tau,1})^2 \; \cdots \; d(A^{\tau(i_k)}, H_{\tau,k-1})^2}{\|A\|_F^2 \; \|A - A_1\|_F^2 \; \cdots \; \|A - A_{k-1}\|_F^2} \\
&= \frac{\sum_{\tau \in \Pi_k} (k!)^2 \text{vol}(\Delta(S))^2}{\|A\|_F^2 \; \|A - A_1\|_F^2 \; \cdots \; \|A - A_{k-1}\|_F^2} \\
&= \frac{(k!)^3 \; \text{vol}(\Delta(S))^2}{\sum_{i=1}^m \sigma_i^2 \; \sum_{i=2}^m \sigma_i^2 \; \cdots \; \sum_{i=k}^m \sigma_i^2} \\
&\leq \frac{(k!)^3 \; \text{vol}(\Delta(S))^2}{\sum_{1 \leq i_1 < i_2 < \ldots < i_k \leq m} \sigma_{i_1}^2 \sigma_{i_2}^2 \cdots \sigma_{i_k}^2} \\
&= \frac{k! \; \text{vol}(\Delta(S))^2}{\sum_{T:|T|=k} \text{vol}(\Delta(T))^2} \qquad \text{(using Lemma 6)} \\
&= k! \; P_S
\end{aligned}$$

□

Now we will show why it suffices to have just the approximate implementation of volume sampling. If we sample subsets S with probabilities $\tilde{P}_S$ instead of P_S, we get an analog of Theorem 5 with a weaker multiplicative approximation.

Proposition 8. *If we sample a subset S of k rows using the k-round adaptive sampling procedure mentioned above, then*

$$\mathsf{E}_S \left[\|A - \pi_S(A)\|_F^2 \right] \leq (k+1)! \; \|A - A_k\|_F^2.$$

Proof. Since we are picking a subset S with probability $\tilde{P}_S$ the expected error is

$$\begin{aligned}
\mathsf{E}_S \left[\|A - \pi_{\text{span}(S)}(A)\|_F^2 \right] &= \sum_{S:|S|=k} \tilde{P}_S \|A - \pi_{\text{span}(S)}(A)\|_F^2 \\
&\leq k! \sum_{S:|S|=k} P_S \|A - \pi_{\text{span}(S)}(A)\|_F^2 \\
&\leq k! \; (k+1) \|A - A_k\|_F^2 \qquad \text{(using Theorem 5)} \\
&= (k+1)! \; \|A - A_k\|_F^2
\end{aligned}$$

□

3 Low-rank approximation with multiplicative error

In this section, we combine adaptive sampling and volume sampling to prove the existence of a small witness and then to derive an efficient algorithm.

3.1 Existence

We now prove Theorem 2.

Proof. From Theorem 5, we know that there exists a subset S_0 of k rows of A such that

$$\|A - \pi_{\text{span}(S_0)}(A)\|_F^2 \leq (k+1)\|A - A_k\|_F^2.$$

Let $V = \text{span}(S_0),\ t = \log(k+1),\ c = 1$ in Corollary 4, we know that there exist subsets $S_1, \ldots, S_t$ of rows with sizes $s_1 = \ldots = s_{t-1} = 2k$ and $s_t = 4k/\epsilon$, respectively, such that

$$\begin{aligned}
\|A - \pi_{\text{span}(V \cup S_1 \cup \ldots \cup S_t),k}(A)\|_F^2 &\leq \left(1 + \frac{\epsilon}{4} + \frac{\epsilon}{8} + \ldots\right) \|A - A_k\|_F^2 + \frac{\epsilon}{2^{t+1}}\ \|E_0\|_F^2 \\
&\leq (1 + \frac{\epsilon}{2})\ \|A - A_k\|_F^2 + \frac{\epsilon}{2^{t+1}}\ \|A - \pi_V(A)\|_F^2 \\
&\leq (1 + \frac{\epsilon}{2})\ \|A - A_k\|_F^2 + \frac{\epsilon}{2^{t+1}}\ (k+1)\|A - A_k\|_F^2 \\
&= (1 + \frac{\epsilon}{2})\ \|A - A_k\|_F^2 + \frac{\epsilon}{2}\ \|A - A_k\|_F^2 \\
&= (1 + \epsilon)\|A - A_k\|_F^2.
\end{aligned}$$

Therefore, for $S = S_0 \cup S_1 \cup \ldots \cup S_t$ we have

$$|S| \leq \sum_{j=0}^{t} |S_j| = k + 2k(\log(k+1) - 1) + \frac{4k}{\epsilon} \leq \frac{4k}{\epsilon} + 2k\log(k+1)$$

and

$$\|A - \pi_{\text{span}(S'),k}(A)\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2.$$

□

3.2 Efficient algorithm

In this section we describe an algorithm that given a matrix $A \in \mathbb{R}^{m \times n}$, finds another matrix $\tilde{A}_k$ of rank at most k such that $\|A - \tilde{A}_k\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2$. The algorithm has two phases. In the first phase, we pick a subset of k rows using the approximate volume sampling procedure described in Subsection 2.3. In the second phase, we use the span of these k rows as our initial subspace and perform $(k+1)\log(k+1)$ rounds of adaptive sampling. The rows chosen are all from the original matrix A.

Linear Time Low-Rank Matrix Approximation

Input: $A \in \mathbb{R}^{m \times n}$, integer $k \leq m$, error parameter $\epsilon > 0$.
Output: $\tilde{A}_k \in \mathbb{R}^{m \times n}$ of rank at most k.

1. Pick a subset S_0 of k rows of A using the approximate volume sampling procedure described in Subsection 2.3. Compute an orthonormal basis $\mathcal{B}_0$ of $\mathrm{span}(S_0)$.
2. Initialize $V = \mathrm{span}(S_0)$. Fix parameters as $t = (k+1)\log(k+1)$, $s_1 = s_2 = \ldots = s_{t-1} = 2k$, and $s_t = 16k/\epsilon$.
3. Pick subsets of rows $S_1, S_2, \ldots, S_t$, using t-round adaptive sampling procedure described in Subsection 2.1. After round j, extend the previous orthonormal basis $\mathcal{B}_{j-1}$ to an orthonormal basis $\mathcal{B}_j$ of $\mathrm{span}(S_0 \cup S_1 \cup \ldots \cup S_j)$.
4. $S = \bigcup_{j=0}^{t} S_j$, and we have an orthonormal basis $\mathcal{B}_t$ of $\mathrm{span}(S)$.
5. Compute $h_1, h_2, \ldots, h_k$, the top k right singular vectors of $\pi_{\mathrm{span}(S)}(A)$.
6. Output matrix $\tilde{A}_k = \pi_{\mathrm{span}(h_1,\ldots,h_k)}(A)$, written in the standard basis.

Here are some details about the implementations of these steps.

In Step 1, we use the k-round adaptive procedure for approximate volume sampling. In the j-th round of this procedure, we sample a row and compute its component v_j orthogonal to the span of the rows picked in rounds $1, 2, \ldots, j-1$. The residual squared lengths of the rows are computed using $\|E_j^{(i)}\|^2 = \|E_{j-1}^{(i)}\|^2 - A^{(i)} \cdot v_j$, and $\|E_j\|_F^2 = \|E_{j-1}\|_F^2 - \|Av_j\|^2$. In the end, we have an orthonormal basis $\mathcal{B}_0 = \{v_1/\|v_1\|, \ldots, v_k/\|v_k\|\}$.

In Step 3, there are $(k+1)\log(k+1)$ rounds of adaptive sampling. In the j-th round, we extend the orthonormal basis from $\mathcal{B}_{j-1}$ to $\mathcal{B}_j$ by Gram-Schmidt orthonormalization. We compute the residual squared lengths of the rows $\|E_j^{(i)}\|^2$, as well as the total, $\|E_j\|_F^2$, by subtracting the contribution $\pi_{\mathrm{span}(B_j \setminus B_{j-1})}(A)$ from the values that they had during the previous round.

Each round in Steps 1 and 3 can be implemented using 2 passes over the matrix: one pass to figure out the sampling distribution, and an another one to sample a row (or a subset of rows) according to this distribution. So Steps 1 and 3 require $2(k+1)\log(k+1) + 2k$ passes.

Finally, in Step 5, we compute $\pi_{\mathrm{span}(S)}(A)$ in terms of basis $\mathcal{B}_t$ using one pass (now we have an $m \times O(k/\epsilon + k^2 \log k)$ matrix), and we compute its top k right singular vectors using SVD. In Step 6, we rewrite them in the standard basis and project matrix A onto their span, which requires one additional pass.

So the total number of passes is $2(k+1)(\log(k+1)+1)$.

Theorem 9. *With probability at least* $3/4$*, the algorithm outputs a matrix* $\tilde{A}_k$ *such that*

$$\|A - \tilde{A}_k\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2.$$

Moreover, the algorithm takes

$$O\left(M\left(\frac{k}{\epsilon} + k^2 \log k\right) + (m+n)\left(\frac{k^2}{\epsilon^2} + \frac{k^3 \log k}{\epsilon} + k^4 \log^2 k\right)\right)$$

time and $O\left(n(\frac{k}{\epsilon} + k^2 \log k)\right)$ *space.*

Proof. We begin with a proof of correctness. After the first phase of approximate volume sampling, using Proposition 8, we have

$$\mathsf{E}_{S_0}\left[\|A - \pi_{\mathrm{span}(S_0)}(A)\|_F^2\right] \le (k+1)!\ \|A - A_k\|_F^2.$$

Now using $V = \mathrm{span}(S_0)$, $c = 1$, $t = (k+1)\log(k+1)$, $s_t = 16k/\epsilon$, $s_{t-1} = \ldots = s_1 = 2k$ in Theorem 4 we get that

$$\begin{aligned}\mathsf{E}_{S_1,\ldots,S_t}\left[\|A - \pi_{\mathrm{span}(S),k}(A)\|_F^2\right] &\le \left(1 + \frac{\epsilon}{16} + \frac{\epsilon}{32} + \ldots\right)\|A - A_k\|_F^2 + \frac{\epsilon}{2^{t+3}}\ \|A - \pi_{\mathrm{span}(S_0)}(A)\|_F^2 \\ &\le (1 + \frac{\epsilon}{8})\ \|A - A_k\|_F^2 + \frac{\epsilon}{8 \cdot 2^t}\ \|A - \pi_{\mathrm{span}(S_0)}(A)\|_F^2.\end{aligned}$$

Now taking expectation over S_0 we have

$$\begin{aligned}\mathsf{E}_{S_0,\ldots,S_t}\left[\|A - \pi_{\mathrm{span}(S),k}(A)\|_F^2\right] &\le (1 + \frac{\epsilon}{8})\ \|A - A_k\|_F^2 + \frac{\epsilon}{8 \cdot 2^t}\ \mathsf{E}_{S_0}\|A - \pi_{\mathrm{span}(S_0)}(A)\|_F^2 \\ &\le (1 + \frac{\epsilon}{8})\ \|A - A_k\|_F^2 + \frac{\epsilon}{8 \cdot 2^t}\ (k+1)!\ \|A - A_k\|_F^2 \\ &\le (1 + \frac{\epsilon}{8})\ \|A - A_k\|_F^2 + \frac{\epsilon}{8 \cdot 2^t}\ (k+1)^{(k+1)}\ \|A - A_k\|_F^2 \\ &\le (1 + \frac{\epsilon}{8})\ \|A - A_k\|_F^2 + \frac{\epsilon}{8}\ \|A - A_k\|_F^2 \\ &= (1 + \frac{\epsilon}{4})\|A - A_k\|_F^2.\end{aligned}$$

This means

$$\mathsf{E}_{S_0,\ldots,S_t}\left[\|A - \pi_{\mathrm{span}(S),k}(A)\|_F^2 - \|A - A_k\|_F^2\right] \le \frac{\epsilon}{4}\|A - A_k\|_F^2.$$

Therefore, using Markov's inequality, with probability at least 3/4 the algorithm gives a matrix $\tilde{A}_k = \pi_{\mathrm{span}(S),k}(A)$ satisfying

$$\|A - \tilde{A}_k\|_F^2 \le (1+\epsilon)\|A - A_k\|_F^2.$$

Now let us analyze its complexity.

Step 1 has k rounds of adaptive sampling. In each round, the matrix-vector multiplication requires $O(M)$ time and storing vector v_j requires $O(n)$ space. So overall, Step 1 takes $O(Mk+nk)$ time, $O(nk)$ space.

Step 3 has $2(k+1)\log(k+1)$ rounds of adaptive sampling. The j-th round (except for the last round), involves Gram-Schmidt orthonormalization of $2k$ vectors in $\mathbb{R}^n$ against an orthonormal basis of size at most $(2j+1)k$, which takes time $O(njk^2)$. Computing $\pi_{\mathrm{span}(B_j \setminus B_{j-1})}(A)$ for updating the values $\|E_j^{(i)}\|^2$ and $\|E_j\|_F^2$ takes time $O(Mk)$. Thus the total time for j-th round is $O(Mk+njk^2)$. In the last round, we pick $O(k/\epsilon)$ rows. The Gram-Schmidt orthonormalization of these $O(k/\epsilon)$ vectors against an orthonormal basis of $O(k^2 \log k)$ vectors takes $O(nk^3 \log k/\epsilon)$ time; storing this basis requires $O(nk/\epsilon + nk^2 \log k)$ space. So overall, Step 3 takes $O\left(Mk^2 \log k + n(k^3 \log k/\epsilon + k^4 \log^2 k)\right)$ time and $O(nk/\epsilon + nk^2 \log k)$ space (to store the basis $\mathcal{B}_t$).

In Step 5, projecting A onto $\mathrm{span}(S)$ takes $O\left(M(k/\epsilon + k^2 \log k)\right)$ time. Now we have $\pi_{\mathrm{span}(S)}(A)$ in terms of our basis $\mathcal{B}_t$ (which is a $m \times O(k^2 \log k + k/\epsilon)$ matrix) and computing its top k right singular vectors takes time $O\left(m(k/\epsilon + k^2 \log k)^2\right)$.

In Step 6, rewriting $h_1, h_2, \ldots, h_k$ in the standard basis takes time $O\left(n(k^3 \log k + k^2/\epsilon)\right)$. And finally, projecting the matrix A onto $\mathrm{span}(h_1, \ldots, h_k)$ takes time $O(Mk)$.

Putting it all together, the algorithm takes

$$O\left(M\left(\frac{k}{\epsilon} + k^2 \log k\right) + (m+n)\left(\frac{k^2}{\epsilon^2} + \frac{k^3 \log k}{\epsilon} + k^4 \log^2 k\right)\right)$$

time and $O\left(n(k/\epsilon + k^2 \log k)\right)$ space, and $O(k \log k)$ passes over the data. □

This algorithm can be made to work with high probability, by running independent copies of the algorithm in each pass and taking the best answer found at the end. The overhead to get a probability of success of $1-\delta$ is $O(\sqrt{\log(1/\delta)})$.

4 Lower-bound for relative low-rank matrix approximation

Here we show a lower bound of $\Omega(k/\epsilon)$ for rank-k approximation using a subset of rows.

Proposition 10. *Given $\epsilon > 0$ and n large enough so that $n\epsilon \geq 2$, there exists an $n \times (n+1)$ matrix A such that for any subset S of its rows with $|S| \leq 1/2\epsilon$,*

$$\|A - \pi_{\mathrm{span}(S),1}(A)\|_F^2 \geq (1+\epsilon)\|A - A_1\|_F^2$$

Proof. Let $e_1, e_2, \ldots, e_{n+1}$ be the standard basis for $\mathbb{R}^{n+1}$, considered as rows. Consider the $n \times (n+1)$ matrix A, whose i-th row is given by $A^{(i)} = e_1 + \epsilon\, e_{i+1}$, for $i = 1, 2, \ldots, n$. The best rank-1 approximation for this is A_1, whose i-th row is given by$A_1^{(i)} = e_1 + \sum_{i=1}^{n} \frac{1}{n} e_{i+1}$. Therefore,

$$\|A - A_1\|_F^2 = \sum_{i=1}^{n} \|A^{(i)} - A_1^{(i)}\|^2 = n\left(\frac{(n-1)^2\epsilon^2}{n^2} + (n-1)\frac{\epsilon^2}{n^2}\right) = (n-1)\epsilon^2.$$

Now let S be any subset of the rows with $|S| = s$. It is easy to see that the best rank-1 approximation for A in the span of S is given by $\pi_{\mathrm{span}(S),1}(A)$, whose i-th row is given by $\pi_{\mathrm{span}(S),1}(A)^{(i)} = e_1 + \frac{\epsilon}{s} \sum_{i \in S} e_{i+1}$, for all i (because it has to be a symmetric linear combination of them). Hence,

$$\begin{aligned}
\|A - \pi_{\mathrm{span}(S),1}(A)\|_F^2 &= \sum_{i \in S} \|A^{(i)} - \pi_{\mathrm{span}(S),1}(A)^{(i)}\|^2 + \sum_{i \notin S} \|A^{(i)} - \pi_{\mathrm{span}(S),1}(A)^{(i)}\|^2 \\
&= s\left(\frac{(s-1)^2\epsilon^2}{s^2} + (s-1)\frac{\epsilon^2}{s^2}\right) + (n-s)\left(s\,\frac{\epsilon^2}{s^2} + \epsilon^2\right) \\
&= \frac{(s-1)^2\epsilon^2}{s} + \frac{(s-1)\epsilon^2}{s} + \frac{n\epsilon^2}{s} + n\epsilon^2 - \epsilon^2 - s\epsilon^2 \\
&= \frac{n\epsilon^2}{s} + n\epsilon^2 - 2\epsilon^2.
\end{aligned}$$

Now if $s \leq \frac{1}{2\epsilon}$ then $\|A - \pi_{\mathrm{span}(S),1}(A)\|_F^2 = (1+2\epsilon)n\epsilon^2 - 2\epsilon^2 \geq (1+\epsilon)n\epsilon^2 \geq (1+\epsilon)\|A - A_1\|_F^2$, for n chosen large enough so that $n\epsilon \geq 2$. □

Now we will try to extend this lower bound for relative rank-k approximation.

Proposition 11. *Given $\epsilon > 0$, k, and n large enough so that $n\epsilon \geq 2k$, there exists a $kn \times k(n+1)$ matrix B such that for any subset S of its rows with $|S| \leq k/2\epsilon$,*

$$\|B - \pi_{\text{span}(S),k}(A)\|_F^2 \geq (1+\epsilon)\|B - B_k\|_F^2.$$

Proof. Consider B to be a $kn \times k(n+1)$ block-diagonal matrix with k blocks, where each of the blocks is equal to A defined as in Proposition 10 above. It is easy to see that

$$\|B - B_k\|_F^2 = k\|A - A_1\|_F^2.$$

Now pick any subset S of rows with $|S| \leq \frac{k}{2\epsilon}$. Let S_i be the subset of rows taken from the i-th block, and let $|S_i| = \frac{k}{2\epsilon_i}$. We know that $\sum_{i=1}^k |S_i| = \sum_{i=1}^k \frac{k}{2\epsilon_i} \leq \frac{k}{2\epsilon}$, and hence $n\epsilon_i \geq n\epsilon \geq 2$.

Therefore,

$$\begin{aligned}
\|B - \pi_{\text{span}(S),k}(B)\|_F^2 &= \sum_{i=1}^k \|A - \pi_{\text{span}(S_i),1}(A)\|_F^2 \\
&\geq \sum_{i=1}^k (1 + \frac{\epsilon_i}{k})\|A - A_1\|_F^2 && \text{(using Proposition 10)} \\
&= (k + \frac{\sum_{i=1}^k \epsilon_i}{k})\|A - A_1\|_F^2 \\
&\geq (k + \frac{k}{\sum_{i=1}^k 1/\epsilon_i})\|A - A_1\|_F^2 && \text{(by A.M.-H.M. inequality)} \\
&\geq (k + k\epsilon)\|A - A_1\|_F^2 \\
&= k(1+\epsilon)\|A - A_1\|_F^2 \\
&= (1+\epsilon)\|B - B_k\|_F^2.
\end{aligned}$$

□

5 Discussion

Our algorithm requires $O(k \log k)$ passes in the streaming model. Can we prove a lower bound of $\Omega(k)$ passes for any algorithm that computes a multiplicative rank-k approximation? Can *exact* volume sampling be implemented efficiently?

It would also be nice to close the gap between the upper bound $O(k/\epsilon + k \log k)$ and the lower bound $\Omega(k/\epsilon)$ on the number of rows whose span "contains" a $(1+\epsilon)$-approximation of rank at most k.

Acknowledgements. We would like to thank Sariel Har-Peled, Prahladh Harsha, Frank McSherry, Luis Rademacher and Grant Wang.

References

[1] D. Achlioptas, F. McSherry, "Fast Computation of Low Rank Approximations." Proceedings of the 33rd Annual Symposium on Theory of Computing, 2001.

[2] C. Aggarwal, C. Procopiuc, J. Wolf, P. Yu, J. Park. "Fast Algorithms for Projected Clustering." Proceedings of SIGMOD, 1999.

[3] Z. Bar-Yosseff. "Sampling Lower Bounds via Information Theory." Proceedings of the 35th Annual Symposium on Theory of Computing, 2003.

[4] W.F. de la Vega, M. Karpinski, C. Kenyon, Y. Rabani. "Approximation schemes for clustering problems." Proceedings of the 35th Annual ACM Symposium on Theory of Computing, 2003.

[5] P. Drineas, A. Frieze, R. Kannan, S. Vempala, V. Vinay. "Clustering in large graphs and matrices." Proceedings of the 10th SODA, 1999.

[6] P. Drineas, R. Kannan. "Pass Efficient Algorithm for approximating large matrices," Proceedings of 14th SODA, 2003.

[7] P. Drineas, R. Kannan, M. Maloney. "Fast Monte Carlo Algorithms for Matrices II: Computing a Low-Rank Approximation to a Matrix." Yale University Technical Report, YALEU/DCS/TR-1270, 2004.

[8] A. Deshpande, L. Rademacher, S. Vempala, G. Wang. "Matrix Approximation and Projective Clustering via Volume Sampling", Proceedings of the 17th ACM-SIAM Symposium on Discrete Algorithms (SODA), 2006.

[9] J. Feigenbaum, S. Kannan, A. McGregor, S. Suri, J. Zhang. "On Graph Problems in a Semi-Streaming Model." Proceedings of the 31st ICALP, 2004.

[10] A. Frieze, R. Kannan, S. Vempala. "Fast Monte-Carlo algorithms for finding low-rank approximations." Journal of the ACM, 51(6):1025-1041, 2004.

[11] S. Guha, N. Koudas, K. Shim. "Data-streams and histograms." Proceedings of 33rd ACM Symposium on Theory of Computing, 2001.

[12] M. Henzinger, P. Raghavan, S. Rajagopalan. "Computing on Data Streams." Technical Note 1998-011, Digital Systems Research Center, Palo Alto, CA, May 1998.

[13] J. Matoušek. "On approximate geometric k-clustering." Discrete and Computational Geometry, pg 61-84, 2000.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de/