THEORY OF COMPUTING, Volume 2 (2006), pp. 225–247
http://theoryofcomputing.org

Matrix Approximation and Projective Clustering via Volume Sampling

Amit Deshpande* Luis Rademacher† Santosh Vempala‡
Grant Wang§

Received: August 11, 2005; revised: September 28, 2006; published: October 12, 2006.

Abstract: Frieze, Kannan, and Vempala (JACM 2004) proved that a small sample of rows of a given matrix A spans the rows of a low-rank approximation D that minimizes $\|A-D\|_F$ within a small additive error, and the sampling can be done efficiently using just two passes over the matrix. In this paper, we generalize this result in two ways. First, we prove that the additive error drops exponentially by iterating the sampling in an adaptive manner (*adaptive sampling*). Using this result, we give a pass-efficient algorithm for computing a low-rank approximation with reduced additive error. Our second result is that there exist k rows of A whose span contains the rows of a multiplicative $(k+1)$-approximation to the best rank-k matrix; moreover, this subset can be found by sampling k-subsets of rows from a natural distribution (*volume sampling*). Combining *volume sampling* with *adaptive sampling* yields the existence of a set of $k+k(k+1)/\varepsilon$ rows whose span contains the rows of a multiplicative $(1+\varepsilon)$-approximation. This leads to a PTAS for the following NP-hard

*Supported by NSF Award CCR-0312339.
†Supported by NSF Award CCR-0312339.
‡Supported by NSF Award CCR-0312339 and a Guggenheim Foundation Fellowship.
§Supported by NSF Award CCR-0312339.

ACM Classification: G.1.3, F.2.1

AMS Classification: 68W25, 65F15

Key words and phrases: algorithms, matrix approximation, projective clustering.

Authors retain copyright to their work and grant Theory of Computing unlimited rights to publish the work electronically and in hard copy. Use of the work is permitted as long as the author(s) and the journal are properly acknowledged. For the detailed copyright statement, see http://theoryofcomputing.org/copyright.html.

© 2006 Amit Deshpande, Luis Rademacher, Santosh Vempala, and Grant Wang DOI: 10.4086/toc.2006.v002a012

projective clustering problem: Given a set P of points in $\mathbb{R}^d$, and integers j and k, find subspaces $F_1, \ldots, F_j$, each of dimension at most k, that minimize $\sum_{p \in P} \min_i d(p, F_i)^2$.

1 Introduction

1.1 Motivation

Let the rows of a matrix be points in a high-dimensional space. It is often of interest to find a low-dimensional representation. The subspace spanned by the top k right singular vectors of the matrix is a good choice for many applications. The problem of efficiently finding an approximation to this subspace has received much attention in the past decade [19, 15, 1, 14, 16]. In this paper, we give new algorithms for this problem and show existence of subspaces lying in the span of a small set of rows with better additive approximation as well as multiplicative approximation. At the heart of our analysis are generalizations of previous sampling schemes [19]. We apply these results to the general problem of finding j subspaces, each of dimension at most k, so as to minimize the sum of squared distances of each point to its nearest subspace, a measure of the "error" incurred by this representation; as a result, we obtain the first polynomial-time approximation scheme for this *projective clustering* problem [5, 32, 6, 3] when j and k are fixed.

The case of $j = 1$, i. e., finding a single k-dimensional subspace, is an important problem in itself and can be solved efficiently (for $j \geq 2$, the problem is NP-hard [30], even for $k = 1$ [15]). The optimal projection is given by the rank-k matrix $A_k = AYY^T$ where the columns of Y are the top k right singular vectors of A and can be computed using the Singular Value Decomposition. Note that among all rank-k matrices D, A_k is the one that minimizes $\|A - D\|_F^2 = \sum_{i,j} (A_{ij} - D_{ij})^2$. The running time of this algorithm, dominated by the SVD computation of an $m \times n$ matrix, is $O(\min\{mn^2, nm^2\})$. Although polynomial, this is still too high for some applications.

For problems on data sets that are too large to store/process in their entirety, one can view the data as a stream of data items arriving in arbitrary order, and the goal is to process a subset chosen judiciously on the fly and then extrapolate from this subset. Motivated by the question of finding a faster algorithm, Frieze et al. [19] showed that any matrix A has k/ε rows whose span contains the rows of a rank-k approximation to A within additive error $\varepsilon \|A\|_F^2$. In fact, the subset of rows can be obtained as independent samples from a distribution that depends only on the norms of the rows. (In what follows, $A^{(i)}$ denotes the ith row of A.)

Theorem 1.1 ([19]). *Let S be a sample of s rows of an $m \times n$ matrix A, each chosen independently from the following distribution: row i is picked with probability*

$$P_i = \frac{\|A^{(i)}\|^2}{\|A\|_F^2} \quad .$$

If $s \geq k/\varepsilon$, then $\mathrm{span}(S)$ *contains the rows of a matrix $\tilde{A}_k$ of rank at most k for which*

$$\mathsf{E}_S(\|A - \tilde{A}_k\|_F^2) \leq \|A - A_k\|_F^2 + \varepsilon \|A\|_F^2 \quad .$$

This can be turned into an efficient algorithm based on sampling [15].[1] The algorithm makes one pass through A to figure out the sampling distribution and another pass to sample and compute the approximation. Its complexity is $O(\min\{m,n\}k^2/\varepsilon^4)$. These results lead us to the following questions: (1) Can the error be reduced significantly by using multiple passes through the data? (2) Can we get multiplicative $(1+\varepsilon)$-approximations? (3) Do these sampling algorithms have any consequences for the general projective clustering problem?

1.2 Our results

We begin with the observation that the additive error term drops *exponentially* with the number of passes when the sampling is done adaptively. Thus, low-rank approximation is a natural problem for which multiple passes through the data are highly beneficial.

The idea behind the algorithm is quite simple. As an illustrative example, suppose the data consists of points along a 1-dimensional subspace of $\mathbb{R}^n$ except for one point. The best rank-2 subspace has zero error. However, one round of sampling will most likely miss the point far from the line. So we use a two-round approach. In the first pass, we get a sample from the squared norm distribution. Then we sample again, but adaptively—we sample with probability proportional to the squared distance to the span of the first sample. We call this procedure *adaptive sampling*. If the lone far-off point is missed in the first pass, it will have a high probability of being chosen in the second pass. The span of the full sample now contains the rows of a good rank-2 approximation. In the theorem below, for a set S of rows of a matrix A, we denote by $\pi_S(A)$ the matrix whose rows are the projection of the rows of A to the span of S.

Theorem 1.2. *Let $S = S_1 \cup \cdots \cup S_t$ be a random sample of rows of an $m \times n$ matrix A where, for $j = 1,\ldots,t$, each set S_j is a sample of s rows of A chosen independently from the following distribution: row i is picked with probability*

$$P_i^{(j)} = \frac{\|E_j^{(i)}\|^2}{\|E_j\|_F^2}$$

where $E_1 = A$, $E_j = A - \pi_{S_1\cup\cdots\cup S_{j-1}}(A)$. Then for $s \geq k/\varepsilon$, $\operatorname{span}(S)$ contains the rows of a matrix $\tilde{A}_k$ of rank at most k such that

$$\mathsf{E}_S(\|A-\tilde{A}_k\|_F^2) \leq \frac{1}{1-\varepsilon}\|A-A_k\|_F^2 + \varepsilon^t\|A\|_F^2 \ .$$

The proof of Theorem 1.2 is given in Section 2.1. The resulting algorithm, described in Section 2.2 uses $2t$ passes through the data. Although the sampling distribution is modified t times, the matrix itself is not changed. This is especially significant when A is sparse as the sparsity of the matrix is maintained.

Theorem 1.2 raises the question of whether we can get a multiplicative approximation instead of an additive approximation. To answer this question, we generalize the sampling approach. For a set S of k points in $\mathbb{R}^n$, let $\Delta(S)$ denote the k-dimensional simplex formed by them along with the origin. We pick a random k-subset S of rows of A with probability proportional to $\operatorname{vol}(\Delta(S))^2$. This procedure, which

[1] Frieze et al. [19] go further to show that there is an $s \times s$ submatrix for $s = \operatorname{poly}(k/\varepsilon)$ from which the low-rank approximation can be computed in $\operatorname{poly}(k, 1/\varepsilon)$ time in an implicit form.

we call *volume sampling*, is a generalization of the earlier sampling approach which picks single rows according to their squared norms.

Theorem 1.3. *Let S be a random k-subset of rows of a given matrix A chosen with probability*

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

Then $\tilde{A}_k$, the projection of A to the span of S, satisfies

$$\mathsf{E}_S(\|A-\tilde{A}_k\|_F^2) \leq (k+1)\|A-A_k\|_F^2 \ .$$

We prove this theorem in Section 1.4. Moreover, the factor of $k+1$ is the best possible for a k-subset (Proposition 3.3). By combining Theorem 1.3 with the adaptive sampling idea from Theorem 1.2, we show that there exist $O(k^2/\varepsilon)$ rows whose span contains the rows of a multiplicative $(1+\varepsilon)$-approximation.

Theorem 1.4. *For any $m \times n$ matrix A, there exist $k+k(k+1)/\varepsilon$ rows whose span contains the rows of a rank-k matrix $\tilde{A}_k$ such that*

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

The existence of a small number of rows containing a good multiplicative approximation is the key ingredient in our last result—a polynomial-time approximation scheme (PTAS) for the general projective clustering problem. This result makes a connection between matrix approximation and projective clustering. A key idea in the matrix approximation work of [14, 15, 19] is that, for any matrix, there is a small subset of its rows whose span contains a good approximation to the row space of the entire matrix. This is similar to the idea of core-sets [2], which have been studied in computing extent measures in computational geometry (and applied to a variant of the projective clustering problem [24]). Roughly speaking, a core-set is a subset of a point-set such that computing the extent measure on the core-set provides an approximation to the extent measure on the entire point set. An extent measure is just a statistic on the point set (for example, the diameter of a point set, or the radius of the minimum enclosing cylinder); typically, it measures the size of the point set or the minimum size of an object that encloses the point set [2].

We state the projective clustering problem using the notation from computational geometry: Let $d(p,F)$ be the orthogonal distance of a point p to a subspace F. Given a set P of n points in $\mathbb{R}^d$, find j subspaces $F_1,\ldots,F_j$, each of dimension k, such that

$$\mathcal{C}(F_1,\ldots,F_j) = \sum_{p\in P} \min_i d(p,F_i)^2 \tag{1.1}$$

is minimized. When subspaces are replaced by flats, the case $k=0$ corresponds to the "j-means problem" in computational geometry.

Theorem 1.4 suggests an enumerative algorithm. The optimal set of k-dimensional subspaces induces a partition $P_1 \cup \cdots \cup P_j$ of the given point set. Each set P_i contains a subset of size $O(k^2/\varepsilon)$ in whose span lies a multiplicative $(1+\varepsilon)$-approximation to the optimal k-dimensional subspace for P_i. So

we consider all possible combinations of j subsets each of size $O(k^2/\varepsilon)$, and a δ-net of k-dimensional subspaces in the span of each subset. The δ-net depends on the points in each subset and is not just a grid, as is often the case. Each possible combination of subspaces induces a partition and we output the best of these. Since the subset size is bounded (and so is the size of the net), this gives a PTAS for the problem (see Section 4) when j and k are taken to be fixed constants.

Theorem 1.5. *Given n points in $\mathbb{R}^d$ and parameters B and ε, in time*

$$d\left(\frac{n}{\varepsilon}\right)^{O(jk^3/\varepsilon)}$$

we can find a solution to the projective clustering problem which is of cost at most $(1+\varepsilon)B$ provided there is a solution of cost B.

1.3 Related work

The work of Frieze et al. [19] and Drineas et al. [15] introduced matrix sampling for fast low-rank approximation. Subsequently, an alternative sampling-based algorithm was given by Achlioptas and McSherry [1]. That algorithm achieves somewhat different bounds (see [1] for a detailed comparison) using only one pass. It does not seem amenable to the multipass improvements presented here. Bar-Yossef [9] has shown that the bounds of these algorithms for one or two passes are optimal up to polynomial factors in $1/\varepsilon$.

These algorithms can also be viewed in the *streaming* model of computation [25]. In this model, we do not have random access to data; the data comes as a stream of data items in arbitrary order and we are allowed one or a few sequential passes over the data. Algorithms for the streaming model have been designed for computing frequency moments [7], histograms [22], etc. and have mainly focused on what can be done in one pass. There has been some recent work on what can be done in multiple passes [14, 18]. The "pass-efficient" model of computation was introduced in [25]. Our multipass algorithms fit this model and relate the quality of approximation to the number of passes. Feigenbaum et al. [18] show such a relationship for computing the maximum unweighted matching in bipartite graphs.

The Lanczos method is an iterative algorithm that is used in practice to compute the Singular Value Decomposition [20, 26]. An exponential decrease in an additive error term has also been proven for the Lanczos method under a different notion of additive error ([20, 26]). However, the exponential decrease in error depends on the gap between singular values. In particular, the following is known for the Lanczos method: after k iterations, each approximate singular value θ_i^2 obeys:

$$\theta_i^2 \geq \sigma_i^2 - c^k C \tag{1.2}$$

where σ_i is the ith singular value. Both constants c and C depend on the gap between singular values; in particular, c is proportional to $(\sigma_2^2 - \sigma_r^2)/(2\sigma_1^2 - \sigma_2^2 - \sigma_r^2)$ and $C = \sigma_1^2 - \sigma_r^2$. The guarantee (1.2) can be transformed into an inequality:

$$\|A - \tilde{A}_k\|_F^2 \leq \|A - A_k\|_F^2 + c^k C \tag{1.3}$$

very similar to Theorem 1.2, but without the multiplicative error term for $\|A - A_k\|_F^2$. However, note that when $\sigma_1^2 = \sigma_2^2$, successive rounds of the Lanczos method fail to converge to σ_i^2. In the Lanczos method,

each iteration can be implemented in one pass over A, whereas our algorithm requires two passes over A in each iteration. Kuczyński and Woźniakowski [27] prove that the Lanczos method, with a randomly chosen starting vector, outputs a vector v with $\tilde{A}_1 = \pi_{\text{span}(v)}(A)$ such that $\|\tilde{A}_1\|_F^2 \geq (1-\varepsilon)\|A_1\|_F^2$ in $\log(n)/\sqrt{\varepsilon}$ iterations. However, this does not imply a multiplicative $(1+\varepsilon)$ error to $\|A-A_1\|_F^2$.

While the idea of volume sampling appears to be new, in [21] it is proved that, using the rows and columns that correspond to the $k \times k$ submatrix of maximal volume, one can compute a rank-k approximation to the original matrix which differs in each entry by at most $(k+1)\sigma_{k+1}$ (leading to much weaker $\sqrt{mn}$ approximations for the Frobenius norm).

The results of our paper connect two previously separate fields: low-rank approximation and projective clustering. Algorithms and systems based on projective clustering have been applied to facial recognition, data-mining, and synthetic data [5, 32, 6], motivated by the observation that no single subspace performs as well as a few different subspaces. It should be noted that the advantage of a low-dimensional representation is not merely in the computational savings, but also the improved quality of retrieval. In [3], Agarwal and Mustafa consider the same problem as in this paper, and propose a variant of the j-means algorithm for it. Their paper has promising experimental results but does not provide any theoretical guarantees. There are theoretical results for special cases of projective clustering, especially the j-means problem (find j points). Drineas et al. [15] gave a 2-approximation to j-means using SVD. Subsequently, Ostrovsky and Rabani [31] gave the first randomized polynomial time approximation schemes for j-means (and also the j-median problem). Matoušek [29] and Effros and Schulman [17] both gave deterministic PTAS's for j-means. Fernandez de la Vega et al. [11] describe a randomized algorithm with a running time of $n(\log n)^{O(1)}$. Using the idea of core-sets, Har-Peled and Mazumdar [23] showed a multiplicative $(1+\varepsilon)$-approximation algorithm that runs in linear time for fixed j, ε. Kumar et al. [28] give a linear-time PTAS that uses random sampling. There is a PTAS for $k=1$ (lines) as well [4]. Other objective functions have also been studied, e. g. sum of distances (j-median when $k=0$, [31, 23]) and maximum distance (j-center when $k=0$, [10]). For general k, Har-Peled and Varadarajan [24] give a multiplicative $(1+\varepsilon)$-approximation algorithm for the maximum distance objective function. Their algorithm runs in time $dn^{O(jk^6 \log(1/\varepsilon)/\varepsilon^5)}$ and is based on core-sets (see [2] for a survey).

1.4 Previous versions of this paper

This paper is a journal version of the paper by the same set of authors presented at the 17th Annual Symposium on Discrete Algorithms (SODA 2006) [12]. This paper contains the same major results, but additionally provides more intuition and more complete proofs. A previous version [33] of this paper by a subset of three of the authors (L. Rademacher, S. Vempala, G. Wang) appeared as an MIT CSAIL technical report. That version contained a subset of the results in this paper. In particular, the notion of volume sampling and related results did not appear in the technical report.

1.5 Notation and preliminaries

Any $m \times n$ real matrix A has a singular value decomposition, that is, it can be written in the form

$$A = \sum_{i=1}^{r} \sigma_i u^{(i)} v^{(i)T}$$

where r is the rank of A and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$ are called the singular values; $\{u^{(1)},\ldots,u^{(r)}\} \in \mathbb{R}^m$, $\{v^{(1)},\ldots,v^{(r)}\} \in \mathbb{R}^n$ are sets of orthonormal vectors, called the left and right singular vectors, respectively. It follows that $A^T u^{(i)} = \sigma_i v^{(i)}$ and $Av^{(i)} = \sigma_i u^{(i)}$ for $1 \leq i \leq r$.

The Frobenius norm of a matrix $A \in \mathbb{R}^{m\times n}$ having elements (a_{ij}) is denoted $\|A\|_F$ and is given by

$$\|A\|_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2 \ .$$

It satisfies $\|A\|_F^2 = \sum_{i=1}^{r} \sigma_i^2$.

For a subspace $V \subseteq \mathbb{R}^n$, let $\pi_{V,k}(A)$ denote the best rank-k approximation (under the Frobenius norm) of A with its rows in V. Let

$$\pi_k(A) = \pi_{\mathbb{R}^n,k}(A) = \sum_{i=1}^{k} \sigma_i u^{(i)} {v^{(i)}}^T$$

be the best rank-k approximation of A. Also $\pi_V(A) = \pi_{V,n}(A)$ is the orthogonal projection of A onto V. When we say "a set (or sample) of rows of A" we mean a set of indices of rows, rather than the actual rows. For a set S of rows of A, let $\mathrm{span}(S) \subseteq \mathbb{R}^n$ be the subspace generated by those rows; we use the simplified notation $\pi_S(A)$ for $\pi_{\mathrm{span}(S)}(A)$ and $\pi_{S,k}(A)$ for $\pi_{\mathrm{span}(S),k}(A)$.

For subspaces $V,W \subseteq \mathbb{R}^n$, their sum is denoted $V+W$ and is given by

$$V+W = \{x+y \in \mathbb{R}^n \ : \ x \in V, y \in W\} \ .$$

The following elementary properties of the operator π_V will be used:

- π_V is linear, that is, $\pi_V(\lambda A+B) = \lambda\,\pi_V(A) + \pi_V(B)$ for any $\lambda \in \mathbb{R}$ and matrices $A,B \in \mathbb{R}^{m\times n}$.
- If $V,W \in \mathbb{R}^n$ are orthogonal linear subspaces, then $\pi_{V+W}(A) = \pi_V(A) + \pi_W(A)$, for any matrix $A \in \mathbb{R}^{m\times n}$.

For a random vector v, its expectation, denoted $\mathsf{E}(v)$, is the vector having as components the expected values of the components of v.

2 Improved approximation via adaptive sampling

2.1 Proof that adaptive sampling works

We will prove Theorem 1.2 in this section. It will be convenient to formulate an intermediate theorem as follows, whose proof is quite similar to one in [19].

Theorem 2.1. *Let $A \in \mathbb{R}^{m\times n}$, and $V \subseteq \mathbb{R}^n$ be a vector subspace. Let $E = A - \pi_V(A)$ and let S be a random sample of s rows of A from a distribution D such that row i is chosen with probability*

$$P_i = \frac{\|E^{(i)}\|^2}{\|E\|_F^2} \ . \tag{2.1}$$

Then, for any nonnegative integer k,

$$\mathsf{E}_S(\|A - \pi_{V+\mathrm{span}(S),k}(A)\|_F^2) \leq \|A-\pi_k(A)\|_F^2 + \frac{k}{s}\|E\|_F^2 \ .$$

Proof. We define vectors $w^{(1)},\ldots,w^{(k)} \in V + \mathrm{span}(S)$ such that $W = \mathrm{span}\{w^{(1)},\ldots,w^{(k)}\}$ and show that W is a good approximation to $\mathrm{span}\{v^{(1)},\ldots,v^{(k)}\}$ in the sense that

$$\mathsf{E}_S(\|A-\pi_W(A)\|_F^2) \leq \|A-\pi_k(A)\|_F^2 + \frac{k}{s}\|E\|_F^2 \quad . \tag{2.2}$$

Recall that $\pi_k(A) = \pi_{\mathrm{span}\{v^{(1)},\ldots,v^{(k)}\}}(A)$, i. e., $\mathrm{span}\{v^{(1)},\ldots,v^{(k)}\}$ is the optimal subspace upon which to project. Proving (2.2) proves the theorem, since $W \subseteq V + \mathrm{span}(S)$.

To this end, define $X_l^{(j)}$ to be a random variable such that for $i = 1,\ldots,m$ and $l = 1,\ldots,s$,

$$X_l^{(j)} = \frac{u_i^{(j)}}{P_i}E^{(i)} = \frac{u_i^{(j)}}{P_i}\left(A^{(i)} - \pi_V(A^{(i)})\right) \text{ with probability } P_i \quad .$$

Note that $X_l^{(i)}$ is a linear function of a row of A sampled from the distribution D. Let $X^{(j)} = \frac{1}{s}\sum_{l=1}^s X_l^{(j)}$, and note that $\mathsf{E}_S(X^{(j)}) = E^T u^{(j)}$.

For $1 \leq j \leq k$, define:

$$w^{(j)} = \pi_V(A)^T u^{(j)} + X^{(j)} \quad . \tag{2.3}$$

Then we have that $\mathsf{E}_S(w^{(j)}) = \sigma_j v^{(j)}$. We seek to bound the second central moment of $w^{(j)}$, i. e., $\mathsf{E}_S(\|w^{(j)} - \sigma_j v^{(j)}\|^2)$. We have that

$$w^{(j)} - \sigma_j v^{(j)} = X^{(j)} - E^T u^{(j)} \quad ,$$

which gives us

$$\begin{aligned}
\mathsf{E}_S(\|w^{(j)} - \sigma_j v^{(j)}\|^2) &= \mathsf{E}_S(\|X^{(j)} - E^T u^{(j)}\|^2) \\
&= \mathsf{E}_S(\|X^{(j)}\|^2) - 2\,\mathsf{E}_S(X^{(j)}) \cdot E^T u^{(j)} + \|E^T u^{(j)}\|^2 \\
&= \mathsf{E}_S(\|X^{(j)}\|^2) - \|E^T u^{(j)}\|^2 \quad .
\end{aligned} \tag{2.4}$$

We evaluate the first term in (2.4),

$$\begin{aligned}
\mathsf{E}_S(\|X^{(j)}\|^2) &= \mathsf{E}_S\left(\left\|\frac{1}{s}\sum_{l=1}^s X_l^{(j)}\right\|^2\right) \\
&= \frac{1}{s^2}\sum_{l=1}^s \mathsf{E}_S(\|X_l^{(j)}\|^2) + \frac{2}{s^2}\sum_{1\leq l_1<l_2\leq s} \mathsf{E}_S(X_{l_1}^{(j)} \cdot X_{l_2}^{(j)}) \\
&= \frac{1}{s^2}\sum_{l=1}^s \mathsf{E}_S(\|X_l^{(j)}\|^2) + \frac{s-1}{s}\|E^T u^{(j)}\|^2 \quad .
\end{aligned} \tag{2.5}$$

In (2.5) we used that $X_{l_1}^{(j)}$ and $X_{l_2}^{(j)}$ are independent. From (2.4) and (2.5) we have that

$$\mathsf{E}_S(\|w^{(j)} - \sigma_j v^{(j)}\|^2) = \frac{1}{s^2}\sum_{l=1}^s \mathsf{E}_S(\|X_l^{(j)}\|^2) - \frac{1}{s}\|E^T u^{(j)}\|^2 \quad . \tag{2.6}$$

The definition of P_i gives us

$$\mathsf{E}_S(\|X_l^{(j)}\|^2) = \sum_{i=1}^{m} P_i \frac{\|u_i^{(j)} E^{(i)}\|}{P_i^2} \leq \|E\|_F^2 \ . \tag{2.7}$$

Thus, we have obtained a bound on the second central moment of $w^{(j)}$:

$$\mathsf{E}_S(\|w^{(j)} - \sigma_j v^{(j)}\|^2) \leq \frac{1}{s}\|E\|_F^2 \ . \tag{2.8}$$

With this bound in hand, we can complete the proof. Let $y^{(j)} = w^{(j)}/\sigma_j$ for $1 \leq j \leq k$, and consider the matrix $F = A\sum_{i=1}^{k} v^{(i)} y^{(i)^T}$. The row space of F is contained in $W = \mathrm{span}\{w^{(1)}, \ldots, w^{(k)}\}$. Therefore, $\|A - \pi_W(A)\|_F^2 \leq \|A - F\|_F^2$. We will use F to bound the error $\|A - \pi_W(A)\|_F^2$.

By decomposing $A - F$ along the left singular vectors $u^{(1)}, \ldots, u^{(r)}$, we can use the inequality (2.8) to bound $\|A - F\|_F^2$:

$$\begin{aligned}
\mathsf{E}_S(\|A - \pi_W(A)\|_F^2) \leq \mathsf{E}_S(\|A - F\|_F^2) &= \sum_{i=1}^{r} \mathsf{E}_S(\|(A-F)^T u^{(i)}\|_F^2) \\
&= \sum_{i=1}^{k} \mathsf{E}_S(\|\sigma_i v^{(i)} - w^{(i)}\|^2) + \sum_{i=k+1}^{r} \sigma_i^2 \\
&\leq \frac{k}{s}\|E\|_F^2 + \|A - \pi_k(A)\|_F^2 \ . \qquad (2.9)
\end{aligned}$$

□

We can now prove Theorem 1.2 inductively using Theorem 2.1.

Proof of Theorem 1.2. We will prove the inequality by induction on t. Theorem 1.1 gives us the base case $t = 1$.

For the inductive step, let $E = A - \pi_{S_1 \cup \cdots \cup S_{t-1}}(A)$. By means of Theorem 2.1 with $s \geq k/\varepsilon$ we have that

$$\mathsf{E}_{S_t}(\|A - \pi_{S_1 \cup \cdots \cup S_t, k}(A)\|_F^2) \leq \|A - \pi_k(A)\|_F^2 + \varepsilon\|E\|_F^2 \ .$$

Combining this inequality with the fact that $\|E\|_F^2 \leq \|A - \pi_{S_1 \cup \cdots \cup S_{t-1}, k}(A)\|_F^2$ we get

$$\mathsf{E}_{S_t}(\|A - \pi_{S_1 \cup \cdots \cup S_t, k}(A)\|_F^2) \leq \|A - \pi_k(A)\|_F^2 + \varepsilon\|A - \pi_{S_1 \cup \cdots \cup S_{t-1}, k}(A)\|_F^2 \ . \tag{2.10}$$

Taking the expectation over $S_1, \ldots, S_{t-1}$, and using the induction hypothesis for $t-1$ gives the result:

$$\begin{aligned}
\mathsf{E}_S(\|A - \pi_{S_1 \cup \cdots \cup S_t, k}(A)\|_F^2) &\leq \|A - \pi_k(A)\|_F^2 + \varepsilon\, \mathsf{E}_{S_1, \ldots, S_{t-1}}\left(\|A - \pi_{S_1 \cup \cdots \cup S_{t-1}, k}(A)\|_F^2\right) &(2.11)\\
&\leq \|A - \pi_k(A)\|_F^2 + \varepsilon\left(\frac{1}{1-\varepsilon}\|A - \pi_k(A)\|_F^2 + \varepsilon^{t-1}\|A\|_F^2\right) &(2.12)\\
&= \frac{1}{1-\varepsilon}\|A - \pi_k(A)\|_F^2 + \varepsilon^t\|A\|_F^2 \ . &(2.13)
\end{aligned}$$

□

2.2 Algorithm

In this section, we present the multipass algorithm for low-rank approximation. We first describe it at a conceptual level and then give the details of the implementation.

Informally, the algorithm will find an approximation to the best rank-k subspace (the span of $v^{(1)},\ldots,v^{(k)}$) by first choosing a sample T of s random rows with probabilities proportional to the squared norms of the rows (as in Theorem 1.1). Then we focus on the space orthogonal to the span of the chosen rows, that is, we consider the matrix $E = A - \pi_T(A)$, which represents the error of our current approximation, and we sample s additional rows with probabilities proportional to the squared norms of the rows of E. We consider the union of this sample with our previous sample, and we continue adding samples in this way, up to the number of passes that we have chosen. Theorem 1.2 gives a bound on the error of this procedure.

Fast Approximate SVD

Input: $A \in \mathbb{R}^{m\times n}$, integers $k \le m$, t, error parameter $\varepsilon > 0$.
Output: A set of k vectors in $\mathbb{R}^n$.

1. Let $S = \emptyset$, $s = k/\varepsilon$.
2. Repeat t times:
   (a) Compute the probabilities $P_i = \|E^{(i)}\|^2/\|E\|_F^2$ for $i = 1,\ldots,m$. (One pass.)
   (b) Let T be a sample of s rows of A according to the distribution that assigns probability P_i to row i. (Another pass.)
   (c) Let $S = S \cup T$.
3. Let $h_1,\ldots,h_k$ be the top k right singular vectors of $\pi_S(A)$.

In what follows, let M be the number of non-zeros of A.

Theorem 2.2. *In $2t$ passes over the data, where in each pass the entries of the matrix arrive in arbitrary order, the algorithm finds vectors $h_1,\ldots,h_k \in \mathbb{R}^n$ such that with probability at least $3/4$ their span V satisfies*

$$\|A - \pi_V(A)\|_F^2 \le \left(1 + \frac{4\varepsilon}{1-\varepsilon}\right)\|A - \pi_k(A)\|_F^2 + 4\varepsilon^t\|A\|_F^2 \quad . \tag{2.14}$$

The running time is $O\left(M\frac{kt}{\varepsilon} + (m+n)\frac{k^2t^2}{\varepsilon^2}\right)$.

Proof. For the correctness, observe that $\pi_V(A)$ is a random variable with the same distribution as $\pi_{S,k}(A)$ as defined in Theorem 1.2. Also, $\|A - \pi_{S,k}(A)\|_F^2 - \|A - \pi_k(A)\|_F^2$ is a nonnegative random variable and Theorem 1.2 gives a bound on its expectation:

$$\mathsf{E}_S(\|A - \pi_{S,k}(A)\|_F^2 - \|A - \pi_k(A)\|_F^2) \le \frac{\varepsilon}{1-\varepsilon}\|A - \pi_k(A)\|_F^2 + \varepsilon^t\|A\|_F^2 \quad . \tag{2.15}$$

Markov's inequality applied to this variable gives that with probability at least $3/4$

$$\|A-\pi_V(A)\|_F^2-\|A-\pi_k(A)\|_F^2 \leq \frac{4\varepsilon}{1-\varepsilon}\|A-\pi_k(A)\|_F^2+4\varepsilon^t\|A\|_F^2 \quad , \tag{2.16}$$

which implies inequality (2.14).

We will now bound the running time. We maintain a basis of the rows indexed by S. In each iteration, we extend this basis orthogonally with a new set Y of vectors, so that it spans the new sample T. The residual squared norm of each row, $\|E^{(i)}\|^2$, as well as the total, $\|E\|_F^2$, are computed by subtracting the contribution of $\pi_T(A)$ from the values that they had during the previous iteration. In each iteration, the projection onto Y needed for computing this contribution takes time $O(Ms)$. In iteration i, the computation of the orthonormal basis Y takes time $O(ns^2i)$ (Gram-Schmidt orthonormalization of s vectors in $\mathbb{R}^n$ with reference to an orthonormal basis of size at most $s(i+1)$). Thus, the total time in iteration i is $O(Ms+ns^2i)$; with t iterations, this is $O(Mst+ns^2t^2)$. At the end of Step 2 we have $\pi_S(A)$ in terms of our basis (an $m\times st$ matrix). Finding the top k singular vectors in Step 3 takes time $O(ms^2t^2)$. Bringing them back to the original basis takes time $O(nkst)$. Thus, the total running time is $O(Mst+ns^2t^2+ms^2t^2+nkst)$ or, in other words, $O\left(Mkt/\varepsilon+(m+n)k^2t^2/\varepsilon^2\right)$. □

3 Volume sampling and multiplicative approximation

We begin with a proof of Theorem 1.3, namely that volume sampling leads to a multiplicative $(k+1)$-approximation (in expectation).

Proof of Theorem 1.3. For every $S\subseteq\{1,\ldots,m\}$, let Δ_S be the simplex formed by $\{A^{(i)} : i\in S\}$ and the origin. We wish to bound $\mathsf{E}_S(\|A-\tilde{A}_k\|_F^2)$ which can be written as follows:

$$\mathsf{E}_S(\|A-\tilde{A}_k\|_F^2)=\frac{1}{\sum_{S,|S|=k}\mathrm{vol}_k(\Delta_S)^2}\sum_{S,|S|=k}\mathrm{vol}_k(\Delta_S)^2\|A-\pi_S(A)\|_F^2 \quad . \tag{3.1}$$

For any $(k+1)$-subset $S=\{i_1,\ldots,i_{k+1}\}$ of rows of A, we can express $\mathrm{vol}_{k+1}(\Delta_S)^2$ in terms of $\mathrm{vol}_k(\Delta_T)^2$ for $T=\{i_1,\ldots,i_k\}$, along with the squared distance from $A^{(i_{k+1})}$ to H_T, where H_T is the linear subspace spanned by $\{A^{(i)} : i\in T\}$:

$$\mathrm{vol}_{k+1}(\Delta_S)^2=\frac{1}{(k+1)^2}\mathrm{vol}_k(\Delta_T)^2 d(A^{(i_{k+1})},H_T)^2 \quad . \tag{3.2}$$

Summing over all subsets S of size $k+1$:

$$\begin{aligned}\sum_{S,|S|=k+1}\mathrm{vol}_{k+1}(\Delta_S)^2 &= \frac{1}{k+1}\sum_{T,|T|=k}\sum_{j=1}^{m}\frac{1}{(k+1)^2}\ \mathrm{vol}_k(\Delta_T)^2\ d(A^{(j)},H_T)^2\\ &= \frac{1}{(k+1)^3}\sum_{T,|T|=k}\mathrm{vol}_k(\Delta_T)^2\sum_{j=1}^{m}d(A^{(j)},H_T)^2\\ &= \frac{1}{(k+1)^3}\sum_{T,|T|=k}\mathrm{vol}_k(\Delta_T)^2\|A-\pi_T(A)\|_F^2 \quad ,\end{aligned} \tag{3.3}$$

where in the last step we noted that $\sum_{j=1}^{m} d(A^{(j)}, H_T)^2 = \|A - \pi_T(A)\|_F^2$. By substituting this equality in (3.1) and applying Lemma 3.1 (proved next) twice, we have:

$$\begin{aligned}
\mathsf{E}_S(\|A-\tilde{A}_k\|_F^2) &= \frac{1}{\sum_{T,|T|=k} \mathrm{vol}_k(\Delta_T)^2} \left((k+1)^3 \sum_{S,|S|=k+1} \mathrm{vol}_{k+1}(\Delta_S)^2 \right) \\
&= \frac{1}{\sum_{T,|T|=k} \mathrm{vol}_k(\Delta_T)^2} \left(\frac{(k+1)}{(k!)^2} \sum_{1\le t_1<\cdots<t_{k+1}\le n} \sigma_{t_1}^2 \cdots \sigma_{t_{k+1}}^2 \right) \\
&\le \frac{1}{\sum_{T,|T|=k} \mathrm{vol}_k(\Delta_T)^2} \left(\frac{(k+1)}{(k!)^2} \sum_{1\le t_1<\cdots<t_k\le r} \sigma_{t_1}^2 \cdots \sigma_{t_k}^2 \sum_{j=k+1}^{r} \sigma_j^2 \right) \qquad (3.4) \\
&= \frac{(k+1)}{\sum_{T,|T|=k} \mathrm{vol}_k(\Delta_T)^2} \left(\sum_{T:|T|=k} \mathrm{vol}_k(\Delta_T)^2 \sum_{j=k+1}^{r} \sigma_j^2 \right) \\
&= (k+1)\|A-A_k\|_F^2 \ .
\end{aligned}$$

□

Lemma 3.1.

$$\sum_{S,|S|=k} \mathrm{vol}_k(\Delta_S)^2 = \frac{1}{(k!)^2} \sum_{1\le t_1<t_2<\cdots<t_k\le r} \sigma_{t_1}^2 \sigma_{t_2}^2 \cdots \sigma_{t_k}^2$$

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

Proof. Let A_S be the sub-matrix of A formed by the rows $\{A^{(i)} : i \in S\}$. Then we know that the volume of the k-simplex formed by these rows is given by $\mathrm{vol}_k(\Delta_S) = \frac{1}{k!}\sqrt{\det(A_S A_S^T)}$. Therefore,

$$\sum_{S,|S|=k} \mathrm{vol}_k(\Delta_S)^2 = \frac{1}{(k!)^2} \sum_{S,|S|=k} \det(A_S A_S^T) = \frac{1}{(k!)^2} \sum_{\substack{B:\text{ principal} \\ k\text{-minor of } AA^T}} \det(B) \ . \qquad (3.5)$$

Let $\det(AA^T - \lambda I) = (-1)^m \lambda^m + c_{m-1}\lambda^{m-1} + \cdots + c_0$ be the characteristic polynomial of AA^T. From basic linear algebra we know that the roots of this polynomial are precisely the eigenvalues of AA^T, i. e., $\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2$ and 0 with multiplicity $(m-r)$. Moreover the coefficient c_{m-k} can be expressed in terms of these roots as:

$$c_{m-k} = (-1)^{m-k} \sum_{1\le t_1<t_2<\cdots<t_k\le r} \sigma_{t_1}^2 \sigma_{t_2}^2 \cdots \sigma_{t_k}^2 \ . \qquad (3.6)$$

But we also know that c_{m-k} is the coefficient of λ^{m-k} in $\det(AA^T - \lambda I)$, which means

$$c_{m-k} = (-1)^{m-k} \sum_{\substack{B:\text{ principal} \\ k\text{-minor of } AA^T}} \det(B) \qquad (3.7)$$

(see e. g., [8]; we prove it next as Proposition 3.2). This gives us our desired result. □

Proposition 3.2. *Let the characteristic polynomial of $M \in \mathbb{R}^{m\times m}$ be*

$$\det(M-\lambda I_m) = \lambda^m + c_{m-1}\lambda^{m-1} + \cdots + c_0 \ .$$

Then

$$c_{m-k} = (-1)^{m-k} \sum_{\substack{B:\ \text{principal} \\ k\text{-minor of } AA^T}} \det(B) \qquad \text{for } 1 \le k \le m \ .$$

Proof. First, it is clear that $c_0 = \det(M)$. Next, let $M' = M - \lambda I$, and S_m be the set of permutations of $\{1,2,\ldots,m\}$. The sign of a permutation $\mathrm{sgn}(\tau)$, for $\tau \in \mathrm{Perm}([m])$, is equal to 1 if it is a product of an even number of transpositions and -1 otherwise. For a subset S of rows, we denote the submatrix of entries $(M_{i,j})_{i,j\in S}$ by M_S.

$$\det(M-\lambda I_m) = \det(M') = \sum_{\tau\in\mathrm{Perm}([m])} \mathrm{sgn}(\tau) M'_{1,\tau(1)} M'_{2,\tau(2)} \cdots M'_{m,\tau(m)} \ . \tag{3.8}$$

The term $c_{m-k}\lambda^{m-k}$ is the sum over τ which fix some set $S \subseteq [m]$ of size $(m-k)$, and the elements $\prod_{i\in S} M'_{i,i}$ contribute $(-1)^{m-k}\lambda^{m-k}$ and the coefficient comes from the constant term in

$$\sum_{\tau\in\mathrm{Perm}([m]-S)} \mathrm{sgn}(\tau) \prod_{i\notin S} M'_{i,\tau(i)} \ .$$

Each term in this sum is the c_0 term of a principal minor of M and so the sum is equal to

$$\sum_{S,|S|=m-k} \det(M_{[m]-S}) \ .$$

Hence

$$c_{m-k} = (-1)^{m-k} \sum_{S,|S|=m-k} \det(M_{[m]-S}) = (-1)^{m-k} \sum_{\substack{B:\ \text{principal} \\ k\text{-minor of } AA^T}} \det(B) \ . \tag{3.9}$$

□

The bound proved in Theorem 1.3 is in fact asymptotically tight:

Proposition 3.3. *Given any $\varepsilon > 0$, there exists a $(k+1)\times(k+1)$ matrix A such that for any k-subset S of rows of A,*

$$\|A - \pi_{S,k}(A)\|_F^2 \ge (1-\varepsilon)\,(k+1)\,\|A-A_k\|_F^2 \ .$$

Proof. The tight example consists of a matrix with $k+1$ rows which are the vertices of a regular k-dimensional simplex lying on the affine hyperplane $\{X_{k+1} = \alpha\}$ in $\mathbb{R}^{k+1}$. Let $A^{(1)}, A^{(2)}, \ldots, A^{(k+1)}$ be the vertices with the point $p = (0,0,\ldots,0,\alpha)$ as their centroid. For α small enough, the best k dimensional subspace for these points is given by $\{X_{k+1} = 0\}$ and

$$\|A-A_k\|_F^2 = (k+1)\alpha^2 \ . \tag{3.10}$$

Consider any k-subset of rows from these, say $S = \{A^{(1)}, A^{(2)}, \ldots, A^{(k)}\}$, and let H_S be the linear subspace spanning them. Then,

$$\|A - \pi_{S,k}(A)\|_F^2 = d(A^{(k+1)}, H_S)^2 \ . \tag{3.11}$$

We claim that for any $\varepsilon > 0$, α can be chosen small enough so that

$$d(A^{(k+1)}, H_S) \geq \sqrt{(1-\varepsilon)}(k+1)\alpha \ . \tag{3.12}$$

Choose α small enough so that $d(p, H_S) \geq \sqrt{(1-\varepsilon)}\alpha$. Now

$$\frac{d(A^{(k+1)}, H_S)}{d(p, H_S)} = \frac{d(A^{(k+1)}, \mathrm{conv}(A^{(1)}, \ldots, A^{(k)}))}{d(p, \mathrm{conv}(A^{(1)}, \ldots, A^{(k)}))} = k+1 \tag{3.13}$$

since the points form a simplex and p is their centroid. The claim follows. Hence,

$$\|A - \pi_{S,k}(A)\|_F^2 = d(A^{(k+1)}, H_S)^2 \geq (1-\varepsilon)\ (k+1)^2\ \alpha^2 = (1-\varepsilon)\ (k+1)\ \|A - A_k\|_F^2 \ . \tag{3.14}$$

□

Next, we prove Theorem 1.4. This theorem follows by interpreting Theorem 1.3 as an existence theorem and applying Theorem 2.1.

Proof of Theorem 1.4. By Theorem 1.3, there exists a k-subset S_1 of rows of A such that

$$\|A - \pi_{S_1}(A)\|_F^2 \leq (k+1)\|A - A_k\|_F^2 \ . \tag{3.15}$$

Now, applying Theorem 2.1 with $V = \mathrm{span}(S_1)$ and $s = k(k+1)/\varepsilon$ we get that a random sample S_2 of the rows of A (according to the specified distribution) satisfies

$$\mathsf{E}_{S_2}(\|A - \pi_{V+\mathrm{span}(S_2),k}(A)\|_F^2) \leq (1+\varepsilon)\|A - A_k\|_F^2 \tag{3.16}$$

so there exists a subset of the rows achieving the expectation. Since $V + \mathrm{span}(S_2) = \mathrm{span}(S_1 \cup S_2)$, and $|S_1 \cup S_2| = k + k(k+1)/\varepsilon$, we have the desired result. □

4 Application: Projective clustering

In this section, we give a polynomial-time approximation scheme for the projective clustering problem described in Section 1.2. We note that a simple multiplicative $(k+1)$-approximation follows from Theorem 1.3 with running time $O(dn^{jk})$. Let $V_1, \ldots, V_j$ be the optimal subspaces partitioning the point set into $P_1 \cup \cdots \cup P_j$, where P_i is the subset of points closest to V_i. Theorem 1.3 tells us that each P_i contains a subset S_i of k points whose span V_i' is a $(k+1)$-approximation, i. e.,

$$\sum_{p \in P_i} d(p, V_i')^2 \leq (k+1) \sum_{p \in P_i} d(p, V_i)^2 \ .$$

We can find the S_i's by simply enumerating all possible subsets of k points, considering j of them at a time, and taking the best of these. This leads to the complexity of dn^{jk}.

Getting a PTAS will be a bit more complicated. Theorem 1.4 implies that there exists a set $\hat{P}_i \subseteq P_i$ of size $k+k(k+1)/\varepsilon$ in whose span lies an approximately optimal k-dimensional subspace W_i. We can enumerate over all combinations of j subsets, each of size $k+k(k+1)/\varepsilon$ to find the $\hat{P}_i$, but we cannot enumerate the infinitely many k-dimensional subspaces lying in the span of $\hat{P}_i$. One natural approach to solve this problem would be to put a finite grid down in a unit ball in the span of $\hat{P}_i$. The hope would be that there are k grid points whose span G is "close" to W_i, since each basis vector for W_i is close to a grid point. However, this will not work; consider a point p very far from the origin. Although the distance between a basis vector and a grid point might be small, the error induced by projecting p onto a grid point is proportional to its distance to the origin, which could be too large.

The problem described above suggests that a grid construction must be dependent on the point set P_i. Our grid construction considers grid points in the span of $\hat{P}_i$, but instead of a uniform grid in a unit ball, we consider grid points at bounded distance from each $p \in \pi_{\mathrm{span}(\hat{P}_i)}(P_i)$, i. e., the points in P_i projected to the span of $\hat{P}_i$. This avoids the problem of points far from the origin, since there are grid points around each point. Note that we only put grid points around projected points. This is because we seek a subspace "close" to W_i, which itself lies in the span of $\hat{P}_i$; W_i and any subspace lying in the span of $\hat{P}_i$ incur the same error for the component of a point orthogonal to the span of $\hat{P}_i$. In Lemma 4.1, we show that there exists a subspace spanned by k points in our grid that is not much worse than W_i. The lemma is stated for a general point set, but we apply it to the projected points in Theorem 1.5.

The algorithm is given below.

Algorithm Cluster

Input: $P \subseteq \mathbb{R}^d$, error parameter $0 < \varepsilon < 1$, and upper bound B on the optimal cost.
Output: A set $\{F_1, \ldots, F_j\}$ of j k-dimensional subspaces.

1. Set $\delta = \frac{\varepsilon\sqrt{B}}{16jk\sqrt{(1+\frac{\varepsilon}{2})n}}$, $R = \sqrt{(1+\frac{\varepsilon}{2})B} + 2\delta k$.

2. For each subset T of P of size $j(k+2k(k+1)/\varepsilon)$:

   (a) For each equipartition $T = T_1 \cup \cdots \cup T_j$:

      i. For each i, construct a δ-net D_i with radius R for the projection of P to the span of T_i.

      ii. For each way of choosing j subspaces $F_1, \ldots, F_j$, where F_i is the span of k points from D_i, compute the cost $\mathcal{C}(F_1, \ldots, F_j)$.

3. Report the subspaces $F_1, \ldots, F_j$ of minimum cost $\mathcal{C}(F_1, \ldots, F_j)$.

In Step 2(a)i, we construct a δ-net D_i. A δ-net D with radius R for S is a set such that for any point q for which $d(q,p) \leq R$, for some $p \in S$, there exists a $g \in D$ such that $d(q,g) \leq \delta$. The size of a δ-net is exponential in the dimension of S. This is why it is crucial that we construct the δ-net for P projected to the span of T_i. By doing so, we reduce the dimension from d to $O(k^2/\varepsilon)$. The correctness of the algorithm relies crucially on the next lemma.

Lemma 4.1. *Let $\delta > 0$. Let P be a point set, with $|P| = n$ and W be a subspace of dimension k. Let D be a δ-net with radius R for P, satisfying*

$$R \geq \sqrt{\sum_{p \in P} d(p,W)^2} + 2\delta k \ .$$

Then there exists a subspace F spanned by k points in D such that:

$$\sum_{p \in P} d(p,F)^2 \leq \sum_{p \in P} d(p,W)^2 + 4k^2 n\delta^2 + 4k\delta \sum_{p \in P} d(p,W) \ . \tag{4.1}$$

Proof. We construct the subspace F in k steps. Let $F_0 = W$. Inductively, in step i, we choose a point p_i and rotate F_{i-1} so that it includes a grid point g_i around p_i. The subspace resulting from the last rotation, F_k, is the subspace F with the bound promised by the lemma. To prove that (4.1) holds, we prove the following inequality for any point $p \in P$ going from F_{i-1} to F_i

$$d(p,F_i) \leq d(p,F_{i-1}) + 2\delta \ . \tag{4.2}$$

Summing over the k steps, squaring, and summing over n points, we have the desired result.

Let $G_1 = \{\vec{0}\}$. G_i will be the span of the grid points $\{g_1, g_2, \ldots, g_{i-1}\}$. We describe how to construct the rotation R_i. Let $p_i \in P$ maximize $\|\pi_{G_i^\perp}(\pi_{F_{i-1}}(p_i))\|$ and let $g_i \in D$ minimize $d(\pi_{F_{i-1}}(p_i), g_i)$. The point p_i is chosen as the furthest point from the origin in the subspace of F_{i-1} orthogonal to G_i. The grid point g_i is the point closest to p_i in F_{i-1}. Consider the plane Z defined by

$$\pi_{G_i^\perp}(g_i),\ \pi_{G_i^\perp}(\pi_{F_{i-1}}(p_i)),\ \text{and}\ \vec{0} \ .$$

Let θ be the angle between $\pi_{G_i^\perp}(g_i)$ and $\pi_{G_i^\perp}(\pi_{F_{i-1}}(p_i))$. Let R_i be the rotation in the plane Z by the angle θ, and define $F_i = R_i F_{i-1}$. Set $G_{i+1} = G_i + \text{span}\{g_i\}$. By choosing the rotation dependent on p_i, we ensure that no point moves more than p_i. This allows us to prove (4.2) for all points p.

Now we prove inequality (4.2). We do so by proving the following inequality by induction on i for any point p:

$$d(\pi_{F_{i-1}}(p), R_i \pi_{F_{i-1}}(p)) \leq 2\delta \ . \tag{4.3}$$

Note that this proves (4.2) by applying the triangle inequality, since:

$$\begin{aligned} d(p,F_i) &\leq d(p,F_{i-1}) + d(\pi_{F_{i-1}}(p), \pi_{F_i}(p)) && (4.4)\\ &\leq d(p,F_{i-1}) + d(\pi_{F_{i-1}}(p), R_i \pi_{F_{i-1}}(p)) \ . && (4.5) \end{aligned}$$

The base case of the inequality, $i = 1$, is trivial. Consider the inductive case; here, we are bounding the distance between $\pi_{F_{i-1}}(p)$ and $R_i \pi_{F_{i-1}}(p)$. It suffices to bound the distance between these two points in the subspace orthogonal to G_i, since the rotation R_i is chosen orthogonal to G_i. That is,

$$d(\pi_{F_{i-1}}(p), R_i \pi_{F_{i-1}}(p)) \leq d(\pi_{G_i^\perp}(\pi_{F_{i-1}}(p)), R_i \pi_{G_i^\perp}(\pi_{F_{i-1}}(p))) \ . \tag{4.6}$$

Now, consider the distance between a point $\pi_{G_i^\perp}(\pi_{F_{i-1}}(p))$ and its rotation, $R_i \pi_{G_i^\perp}(\pi_{F_{i-1}}(p))$. This distance is maximized when $\|\pi_{G_i^\perp}(\pi_{F_{i-1}}(p))\|$ is maximized, so we have, by construction, that the maximum value is achieved by p_i:

$$d(\pi_{G_i^\perp}(\pi_{F_{i-1}}(p)), R_i \pi_{G_i^\perp}(\pi_{F_{i-1}}(p))) \leq d(\pi_{G_i^\perp}(\pi_{F_{i-1}}(p_i)), R_i \pi_{G_i^\perp}(\pi_{F_{i-1}}(p_i))) \ . \tag{4.7}$$

By the triangle inequality we have:

$$d(\pi_{G_i^\perp}(\pi_{F_{i-1}}(p_i)), R_i\,\pi_{G_i^\perp}(\pi_{F_{i-1}}(p_i))) \le d(\pi_{G_i^\perp}(\pi_{F_{i-1}}(p_i)), \pi_{G_i^\perp}(g_i)) + d(\pi_{G_i^\perp}(g_i), R_i\,\pi_{G_i^\perp}(\pi_{F_{i-1}}(p_i))) \ .$$

To bound the first term, $d(\pi_{G_i^\perp}(\pi_{F_{i-1}}(p_i)), \pi_{G_i^\perp}(g_i))$, note that

$$d(\pi_{G_i^\perp}(\pi_{F_{i-1}}(p_i)), \pi_{G_i^\perp}(g_i)) \le d(\pi_{F_{i-1}}(p_i), g_i) \ .$$

We show that $\pi_{F_{i-1}}(p_i)$ is within a ball of radius R around p_i; this implies

$$d(\pi_{F_{i-1}}(p_i), g_i) \le \delta \tag{4.8}$$

by construction of the δ-net around p_i. We have:

$$\begin{aligned} d(p_i, \pi_{F_{i-1}}(p_i)) &\le d(p_i, F_0) + \sum_{j=1}^{i-2} d(\pi_{F_j}(p_i), \pi_{F_{j+1}}(p_i)) \\ &\le \sqrt{\sum_{p\in P} d(p,W)^2} + \sum_{j=1}^{i-2} d(\pi_{F_j}(p_i), R_{j+1}\,\pi_{F_j}(p_i)) \\ &\le \sqrt{\sum_{p\in P} d(p,W)^2} + 2\delta k \le R \ . \end{aligned}$$

The third line uses the induction hypothesis.

Now we bound the second term, $d(\pi_{G_i^\perp}(g_i), R_i\,\pi_{G_i^\perp}(\pi_{F_{i-1}}(p_i)))$. Note that $R_i\,\pi_{G_i^\perp}(\pi_{F_{i-1}}(p_i))$ is just a rescaling of $\pi_{G_i^\perp}(g_i)$ and that $\|\,\pi_{G_i^\perp}(\pi_{F_{i-1}}(p^*))\| = \|R_i\,\pi_{G_i^\perp}(\pi_{F_{i-1}}(p^*))\|$, since rotation preserves norms. The bound on the first term implies that $\|\,\pi_{G_i^\perp}(g_i)\| \ge \|\,\pi_{G_i^\perp}(\pi_{F_{i-1}}(p^*))\| - \delta$, so

$$d(\pi_{G_i^\perp}(g_i), R_i\,\pi_{G_i^\perp}(\pi_{F_{i-1}}(p_i))) \le \delta \ . \tag{4.9}$$

Combining (4.8) and (4.9), we have proved (4.3). □

Now, we are ready to prove Theorem 1.5, which includes the correctness of the algorithm.

Proof of Theorem 1.5. Assume that the optimal solution is of value at most B. Let $V_1,\ldots,V_j$ be the optimal subspaces, and let $P_1,\ldots,P_j$ be the partition of P such that P_i is the subset of points closest to V_i. Theorem 1.4 implies that there exists $S_i \subseteq P_i$ of size at most $k + 2k(k+1)/\varepsilon$ such that there is a k-dimensional subspace W_i in the span of S_i with

$$\sum_{p\in P_i} d(p,W_i)^2 \le \left(1+\frac{\varepsilon}{2}\right) \sum_{p\in P_i} d(p,V_i)^2 \le \left(1+\frac{\varepsilon}{2}\right) B \ . \tag{4.10}$$

Consider $\pi_{\mathrm{span}(S_i)}(P_i)$, the projection of P_i to $\mathrm{span}(S_i)$. We want to apply Lemma 4.1 to $\pi_{\mathrm{span}(S_i)}(P_i)$ and W_i with radius R and δ as in the algorithm. Note that the optimal solution is of value at most B, so we have that:

$$\sqrt{\sum_{p\in\pi_{\mathrm{span}(S_i)}(P_i)} d(p,W_i)^2} + 2\delta k \le \sqrt{\sum_{p\in P_i} d(p,W_i)^2} + 2\delta k \le \sqrt{\left(1+\frac{\varepsilon}{2}\right) B} + 2\delta k = R \ . \tag{4.11}$$

Let F_i be the subspace spanned by k points from the δ-net D_i for $\pi_{\mathrm{span}(S_i)}(P_i)$ promised by Lemma 4.1. For every i, we have that:

$$\begin{aligned}\sum_{p\in\pi_{\mathrm{span}(S_i)}(P_i)} d(p,F_i)^2 &\le \sum_{p\in\pi_{\mathrm{span}(S_i)}(P_i)} d(p,W_i)^2 + 4k^2n\delta^2 + 4k\delta \sum_{p\in\pi_{\mathrm{span}(S_i)}(P_i)} d(p,W_i)\\ &\le \sum_{p\in\pi_{\mathrm{span}(S_i)}(P_i)} d(p,W_i)^2 + \frac{\varepsilon}{4j}B + 4k\delta\left(n\sum_{p\in\pi_{\mathrm{span}(S_i)}(P_i)} d(p,W_i)^2\right)^{1/2}\\ &\le \sum_{p\in\pi_{\mathrm{span}(S_i)}(P_i)} d(p,W_i)^2 + \frac{\varepsilon}{2j}B\ . \end{aligned} \tag{4.12}$$

Now, for any point $p\in P_i$, we can decompose the squared distance from p to F_i as follows:

$$d(p,F_i)^2 = d(\pi_{\mathrm{span}(S_i)}(p),F_i)^2 + d(\pi_{\mathrm{span}(S_i)^\perp}(p),F_i)^2\ .$$

The same decomposition can be done for $d(p,W_i)^2$. Now, since F_i and W_i both lie in the span of S_i, we have the following for any point $p\in P_i$: $d(\pi_{\mathrm{span}(S_i)^\perp}(p),F_i)^2 = d(\pi_{\mathrm{span}(S_i)^\perp}(p),W_i)^2$. Applying this to (4.12) and (4.10), we have:

$$\sum_{p\in P_i} d(p,F_i)^2 \le \left(1+\frac{\varepsilon}{2}\right)\sum_{p\in P_i} d(p,V_i)^2 + \frac{\varepsilon}{2j}B\ .$$

Let $S=\bigcup_i S_i$. The algorithm will enumerate S in Step 2a, and it will enumerate the partition $S_1\cup\cdots\cup S_j$ in Step 2a. In Step 2(a)i, the algorithm will, for each i, construct a δ-net D_i for $\pi_{\mathrm{span}(S_i)}(P)$. Lastly, in Step 2(a)ii it will consider the subspaces $F_1,\ldots,F_j$ whose existence is proven above. The cost associated with this solution is:

$$\mathcal{C}(F_1,\ldots,F_j) \le \sum_{i=1}^{j}\sum_{p\in P_i} d(p,F_i)^2 \le \left(1+\frac{\varepsilon}{2}\right)\sum_{i=1}^{j}\sum_{p\in P_i} d(p,V_i)^2 + \frac{\varepsilon}{2}B = (1+\varepsilon)B\ .$$

The number of subsets of size $k+2k(k+1)/\varepsilon$ enumerated by the algorithm is at most $\binom{n}{2(k+1)^2/\varepsilon}^j$. A δ-net D with radius R for a point set Q of dimension d is implemented by putting a box with side length $2R$ of grid width $\delta/\sqrt{d}$ around each point in Q. Let X be the set of grid points in the box around a point p. The number of subspaces in each δ-net D_i is therefore at most the number of j subspaces that one can choose for a partition $T_1\cup\cdots\cup T_j$ is $(n|X|)^{jk}$. The computation for projecting points, finding a basis, and determining the cost of a candidate family of subspaces takes time $O(ndjk)$. The cardinality of X is (for $\varepsilon<1$):

$$|X| = \left(\frac{2R}{\delta/\sqrt{2(k+1)^2/\varepsilon}}\right)^{2(k+1)^2/\varepsilon} \le \left(O\left(jk\sqrt{\frac{n}{\varepsilon}}\right)\right)^{2(k+1)^2/\varepsilon}\ . \tag{4.13}$$

Therefore, the running time of the algorithm is at most $O(ndjk)\,n^{2j(k+1)^2/\varepsilon}\,(n|X|)^{jk} = d\left(\frac{n}{\varepsilon}\right)^{O(jk^3/\varepsilon)}$. □

5 Conclusions and subsequent work

Theorem 1.4 was further improved in [13] to show that for any real matrix, there exist $O(k/\varepsilon + k\log k)$ rows whose span contains the rows of a multiplicative $(1+\varepsilon)$-approximation to the best rank-k matrix. Using this subset of $O(k/\varepsilon + k\log k)$ rows, the exponent in the running time of the projective clustering algorithm decreases from $O(jk^3/\varepsilon)$ to $O(jk^2/\varepsilon)$. It would be interesting to know if we can compute this set of $O(k/\varepsilon + k\log k)$ rows efficiently in a small number of passes. It would also be nice to see other applications of volume sampling.

In a recent result, Sarlos [34] proved that, using random linear combinations of rows instead of a subset of rows, we can compute a multiplicative $(1+\varepsilon)$-approximation using only 2 passes.

References

[1] * D. ACHLIOPTAS AND F. MCSHERRY: Fast computation of low rank matrix approximations. In *Proc. 33rd STOC.*, pp. 611–618. ACM Press, 2001. [STOC:380752.380858]. 1.1, 1.3

[2] * P. AGARWAL, S. HAR-PELED, AND K. VARADARAJAN: Geometric approximations via core-sets. *Combinatorial and Computational Geometry - MSRI Publications*, 52:1–30, 2005. 1.2, 1.3

[3] * P. AGARWAL AND N. MUSTAFA: k-means projective clustering. In *Proc. Principles of Database Systems (PODS'04)*, pp. 155–165. ACM Press, 2004. [PODS:1055558.1055581]. 1.1, 1.3

[4] * P. AGARWAL, C. PROCOPIUC, AND K. VARADARAJAN: Approximation algorithms for a k-line center. *Algorithmica*, 42:221–230, 2005. [Algorithmica:p321725768012505]. 1.3

[5] * R. AGARWAL, J. GEHRKE, D. GUNOPULOS, AND P. RAGHAVAN: Automatic subspace clustering of high dimensional data for data mining applications. *Data Mining and Knowledge Discovery*, 11:5–33, 2005. [Springer:m18077t9134v55n2]. 1.1, 1.3

[6] * C. AGGARWAL, C. PROCOPIUC, J. WOLF, P. YU, AND J. PARK: Fast algorithms for projected clustering. In *Proc. Conf. on Management of Data (SIGMOD'99)*, pp. 61–72. ACM Press, 1999. [SIGMOD:304182.304188]. 1.1, 1.3

[7] * N. ALON, Y. MATIAS, AND M. SZEGEDY: The space complexity of approximating the frequency moments. *J. Computer and System Sciences*, 58:137–147, 1999. [JCSS:10.1006/jcss.1997.1545]. 1.3

[8] * M. ARTIN: *Algebra*. Prentice-Hall, 1991. 3

[9] * Z. BAR-YOSSEF: Sampling lower bounds via information theory. In *Proc. 35th STOC.*, pp. 335–344. ACM Press, 2003. [STOC:780542.780593]. 1.3

[10] * M. BĂDOIU, S. HAR-PELED, AND P. INDYK: Approximate clustering via core-sets. In *Proc. 34th STOC.*, pp. 250–257. ACM Press, 2002. [STOC:509907.509947]. 1.3

[11] * W. FERNANDEZ DE LA VEGA, M. KARPINSKI, C. KENYON, AND Y. RABANI: Approximation schemes for clustering problems. In *Proc. 35th STOC.*, pp. 50–58. ACM Press, 2003. [STOC:780542.780550]. 1.3

[12] * A. DESHPANDE, L. RADEMACHER, S. VEMPALA, AND G. WANG: Matrix approximation and projective clustering via volume sampling. In *Proc. 17th ACM-SIAM Symp. on Discrete Algorithms (SODA'06)*, pp. 1117–1126. SIAM, 2006. [SODA:1109557.1109681]. 1.4

[13] * A. DESHPANDE AND S. VEMPALA: Adaptive sampling and fast low-rank matrix approximation. In *Proc. 10th Internat. Workshop on Randomization and Computation (RANDOM'06)*, pp. 292–303. Springer, 2006. 5

[14] * P. DRINEAS AND R. KANNAN: Pass efficient algorithms for approximating large matrices. In *Proc. 14th ACM-SIAM Symp. on Discrete Algorithms (SODA'03)*, pp. 223–232. SIAM, 2003. [SODA:644108.644147]. 1.1, 1.2, 1.3

[15] * P. DRINEAS, R. KANNAN, A. FRIEZE, S. VEMPALA, AND V. VINAY: Clustering large graphs via the singular value decomposition. *Machine Learning*, 56:9–33, 2004. [ML:u424k6nn6k622788]. 1.1, 1.1, 1.2, 1.3, 1.3

[16] * P. DRINEAS, R. KANNAN, AND M. MAHONEY: Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix. *SIAM J. on Computing*, 36:158–183, 2006. [SICOMP:10.1137/S0097539704442696]. 1.1

[17] * M. EFFROS AND L. J. SCHULMAN: Rapid near-optimal VQ design with a deterministic data net. In *Proc. Int. Symp. on Information Theory*, p. 298. IEEE, 2004. [ISIT:10.1109/ISIT.2004.1365336]. 1.3

[18] * J. FEIGENBAUM, S. KANNAN, A. MCGREGOR, S. SURI, AND J. ZHANG.: On graph problems in a semi-streaming model. *Theoretical Computer Science*, 348:207–216, 2005. [TCS:10.1016/j.tcs.2005.09.013]. 1.3

[19] * A. FRIEZE, R. KANNAN, AND S. VEMPALA: Fast Monte Carlo algorithms for finding low-rank approximations. *Journal of the ACM*, 51:1025–1041, 2004. [JACM:1039488.1039494]. 1.1, 1.1, 1, 1.2, 1.3, 2.1

[20] * G. H. GOLUB AND C. F. VAN LOAN: *Matrix Computations*. Johns Hopkins, third edition, 1996. 1.3

[21] * S. A. GOREINOV AND E. E. TYRTYSHNIKOV: The maximal-volume concept in approximation by low-rank matrices. *Contemporary Mathematics*, 280:47–51, 2001. 1.3

[22] * S. GUHA, N. KOUDAS, AND K. SHIM: Approximation and streaming algorithms for histogram construction problems. *ACM Transactions on Database Systems*, 31:396–438, 2006. [TODS:1132863.1132873]. 1.3

[23] * S. HAR-PELED AND S. MAZUMDAR: On coresets for k-means and k-median clustering. In *Proc. 36th STOC.*, pp. 291–300. ACM Press, 2004. [STOC:1007352.1007400]. 1.3

[24] * S. HAR-PELED AND K. VARADARAJAN: Projective clustering in high dimensions using core-sets. In *Proc. 18th Symp. on Comp. Geometry (SOCG)*, pp. 312–318. ACM Press, 2002. [SOCG:513400.513440]. 1.2, 1.3

[25] * M. HENZINGER, P. RAGHAVAN, AND S. RAJAGOPALAN: Computing on data streams. *External Memory Algorithms – DIMACS Series in Discrete Mathematics and Computer Science*, 50:107–118, 1999. 1.3

[26] * D. KEMPE AND F. MCSHERRY: A decentralized algorithm for spectral analysis. In *Proc. 36th STOC.*, pp. 561–568. ACM Press, 2004. [STOC:1007352.1007438]. 1.3

[27] * J. KUCZYŃSKI AND H. WOŹNIAKOWSKI: Estimating the largest eigenvalues by the power and Lanczos algorithms with a random start. *SIAM J. Matrix Anal. Appl.*, 13:1094–1122, 1992. [SIMAX:10.1137/0613066]. 1.3

[28] * A. KUMAR, Y. SABHARWAL, AND S. SEN.: A simple linear time $(1+\varepsilon)$-approximation algorithm for k-means clustering in any dimensions. In *Proc. 45th FOCS*, pp. 454–462. IEEE Computer Society Press, 2004. [FOCS:10.1109/FOCS.2004.7]. 1.3

[29] * J. MATOUŠEK: On approximate geometric k-clustering. *Discrete and Computational Geometry*, 24:61–84, 2000. [Springer:xawxbau59gtjtvv6]. 1.3

[30] * N. MEGIDDO AND A. TAMIR: On the complexity of locating linear facilities in the plane. *Operations Research Letters*, 13:194–197, 1982. [Elsevier:10.1016/0167-6377(82)90039-6]. 1.1

[31] * R. OSTROVSKY AND Y. RABANI: Polynomial time approximation schemes for geometric min-sum median clustering. *Journal of the ACM*, 49:139–156, 2002. [JACM:506147.506149]. 1.3

[32] * C. PROCOPIUC, P. AGARWAL, T. MURALI, AND M. JONES: A Monte Carlo algorithm for fast projective clustering. In *Proc. Conf. on Management of Data (SIGMOD'02)*, pp. 418–427. ACM Press, 2002. [SIGMOD:564691.564739]. 1.1, 1.3

[33] * L. RADEMACHER, S. VEMPALA, AND G. WANG: Matrix approximation and projective clustering. Technical Report 2005–018, MIT CSAIL, 2005. 1.4

[34] * T. SARLÓS: Improved approximation algorithms for large matrices via random projection. In *Proc. 47th FOCS*, pp. 143–152. IEEE Computer Society Press, 2006. 5

AUTHORS

Amit Deshpande
2-342
77 Mass. Ave.,
Cambridge, MA 02139, USA.
amitd@mit.edu
http://www.mit.edu/~amitd/

Luis Rademacher
2-331
77 Mass. Ave.
Cambridge, MA 02139, USA.
lrademac@math .mit.edu
http://www-math.mit.edu/~lrademac/

Santosh Vempala
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA.
vempala@cc.gatech.edu
http://www-math.mit.edu/~vempala/

Grant Wang
Yahoo!
701 First Avenue
Sunnyvale, CA 94089, USA.
gjw@alum.mit.edu
http://theory.csail.mit.edu/~gjw/

ABOUT THE AUTHORS

AMIT DESHPANDE is a graduate student in Applied Math at MIT. He did his undergraduate studies at Chennai Mathematical Institute (CMI) in India. Apart from math and theory, he enjoys north Indian classical music.

LUIS RADEMACHER is a Ph. D. candidate in the Department of Mathematics at MIT, supervised by Santosh Vempala. His research interests include game theory, matrix approximation, computational lower bounds, and the intersection between geometry and algorithms. He grew up in Chile and enjoys music as a hobby.

SANTOSH VEMPALA is a professor in the College of Computing and director of the newly formed Algorithms and Randomness Center at Georgia Tech. His research interests, ironically, are in algorithms, randomness, and geometry. He graduated from CMU in 1997 following the advice of Avrim Blum and was at MIT till 2006 except for a year as a Miller fellow at UC Berkeley. He gets unreasonably excited when a phenomenon that appears complex from one perspective turns out to be simple from another.

GRANT WANG graduated from MIT in August 2006 with a Ph. D. in computer science. His advisor was Santosh Vempala. He graduated from Cornell University with a B. S. in Computer Science in 2001. His research interests are in algorithms, machine learning, and data mining. As of September 2006, he is working at Yahoo!