# Matrix Approximation and Projective Clustering via Volume Sampling*

Amit Deshpande† Luis Rademacher† Santosh Vempala† Grant Wang†

**Abstract**

Frieze et al. [17] proved that a small sample of rows of a given matrix $A$ contains a low-rank approximation $D$ that minimizes $||A - D||_F$ to within small additive error, and the sampling can be done efficiently using just two passes over the matrix [12]. 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. Using this result, we give a pass-efficient algorithm for computing low-rank approximation with reduced additive error. Our second result is that using a natural distribution on subsets of rows (called *volume* sampling), there exists a subset of $k$ rows whose span contains a factor $(k+1)$ relative approximation and a subset of $k + k(k+1)/\epsilon$ rows whose span contains a $1+\epsilon$ relative approximation. The existence of such a small certificate for multiplicative low-rank approximation 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

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 $k$ singular vectors of the matrix is a good choice for many applications. The problem of finding an approximation to this subspace very quickly has received much attention in the past decade [17, 12, 1, 13, 14]. In this paper, we give new algorithms for this problem and show small certificates for better additive approximation as well as relative approximation. At the heart of our analysis is a natural sampling scheme and a small certificate for relative approximation. 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, 29, 6, 4] when $j, k$ 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 [27], even for $k = 1$ [12]). 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 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. [17] showed that any matrix $A$ has a subset of $k/\epsilon$ 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. The approximation is additive. (In what follows, $A^{(i)}$ denotes the $i$th row of $A$ considered as a column vector.)

THEOREM 1.1. ([17]) *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*

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

This can be turned into an efficient algorithm based on sampling [12][1]. The algorithm makes one pass through $A$ to figure out the sampling distribution and another pass to sample and compute the approximation. Its

*An early version of this paper appeared as a technical report [30].

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

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

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.

We begin with the observation that the additive error term drops *exponentially* with the number of passes, while introducing a relative error. 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 length distribution and find a rank-2 subspace using it. Then we sample again, but adaptively – we sample with probability proportional to the squared distance to the first subspace. 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 a good rank 2 approximation. In the 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 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*

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

*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*

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

The proof of Theorem 1.2 is given in Section 2. The resulting algorithm, described in Section 2.1 uses $2t$ passes through the data. 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).

Theorem 1.2 raises the question of whether we can get a purely multiplicative approximation. To answer this question, we generalize the sampling approach. For a subset $S$ of $k$ points, let $\Delta(S)$ denote the $k$-dimensional simplex formed by them along with origin. We pick a random subset $S$ of size $k$ with probability proportional to $\mathrm{vol}(\Delta(S))^2$. This procedure, which we call *volume sampling*, is a generalization of the earlier sampling approach which picks single rows according to their squared lengths.

THEOREM 1.3. *Let $S$ be a random subset of $k$ 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*

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

We prove this theorem in Section 1.4. Moreover, the factor of $k+1$ is the best possible for a $k$-point subset (Proposition 4). By combining Theorem 1.3 with the adaptive sampling idea from Theorem 1.2, we show that there exists a subset of the rows whose span contains a $(1+\epsilon)$ relative matrix approximation.

THEOREM 1.4. *For any $m \times n$ matrix $A$, there exists a subset of $k + k(k+1)/\epsilon$ 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 existence of a small number of rows containing a good relative 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 [13, 12, 17] is that, for any matrix, there is a small subset of the rows whose span contains a good approximation to the rowspace of the entire matrix. This is similar to the idea of coresets [3], which have been studied in computing extent measures in computational geometry (and applied to a variant of the projective clustering problem [22]). Roughly speaking, a coreset is a subset of a point-set such that computing the extent measure on the coreset provides an approximation to the extent measure on the entire point set.

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 of $n$ points $P$ in $\mathbb{R}^d$, find a set of $j$ $k$-dimensional subspaces $F_1,\ldots,F_j$ such that $\sum_{p\in P} \min_i d(p,F_i)^2$ is minimized. When subspaces are replaced by flats, the case $k=0$ corresponds to the $j$-means problem.

Theorem 1.4 suggests an enumerative algorithm. The optimal set of $k$-dimensional subspaces induces a partition $P_1, \ldots, P_j$ of the given point set. Each set $P_i$ contains a subset of size $O(k^2/\epsilon)$ in whose span lies a $(1+\epsilon)$ approximation to the optimal $k$-dimensional subspace for $P_i$. So we consider all possible combinations of $j$ subsets each of size $O(k^2/\epsilon)$, and a $\delta$-net of $k$-dimensional subspaces in the span of each subset. The $\delta$-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 5) when $k, j$ are taken to be fixed constants.

THEOREM 1.5. *Given $n$ points in $\mathbb{R}^d$ and parameters $B$ and $\epsilon$, 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 $B$.*

**1.2 Related work.** The work of Frieze et al. [17] and Drineas et al. [12] 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 [10] 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 [23]. 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 [20], 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 [13, 16]. The "pass-efficient" model of computation was introduced in [23]. Our multipass algorithms fit this model and reveal the tradeoff between approximation and the number of passes. Feigenbaum, et. al [16] show such a tradeoff for computing the maximum unweighted matching in bipartite graphs.

The exponential decrease in the error term that we see in our adaptive sampling result was also observed in orthogonal iteration and Lanczos method [18, 24], but under a somewhat different notion of error.

While the idea of volume sampling appears to be new, in [19] 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 approximations for the Frobenius and 2-norms).

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, 29, 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 [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 $j$ points). Drineas et al. [12] gave a 2-approximation to $j$-means using SVD. Subsequently, Ostrovsky and Rabani [28] gave the first randomized polynomial time approximation schemes for $j$-means (and also the $j$-median problem). Matoušek [26] and Effros and Schulman [15] both gave deterministic PTAS's for $j$-means. Fernandez de la Vega et al. [11] 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 [21] showed a $(1+\epsilon)$ approximation algorithm that runs in linear time for fixed $j, \epsilon$. Kumar et al. [25] 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 $k = 0$, [28, 21]) and maximum distance ($j$-center when $k = 0$, [9]). For general $k$, Har-Peled and Varadarajan [22] give a $(1+\epsilon)$ approximation algorithm for the maximum distance objective function. 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.** Any matrix $A$ has a singular value decomposition, that is, it can be written as $A = \sum_{i=1}^{r} \sigma_i u^{(i)} v^{(i)^T}$ 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. The sets of vectors, $\{u^{(1)}, \ldots, u^{(r)}\} \in \mathbb{R}^m$ and $\{v^{(1)}, \ldots, v^{(r)}\} \in \mathbb{R}^n$, are orthonormal and are called the left and right singular vectors, respectively.

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 $A_k = \pi_{\mathbb{R}^n,k}(A) = \sum_{i=1}^k \sigma_i u^{(i)} {v^{(i)}}^T$ 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 $\mathrm{span}(S) \subseteq \mathbb{R}^n$ 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 $\pi_{\mathrm{span}(S),k}(A)$.

## 2 Improved approximation via adaptive sampling

We will prove Theorem 1.2 in this section. It will be convenient to formulate an intermediate theorem, whose proof is quite similar to one in [17].

THEOREM 2.1. *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*

$$P_i \geq c\|E^{(i)}\|^2/\|E\|_F^2. \tag{2.1}$$

*Then, for any nonnegative integer $k$,*

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

*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, $\mathsf{E}_S(w^{(j)}) = \pi_V(A)^T u^{(j)} + E^T u^{(j)} = \sigma_j v^{(j)}$. Now we will bound $\mathsf{E}_S(\|w^{(j)} - \sigma_j v^{(j)}\|^2)$. 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:

$$\|w^{(j)} - \sigma_j v^{(j)}\|^2 = \left\|\frac{1}{s}\sum_{i=1}^s \frac{u_{r_i}^{(j)}}{P_{r_i}} E^{(r_i)}\right\|^2 - \\ - \frac{2}{s}\sum_{i=1}^s \frac{u_{r_i}^{(j)}}{P_{r_i}} E^{(r_i)} \cdot (E^T u^{(j)}) + \|E^T u^{(j)}\|^2. \tag{2.2}$$

Observe that

$$\mathsf{E}_S\left(\frac{u_{r_i}^{(j)}}{P_{r_i}} E^{(r_i)}\right) = \sum_{i=1}^m P_i \frac{u_i^{(j)}}{P_i} E^{(i)} = E^T u^{(j)}, \tag{2.3}$$

which implies that

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

Using this, apply $\mathsf{E}_S$ to Equation (2.2) to get:

$$\mathsf{E}_S(\|w^{(j)} - \sigma_j v^{(j)}\|^2) \\ = \mathsf{E}_S\left(\left\|\frac{1}{s}\sum_{i=1}^s \frac{u_{r_i}^{(j)}}{P_{r_i}} E^{(r_i)}\right\|^2\right) - \|E^T u^{(j)}\|^2. \tag{2.4}$$

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

$$\mathsf{E}_S\left(\left\|\frac{1}{s}\sum_{i=1}^s \frac{u_{r_i}^{(j)}}{P_{r_i}} E^{(r_i)}\right\|^2\right) \\ = \frac{1}{s^2}\sum_{i=1}^s \mathsf{E}_S\left(\frac{\|u_{r_i}^{(j)} E^{(r_i)}\|^2}{P_{r_i}^2}\right) + \\ + \frac{2}{s^2}\sum_{1\leq i<l\leq s} \mathsf{E}_S\left(\frac{u_{r_i}^{(j)} E^{(r_i)}}{P_{r_i}} \cdot \frac{u_{r_l}^{(j)} E^{(r_l)}}{P_{r_l}}\right) \tag{2.5}$$

where

$$\sum_{i=1}^s \mathsf{E}_S\left(\frac{\|u_{r_i}^{(j)} E^{(r_i)}\|^2}{P_{r_i}^2}\right) = \sum_{i=1}^s \sum_{l=1}^m P_l \frac{\|u_l^{(j)} E^{(l)}\|^2}{P_l^2} \\ = s\sum_{l=1}^m \frac{\|u_l^{(j)} E^{(l)}\|^2}{P_l} \tag{2.6}$$

and, using the independence of the $r_i$'s and Equation (2.3),

$$\sum_{1\leq i<l\leq s} \mathsf{E}_S\left(\frac{u_{r_i}^{(j)} E^{(r_i)}}{P_{r_i}} \cdot \frac{u_{r_l}^{(j)} E^{(r_l)}}{P_{r_l}}\right) \\ = \sum_{1\leq i<l\leq s} \mathsf{E}_S\left(\frac{u_{r_i}^{(j)} E^{(r_i)}}{P_{r_i}}\right) \cdot \mathsf{E}_S\left(\frac{u_{r_l}^{(j)} E^{(r_l)}}{P_{r_l}}\right) \\ = \frac{s(s-1)}{2}\|E^T u^{(j)}\|^2. \tag{2.7}$$

The substitution of Equations (2.6) and (2.7) in (2.5) gives

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

Using this in Equation (2.4) we have

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

and, using the hypothesis for $P_i$ (Equation (2.1)), remembering that $u^{(j)}$ is a unit vector and discarding the second term we conclude that

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

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 $\pi_W$ is the projection operator onto the subspace of all matrices with row space in $W$ with respect to the Frobenius norm. Thus,

$$\|A - \pi_W(A)\|_F^2 \leq \|A - \hat{F}\|_F^2. \tag{2.9}$$

Moreover,

$$\begin{aligned} \|A - \hat{F}\|_F^2 &= \sum_{i=1}^{r} \|(A - \hat{F})^T u^{(i)}\|^2 \\ &= \sum_{i=1}^{k'} \|\sigma_i v^{(i)} - w^{(i)}\|^2 + \sum_{i=k'+1}^{r} \sigma_i^2. \end{aligned} \tag{2.10}$$

Taking expectation and using (2.8) we get

$$\begin{aligned} \mathsf{E}_S(\|A - \hat{F}\|_F^2) &\leq \sum_{i=k+1}^{n} \sigma_i^2 + \frac{k}{cs}\|E\|_F^2 \\ &= \|A - A_k\|_F^2 + \frac{k}{cs}\|E\|_F^2. \end{aligned}$$

This and Equation (2.9) give

$$\mathsf{E}_S(\|A - \pi_W(A)\|_F^2) \leq \|A - A_k\|_F^2 + \frac{k}{cs}\|E\|_F^2. \tag{2.11}$$

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

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

and, combining this with Equation (2.11), we conclude that

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

□

We can now prove Theorem 1.2 inductively using Theorem 2.1.

*Proof.* (of Theorem 1.2) We will prove the slightly stronger result

$$\begin{aligned} &\mathsf{E}_S(\|A - \pi_{S,k}(A)\|_F^2) \\ &\quad \leq \frac{1 - (\frac{k}{cs})^t}{1 - \frac{k}{cs}} \|A - A_k\|_F^2 + \left(\frac{k}{cs}\right)^t \|A\|_F^2 \end{aligned} \tag{2.12}$$

by induction on $t$. The case $t = 1$ is precisely Theorem 1.1.

For the inductive step, let $V = \text{span}(S_1 \cup \cdots \cup S_{t-1})$ and $E = A - \pi_{S_1 \cup \cdots \cup S_{t-1}}(A)$. Applying Theorem 2.1 with $V$ and $E$, we have that,

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

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

$$\begin{aligned} &\mathsf{E}_{S_t}(\|A - \pi_{S_1 \cup \cdots \cup S_t,k}(A)\|_F^2) \\ &\quad \leq \|A - A_k\|_F^2 + \frac{k}{cs}\|A - \pi_{S_1 \cup \cdots \cup S_{t-1},k}(A)\|_F^2. \end{aligned} \tag{2.13}$$

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

$$\begin{aligned} &\mathsf{E}_S(\|A - \pi_{S_1 \cup \cdots \cup S_t,k}(A)\|_F^2) \leq \|A - A_k\|_F^2 + \\ &\quad + \frac{k}{cs} \mathsf{E}_{S_1,\ldots,S_{t-1}}\left(\|A - \pi_{S_1 \cup \cdots \cup S_{t-1},k}(A)\|_F^2\right) \end{aligned} \tag{2.14}$$

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

### 2.1 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 1.1). 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 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 1.2 gives a bound on the error of this procedure. We next give the details of the algorithm; the proof of correctness and running time follow.

**Fast SVD**

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

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 $\|E^{(i)}\|^2/\|E\|_F^2$ to row $i$.
   (c) Let $S = S \cup T$.
3. Let $h_1, \ldots, h_k$ be the top $k$ right singular vectors of $\pi_S(A)$.

THEOREM 2.2. *Algorithm* **Fast SVD** *finds vectors* $h_1 \ldots, h_k \in \mathbb{R}^n$ *such that with probability at least* $3/4$

*their span $V$ satisfies*
(2.15)
$$\|A-\pi_V(A)\|_F^2 \le \left(1+\frac{4\epsilon}{1-\epsilon}\right)\|A-A_k\|_F^2+4\epsilon^t\|A\|_F^2.$$
*The running time is* $O\left(M\frac{kt}{\epsilon}+(m+n)\frac{k^2t^2}{\epsilon^2}\right)$.

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

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

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

$$\|A-\pi_V(A)\|_F^2-\|A-A_k\|_F^2 \le \frac{4\epsilon}{1-\epsilon}\|A-A_k\|_F^2+4\epsilon^t\|A\|_F^2,$$

which implies inequality (2.15).

We will now bound the running time. We maintain a basis of the rows in $S$. 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, $O\left(M\frac{kt}{\epsilon}+(m+n)\frac{k^2t^2}{\epsilon^2}\right)$. □

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

## 3 Volume sampling and relative approximation

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

*Proof.* (of Theorem 1.3.) For every $S\subseteq[m]$, let $\Delta_S$ be the simplex formed by $\{A^{(i)}|i\in S\}$ and the origin, and let $H_S$ be the linear subspace spanned by these rows.

$$\begin{aligned}\sum_{S,|S|=k+1}&\mathrm{vol}_{k+1}(\Delta_S)^2\\ &=\frac{1}{k+1}\sum_{S,|S|=k}\sum_{j=1}^m\frac{1}{(k+1)^2}\ \mathrm{vol}_k(\Delta_S)^2\ d(A^{(j)},H_S)^2\\ &=\frac{1}{(k+1)^3}\sum_{S,|S|=k}\mathrm{vol}_k(\Delta_S)^2\ \sum_{j=1}^m d(A^{(j)},H_S)^2.\end{aligned}$$

Now using Lemma 3.1 (proved next), we can rewrite the LHS above as:

$$\sum_{S,|S|=k+1}\mathrm{vol}_{k+1}(\Delta_S)^2 = \frac{1}{\left((k+1)!\right)^2}\sum_{1\le t_1<\ldots<t_{k+1}\le n}\sigma_{t_1}^2\cdots\sigma_{t_{k+1}}^2.$$

Therefore, we have:

$$\frac{1}{\left((k+1)!\right)^2}\sum_{1\le t_1<\ldots<t_{k+1}\le n}\sigma_{t_1}^2\cdots\sigma_{t_{k+1}}^2 = \frac{1}{(k+1)^3}\sum_{S,|S|=k}\mathrm{vol}_k(\Delta_S)^2\sum_{j=1}^m d(A^{(j)},H_S)^2$$

which means that

$$\begin{aligned}\sum_{S,|S|=k}&\mathrm{vol}_k(\Delta_S)^2\|A-\pi_S(A)\|_F^2\\ &=\frac{k+1}{(k!)^2}\sum_{1\le t_1<\ldots<t_{k+1}\le n}\sigma_{t_1}^2\ldots\sigma_{t_{k+1}}^2\\ &\le\frac{k+1}{(k!)^2}\sum_{1\le t_1<\ldots<t_k\le n}\sigma_{t_1}^2\ldots\sigma_{t_k}^2\sum_{j=k+1}^n\sigma_j^2\\ &=\left(\sum_{S,|S|=k}\mathrm{vol}_k(\Delta_S)^2\right)(k+1)\|A-A_k\|_F^2.\end{aligned}$$

Therefore

$$\frac{1}{\sum_{S,|S|=k}\mathrm{vol}_k(\Delta_S)^2}\sum_{S,|S|=k}\mathrm{vol}_k(\Delta_S)^2\|A-\pi_S(A)\|_F^2 \le (k+1)\|A-A_k\|_F^2.$$

The coefficient of $\|A-\pi_S(A)\|_F^2$ on the LHS is precisely the probability with which $S$ is chosen by volume sampling. Hence, $\mathsf{E}(\|A-\pi_S(A)\|_F^2)\le(k+1)\|A-A_k\|_F^2$. □

LEMMA 3.1.
$$\sum_{S,|S|=k}\mathrm{vol}_k(\Delta_S)^2=\frac{1}{(k!)^2}\sum_{1\le t_1<t_2<\ldots<t_k\le n}\sigma_{t_1}^2\sigma_{t_2}^2\ldots\sigma_{t_k}^2$$
*where* $\sigma_1,\sigma_2,\ldots,\sigma_r>0=\sigma_{r+1}=\cdots=\sigma_n$ *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,

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

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., $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ and 0 with multiplicity $(m-n)$. Moreover the coefficient $c_{m-k}$ can be expressed in terms of these roots as:

$$c_{m-k} = (-1)^{m-k} \sum_{1 \le t_1 < t_2 < \ldots < t_k \le n} \sigma_{t_1}^2 \sigma_{t_2}^2 \ldots \sigma_{t_k}^2$$

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

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

(see e.g., [8]; we state it next as Fact 3.1; a proof appears in the journal version of this paper). This gives us our desired result. $\square$

FACT 3.1. *Let the characteristic polynomial of $M \in \mathbb{R}^{m\times m}$ be $\det(M - \lambda I_m) = (-1)^m\lambda^m + c_{m-1}\lambda^{m-1} + \ldots + c_0$. Then*

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

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

PROPOSITION 4. *Given any $\epsilon > 0$, there exists a $(k+1)\times(k+1)$ matrix $A$ such that for any subset $S$ of $k$ rows of $A$, $\|A - \pi_{S,k}(A)\|_F^2 \ge (1-\epsilon)\ (k+1)\ \|A - A_k\|_F^2$.*

The proof of this proposition, based on an explicit matrix with $k+1$ rows, appears in the journal version of this paper.

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 subset $S_1$ of $k$ rows of $A$ such that $||A - \pi_{S_1}(A)||_F^2 \le (k+1)||A - A_k||_F^2$. Now, applying Theorem 2.1 with $V = \mathrm{span}(S_1)$ and $s = k(k+1)/\epsilon$ we get that a random sample $S_2$ of the rows of $A$ (according to the specified distribution) satisfies $\mathsf{E}_{S_2}(||A - \pi_{V+\mathrm{span}(S_2),k}||_F^2) \le (1+\epsilon)||A - A_k||_F^2$ 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 $|S_1 \cup S_2| = k + k(k+1)/\epsilon$, we have the desired result. $\square$

## 5 Application: projective clustering

In this section, we give a polynomial-time approximation scheme for the projective clustering problem described in Section 1.1. We note that a simple $k+1$ 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 $P_1, \ldots, P_j$, 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., $\sum_{p\in P_i} d(p, V_i')^2 \le (k+1)\sum_{p\in P_i} d(p, V_i)^2$. 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 subset $\hat{P}_i \subseteq P_i$ of size $k+k(k+1)/\epsilon$ in whose span lies an approximately optimal $k$-dimensional subspace $W_i$. We can enumerate over all combinations of $j$ subsets, each of size $k + k(k+1)/\epsilon$ to find the $\hat{P}_i$, but we cannot enumerate the infinitely many $k$-dimensional subspaces lying in the span of $\hat{P}_i$. 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.

Our grid construction also 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 5.1, 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.

**Algorithm Cluster**

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

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 $j(k + 2k(k+1)/\epsilon)$:
   (a) For each equipartition of $(T_1, \ldots, T_j)$ of $T$ into $j$ parts:
      i. For each $i$, construct a $\delta$-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 $\delta$-net $D_i$. A $\delta$-net $D$ with radius $R$ for $S$ is a set such that for any point $q$ for which $d(q,p) \le R$, for some $p \in S$, there exists a $g \in D$ such that $d(q,g) \le \delta$. The size of a $\delta$-net is exponential in the dimension of $S$. This is why it is crucial that we construct the $\delta$-net for $P$ projected to the span of $T_i$. By doing so, we reduce the dimension from $d$ to $O(k^2/\epsilon)$. The correctness of the algorithm relies crucially on the next lemma.

LEMMA 5.1. *Let $\delta > 0$. Let $P$ be a point set, with $|P| = n$ and $W$ be a subspace of dimension $k$ such that $\sum_{p\in P} d(p,W)^2 \le \alpha$. Let $D$ be a $\delta$-net with radius $\sqrt{\alpha}+2\delta k$ for $P$. Then there exists a subspace $F$ spanned by $k$ points in $D$ such that:*

$$\sum_{p\in P} d(p,F)^2 \le \sum_{p\in P} d(p,W)^2 + 4k^2 n\delta^2 + 4k\delta \sum_{p\in P} d(p,W). \tag{5.16}$$

*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 (5.16) holds, we prove the following inequality for any point $p \in P$ going from $F_{i-1}$ to $F_i$

$$d(p, F_i) \le d(p, F_{i-1}) + 2\delta. \tag{5.17}$$

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

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

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 $\theta$ 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 $\theta$, and define $F_i = R_i F_{i-1}$. Set $G_{i+1} = G_i + \mathrm{span}\{g_i\}$.

Now we prove inequality (5.17). We do so by proving the following inequality by induction on $i$ for any point $p$:

$$d(\pi_{F_{i-1}}(p), R_i\pi_{F_{i-1}}(p)) \le 2\delta. \tag{5.18}$$

Note that this proves (5.17) by applying the triangle inequality, since:

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

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,

$$\begin{aligned} &d(\pi_{F_{i-1}}(p), R_i\pi_{F_{i-1}}(p)) \\ &\qquad \le d(\pi_{G_i^\perp}(\pi_{F_{i-1}}(p)), R_i\pi_{G_i^\perp}(\pi_{F_{i-1}}(p))). \end{aligned}$$

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 $p_i$:

$$\begin{aligned} &d(\pi_{G_i^\perp}(\pi_{F_{i-1}}(p)), R_i\pi_{G_i^\perp}(\pi_{F_{i-1}}(p))) \\ &\qquad \le d(\pi_{G_i^\perp}(\pi_{F_{i-1}}(p_i)), R_i\pi_{G_i^\perp}(\pi_{F_{i-1}}(p_i))). \end{aligned}$$

By the triangle inequality we have:

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

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

$$d(\pi_{F_{i-1}}(p_i), g_i) \le \delta \tag{5.19}$$

by construction of the $\delta$-net around $p_i$. We have:

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

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

$$d(\pi_{G_i^\perp}(g_i), R_i\pi_{G_i^\perp}(\pi_{F_{i-1}}(p_i))) \le \delta. \tag{5.20}$$

Combining (5.19) and (5.20), we have proved (5.18). □

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 $k + 2k(k+1)/\epsilon$ such that there is a $k$-dimensional subspace $W_i$ in the span of $S_i$ with

$$\sum_{p\in P_i} d(p, W_i)^2 \le \left(1 + \frac{\epsilon}{2}\right) \sum_{p\in P_i} d(p, V_i)^2. \tag{5.21}$$

Consider $\pi_{\mathrm{span}(S_i)}(P_i)$, the projection of $P_i$ to $\mathrm{span}(S_i)$. Apply Lemma 5.1 to $\pi_{\mathrm{span}(S_i)}(P_i)$ and $W_i$ with $\alpha = (1 + \epsilon/2)B$ and radius $R$ and $\delta$ as in the algorithm. Note that the optimal solution is of value at most $B$, so we have that:

$$\sum_{p\in\pi_{\mathrm{span}(S_i)}(P_i)} d(p, W_i)^2 \le \sum_{p\in P_i} d(p, W_i)^2 \le \left(1 + \frac{\epsilon}{2}\right) B.$$

Let $F_i$ be the subspace spanned by $k$ points from the $\delta$-net $D_i$ for $\pi_{\mathrm{span}(S_i)}(P_i)$ promised by Lemma 5.1. For every $i$, we have that:

$$\sum_{p\in\pi_{\mathrm{span}(S_i)}(P_i)} d(p, F_i)^2 \le \sum_{p\in\pi_{\mathrm{span}(S_i)}(P_i)} d(p, W_i)^2 + \frac{\epsilon}{2j} B. \tag{5.22}$$

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 (5.22) and (5.21), we have:

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

Let $S = \cup_i S_i$. The algorithm will enumerate $S$ in Step 2a, and it will enumerate the partition $(S_1 \ldots S_j)$ in Step 2a. In Step 2(a)i, the algorithm will, for each $i$, construct a $\delta$-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:

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

The number of subsets of size $k + 2k(k+1)/\epsilon$ enumerated by the algorithm is at most $\binom{n}{2(k+1)^2/\epsilon}^j$. A $\delta$-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 $\delta$-net $D_i$ is therefore at most $(n|X|)^k$, so the number of $j$ subspaces that one can choose for a partition $(T_1 \ldots T_j)$ is $(n|X|)^{jk}$. 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 (for $\epsilon < 1$):

$$\begin{aligned} |X| &= \left(\frac{2R}{\delta/\sqrt{2(k+1)^2/\epsilon}}\right)^{2(k+1)^2/\epsilon} \\ &\le \left(O\left(jk\sqrt{\frac{n}{\epsilon}}\right)\right)^{2(k+1)^2/\epsilon}. \end{aligned}$$

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

## 6 Conclusion

Theorem 1.4 shows that there exists a subset of $O(k^2/\epsilon)$ rows in whose span lies a $(1+\epsilon)$ relative approximation to the best rank-$k$ matrix. Many questions arise from this. Can we find this subset and the matrix with multiple passes through the data or in nearly linear time? Is Theorem 1.4 tight, i.e. how many rows are necessary to obtain a $(1+\epsilon)$ relative approximation?

For the projective clustering problem, the exponent in the running time is $jk^3/\epsilon$ for a $(1+\epsilon)$ approximation but only $jk$ for a factor $k+1$ approximation. This suggests that it might be possible to improve the exponent in the PTAS running time.

## 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. Artin. *Algebra*, Prentice-Hall, 1991.

[9] M. Bădoiu, S. Har-Peled, P. Indyk, *Approximate clustering via core-sets*. Proceedings of 34th Annual Symposium on Theory of Computing, 2002.

[10] Z. Bar-Yosseff, *Sampling lower bounds via information theory*. Proceedings of the 35th Annual Symposium on Theory of Computing, 2003.

[11] 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.

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

[13] P. Drineas, R. Kannan, *Pass Efficient algorithm for approximating large matrices*, Proceedings of 14th SODA, 2003.

[14] 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.

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

[16] J. Feigenbaum, S. Kannan, A. McGregor, S. Suri, J. Zhang, *On graph problems in a semi-streaming model*. Proceedings of the 31st ICALP, 2004.

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

[18] G. Golub, C. Van Loan, *Matrix Computations*. John Hopkins University Press, third edition, 1996.

[19] S. A. Goreinov, E. E. Tyrtyshnikov, *The maximal-volume concept in approximation by low-rank matrices* Contemporary Mathematics, Vol. 280 (2001), 47–51.

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

[21] 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 Computing, 2004.

[22] S. Har-Peled, K. Varadarajan, *Projective clustering in high dimensions using core-sets*. Proceedings of Symposium on Computational Geometry, 2002.

[23] M. Henzinger, P. Raghavan, S. Rajagopalan, *Computing on data streams*. Technical Note 1998-011, Digital Systems Research Center, Palo Alto, CA, May 1998.

[24] D. Kempe, F. McSherry, *A decentralized algorithm for spectral analysis*. Proceedings of the 36th Annual Symposium on Theory of Computing, 2004.

[25] 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.

[26] J. Matoušek, *On approximate geometric k-clustering*. Discrete and Computational Geometry, 61–84, 2000.

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

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

[29] C. Procopiuc, P. Agarwal, T. Murali, M. Jones, *A Monte Carlo algorithm for fast projective clustering*. Proceedings of SIGMOD, 2002.

[30] L. Rademacher, S. Vempala, G. Wang, *Matrix approximation and projective clustering via iterative sampling*. MIT-LCS-TR-983, 2005.