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 $O(k/\varepsilon)$ rows of the matrix, one can find a low-rank approximation with additive error $\varepsilon||A||_F^2$. Our first result shows that with adaptive sampling in t rounds and $O(k/\varepsilon)$ 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 $O(k^2/\varepsilon)$ rows such that their span contains a rank-k approximation with *multiplicative* $(1 + \varepsilon)$ 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 $||A - D||_F^2 = \sum_{i,j} ||A_{ij} - D_{ij}||_F^2$.

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

The running time of 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.

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

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

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 $O(\min\{m, n\}k^2/\varepsilon^4)$. 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 $(1+\varepsilon)$ 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.

Theorem 2. *Let a random sample $S = (S_1, \ldots, S_t)$ 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*

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

distribution: row i is picked with probability

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

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

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

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. The space used is $O\big((m+n)kt/\varepsilon\big)$.

This theorem 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 $\varepsilon||A||_F^2$ 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 $(1+\varepsilon)$ 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 $4k^2/\varepsilon$ rows in whose span lies a rank-k matrix $\tilde{A}_k$ such that*

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

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 $d(p, F)$ is 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 the theorem, a subset of size $O(k^2/\varepsilon)$ in whose span lies a $(1+\varepsilon)$ approximation to the optimal k-dimensional subspace for this set P_i. So we consider all possible combinations of j subsets each of size $O(k^2/\varepsilon)$, 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 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 $O(k^2/\varepsilon)$ 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 $1/\varepsilon$.

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]. In [12], the "pass-efficient" model of computation is introduced. 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. 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 using SVD. Ostrovsky and Rabani [24] gave the first polynomial time approximation schemes for this problem (and also the j-median problem). Matoušek [22] and Effros and Schulman [14] gave deterministic PTAS's. Fernandez de la Vega et al. [10] describe a randomized algorithm with a running time of $O(n(\log n)^{O(1)})$. Using the idea of core-sets, Har-Peled and Mazumdar [18] showed a $(1+\varepsilon)$ 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 $(1+\varepsilon)$ approximation algorithm for the maximum distance objective. Their algorithm runs in time $dn^{O(jk^6 \log(1/\varepsilon)/\varepsilon^5)}$ and is based on core-sets (see [3] for a survey).

1.3 Notation

Let $A \in \mathbb{R}^{m\times n}$. Let $A^{(i)}$ denote the ith row of A. 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; $\{u^{(1)}, \ldots, u^{(r)}\}$, $\{v^{(1)}, \ldots, v^{(r)}\}$ 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 $1 \leq i \leq r$.

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. For a set S of rows of A, let $\text{span}(S) \subseteq \mathbb{R}^n$ be the subspace generated by them; we use $\pi_{\text{span}(S)}(A)$ and $\pi_{S,k}(A)$ for $\pi_{\text{span}(S),k}(A)$.

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*

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

Then, for any nonnegative integer k,

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

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, $\mathbb{E}_S(w^{(j)}) = \pi_V(A)^T u^{(j)} + E^T u^{(j)} = \sigma_j v^{(j)}$ and

$$\begin{aligned}
\mathbb{E}_S(\|w^{(j)} - \sigma_j v^{(j)}\|^2) &= \mathbb{E}_S\left(\|\frac{1}{s}\sum_{i=1}^{s}\frac{u_{r_i}^{(j)}}{P_{r_i}}E^{(r_i)} - E^T u^{(j)}\|^2\right) \\
&= \mathbb{E}_S\left(\|\frac{1}{s}\sum_{i=1}^{s}\frac{u_{r_i}^{(j)}}{P_{r_i}}E^{(r_i)}\|^2\right) - \|E^T u^{(j)}\|^2 \\
&= \frac{1}{s^2}\sum_{i=1}^{s}\mathbb{E}_S\left(\frac{\|u_{r_i}^{(j)}E^{(r_i)}\|^2}{P_{r_i}^2}\right) + \frac{1}{s^2}\sum_{1\le i<l\le s}\mathbb{E}_S\left(\frac{u_{r_i}^{(j)}E^{(r_i)}}{P_{r_i}}\frac{u_{r_l}^{(j)}E^{(r_l)}}{P_{r_l}}\right) - \|E^T u^{(j)}\|^2 \\
&= \frac{1}{s}\sum_{i=1}^{n}\frac{||u_i^{(j)}E^{(i)}||^2}{P_i^2} - \frac{1}{s}\|E^T u^{(j)}\|^2 \\
&\le \frac{1}{cs}\|E\|_F^2.
\end{aligned} \tag{1}$$

Let $\hat{y}^{(j)} = \frac{1}{\sigma_j}w^{(j)}$, $W = \text{span}\{\hat{y}^{(1)}, \ldots, \hat{y}^{(k)}\}$, 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,

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

Moreover,

$$\|A - \hat{F}\|_F^2 = \sum_{i=1}^{r}\|(A - \hat{F})^T u^{(i)}\|^2 = \sum_{i=1}^{k}\|\sigma_i v^{(i)} - w^{(i)}\|^2 + \sum_{i=k+1}^{r}\sigma_i^2. \tag{3}$$

Taking expectation and using (1) we get

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

This, (2) and the fact that $W \subseteq V + \text{span}(S)$ imply the desired inequality. □

We can now prove Theorem 2 inductively using Theorem 5.

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

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

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

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

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

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

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

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

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

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.

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

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

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 $\pi_S(A)$.

The following argument proves the correctness of the algorithm: $\|A-\pi_V(A)\|_F^2 - \|A-\pi_k(A)\|_F^2$ is a nonnegative random variable and its expectation is bounded by Theorem 2; Markov's inequality applied to this variable and the choice of parameters give inequality (4).

Let M be the number of non-zeros of A. We maintain a basis of 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 $O(ns^2i)$ (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 $O(Ms + ns^2i)$; 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, $O\left(M\frac{kt}{\epsilon} + (m+n)\frac{k^2t^2}{\epsilon^2}\right)$.

4 Existence of a subset with multiplicative $(1+\varepsilon)$ 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 6. *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, if we let b be a random row of A picked according to the distribution that assigns probability $P_i = ||A^{(i)}||^2/\|A\|_F^2$ to row i and 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, similar to Equation (1) of the proof of Theorem 5 (with $c = s = 1$ and $E = A$)

$$\mathbb{E}\left(\left\|w - \sigma_1 v^{(1)}\right\|^2\right) = \frac{\|u_b^{(1)} A^{(b)}\|^2}{P_b} - \|A^T u^{(1)}\|^2 \leq \|A\|_F^2 - \sigma_1^2 = \|A - \pi_1(A)\|_F^2.$$

Therefore, as in Equations (2) and (3),

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

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 of rows of A of size $(s+1)k$ such that

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

Let $v^{(1)}$ be top right singular vector of A. By Lemma 6, 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 are s rows of A, $S = (r_1, \ldots, r_s)$ such that

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

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, and so an ε multiplicative error at an earlier stage (e.g., for the first stage) is already too large.

Instead, we will use the sampling idea backwards. Suppose inductively that for some k and any matrix A and $k \geq k' \geq 1$, 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 $s+1$ rows of C such that

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

and hence,

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

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' we get a set S' of $(s+1)k$ rows such that

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

Therefore, we have a set S of $(s+1)(k+1)$ rows of A for which

$$\begin{aligned}
\|A - \pi_{S,k+1}(A)\|_F^2 &\leq \|A - \pi_{\mathrm{span}(b)}(A) - \pi_{\mathrm{span}(S')\cap\mathrm{span}(b)^\perp,k}(A)\|_F^2 \\
&\leq \|A' - \pi_{S',k}(A')\|_F^2 \\
&\leq \left(1+\frac{2}{s}\right)^k \|A' - \pi_{V'}(A')\|_F^2 \\
&\leq \left(1+\frac{2}{s}\right)^k \|A - \pi_{\mathrm{span}(b)}(A) - \pi_{V'}(A)\|_F^2 \\
&\leq \left(1+\frac{2}{s}\right)^k \|A - \pi_{V'\oplus\mathrm{span}(b)}(A)\|_F^2 \\
&\leq \left(1+\frac{2}{s}\right)^{k+1} \|A - \pi_V(A)\|_F^2.
\end{aligned}$$

The choice of $s = \frac{4k}{\varepsilon}$ 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 $P_1, \ldots, P_j$. Theorem 3 states that there exists a subset $\hat{P}_i \subseteq P_i$ of size $4k^2/\varepsilon$ in whose span lies an approximately optimal k-dimensional subspace W_i. We can enumerate over all combinations of j subsets to find the $\hat{P}_i$, but we cannot enumerate the infinitely many k-dimensional subspaces lying in the span of $\hat{P}_i$. Instead, we construct a δ-net D_i for $\hat{P}_i$ and enumerate all subspaces spanned by any k points in 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 $d(q,g) \leq \delta$.

The algorithm is given below.

Algorithm Cluster

Input: $P \subseteq \mathbb{R}^d$, error parameter $0 < \varepsilon < 1$, and B.
Output: A set of j k-dimensional subspaces $F_1 \ldots F_j$, such that $\mathcal{C}(\{F_1 \ldots F_j\}) \leq (1+\varepsilon)B$ provided a solution of cost B exists.

1. Set $\delta = \frac{\varepsilon\sqrt{B}}{16k\sqrt{(1+\frac{\varepsilon}{2})n}}$, $R = \sqrt{(1+\frac{\varepsilon}{2})B} + 2\delta k$.
2. For each subset S of P of size $8jk^2/\varepsilon$
   (a) For each partition of $(S_1 \ldots S_j)$ of S into j parts of size $8k^2/\varepsilon$
      i. (δ-net) For each S_i, construct a δ-net D_i.
      ii. For each S_i, construct a subspace $F_i = \mathrm{span}\{g_1 \ldots g_k\}$, such that $g_l \in D_i$, for $l = 1, \ldots, k$.
         A. Compute the cost $\mathcal{C}(\{F_1 \ldots F_j\})$.
3. Report the subspaces $F_1 \ldots F_j$ of minimum cost $\mathcal{C}(\{F_1 \ldots F_j\})$.

Note that the δ-net D_i depends on S_i and the input parameter B. If we used a grid that was independent of S_i, the size of the grid would be larger and the error of the best subspace (as compared to W_i) in the grid would depend on the Frobenius norm of the points. We avoid these difficulties by only considering grid points at distance at most R from points $p \in S_i$. Consider when $S_i = \hat{P}_i$. Here, $(1 + \frac{\varepsilon}{2})B$ is an upper bound on the error incurred by W_i, since B is an upper bound on the cost of the optimal solution. Therefore, W_i passes through the ball of radius R around each point p. Having a sufficiently fine grid in each ball ensures that there is a subspace spanned by points in D_i which is approximately as good as W_i. We state this precisely in the next lemma.

Lemma 7. *Let $\hat{P} \subseteq P$, and let W be a subspace of dimension k lying in the span of $\hat{P}$ such that*

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

Let D be a δ-net for $\hat{P}$ with radius $R = \sqrt{\alpha} + 2\delta k$ and gap δ. Then, there exist $g_1 \dots g_k \in D$ such that $F = \text{span}\{g_1 \dots g_k\}$ and

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

Proof. We construct F by modifying W in k steps to include the points $g_1, \dots, g_k$. Let $F_0 = W$. In each step, we apply a rotation R_i to F_{i-1} to obtain F_i, so that it includes $g_1, \dots, g_i$. We set $F = F_k$. We prove the following inequality for any point p for going from F_{i-1} to F_i:

$$d(p, F_i) \leq d(p, F_{i-1}) + 2\delta. \tag{5}$$

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

Let $G_1 = \{\vec{0}\}$. We describe how to construct the rotation R_i. Let $p^* \in \hat{P}$ maximize

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

and let $g_i \in D$ minimize

$$d(\pi_{F_{i-1}}(p^*), g_i).$$

Consider the plane Z defined by $\pi_{G_i^\perp}(g_i), \pi_{G_i^\perp}(\pi_{F_{i-1}}(p^*))$, and $\vec{0}$. Let θ be the angle between $\pi_{G_i^\perp}(g_i)$ and $\pi_{G_i^\perp}(\pi_{F_{i-1}}(p^*))$. Let R_i be the rotation in the plane Z by the angle θ, and define $F_i = R_i F_{i-1}$. Set $G_{i+1} = G_i \oplus \text{span}\{g_i\}$.

Now we prove inequality (5). We do so by proving the following inequality by induction on i for any point p:

$$d(\pi_{F_{i-1}}(p), R_i \pi_{F_{i-1}}(p)) \leq 2\delta. \tag{6}$$

Note that this proves (5) by applying the triangle inequality. The base case of the inequality ($i = 1$) is trivial. For the inductive case, we have that for any point p:

$$\begin{aligned}
d(\pi_{F_{i-1}}(p), R_i \pi_{F_{i-1}}(p)) &= d(\pi_{G_i}(\pi_{F_{i-1}}(p)) + \pi_{G_i^\perp}(\pi_{F_{i-1}}(p)), R_i \pi_{G_i}(\pi_{F_{i-1}}(p)) - R_i \pi_{G_i^\perp}(\pi_{F_{i-1}}(p))) \\
&= d(\pi_{G_i^\perp}(\pi_{F_{i-1}}(p)), R_i \pi_{G_i^\perp}(\pi_{F_{i-1}}(p))) \qquad (7) \\
&\leq d(\pi_{G_i^\perp}(\pi_{F_{i-1}}(p^*)), R_i \pi_{G_i^\perp}(\pi_{F_{i-1}}(p^*))) \qquad (8) \\
&\leq d(\pi_{G_i^\perp}(\pi_{F_{i-1}}(p^*)), \pi_{G_i^\perp}(g_i)) + d(R_i \pi_{G_i^\perp}(\pi_{F_{i-1}}(p^*)), \pi_{G_i^\perp}(g_i)) \\
&\leq 2\delta. \qquad (9)
\end{aligned}$$

The following justifies steps 7, 8, and 9.

(7) $\pi_{G_i}(\pi_{F_{i-1}}(p)) = R_i\pi_{G_i}(\pi_{F_{i-1}}(p))$, because the rotation R_i is chosen orthogonal to G_i.

(8) By construction: For all points p, $||\pi_{G_i^\perp}(\pi_{F_{i-1}}(p^*))|| \geq ||\pi_{G_i^\perp}(\pi_{F_{i-1}}(p))||$.

(9) To bound the first quantity, first note that $d(\pi_{G_i^\perp}(\pi_{F_{i-1}}(p^*)), \pi_{G_i^\perp}(g_i)) \leq d(\pi_{F_{i-1}}(p^*), g_i)$. We show that $d(p^*, F_{i-1}) \leq \sqrt{\alpha} + 2k\delta$; this will show that $d(\pi_{F_{i-1}}(p^*), g_i) \leq \delta$ since $\pi_{F_{i-1}}(p^*)$ is in the ball of radius R around p^*. We have:

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

The third line uses the induction hypothesis.

To bound the second quantity, note that $\pi_{G_i^\perp}(g_i)$ and $R_i\pi_{G_i^\perp}(\pi_{F_{i-1}}(p^*))$ are on the same line. Since rotation preserves norms, we have $||\pi_{G_i^\perp}(\pi_{F_{i-1}}(p^*))|| = ||R_i\pi_{G_i^\perp}(\pi_{F_{i-1}}(p^*))||$. Since $||\pi_{G_i^\perp}(g_i)|| \geq ||\pi_{G_i^\perp}(\pi_{F_{i-1}}(p^*))|| - \delta$, we have the desired bound.

□

Now, we are ready to prove Theorem 4.

Proof. 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 the points of P_i of size at most $8k^2/\varepsilon$ such that there is a subspace W_i in the span of S_i with

$$\sum_{p \in P_i} d(p, W_i)^2 \leq (1 + \frac{\varepsilon}{2}) \sum_{p \in P_i} d(p, V_i)^2. \tag{10}$$

Let $S = \cup_i S_i$. The algorithm will enumerate some set $S' \supset S$ of size $8jk^2/\varepsilon$ in Step 2. Furthermore, it will consider a partition $\{S'_1 \ldots S'_j\}$ such that $S'_i \supset S_i$ in Step 2a.

We can now apply Lemma 7. For each S'_i, we have that there exists a subspace $F_i = \text{span}\{g_1 \ldots g_k\}$, with $g_i \in D_i$, such that:

$$\sum_{p \in P_i} d(p, F_i)^2 \leq \sum_{p \in P_i} d(p, W_i)^2 + 4k^2 n_i \delta^2 + 4k\delta \sum_{p \in P} d(p, W_i).$$

Now, assume a solution of cost B exists, i.e. $B \geq \text{OPT}$. Applying (10) and summing over all the points in P, we have the desired result.

The running time analysis follows by the following bounds. The number of subsets of size $8jk^2/\varepsilon$ is at most $n^{8jk^2/\varepsilon}$. The number of equipartitions of a set of size $8jk^2/\varepsilon$ into j parts is at most $j^{8jk^2/\varepsilon}$. The number of subspaces in each δ-net D_i is at most $\left(\frac{8k^2}{\varepsilon}|D_i|\right)^k$. The number

of subspaces that one can choose for each S_i is $\left(\frac{8k^2}{\varepsilon}|D_i|\right)^{jk}$. The computation of the cost of a candidate family of subspaces takes time $O(ndjk)$. The size of each δ-net is (remember that $\epsilon < 1$):

$$|D_i| = \left(\frac{2R}{2\delta/\sqrt{\frac{8k^2}{\epsilon}}}\right)^{8k^2/\epsilon} \leq \left(O\left(\sqrt{\frac{n}{\epsilon^3}}\right)\right)^{8k^2/\epsilon}.$$

Therefore, the running time of the algorithm is at most

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

□

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 $(1+\varepsilon)$-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.