Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2005-018
MIT-LCS-TR-983

March 29, 2005

Matrix Approximation and Projective Clustering via Iterative Sampling

Luis Rademacher, Santosh Vempala, Grant Wang

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Matrix Approximation and Projective Clustering via Iterative Sampling

Luis Rademacher* Santosh Vempala* Grant Wang*

Abstract

We present two new results for the problem of approximating a given real $m \times n$ matrix A by a rank-k matrix D, where $k < \min\{m, n\}$, so as to minimize $||A - D||_F^2$. It is known that by sampling $O(k/\epsilon)$ rows of the matrix, one can find a low-rank approximation with additive error $\epsilon||A||_F^2$. Our first result shows that with adaptive sampling in t rounds and $O(k/\epsilon)$ samples in each round, the additive error drops exponentially as ϵ^t; the computation time is nearly linear in the number of nonzero entries. This demonstrates that multiple passes can be highly beneficial for a natural (and widely studied) algorithmic problem. Our second result is that there *exists* a subset of $O(k^2/\epsilon)$ rows such that their span contains a rank-k approximation with *multiplicative* $(1 + \epsilon)$ error (i.e., the sum of squares distance has a small "core-set" whose span determines a good approximation). This existence theorem leads to a PTAS for the following projective clustering problem: Given a set of points P in $\mathbb{R}^d$, and integers k, j, find a set of j 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

Given data consisting of points in high-dimensional space, it is often of interest to find a low-dimensional representation. In this paper, we consider the general problem of finding one or more (up to j) subspaces, each of dimension at most k, and representing each point by its orthogonal projection to the nearest subspace. Our goal will be to minimize the sum of squared distances of each point to its nearest subspace, a measure of the "error" incurred by this representation.

This problem has been called *projective clustering*, since the j subspaces induce a partition of the point set. Algorithms and systems based on projective clustering have been applied to facial recognition, data-mining, and synthetic data [5, 25, 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. We discuss related theoretical work in Section 1.2.

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 [23], even for $k = 1$ [11]). Viewing the points as the rows of an $m \times n$ matrix A, we find the top k right singular vectors of this matrix via the Singular Value Decomposition (SVD). The projection itself 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. 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

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

this algorithm, dominated by the SVD computation, 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 or expensive to store/process in their entirety, one can view the data as a stream and the goal is to store/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. [16] showed that any matrix A has a subset of k/ϵ rows whose span contains an approximately optimal rank k approximation to A. In fact, the subset of rows can be obtained as independent samples from a distribution that depends only on the lengths of the rows. (In what follows, $A^{(i)}$ denotes the ith row of A, as a column vector.)

Theorem 1 ([16]). *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*

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

If $s \geq k/c\epsilon$, then the span of S contains a matrix $\tilde{A}_k$ of rank at most k for which

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

This can be turned into an efficient algorithm based on sampling [11] [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 $O(\min\{m, n\}k^2/\epsilon^4)$. These results lead to the following questions: (1) Can the error be reduced significantly by using multiple passes through the data? (2) Can we get multiplicative $(1+\epsilon)$ approximations? (3) Do these sampling algorithms have any consequences for the general projective clustering problem?

1.1 Our results

Our first result is that the additive error term drops *exponentially* with the number of passes. 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 consider the following two-round approach. In the first pass, we get a sample and find a rank 2 approximation using it. Then we sample again, but this time with probability proportional to the error of the approximation. If the lone far-off point is missed in the first pass, it will have a very high probability of being chosen in the second pass. The span of the full sample now contains a very good rank 2 approximation. In the general theorem below, for a set of rows S 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.

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

Theorem 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*

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

where $E_1 = A$, $E_j = A - \pi_{S_1 \cup \cdots \cup S_{j-1}}(A)$ and c is a constant. Then for $s \geq k/c\epsilon$, the span of S contains a matrix $\tilde{A}_k$ of rank k such that

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

The proof of Theorem 2 is given in Section 2. The resulting algorithm, described in Section 3 uses $2t$ passes through the data and $O(Mst + (m+n)s^2t^2)$ computation time where M is the number of nonzeros in A. Although the sampling distribution is modified t times, the matrix itself is not changed and so its sparsity is maintained. The algorithm fits the streaming model in that the entries of A can arrive in any order (see Section 1.2). The space used is $O\big((m+n)kt/\epsilon\big)$.

Theorem 2 implies that for any matrix A, there *exists* a subset of kt/ϵ rows whose span contains a rank-k matrix whose error is within an additive $\epsilon^t||A||_F^2$ of the best rank-k matrix. Can this be improved? In particular, is there a small subset of rows whose span contains a rank-k matrix whose error is within a $(1+\epsilon)$ multiplicative factor of the error of the best possible rank-k approximation? Our next theorem answers this question affirmatively.

Theorem 3. *For any matrix A, there exists a subset of $4k^2/\epsilon$ rows in whose span lies a rank-k matrix $\tilde{A}_k$ such that*

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

The proof of this theorem also uses iterative sampling, albeit in a "backwards" manner and yields a simple sampling-based algorithm for finding such an approximation only for $k = 1$. The proof for general k is by induction on k, and uses the sampling algorithm for $k = 1$ to extend the (approximately) best $(k-1)$-dimensional subspace to an approximately best k-dimensional subspace. Although this existence result does not imply an algorithm faster than the SVD for finding such an approximation, it will be the key ingredient in our last result—a polynomial-time approximation scheme (PTAS) for the general projective clustering problem ($j \geq 2$).

We restate the 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 of n points P in $\mathbb{R}^d$, find a set of j k-dimensional subspaces $F_1, \ldots, F_j$ such that

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

is minimized. When subspaces are replaced by flats, the case $k = 0$ corresponds to the j-means problem (with the sum of squares objective function).

Theorem 3 suggests an enumerative algorithm. The optimal set of k-dimensional subspaces induces a partition $P_1, \ldots, P_j$ of the given point set. In each set P_i, there is, by Theorem 3, a subset of size $O(k^2/\epsilon)$ in whose span lies a $(1+\epsilon)$ approximation to the optimal k-dimensional subspace for this set P_i. So we consider all possible combinations of j subsets each of size $O(k^2/\epsilon)$,

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 simply output the best. Since the subset size is bounded (and so is the size of the net), this gives a PTAS for the problem (see Section 5).

Theorem 4. *Given n points in $\mathbb{R}^d$ and parameters B and ϵ, in time*

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

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

Our technique can also be viewed as an extension of the idea of *core-sets*. Roughly stated, a core-set is a small subset of the data which captures a near-optimal solution to the entire set. Theorem 3 states that there is a core-set of size $O(k^2/\epsilon)$ for minimizing the squared error to a rank k subspace. Unlike proofs of core-sets for other problems, our proof relies on the probabilistic method along with the properties of the SVD. Finally, our result can be extended to finding affine subspaces instead of linear subspaces.

1.2 Related work

Following the work of [16] and [11] which introduced matrix sampling for fast low-rank approximation, Achlioptas and McSherry [1] gave an alternative sampling-based algorithm for the problem. Their algorithm achieves similar bounds (see [1] for a detailed comparison) using only one pass. It does not seem amenable to the multipass improvements presented here. Subsequently, Bar-Yossef [9] has shown that the bounds of these algorithms for one or two passes are optimal up to polynomial factors in $1/\epsilon$.

These algorithms can also be viewed in the *streaming* model of computation [20]. In this model, we do not have random access to data; the data comes as a stream 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 [17], 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 [12, 15]. The "pass-efficient" model of computation was introduced in [20]. Our multipass algorithm fits this model and investigates the tradeoff between approximation and the number of passes. Feigenbaum, et. al [15] show such a tradeoff for computing the maximum unweighted matching in bipartite graphs.

The results of our paper connect two previously separate fields — low-rank approximation and projective clustering. As mentioned earlier, projective clustering has been used in various contexts [5, 25, 6]. In [4], the authors 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 ($k = 0$, find j 0-dimensional affine subspaces, i.e., points). Drineas et al. [11] gave a 2-approximation to j-means using SVD. Subsequently, Ostrovsky and Rabani [24] gave the first randomized polynomial time approximation schemes for j-means (and also the j-median problem). Matoušek [22] and Effros and Schulman [14] both gave deterministic PTAS's for j-means. Fernandez de la Vega et al. [10] describe a randomized algorithm with a running time of $O(n(\log n)^{O(1)})$.z Using the idea of core-sets, Har-Peled and Mazumdar [18] showed a $(1 + \epsilon)$

approximation algorithm that runs in linear time for fixed j, ϵ. Kumar et al. [21] give a linear-time PTAS that uses random sampling. There is a PTAS for $k = 1$ (lines) as well [2]. Other objective functions have also been studied, e.g. sum of distances (j-median when $k = 0$, [24, 18]) and maximum distance (j-center when $k = 0$, [8]). For general k, Har-Peled and Varadarajan [19] give a $(1 + \epsilon)$ approximation algorithm for the maximum distance objective. Their algorithm runs in time $dn^{O(jk^6 \log(1/\epsilon)/\epsilon^5)}$ and is based on core-sets (see [3] for a survey).

1.3 Notation and Preliminaries

Let $A \in \mathbb{R}^{m \times n}$. Let $A^{(i)}$ denote the ith row of A, seen as a column vector. Any matrix accepts 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 $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$ 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 $A v^{(i)} = \sigma_i u^{(i)}$ 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

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

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 $\text{span}(S) \subseteq \mathbb{R}^n$ be the subspace generated those rows; we use the simplified notation $\pi_S(A)$ for $\pi_{\text{span}(S)}(A)$ and $\pi_{S,k}(A)$ for $\pi_{\text{span}(S),k}(A)$.

For subspaces $V, W \subseteq \mathbb{R}^n$, the sum of them 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 $\mathbb{E}(V)$, is the vector having as components the expected values of the components of v.

2 A random sample contains a good approximation

We will prove Theorem 2 in this section. It will be convenient to formulate an intermediate theorem as follows.

Theorem 5. *Let $A \in \mathbb{R}^{m\times n}$. Let $V \subseteq \mathbb{R}^n$ be a vector subspace. Let $E = A - \pi_V(A)$. For a fixed $c \in \mathbb{R}$, let S be a random sample of s rows of A from a distribution such that row i is chosen with probability*

$$P_i \geq c\frac{\|E^{(i)}\|^2}{\|E\|_F^2}. \tag{1}$$

Then, for any nonnegative integer k,

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

Proof. For $S = (r_i)_{i=1}^s$ a sample of rows of A and $1 \leq j \leq r$, let

$$w^{(j)} = \pi_V(A)^T u^{(j)} + \frac{1}{s}\sum_{i=1}^{s} \frac{u_{r_i}^{(j)}}{P_{r_i}} E^{(r_i)}.$$

Then, $\mathbb{E}_S(w^{(j)}) = \pi_V(A)^T u^{(j)} + E^T u^{(j)} = \sigma_j v^{(j)}$. Now we will bound $\mathbb{E}_S(\|w^{(j)} - \sigma_j v^{(j)}\|^2)$. Use the definition of $w^{(j)}$ to get

$$w^{(j)} - \sigma_j v^{(j)} = \frac{1}{s}\sum_{i=1}^{s} \frac{u_{r_i}^{(j)}}{P_{r_i}} E^{(r_i)} - E^T u^{(j)}.$$

Apply the norm squared to each side and expand the left hand side:

$$\|w^{(j)} - \sigma_j v^{(j)}\|^2 = \left\|\frac{1}{s}\sum_{i=1}^{s} \frac{u_{r_i}^{(j)}}{P_{r_i}} E^{(r_i)}\right\|^2 - \frac{2}{s}\sum_{i=1}^{s} \frac{u_{r_i}^{(j)}}{P_{r_i}} E^{(r_i)} \cdot (E^T u^{(j)}) + \|E^T u^{(j)}\|^2. \tag{2}$$

Observe that

$$\mathbb{E}_S\left(\frac{u_{r_i}^{(j)}}{P_{r_i}} E^{(r_i)}\right) = \sum_{i=1}^{m} P_i \frac{u_i^{(j)}}{P_i} E^{(i)} = E^T u^{(j)}, \tag{3}$$

which implies that

$$\mathbb{E}_S\left(\frac{2}{s}\sum_{i=1}^{s} \frac{u_{r_i}^{(j)}}{P_{r_i}} E^{(r_i)} \cdot (E^T u^{(j)})\right) = 2\|E^T u^{(j)}\|^2.$$

Using this, apply $\mathbb{E}_S$ to Equation (2) to get:

$$\mathbb{E}_S(\|w^{(j)} - \sigma_j v^{(j)}\|^2) = \mathbb{E}_S\left(\left\|\frac{1}{s}\sum_{i=1}^{s} \frac{u_{r_i}^{(j)}}{P_{r_i}} E^{(r_i)}\right\|^2\right) - \|E^T u^{(j)}\|^2 \tag{4}$$

Now, from the left hand side, and expanding the norm squared,

$$\begin{aligned}\mathbb{E}_S\left(\left\|\frac{1}{s}\sum_{i=1}^{s} \frac{u_{r_i}^{(j)}}{P_{r_i}} E^{(r_i)}\right\|^2\right) &= \frac{1}{s^2}\sum_{i=1}^{s} \mathbb{E}_S\left(\frac{\|u_{r_i}^{(j)} E^{(r_i)}\|^2}{P_{r_i}^2}\right) + \\ &\quad + \frac{2}{s^2}\sum_{1\leq i<l\leq s} \mathbb{E}_S\left(\frac{u_{r_i}^{(j)} E^{(r_i)}}{P_{r_i}} \cdot \frac{u_{r_l}^{(j)} E^{(r_l)}}{P_{r_l}}\right)\end{aligned} \tag{5}$$

where

$$\sum_{i=1}^{s} \mathbb{E}_S \left(\frac{\|u_{r_i}^{(j)} E^{(r_i)}\|^2}{P_{r_i}^2} \right) = \sum_{i=1}^{s} \sum_{l=1}^{m} P_l \frac{\|u_l^{(j)} E^{(l)}\|^2}{P_l^2} = s \sum_{l=1}^{m} \frac{\|u_l^{(j)} E^{(l)}\|^2}{P_l} \tag{6}$$

and, using the independence of the r_i's and Equation (3),

$$\begin{aligned} \sum_{1 \le i < l \le s} \mathbb{E}_S \left(\frac{u_{r_i}^{(j)} E^{(r_i)}}{P_{r_i}} \cdot \frac{u_{r_l}^{(j)} E^{(r_l)}}{P_{r_l}} \right) &= \sum_{1 \le i < l \le s} \mathbb{E}_S \left(\frac{u_{r_i}^{(j)} E^{(r_i)}}{P_{r_i}} \right) \cdot \mathbb{E}_S \left(\frac{u_{r_l}^{(j)} E^{(r_l)}}{P_{r_l}} \right) \\ &= \frac{s(s-1)}{2} \|E^T u^{(j)}\|^2. \end{aligned} \tag{7}$$

The substitution of Equations (6) and (7) in (5) gives

$$\mathbb{E}_S \left(\left\| \frac{1}{s} \sum_{i=1}^{s} \frac{u_{r_i}^{(j)}}{P_{r_i}} E^{(r_i)} \right\|^2 \right) = \frac{1}{s} \sum_{i=1}^{m} \frac{\|u_i^{(j)} E^{(i)}\|^2}{P_i} + \frac{s-1}{s} \|E^T u^{(j)}\|^2.$$

Using this in Equation (4) we have

$$\mathbb{E}_S(\|w^{(j)} - \sigma_j v^{(j)}\|^2) = \frac{1}{s} \sum_{i=1}^{m} \frac{\|u_i^{(j)} E^{(i)}\|^2}{P_i} - \frac{1}{s} \|E^T u^{(j)}\|^2,$$

and, using the hypothesis for P_i (Equation (1)), remembering that $u^{(j)}$ is a unit vector and discarding the second term we conclude

$$\mathbb{E}_S(\|w^{(j)} - \sigma_j v^{(j)}\|^2) \le \frac{1}{cs} \|E\|_F^2. \tag{8}$$

Let $\hat{y}^{(j)} = \frac{1}{\sigma_j} w^{(j)}$ for $j = 1, \ldots, r$, let $k' = \min\{k, r\}$ (think of k' as equal to k, this is the interesting case), let $W = \text{span}\{\hat{y}^{(1)}, \ldots, \hat{y}^{(k')}\}$, and $\hat{F} = A \sum_{t=1}^{k'} v^{(t)} \hat{y}^{(t)T}$. We will bound the error $\|A - \pi_W(A)\|_F^2$ using $\hat{F}$. Observe that the row space of $\hat{F}$ is contained in W and π_W is the projection operator onto the subspace of all matrices with row space in W with respect to the Frobenius norm. Thus,

$$\|A - \pi_W(A)\|_F^2 \le \|A - \hat{F}\|_F^2. \tag{9}$$

Moreover,

$$\|A - \hat{F}\|_F^2 = \sum_{i=1}^{r} \|(A - \hat{F})^T u^{(i)}\|^2 = \sum_{i=1}^{k'} \|\sigma_i v^{(i)} - w^{(i)}\|^2 + \sum_{i=k'+1}^{r} \sigma_i^2. \tag{10}$$

Taking expectation and using (8) we get

$$\mathbb{E}_S(\|A - \hat{F}\|_F^2) \le \sum_{i=k+1}^{n} \sigma_i^2 + \frac{k}{cs} \|E\|_F^2 = \|A - \pi_k(A)\|_F^2 + \frac{k}{cs} \|E\|_F^2.$$

This and Equation (9) give

$$\mathbb{E}_S(\|A - \pi_W(A)\|_F^2) \le \|A - \pi_k(A)\|_F^2 + \frac{k}{cs} \|E\|_F^2. \tag{11}$$

Finally, the fact that $W \subseteq V + \text{span}(S)$ and $\dim(W) \leq k$ imply that

$$\|A - \pi_{V+\text{span}(S),k}(A)\|_F^2 \leq \|A - \pi_W(A)\|_F^2,$$

and, combining this with Equation (11), we conclude

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

□

We can now prove Theorem 2 inductively using Theorem 5.

Proof. (of Theorem 2). We will prove the slightly stronger result

$$\mathbb{E}_S(\|A - \pi_{S,k}(A)\|_F^2) \leq \frac{1 - (\frac{k}{cs})^t}{1 - \frac{k}{cs}}\|A - \pi_k(A)\|_F^2 + \left(\frac{k}{cs}\right)^t \|A\|_F^2$$

by induction on t. The case $t = 1$ is precisely Theorem 1.

For the inductive step, let $E = A - \pi_{S_1 \cup \cdots \cup S_{t-1}}(A)$. By means of Theorem 5 we have that,

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

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

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

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

$$\mathbb{E}_S(\|A - \pi_{S_1 \cup \cdots \cup S_t,k}(A)\|_F^2) \leq \|A - \pi_k(A)\|_F^2 + \frac{k}{cs}\mathbb{E}_{S_1,\ldots,S_{t-1}}\left(\|A - \pi_{S_1 \cup \cdots \cup S_{t-1},k}(A)\|_F^2\right)$$

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

3 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 density proportional to the squared norm of each row (as in Theorem 1). Then we focus ourselves on the space orthogonal to the span of the chosen rows, that is, we consider the matrix $E = A - \pi_T(A)$, which represents in some sense the error of our current approximation, and we sample s additional rows with density proportional to the squared norm 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 2 gives a bound on the error of this procedure.

Fast SVD

Input: $A \in \mathbb{R}^{m \times n}$, integers $k \leq m$, t, error parameter $\epsilon > 0$.
Output: $h_1 \ldots, h_k \in \mathbb{R}^n$ such that with probability at least $3/4$ their span V satisfies

$$\|A - \pi_V(A)\|_F^2 \leq \left(1 + \frac{4\epsilon}{1-\epsilon}\right) \|A - \pi_k(A)\|_F^2 + 4\epsilon^t \|A\|_F^2. \tag{12}$$

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

Let M be the number of non-zeros of A.

Theorem 6. *Algorithm* **Fast SVD** *is correct and has running time* $O\left(M\frac{kt}{\epsilon} + (m+n)\frac{k^2t^2}{\epsilon^2}\right)$.

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 2. Also, $\|A - \pi_{S,k}(A)\|_F^2 - \|A - \pi_k(A)\|_F^2$ is a nonnegative random variable and Theorem 2 gives a bound on its expectation:

$$\mathbb{E}_S(\|A - \pi_{S,k}(A)\|_F^2 - \|A - \pi_k(A)\|_F^2) \leq \frac{\epsilon}{1-\epsilon}\|A - \pi_k(A)\|_F^2 + \epsilon^t \|A\|_F^2.$$

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

$$\|A - \pi_V(A)\|_F^2 - \|A - \pi_k(A)\|_F^2 \leq \frac{4\epsilon}{1-\epsilon}\|A - \pi_k(A)\|_F^2 + 4\epsilon^t \|A\|_F^2.$$

which implies inequality (12).

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 of vectors Y, so that it spans the new sample T. The residual squared length 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 $O(ns^2 i)$ (Gram-Schmidt orthonormalization of s vectors in $\mathbb{R}^n$ against an orthonormal basis of size at most $s(i+1)$). Thus, the total time in iteration i is $O(Ms + ns^2 i)$; with t iterations, this is $O(Mst + ns^2 t^2)$. 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 $O(ms^2 t^2)$. Bringing them back to the original basis takes time $O(nkst)$. Thus, the total running time is $O(Mst + ns^2 t^2 + ms^2 t^2 + nkst)$ or, in other words, $O\left(M\frac{kt}{\epsilon} + (m+n)\frac{k^2t^2}{\epsilon^2}\right)$. □

4 Existence of a subset with multiplicative $(1+\epsilon)$ error

In this section, we will prove Theorem 3. The first step is to show that for any matrix A, there is a row a such that the span of a is a factor 2 approximation to the best rank-1 subspace.

Lemma 7. *In any matrix A, there is a row a such that*

$$\|A - \pi_{\text{span}(a)}(A)\|_F^2 \leq 2\|A - \pi_1(A)\|_F^2.$$

Proof. As in the proof of Theorem 5, let b be (the index of) a random row of A picked according to the distribution that assigns probability $P_i = ||A^{(i)}||^2/\|A\|_F^2$ to row i. Define the random vector

$$w = \frac{u_b^{(1)}}{P_b} A^{(b)}$$

where $u^{(1)}$ is the top left singular vector of A. Then, we have $\mathbb{E}(w) = \sigma_1 v^{(1)}$ and, by expanding $\left\|w - \sigma_1 v^{(1)}\right\|^2$, we also have

$$\mathbb{E}\left(\left\|w - \sigma_1 v^{(1)}\right\|^2\right) = \mathbb{E}\left(\left\|\frac{u_b^{(1)} A^{(b)}}{P_b}\right\|^2 - 2\sigma_1 \frac{u_b^{(1)} A^{(b)^T} v^{(1)}}{P_b} + \sigma_1^2\right)$$

and, writing the expectation as a sum and using the definition of P_b,

$$\begin{aligned}
&= \left(\sum_{b=1}^{m} \|A\|_F^2 \frac{\|u_b^{(1)} A^{(b)}\|^2}{\|A^{(b)}\|^2}\right) - \sigma_1^2 \\
&= \|A\|_F^2 - \sigma_1^2 \\
&= \|A - \pi_1(A)\|_F^2.
\end{aligned}$$

Therefore, as in Equations (9) and (10),

$$\mathbb{E}(\|A - \pi_{\text{span}(b)}(A)\|_F^2) \leq \mathbb{E}\left(\left\|w - \sigma_1 v^{(1)}\right\|^2\right) + \sum_{i=2}^{r} \sigma_i^2 \leq 2\|A - \pi_1(A)\|_F^2.$$

and hence there exists a row that proves the lemma. □

Proof. (of Theorem 3.) We will prove by induction on k that for integers s, k and $A \in \mathbb{R}^{m \times n}$ there exists a subset S of rows of A of size $(s+1)k$ such that

$$\|A - \pi_{S,k}(A)\|_F^2 \leq \left(1 + \frac{2}{s}\right)^k \|A - \pi_k(A)\|_F^2.$$

By Lemma 7, there is a row b of A such that

$$\|A - \pi_{\text{span}(b)}(A)\|_F^2 \leq 2\|A - \pi_1(A)\|_F^2.$$

For $k = 1$, apply Theorem 5 with $V = \text{span}\,(b)$; we get that there is a set S of s indices of rows of A such that

$$\begin{aligned}\|A - \pi_{S\cup\{b\},1}(A)\|_F^2 &\leq \|A - \pi_1(A)\|_F^2 + \frac{1}{s}\|A - \pi_{\text{span}(b)}(A)\|_F^2 \\ &\leq \left(1 + \frac{2}{s}\right)\|A - \pi_1(A)\|_F^2.\end{aligned}$$

Now, to extend this to higher k, one might consider the approach of finding such a sample to approximate the best rank 1 subspace, projecting orthogonal to it and repeating. This does not work. The reason is that the error incurred by a higher rank approximation could be much smaller: an ϵ multiplicative error at an earlier stage (e.g., for the first stage, multiplicative *with respect to* $\|A - \pi_1(A)\|_F^2$, the best rank-1 approximation) is already too large (given that at the end we want a multiplicative error *with respect to* $\|A - \pi_k(A)\|_F^2$, the best rank-k approximation).

Instead, for proving the inductive step, we will use the sampling idea backwards. In our context, finding a rank-$(k+1)$ approximation to A is equivalent to finding a $(k+1)$-dimensional subspace V onto which to project, so that $\pi_V(A)$ is the approximation. If we think of our problem as finding such a subspace, then the quality of subspace V is given by $\|A - \pi_V(A)\|_F^2$. We want to find a $(k+1)$-dimensional subspace in the span of $(s+1)(k+1)$ rows of A that is worse than the best possible only by a multiplicative factor of $(1+2/s)^{k+1}$.We will first show the existence of $s+1$ rows such that if we replace $v^{(1)}$ in the best rank-$k+1$ subspace (generated by $v^{(1)}, \ldots, v^{(k+1)}$) with a vector in the span of those rows, then the resulting $k+1$-dimensional subspace is worse that the best by at most a $(1 + 2/s)$ multiplicative factor. Then we will focus on our problem projected onto the orthogonal complement of span(b) to get from the inductive hypothesis that the best k-dimensional subspace it this restricted problem (which happens to be generated by $v^{(2)}, \ldots, v^{(k+1)}$) can be replaced by another k-dimensional subspace in the span of $(s+1)k$ rows, which is worse that the best by a multiplicative factor of at most $(1+2/s)^k$. The combination of these two replacements will give the desired result.

Suppose inductively that for some k and any matrix A and $1 \leq k' \leq k$, there exists a subset of rows of A of size $(s+1)k'$ such that

$$\|A - \pi_{S,k'}(A)\|_F^2 \leq \left(1 + \frac{2}{s}\right)^{k'}\|A - \pi_{k'}(A)\|_F^2.$$

To prove this for $k+1$, let V be the optimal rank-$(k+1)$ subspace. Let V' be the subspace of V orthogonal to $v^{(1)}$. Project the rows of A orthogonal to V' to get a matrix C, i.e.,

$$C = A - \pi_{V'}(A).$$

Applying the hypothesis with $k' = 1$ to C, we get that there exists a vector b in the span of a set S'' of $s+1$ rows of C such that

$$\|C - \pi_{\text{span}(b)}(C)\|_F^2 \leq \left(1 + \frac{2}{s}\right)\|C - \pi_1(C)\|_F^2 = \left(1 + \frac{2}{s}\right)\|C - \pi_{\text{span}(v^{(1)})}(C)\|_F^2. \tag{13}$$

Notice that V' and span(b) are orthogonal subspaces, this implies that

$$\|A - \pi_{V'+\text{span}(b)}(A)\|_F^2 = \|A - \pi_{V'}(A) - \pi_{\text{span}(b)}(A)\|_F^2.$$

This combined with the fact that $\pi_{\text{span}(b)}(A) = \pi_{\text{span}(b)}(C)$ and the definition of C gives

$$\|A - \pi_{V'+\text{span}(b)}(A)\|_F^2 = \|C - \pi_{\text{span}(b)}(C)\|_F^2.$$

We can now apply Equation (13) to get

$$\|A - \pi_{V'+\text{span}(b)}(A)\|_F^2 \leq \left(1 + \frac{2}{s}\right) \|C - \pi_{\text{span}(v^{(1)})}(C)\|_F^2.$$

Use the definition of C again and the fact that $\pi_{\text{span}(v^{(1)})}(C) = \pi_{\text{span}(v^{(1)})}(A)$ to conclude

$$\begin{aligned} \|A - \pi_{V'+\text{span}(b)}(A)\|_F^2 &\leq \left(1 + \frac{2}{s}\right) \|A - \pi_{V'}(A) - \pi_{\text{span}(v^{(1)})}(A)\|_F^2 \\ &= \left(1 + \frac{2}{s}\right) \|A - \pi_V(A)\|_F^2. \end{aligned} \tag{14}$$

Now project A to the subspace orthogonal to b to get a matrix $A' = A - \pi_{\text{span}(b)}(A)$. Applying the inductive hypothesis to A' and $k' = k$, we get a set S' of $(s+1)k$ rows such that (remember that V' is k-dimensional)

$$\|A' - \pi_{S',k}(A')\|_F^2 \leq \left(1 + \frac{2}{s}\right)^k \|A' - \pi_k(A')\|_F^2 \leq \left(1 + \frac{2}{s}\right)^k \|A' - \pi_{V'}(A')\|_F^2. \tag{15}$$

The combination of the rows that we got from the two applications of the induction hypothesis gives us a set $S = S' \cup S''$ of $(s+1)(k+1)$ rows of A, and we will see now that this set proves the induction for $k+1$.

Note that $\pi_{\text{span}(b)}(A) - \pi_{S',k}(A')$ is a matrix of rank at most $k+1$ whose row-space is contained in span (S). This implies

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

using the definition of A':

$$= \|A' - \pi_{S',k}(A')\|_F^2,$$

and using Equation (15):

$$\leq \left(1 + \frac{2}{s}\right)^k \|A' - \pi_{V'}(A')\|_F^2$$

¿From this, the fact that $\pi_{V'}$ as a projection satisfies $\|A' - \pi_{V'}(A')\|_F^2 \leq \|A' - B\|_F^2$ for any matrix $B \in \mathbb{R}^{m \times n}$ having row-space in V' gives us:

$$\|A - \pi_{S,k+1}(A)\|_F^2 \leq \left(1 + \frac{2}{s}\right)^k \|A' - \pi_{V'}(A)\|_F^2,$$

using the definition of A' again:

$$\leq \left(1 + \frac{2}{s}\right)^k \|A - \pi_{\text{span}(b)}(A) - \pi_{V'}(A)\|_F^2,$$

using the orthogonality of span(b) and V':

$$= \left(1 + \frac{2}{s}\right)^k \|A - \pi_{V'+\text{span}(b)}(A)\|_F^2,$$

and using Equation (14),

$$\leq \left(1 + \frac{2}{s}\right)^{k+1} \|A - \pi_V(A)\|_F^2.$$

This finishes the induction. The choice of $s = \frac{4k}{\epsilon}$ gives us the theorem. □

5 Application: projective clustering

In this section, we give an approximation algorithm for the projective clustering problem described in Section 1.1. The algorithm is motivated by Theorem 3. Let $V_1, \ldots, V_j$ be the optimal subspaces partitioning the point set into $P_1, \ldots, P_j$, where P_i is the subset of points closest to V_i. Theorem 3 states that there exists a subset $\hat{P}_i \subseteq P_i$ of size $4k^2/\epsilon$ in whose span lies an approximately optimal k-dimensional subspace W_i. We can enumerate over all combinations of j subsets, each of size $4k^2/\epsilon$ to find the $\hat{P}_i$, but we cannot enumerate the infinitely many k-dimensional subspaces lying in the span of $\hat{P}_i$.

Thus, we cannot hope to find W_i given $\hat{P}_i$. A natural approach to solve this problem would be to put a finite grid on the points in the span of $\hat{P}_i$. The hope would be that there are k grid points $g_1, \ldots, g_k$ whose span G is "close" to W_i, since each basis vector for W_i is close to a grid point. Although G may be "close" to W_i, there may be a point $p \in P_i$ for which $d(p, G)^2 \gg d(p, W_i)^2$, i.e. G incurs a much greater error than W_i. Indeed, consider a point $p \in P_i$ whose distance to the origin is very large. Although the distance between a basis vector and a grid point might be small, the error induced by projecting a point onto the grid point will be proportional to its distance to the origin.

The problem described above implies that a grid construction must be dependent on the point set P_i. Our grid construction does depend on P_i; we consider grid points in balls around a subset of P_i. The grid is laid down in the span of $\hat{P}_i$ (actually, a subspace of slightly higher dimension) to reduce the size of the grid, which is exponential in the dimension of the points. In Lemma 8, we show that there is a subspace F_i that is the span of k grid points such that $\sum_{p \in P_i} d(p, F_i)^2$ is not much worse than $\sum_{p \in P_i} d(p, W_i)^2$. This construction avoids the problem that points far away from the origin will have error proportional to their distance to the origin, since there are grid points close to such points. Although we find F_i in the span of $\hat{P}_i$, it can be brought back up to $\mathbb{R}^d$, where its error with respect to W_i is the same as in $\hat{P}_i$. This is because W_i lies in the span of $\hat{P}_i$.

The algorithm is given below.

Algorithm Cluster

Input: $P \subseteq \mathbb{R}^d$, error parameter $0 < \epsilon < 1$, and B.
Output: A set of j k-dimensional subspaces $F_1 \dots F_j$, such that

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

is at most $(1+\epsilon)B$ provided a solution of cost B exists.

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

2. For each subset T of P of size $8jk^2/\epsilon + jk$:

   (a) For each equipartition of $(T_1 \dots T_j)$ of T into j parts:

      i. Construct a δ-net D_i with radius R for each T_i in the span of T_i.

      ii. For each way of choosing j subspaces $F_1, \dots, F_j$, where F_i is the span of k points from D_i:

         A. Compute the cost $\mathcal{C}(\{F_1 \dots F_j\})$.

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

In Step 2(a)i, we construct a δ-net D_i. A δ-net D with radius R for a point set S is a set of points such that for any point q within distance R of some $p \in S$, there exists a $g \in D$ such that $d(q, g) \leq \delta$. To construct a δ-net D for a point set S with radius R, we simply put a box of side length $2R$ around each point $p \in S$. Each point at grid length $\delta/\sqrt{d}$ in each box is in D, where d is the dimension of the points in S. The size of the δ-net is thus exponential in the dimension of the points in S. This is why it is crucial that we constuct the δ-net for T_i in the span of T_i; if we were to simply create a δ-net for T_i in the original $\mathbb{R}^d$ space, the size of the net would be exponential in d! By constructing the δ-net D_i in the span of T_i in Step 2(a)i, the size of D_i is instead exponential in just $8k^2/\epsilon + k$, the number of points in T_i.

The correctness of the algorithm relies crucially on the next lemma.

Lemma 8. *Let P be a point set, and let W be a subspace of dimension k such that*

$$\sum_{p \in P} d(p, W)^2 \leq \alpha.$$

There exists a set $C \subseteq P$ of k points such that there is a subspace F with the following properties:

1. *F is the span of k points from a δ-net D with radius $\sqrt{\alpha} + 2\delta k$ for C.*

2. *F is not too far from W:*

$$\sum_{p \in P} d(p, F)^2 \leq \sum_{p \in P} d(p, W)^2 + 4k^2 n\delta^2 + 4k\delta \sum_{p \in P} d(p, W). \tag{16}$$

Proof. We simultaneously construct F and choose the k points $p_1, \ldots, p_k$ in C in k steps. Let $F_0 = W$. Inductively, in step i, we choose a point p_i to put in C 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 (16) holds, we prove the following inequality for any point $p \in P$ going from F_{i-1} to F_i

$$d(p, F_i) \leq d(p, F_{i-1}) + 2\delta. \tag{17}$$

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

$$||\pi_{G_i^\perp}(\pi_{F_{i-1}}(p))||$$

and let $g_i \in D$ minimize

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

Put p_i in C. Consider the plane Z defined by $\pi_{G_i^\perp}(g_i), \pi_{G_i^\perp}(\pi_{F_{i-1}}(p_i))$, 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 $G_{i+1} = G_i + \text{span}\{g_i\}$.

Now we prove inequality (17). We do so by proving the following inequality by induction on i for any point p:

$$d(\pi_{F_{i-1}}(p), R_i \pi_{F_{i-1}}(p)) \leq 2\delta. \tag{18}$$

Note that this proves (17) by applying the triangle inequality, since:

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

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,

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

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:

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

By the triangle inequality we have:

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

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

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

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

$$d(\pi_{F_{i-1}}(p_i), g_i) \leq \delta \tag{19}$$

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

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

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 $||\pi_{G_i^\perp}(g_i)|| \geq ||\pi_{G_i^\perp}(\pi_{F_{i-1}}(p^*))|| - \delta$, so

$$d(\pi_{G_i^\perp}(g_i), R_i\pi_{G_i^\perp}(\pi_{F_{i-1}}(p_i))) \leq \delta. \tag{20}$$

Combining (19) and (20), we have proved (18). □

Now, we are ready to prove Theorem 4, which proves the correctness of the algorithm.

Proof. 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 set of points closest to V_i. Let $n_i = |P_i|$, with $\sum_i n_i = n$. By Theorem 3, there exists a subset S_i of P_i of size at most $8k^2/\epsilon$ such that there is a subspace W_i in the span of S_i with

$$\sum_{p \in P_i} d(p, W_i)^2 \leq \left(1 + \frac{\epsilon}{2}\right) \sum_{p \in P_i} d(p, V_i)^2. \tag{21}$$

Consider each subset P_i and its corresponding subspace W_i. Now, apply Lemma 8 to P_i and W_i using R, δ as in the algorithm. Since the optimal solution is of value at most B, we have that, for every i:

$$\sum_{p \in P_i} d(p, W_i)^2 \leq \left(1 + \frac{\epsilon}{2}\right) B.$$

Let C_i, F_i be the set of k points and subspace, respectively, promised by the lemma. F_i is the span of k points from a δ-net of C_i and obeys the following inequality:

$$\sum_{p \in P_i} d(p, F_i)^2 \leq (1 + \frac{\epsilon}{2}) \sum_{p \in P_i} d(p, V_i)^2 + \frac{\epsilon}{4j}B + \frac{\epsilon}{4j}B.$$

Let $T = \bigcup_i S_i \cup C_i$. We have that $|T| = 8jk^2/\epsilon + jk$. The algorithm will enumerate some set $T' \supseteq T$ in Step 2. Furthermore, it will consider a partition $\{T_1' \ldots T_j'\}$ such that $T_i' \supseteq S_i \cup C_i$ in Step 2a.

Let D'_i be the δ-net for T'_i projected to the span of T'_i. Since the algorithm enumerates over all subspaces lying in the span of D'_i, the algorithm will consider the subspaces $F_1, \ldots, F_j$ whose existence is proven above. The cost associated with $F_1, \ldots, F_j$ is:

$$\begin{aligned}
\mathcal{C}(\{F_1 \ldots F_j\}) &\leq \sum_{i=1}^{j} \sum_{p \in P_i} d(p, F_i)^2 \\
&\leq \left(1 + \frac{\epsilon}{2}\right) \sum_{i=1}^{j} \sum_{p \in P_i} d(p, V_i)^2 + \frac{\epsilon}{4} B + \frac{\epsilon}{4} B \\
&\leq (1 + \epsilon) B.
\end{aligned}$$

The running time analysis follows by the following bounds. The number of subsets of size $8jk^2/\epsilon + jk$ is at most $n^{8jk^2/\epsilon + jk}$. The number of equipartitions of a set of size $8jk^2/\epsilon + jk$ into j parts is at most $j^{8jk^2/\epsilon + jk}$. Recall that the δ-net D for a point set T of dimension d is implemented by putting a box with side length $2R$ of grid width $\delta/\sqrt{d}$ around each point in T. 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 $\left(\left(8k^2/\epsilon + k\right)|X|\right)^k$, so the number of j subspaces that one can choose for a partition $(T_1 \ldots T_j)$ is $\left(\left(8k^2/\epsilon + k\right)|X|\right)^{jk}$. The computation of projecting points, finding a basis, and determining the cost of a candidate family of subspaces takes time $O(ndjk)$. The cardinality of X is (remember that $\epsilon < 1$):

$$|X| = \left(\frac{2R}{\delta/\sqrt{8k^2/\epsilon + k}}\right)^{8k^2/\epsilon + k} \leq \left(O\left(jk\sqrt{\frac{n}{\epsilon}}\right)\right)^{8k^2/\epsilon + k}.$$

Therefore, the running time of the algorithm is at most

$$O(ndjk)\, n^{8jk^2/\epsilon + jk} j^{8jk^2/\epsilon + jk} \left(\left(8k^2/\epsilon + k\right)|X|\right)^{jk} = d\left(\frac{n}{\epsilon}\right)^{O(k^3 j/\epsilon)}.$$

□

References

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

[2] P. Agarwal, C. Procopiuc, K. Varadajan. "Approximation Algorithms for k-line center." Proceedings of European Symposium on Algorithms, 2002.

[3] P. Agarwal, S. Har-Peled, K. Varadajan. "Geometric Approximations via Coresets." Manuscript, 2004. `http://valis.cs.uiuc.edu/~sariel/papers/04/survey/`.

[4] P. Agarwal, N. Mustafa. "k-Means Projective Clustering." Proceedings of PODS, 2004.

[5] R. Agarwal, J. Gehrke, D. Gunopulos, P. Raghavan. "Automatic subspace clustering of high dimensional data for data mining applications." Proceedings of SIGMOD, 1998.

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

[7] N. Alon, Y. Matias, M. Szegedy, "The space complexity of approximating the frequency moments." Journal of Computer and System Sciences, 58(1):137-147, Feb. 1999.

[8] M. Bădoiu, S. Har-Peled, P. Indyk. "Approximate Clustering via Core-Sets." Proceedings of 34th Annual Symposium on Theory of Computing, 2002.

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

[10] 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.

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

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

[13] 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.

[14] M. Effros, L. J. Schulman, "Deterministic clustering with data nets," ECCC TR04-050, 2004.

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

[16] A. Frieze, R. Kannan, S. Vempala. "Fast Monte-Carlo algorithms for finding low-rank approximations." Proceedings of 39th FOCS, 1998.

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

[18] S. Har-Peled, S. Mazumdar. "Coresets for k-means and k-median clustering and their applications." Proceedings of the 36th Annual Symposium on Theory of Comptuing, 2004.

[19] S. Har-Peled, K. Varadarajan. "Projective Clustering in High Dimensions using Core-Sets." Proceedings of Symposium on Computation Geometry, 2002.

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

[21] A. Kumar, Y. Sabharwal, S. Sen. "A simple linear time $(1+\epsilon)$-approximation algorithm for k-means clustering in any dimensions." Proceedings of the 45th Annual IEEE Foundations of Computer Science, 2004.

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

[23] N. Megiddo, A. Tamir. "On the complexity of locating linear facilities in the plane." Operations Research Letters, 1 (1982), 194-197.

[24] R. Ostrovsky, Y. Rabani. "Polynomial time approximation schemes for geometric clustering problems." Journal of the ACM, 49(2):139-156, March, 2002.

[25] C. Procopiuc, P. Agarwal, T. Murali, M. Jones. "A Monte Carlo Algorithm for Fast Projective clustering." Proceedings of SIGMOD, 2002.