Faster coreset construction for subspace and projective clustering

Rameshwar Pratap\textsuperscript{1} and Sandeep Sen\textsuperscript{2}

1 TCS Innovation Labs, India
rameshwar.pratap@gmail.com
2 IIT Delhi, India
ssen@cse.iitd.ernet.in

Abstract
We present randomized coreset constructions for subspace and projective clustering. If $A$ is the input matrix, then our construction relies on projecting $A$ on the approximation of the first few right singular vectors of $A$. We are able to achieve a faster algorithm, as compared to the corresponding deterministic algorithm for the same problem by Fledman et. al.\textsuperscript{9}, while maintaining a desired accuracy. We also complement our theoretical result by supporting experiments.

Digital Object Identifier 10.4230/LIPIcs...

1 Introduction

Recent years have witnessed a dramatic increase in our ability to collect data from various sources like social media platforms, mobile applications, sensors, finance, and biology. This data flood has surpassed our ability to understand, analyse, and process them. Big data is a new terminology that has become quite popular to identify such datasets that are difficult to analyse with the current available technologies.

One possible approach to manage such large volume of the datasets is to keep a succinct summary of the data sets such that it approximately preserves the required properties of the original datasets. This notion was initially formalised by Agrawal et. al.\textsuperscript{1}, in the context of approximating various descriptors of the extent of a point set. They derived $\epsilon$-approximation algorithms for computing smallest enclosing ball, computing diameter, width etc, and coined the term coreset for such summaries of the given input points. Intuitively, a coreset can be considered as a semantic compression of the dataset. Consider a set $Q$ (possibly of infinite size) of query shapes (for example: subspaces, set of points, set of lines etc), then for every shape $q \in Q$, the sum of distances from $q$ to the input points, and the sum of distances from $q$ to the points in the coreset, is approximately the same. If the query set belongs to some particular candidate query set, then such coreset is called as weak coreset (see a survey on weak coreset \textsuperscript{12}); and if the coreset approximates the distances from all possible (potentially infinite) query shapes, then it is called as strong coreset. In many scenarios (including both weak and strong coreset) coresets could be a weighted subset of the input point set - weights represent the relative importance of the points.

Coreset constructions have been studied extensively for various data analysis task. Their construction and analysis techniques mostly includes geometric approximations and linear algebraic approach. Geometric coreset constructions are usually done in an iterative fashion. To start with, some weights are assigned to each point, and in each iteration their weights are refined as per the significance of the points, with respect to the problem. However, in linear algebraic approach, dimensionality reduction step of the coreset construction includes projecting points in a low dimension space such that the original geometry of points is also preserved in a low dimensional space. These projection techniques includes SVD
decomposition, random projections, row/column subset selections, or any combinations of these (see \[2\], \[3\]). See a recent survey article by Jeff M. Phillips \[13\] on coreset and sketches covering the above, and an earlier survey by Agarwal et al. \[2\] on coreset construction for geometric problems.

### 1.1 Our contribution

In this work, we present randomized coreset constructions for subspace and projective clustering. Informally, in $j$-subspace clustering, given a set of $n$ points in $d$-dimensional space, the goal is to find a $j$-dimensional subspace such that it minimizes the sum of square distances over every $j$-dimensional subspace. And in the problem of linear (affine) $(k, j)$-subspace clustering (projective clustering), the goal is to find a closed set $C$ which is the union of $k$ linear (affine) subspaces each of dimension $j$, such that it minimizes the sum of square distances over every possible choice of $C$ (see Definitions \[5\] \[6\] for formal definitions).

Feldman et al. \[9\] presented a deterministic coreset construction for subspace and projective clustering. They build up their coreset by projecting the points on a few first right singular vectors of the input matrix. However, the main drawback of their construction is that it requires to compute the SVD of the given input matrix $A \in \mathbb{R}^{n \times d}$. The complexity of computing SVD is $nd \min \{n, d\}$, which may not be feasible for large values of $n$ and $d$. In this work, we randomize the coreset construction of \[9\]. More precisely, we want to come up with randomized algorithms that output a coreset which satisfies the required properties with high probability, and it does not require computing the SVD of $A$. The central idea of our approach is that using a relative error low rank approximation algorithm, we can obtain a good approximation of the right singular vectors of $A$; and by projecting the points on these approximated right singular vectors of $A$, we develop our coresets.

Algorithm 1 is a randomized algorithm for constructing coreset for subspace clustering. We state our result for subspace clustering as follows and details are described in Theorem 4.

In the following theorems, $\text{nnz}(A)$ denotes the number of non-zero entries of $A$.

#### Theorem 1

Let $A \in \mathbb{R}^{n \times d}$ and $j$ be an integer less than $d - 1$; $\epsilon \in (0, \epsilon')$, with $\epsilon' = \min \left\{ \frac{1}{3j}, \frac{3d}{2d - j} \right\}$. Then there is a randomized algorithm which runs in the expected time $O(n \text{nnz}(A)(\frac{n}{\epsilon^2 m} + m \log m) + (n + d)(\frac{n}{\epsilon^2 m} + m \log m)^2 + nd \text{nnz}(A))$, and outputs a matrix $\tilde{A} \in \mathbb{R}^{m \times m}$ such that for every $j$-dimensional subspace $L$, the following holds with high probability:

$$
(1 - \epsilon)\text{dist}^2(A, L) \leq \text{dist}^2(\tilde{A}, L) \leq (1 + \epsilon)\text{dist}^2(A, L).
$$

Here, $m = O(\frac{n}{\epsilon^2})$; $\text{dist}^2(A, L)$ is the sum of square distances from the rows of $A$ to the subspace $L$; and $\tilde{A}$ is our randomized coreset for $j$-subspace clustering problem.

#### Remark

Our results do not hold for all $\epsilon \in (0, 1)$, but for the smaller values of $\epsilon$. This is purely technical and will be clear later in the proofs.

Algorithm 2 is a randomized algorithm for the dimensionality reduction step of coreset construction for projective clustering. We state our result for projective clustering as follows:

#### Theorem 2

Let $A \in \mathbb{R}^{n \times d}$ and $j, k$ be two integers less than $(d - 1)$, and $(n - 1)$, respectively. Let $\epsilon \in (0, \epsilon')$ with $\epsilon' = \min \left\{ \frac{2}{\sqrt{3j} \sqrt{\frac{26d^2}{\epsilon^2(j + 1)}}, \frac{2}{\sqrt{3j} \sqrt{\frac{26d^2}{\epsilon^2(j + 1)}}} \right\}$, and $j^* = k(j + 1)$. Then there is a randomized algorithm which runs in the expected time $O(n \text{nnz}(A)(\frac{n}{\epsilon^2 m^*} + m^* \log m^*) + (n + d)(\frac{n}{\epsilon^2 m^*} + m^* \log m^*)^2 + nd \text{nnz}(A))$, and outputs a matrix $A^* \in \mathbb{R}^{n \times m^*}$ such that for every non-empty closed set $C$, which is the union of $k$ linear (affine) subspaces each of dimension at most $j$, the following holds with high probability:

$$
|\text{dist}^2(A^*, C) - \text{dist}^2(A, C)| \leq \epsilon\text{dist}^2(A, C).
$$
Where, $m^* = O\left(\frac{d^3}{\epsilon^4}\right)$; and $\text{dist}^2(A, C)$ denotes the sum of square distances from each row of $A$ to its closest point in $C$.

**Remark.** We are able to achieve a faster coreset construction because we develop our coreset by projecting points on a good approximation of right singular vectors of $A$, which can be obtained quickly due to the low rank approximation algorithm of [14], rather than projecting points on the right singular vectors of $A$ (as in [9]). A detailed comparison of the running times is presented later in Sections 3, 4 for subspace and projective clustering, an empirical comparison is also presented in Section 5.

Theorem 2 discusses about the dimensionality reduction step of coreset construction. Further by using techniques from [9, 15], for the points that are on integer grid, and having largest coordinate bounded by a polynomial in $n$ and $d$, we can achieve a sublinear size coreset for projective clustering.

Our work closely resembles the work of Cohen et al. [3] who defined the notion of projection cost preserving sketch which is “typically” the dimensionality reduction step of our coreset construction. For subspace clustering case, their dimensionality reduction is tighter than ours by a constant factor. While our subspace clustering result is used as a base case for the projective clustering result, on the contrary, it is not clear that how the subspace clustering result of [3] can be extended for the projective clustering case.

### 1.2 Related work

Coreset construction has been studied extensively for the problem of $j$-subspace clustering. However, we will discuss a few of them which appear closer to our work. Feldman et al. [6] developed a strong coreset whose size is exponential in $d, j$, logarithmic in $n$, and their coreset construction requires $O(n)$ time. Feldman et al. [8] improved the result of [6] and developed a coreset of size logarithmic in $n$, linear in $d$, and exponential in $j$. However, the construction requires $O(ndj)$ time. Feldman et al. [9] presented a novel coreset construction for subspace and projective clustering. They showed that the sum of square Euclidean distance from $n$ rows of $A \in \mathbb{R}^{n \times d}$ to any $j$-dimensional subspace can be approximated up to $(1 + \epsilon)$ factor, with an additive constant which is the sum of a few last singular values of $A$, by projecting the points on the first $O(\frac{d}{\epsilon})$ right singular vectors of $A$. Recently, Cohen et al. [3] improved the construction of [9] by using only first $\left\lceil \frac{d}{\epsilon} \right\rceil$ right singular vectors, which is an improvement over [9] by a constant factor.

Sariel Har-Peled [10] showed that for projective clustering problem there cannot be any strong coreset of size sublinear in $n$- even for family of pair of planes in $\mathbb{R}^3$. However, Varadarajan et al. [15] showed a sublinear size coreset for projective clustering for a restricted setting when points are on an integer grid, and the largest coordinate of any point is bounded by a polynomial in $n$ and $d$.

### 1.3 Organization of the paper

In Section 2 we begin with the necessary background and give an overview of the problem. In Section 3 we discuss about the coreset construction for subspace clustering problem. In Section 4 we present the coreset construction for projective clustering problem. In Section 5 we complement our theoretical results of Section 3 with encouraging experimental results. In Section 6 we conclude our discussion, and state some possible extensions of the work.
2 Preliminaries

| Notations |
|----------------|
| $A = U\Sigma V^T$ | columns of $U, V$ are orthonormal and called as left and right singular vectors of $A$; $[\Sigma]$ is a diagonal matrix having the corresponding singular values |
| $A^{(m)} = U^{(m)} \Sigma V^T$ | $\Sigma^{(m)}$ is the diagonal having the $m$ largest entries of $\Sigma$, and 0 otherwise |
| $[X]_{d \times j}$ | $j$ orthonormal columns represent a $j$-dimensional subspace $L$ in $\mathbb{R}^d$ |
| $[X^⊥]_{d \times (d-j)}$ | a $(d-j)$ dimensional subspace $L^⊥$ orthogonal to the subspace $L$ |
| $\pi_{V,k}(A)$ | best rank-$k$ approximation of $A$ with its rows projected on the row span of $V$ |
| $A_k$ | the best rank-$k$ approximation of $A$ |

In the following paragraph, we briefly discuss about singular value decomposition of a matrix, and its basic properties. Then, in the following couple of paragraphs, we discuss about the distance of a point, and sum of square distances of the rows of matrix - from a subspace and a closed set.

The Singular Value Decomposition: A matrix $A \in \mathbb{R}^{n \times d}$ of rank at most $r$ can be written due to its SVD decomposition as $A = \Sigma = \sum_{i=1}^{r} \sigma_i u^{(i)} v^{(i)T}$. Here, $u^{(i)}$ and $v^{(i)}$ are $i$-th orthonormal columns of $U$ and $V$ respectively, and $\sigma_1 \geq \sigma_2 \ldots \sigma_{r} \geq 0$. Also, $u^{(i)T} A = \sigma_i v^{(i)}$, and $A v^{(i)} = \sigma_i u^{(i)}$ for $1 \leq i \leq r$. Further, the matrix $A_k$ that minimizes $\|A - B\|_F$ among all matrices $B$ (of rank at most $k$) is given by $A_k = \sum_{i=1}^{k} v^{(i)} u^{(i)T}$, i.e., by projecting $A$ on the first $k$ right singular vectors of $A$. Further, by standard linear algebra, we have $\|A_k\|_F^2 = \sum_{i=1}^{k} \sigma_i^2$, and $\|A - A_k\|_F^2 = \sum_{i=k+1}^{n} \sigma_i^2$.

$l_2$ distances to a subspace: Let $L$ be a $j$-dimensional subspace in $\mathbb{R}^d$ represented by an orthonormal matrix $X \in \mathbb{R}^{n \times d}$. Then, given a point $p \in \mathbb{R}^d$, $||p^T X||_2^2$ is the square of the length of projection of the point $p$ on the subspace $L$. Similarly, given a matrix $A \in \mathbb{R}^{n \times d}$, $||AX||_F^2$ is the sum of square of the length of projection of the points (rows) of $A$ on the subspace $L$. Now, let $L^⊥$ be the orthogonal complement of $L$ represented by an orthonormal matrix $X^⊥ \in \mathbb{R}^{n \times (d-j)}$. Then, $||AX^⊥||_F^2$ is the sum of square of distances of the points of $A$ from the subspace $L$.

$l_2$ distance to a closed set: Let $S \in \mathbb{R}^d$ be a closed set and $p$ be a point in $\mathbb{R}^d$. We define the $l_2$ distance between $p$ and $S$ by $\text{dist}^2(p, S) := \min_{s \in S} \text{dist}^2(p, s)$, i.e., the smallest distance between $p$ and any element $s \in S$. If $S$ consist of union of $k$-dimensional subspaces $L_1, \ldots, L_k$, then $\text{dist}^2(p, S)$ denotes the distance from $p$ to the closest set $S$. Similarly, given a matrix $A \in \mathbb{R}^{n \times d}$, $\text{dist}^2(A, S) := \min_{i=1}^{n} \text{dist}^2(A_r, S)$. Here, $A_r$ denotes the $i$th row of $A$.

In the following, we state some facts from elementary linear algebra which are required for deriving the correctness of our result.

- **Fact 1.** Let $A \in \mathbb{R}^{n \times d}$ and $U \Sigma V^T$ be the SVD of $A$. Then, the first $j$ columns of $V$ span a subspace that minimizes the sum of square distances from all $j$-dimensional subspaces, and this sum is $\sum_{i=j+1}^{d} \sigma_i^2$. Thus, for any $j$-dimensional subspace $X$, we have $||AX||_F^2 \geq \sum_{i=j+1}^{d} \sigma_i^2$.

- **Fact 2.** Let $A \in \mathbb{R}^{n \times d}$, and $X \in \mathbb{R}^{d \times k}$ be an orthonormal matrix. Then, by elementary linear algebra we can have, $||A - AX X^T||_F^2 = ||A||_F^2 - ||AX||_F^2$.

- **Fact 3.** Let $AX$ be the projection of points of $A$ on the $j$-dimensional subspace $L$ represented by a matrix $X$. We can also write the projection of the points in the rows of $A$ to $L$ as $AX X^T$, these projected points are still $d$-dimensional, but lie within the $j$-dimensional subspace. Further, $||AX||_F^2 = ||AX X^T||_F^2$.

- **Fact 4 (Cauchy–Schwarz inequality).** Let $A \in \mathbb{R}^{p \times q}$ and $B \in \mathbb{R}^{q \times r}$ be the two matrices, then due to Cauchy-Schwarz inequality, we have $||AB||_F^2 \leq ||A||_F^2 ||B||_F^2$. 
In the following, we state the definitions of subspace and projective clustering.

Definition 5 (Subspace clustering). Let $A \in \mathbb{R}^{n \times d}$ and $j$ be an integer less than $d$. Then, the problem of $j$-subspace clustering is to find a $j$-dimensional subspace $L$ of $\mathbb{R}^d$ that minimizes the dist$^2(A, L)$.

Definition 6 (linear (affine) $(k, j)$-subspace clustering - projective clustering). Let $A \in \mathbb{R}^{n \times d}$, $j$ be an integer less than $d$, and $k$ be an integer less than $n$. Then, the problem of linear (affine) $(k, j)$-subspace clustering is to find a closed set $C$, which is the union of $k$ linear (affine) subspaces $\{L_1, \ldots, L_k\}$ each of dimension at most $j$, such that it minimizes the dist$^2(A, C)$, over every possible choice of $C$.

Definition 7 (Coreset for $C$-clustering [11]). Let $C$ be a family of closed set in $\mathbb{R}^d$. Let $A \in \mathbb{R}^{n \times d}$, $r$ be an integer less than $n$, and $\epsilon > 0$. We define $S \in \mathbb{R}^{r \times d}$ with a set of weights $\{w_i\}_{i=1}^r > 0$ associated with its rows (let we denote the $i$ th row of $S$ by $S_i$), is an $\epsilon$-coreset for $C$-clustering if for every $C \in C$, the following condition satisfies:

$$(1 - \epsilon)\text{dist}^2(A, C) \leq \sum_{i=1}^r w_i \text{dist}^2(S_i, C) \leq (1 + \epsilon)\text{dist}^2(A, C).$$

In the following, we state two technical lemmas from the subspace clustering result of [9], which we will be required for our analysis purpose:

Lemma 8 (Lemma 6.1 of [9]). Let $A \in \mathbb{R}^{n \times d}$ matrix, and let $X \in \mathbb{R}^{d \times j}$ be a matrix having orthonormal columns. Let $\epsilon \in (0, 1]$ and $m = \min\{j + \lceil \frac{\epsilon}{2} \rceil - 1, d - 1\}$. Then, $0 \leq ||AX||_F^2 - ||A^{(m)}X||_F^2 \leq \epsilon^2 \sum_{i=m+1}^d \sigma_i^2$.

Lemma 9 (Lemma 6.2 of [9]). Let $A \in \mathbb{R}^{n \times d}$, and let $X^\perp \in \mathbb{R}^{d \times (d-j)}$ be a matrix having orthonormal columns. Let $\epsilon \in (0, 1]$ and $m = \min\{j + \lceil \frac{\epsilon}{2} \rceil - 1, d - 1\}$. Then, $||AX^\perp||_F^2 \leq ||A^{(m)}X^\perp||_F^2 + ||A - A^{(m)}||_F^2 \leq (1 + \epsilon)||AX^\perp||_F^2$.

Theorem 3 (Relative low-rank approximation by Sarlós [14]). Let $A \in \mathbb{R}^{n \times d}$, and $\pi(.)$ denote the projection operators stated in the notation table. If $\epsilon \in (0, 1]$ and $S$ is an $(r \times n)$ Johnson-Lindenstrauss matrix with i.i.d. zero-mean ±1 entries and $r = \theta((\frac{1}{\epsilon} + k \log k) \log \frac{1}{\delta})$, then with probability at least $1 - \delta$ it holds that

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

Further, computing the singular vectors spanning $\pi_{S,A,k}(A)$ in two passes over the data requires $O(nnz(A)r + (n + d)r^2)$ time, where $nnz(A)$ denotes the number of non-zeroes in $A$.

For our analysis, we will use a weak triangle inequality which can be stated as follows:

Lemma 10 (Lemma 7.1 of [9]). For any $0 < \epsilon < 1$, a closed set $C$, and two points $p, q \in \mathbb{R}^d$,

$$|\text{dist}^2(p, C) - \text{dist}^2(q, C)| \leq \frac{12||p - q||^2}{\epsilon} + \frac{\epsilon}{2}\text{dist}^2(p, C).$$

### 3 Faster coreset construction for subspace clustering

Here, we present a randomized algorithm for the coreset construction for $j$-subspace clustering. We state our coreset as a randomized coreset as it satisfies the required coreset properties with high probability. As discussed in the introduction, we randomize the coreset construction.

---

1 Two passes are required as we first multiply $A$ on the right with a Johnson-Lindenstrauss matrix $S$, and then we project the rows of $A$ again onto the row span of $SA$. 
of \cite{9}. Their construction relies on computing the SVD of the input matrix \( A \), and projecting the input points (rows of \( A \)) on the first few right singular vectors of \( A \). In order to randomize the construction of \cite{9}, we use an algorithm by Sarlós \cite{14} for approximating the right singular vectors of \( A \). Further, using analysis techniques similar to \cite{9,3} on these approximated right singular vectors of \( A \), we develop our coreset. We present our algorithm as follows:

1. **Input**: Input matrix \( A \in \mathbb{R}^{n \times d} \), an integer \( 1 \leq j < d - 1 \), error probability \( \delta \in (0, 1) \), \( \epsilon \in (0, \epsilon') \), where \( \epsilon' = \min \left( \frac{1}{\sqrt{5}}, \frac{3}{6(d - j + 1)} \right) \).
2. **Result**: Randomized coreset for \( j \)-subspace clustering problem.
3. Compute a Johnson-Lindenstrauss matrix \( [S]_{r \times n} \) having i.i.d. \pm 1 entries and zero-mean, with \( r = O((\frac{n}{\epsilon^3} + m \log m) \log \frac{1}{\delta}) \), and \( m = \min\{j + \lceil \frac{1}{\epsilon^2} \rceil - 1, d - 1\} \).
4. Compute the projection matrix \( \pi_{SA}(A) \) (see Notation.)
5. Compute the SVD of \( \pi_{SA}(A) \), let \( R' \in \mathbb{R}^{d \times m} \) be the first \( m \) right singular vectors of \( \pi_{SA}(A) \).
6. Let us denote \( A'R'R^T \) by \( \tilde{A} \). If the SVD of \( \tilde{A} \) is \( \tilde{U}\tilde{\Sigma}\tilde{V}^T \), then output \( \tilde{U}\tilde{\Sigma}\tilde{V}^T \).

**Algorithm 1**: Algorithm for computing randomized coreset for \( j \)-subspace clustering.

The following theorem discusses about the bounds achieved by Algorithm 1.

**Theorem 4.** Let \( X \in \mathbb{R}^{d \times j} \) be an orthonormal matrix representing a subspace \( L \), let \( X^\perp \in \mathbb{R}^{d \times (d-j)} \) be the orthonormal matrix representing the subspace orthogonal to \( L \), and \( \epsilon \in (0, \epsilon') \), with \( \epsilon' = \min \left( \frac{1}{\sqrt{5}}, \frac{3}{6(d - j + 1)} \right) \). Then in Algorithm 1, the following holds true with probability at least \( 1 - \delta \):

\[
(1 - \epsilon)||AX^\perp||_F^2 \leq ||\tilde{A}X^\perp||_F^2 \leq (1 + \epsilon)||AX^\perp||_F^2.
\]

**Expected time bound of Algorithm 1** is \( O(nnz(A)(\frac{n}{\epsilon^3} + m \log m) + (n + d)(\frac{n}{\epsilon^3} + m \log m)^2 + ndm) \).

**Proof.** First using a result of Sarlós \cite{14} we get a rank \( m \) approximation of \( A \). If \( S \) is an \((r \times n)\) JL matrix as described in Algorithm 1 then due to the result of \cite{14} (also stated in preliminaries, see Theorem 3), the following holds true with probability at least \( 1 - \delta \):

\[
||A - \pi_{SA,m}(A)||_F \leq (1 + \epsilon)||A - A_m||_F.
\]

Here, \( A_m \) is the best \( m \) rank approximation of \( A \). If \( V' \) is the matrix having the first \( m \) right singular vectors of \( A \), then \( A_m = AV'V'^T \). Let \( R' \) be the matrix having the first \( m \) right singular vectors of \( \pi_{SA}(A) \), and let we denote \( A'R'R^T \) by \( \tilde{A} \), then by Equation 1, the following holds true with probability at least \( 1 - \delta \):

\[
||A - \tilde{A}||_F \leq (1 + \epsilon)||A - A_m||_F,
\]

\[
||A - \tilde{A}||_F^2 \leq (1 + \epsilon)^2||A - AV'V'^T||_F^2,
\]

\[
||A||_F^2 - ||\tilde{A}||_F^2 \leq (1 + \epsilon)^2 \left( ||A||_F^2 - ||AV'V'^T||_F^2 \right)
\]

from Fact 2.

\[
||AV'V'^T||_F^2 - ||\tilde{A}||_F^2 \leq (\epsilon^2 + 2\epsilon)||A - AV'V'^T||_F^2
\]

\[
||AV'V'^T||_F^2 - ||\tilde{A}||_F^2 \leq 3\epsilon||A - AV'V'^T||_F^2
\]

Now, Equation 2 along with Lemma 11 completes a proof of the theorem.

We give an expected time bound of Algorithm 1. Time required for execution of line number 3, 4, 5 is \( O(nnz(A)(\frac{n}{\epsilon^3} + m \log m) + (n + d)(\frac{n}{\epsilon^3} + m \log m)^2) \) due to \cite{14}. Further, line number 5 require time for the two subroutines- firstly for projecting \( A \) on \( R' \), which requires \( O(ndm) \) time due to an elementary matrix multiplication; and secondly for computing
We now give a lower bound on the expression:

\[ \text{that for Equation 2.} \]

Thus, the total expected running time of Algorithm 1 is \( O(nm^2) \) time. Note that in line number 6, we compute the SVD of \( AR' \) to get the coreset, we can obtain the same coreset from the SVD of \( AR' \) also, due to Fact [3].

\[ \text{Equality 10 follows from pythagoras theorem; Inequality 11 holds from Lemma 8; Inequality 12 holds from Fact 1; Inequality 6 holds from Fact 4; Inequality 7 holds from Equation 2.} \]

**Proof.** We first show an upper bound on the expression:

\[ \| \tilde{AX}^\perp \|^2_F \leq (1 + \epsilon) \| AX^\perp \|^2_F. \]

Equality [3] follows from pythagoras theorem; Inequality [4] holds from Lemma [8]; Inequality [5] holds from Fact [4]; Inequality [6] holds from Fact [3]; Inequality [7] holds from Equation [2].

Further, as \( \epsilon \leq \frac{1}{3d} \), then in Equation [8], we have the following:

\[ \| \tilde{AX}^\perp \|^2_F \leq (1 + \epsilon) \| AX^\perp \|^2_F. \]

We now give a lower bound on the expression:

\[ \| \tilde{AX}^\perp \|^2_F - \| AX^\perp \|^2_F \]

\[ \| \tilde{AX}^\perp \|^2_F - \| AX^\perp \|^2_F - \| A \|^2_F + \| AX \|^2_F \]

\[ \| AV'\tilde{V'}^T X \|^2_F - \| \tilde{AX} \|^2_F - \| AX \|^2_F - \| A \|^2_F - \| A \|^2_F + \| AX \|^2_F \]

\[ \geq \| AV'\tilde{V'}^T X \|^2_F - \| \tilde{AX} \|^2_F + \| AX \|^2_F - \| A \|^2_F \]

\[ \| \tilde{AX} \|^2_F - \| A \|^2_F \]

\[ \geq -3\epsilon \| A - A\|^2_F + \| A\|^2_F - \| A \|^2_F \]

\[ = -(1 + 3\epsilon) \| A - A\|^2_F \]

Equality [10] follows from pythagoras theorem; Inequality [11] holds from Lemma [8]; Inequality [12] holds from Equation [2]. The remaining proof is presented in the appendix, where we show that for \( \epsilon \in (0, \epsilon') \) with \( \epsilon' < \frac{3j}{3d-3j} + 1 \), we have the following:

\[ \| \tilde{AX}^\perp \|^2_F - \| AX^\perp \|^2_F \geq -\epsilon \| AX^\perp \|^2_F. \]

Thus, we are done with Equations [9] and [14].
Faster coreset construction for subspace and projective clustering

Remark. In Algorithm 1, \( \tilde{\Sigma}V^T \) is a randomized coreset for subspace clustering—since \( \tilde{U} \) is an orthonormal matrix, we have \( ||\tilde{U}\tilde{\Sigma}V^TX||_F^2 = ||\tilde{\Sigma}V^TX||_F^2 \), and further by Theorem 4, \( \tilde{\Sigma}V^T \) approximately preserve the distances from every \( j \)-dimension subspace \( \epsilon \)-h.p.

Remark. We will compare and contrast our result with that of [9]. As stated earlier, coreset of [9] obtained by projecting \( \mathcal{A} \) on the first few right singular values of \( \mathcal{A} \), which is of complexity \( \min\{n^2d, nd^2\} \). While we develop our coreset by projecting \( \mathcal{A} \) on orthogonal vectors which are good approximation of right singular vectors of \( \mathcal{A} \). We obtain them by computing the right singular vectors of the projection matrix \( \pi_{\mathcal{A}}(\mathcal{A}) \), which (due to [1]) have complexity \( O(\min\{n\Sigma(\mathcal{A})(\frac{m}{\epsilon^2} + m\log m) + (n + d)(\frac{m}{\epsilon^2} + m\log m)^2\}) \), where \( m = O(\frac{1}{\epsilon^2}) \).

Remark. The reason for choosing algorithm of [14] for low rank approximation, over other similar work such as [5], as to the best of our knowledge, it is simple and easy to implement; require fewer passes over the input data; and have a less dependency on the \( \epsilon \).

4 Faster coreset construction for projective clustering

In this section, we present a randomized coreset construction for the problem of projective clustering. This is essentially a generalization of the result of Section 3. To be precise, let \( L_1, \ldots, L_k \) be a set of \( k \) subspaces each of dimension at most \( j \), and let \( \mathcal{C} \) be a closed set containing union of them, then our randomized coreset approximately preserve the distances from every such closed set \( \mathcal{C} \), with high probability. Our main contribution is the dimensionality reduction step of the coreset construction which is presented in Algorithm 2.

Feldman et al. presented the dimensionality reduction step of deterministic coreset construction for projective clustering (see Theorem 8.1 of [9]). Their construction relies on projecting the input points (row of matrix \( A \)) on the first \( O\left(\frac{k(j+1)}{\epsilon^2}\right) \) right singular vectors of \( A \). We randomize their construction by using the result of [14], and able to achieve a faster construction of this step. Let \( L^\ast \) be a \( j^\ast \) (with, \( j^\ast = k(j + 1) \)) dimensional subspace containing a closed set \( \mathcal{C} = L_1 \cup \ldots \cup L_k \). Then, the following algorithm outputs a matrix \( A^\ast \in \mathbb{R}^{n \times O\left(\frac{j^\ast}{\epsilon^2}\right)} \) whose dimension is much smaller than \( d \) (independent of \( d \)), and \( \epsilon \)-h.p. it approximately preserve the distances from every closed set \( \mathcal{C} \).

Algorithm 2: Algorithm for dimensionality reduction for projective clustering.

The following theorem establishes the correctness of the above algorithm.

Theorem 5. In Algorithm 2, for any non-empty closed set \( \mathcal{C} \) which is contained in a \( j^\ast \)-dimensional subspace, \( \epsilon \in (0, \epsilon') \) with \( \epsilon' = \min\left\{ \frac{2}{3j^\ast}, \sqrt{\frac{20j^\ast}{d-j^\ast-0.6}} \right\} \), the following is true with probability at least \( 1 - \delta \):

\[ |\text{dist}^2(A^\ast, \mathcal{C}) - \text{dist}^2(A, \mathcal{C})| \leq \epsilon \text{dist}^2(A, \mathcal{C}). \]
**Expected time bound of Algorithm 2** is \(O(\text{nnz}(A) (\frac{2n^*}{\epsilon} + m^* \log m^*) + (n+d)(\frac{2n^*}{\epsilon} + m^* \log m^*)^2 + ndm^*)\).

**Proof.** Let \([X^*]_{d \times j^*}\) be a matrix with orthonormal columns whose span is \(L^*\), and let \(L^\perp\) be the orthogonal complement of \(L^*\) spanned by \([X^\perp]_{d \times (d-j^*)}\). In our analysis, we use the equality, \(\text{dist}^2(A, C) = ||AX^\perp||_F^2 + \text{dist}^2(AX^*X^T, C)\), which holds true due to the Pythagoras theorem. We have

\[
|\text{dist}^2(A^*, C) - \text{dist}^2(A, C)| = \left|\left(||A^*X^\perp||_F^2 + \text{dist}^2(A^*X^*X^T, C)\right) - \left(||AX^\perp||_F^2 + \text{dist}^2(AX^*X^T, C)\right)\right|
\]

\[
= \left(||A^*X^\perp||_F^2 - ||AX^\perp||_F^2 + \text{dist}^2(A^*X^*X^T, C) - \text{dist}^2(AX^*X^T, C)\right) = \left(||A^*X^\perp||_F^2 - ||AX^\perp||_F^2\right) + \left(\text{dist}^2(A^*X^*X^T, C) - \text{dist}^2(AX^*X^T, C)\right)
\]

We have two terms to bound in the above expression. The first term can be upper bounded using a similar analysis as of Lemma 11 (see Equations 8-13), which holds true with probability at least \(1 - \delta\). (In Equations 8-13 we replace \(j\) by \(j^*\), \(m\) by \(m^*\), \(\epsilon\) by \(\frac{\epsilon}{26}\).) In the remaining proof of the theorem, we denote \(\|A - A^{(m)}\|_F^2\) by \(\Delta^*\).

\[
\left(||A^*X^\perp||_F^2 - ||AX^\perp||_F^2\right) \leq \max\left\{\frac{\epsilon^2}{26}||AX^\perp||_F^2 + (3j^*\frac{\epsilon^2}{26} - 1)\Delta^*, (1 + \frac{3\epsilon^2}{26})\Delta^*\right\} \quad (15)
\]

In order to bound the second term \(\text{dist}^2(A^*X^*X^T, C) - \text{dist}^2(AX^*X^T, C)\), we use a triangle inequality from Lemma 10. For any \(\epsilon \in (0, 1)\) and from Lemma 10, we have

\[
\left|\text{dist}^2(A^*X^*X^T, C) - \text{dist}^2(AX^*X^T, C)\right| \leq \frac{12}{\epsilon^2} \left(||A^*X^\perp||_F^2 + 3j^*\frac{\epsilon^2}{26}\Delta^*\right) + \frac{\epsilon}{2} \text{dist}^2(AX^*X^T, C) \quad (16)
\]

Inequality 16 holds due to Lemma 12. Thus, we have

\[
\left|\text{dist}^2(A^*X^*X^T, C) - \text{dist}^2(AX^*X^T, C)\right| \leq \frac{12}{\epsilon^2} \left(||AX^\perp||_F^2 + 3j^*\frac{\epsilon^2}{26}\Delta^*\right) + \frac{\epsilon}{2} \text{dist}^2(AX^*X^T, C) \quad (17)
\]

We have two cases in Equation 15 and we consider them one by one.

**Case 1:** \(||A^*X^\perp||_F^2 - ||AX^\perp||_F^2\) \(\leq \frac{\epsilon^2}{26}||AX^\perp||_F^2 + (3j^*\frac{\epsilon^2}{26} - 1)\Delta^*\).

The above expression in conjunction with Equation 17 give us the following:

\[
\left|\text{dist}^2(A^*, C) - \text{dist}^2(A, C)\right| \leq \left(1 + \frac{12}{\epsilon^2}\right) \frac{\epsilon^2}{26} ||AX^\perp||_F^2 + \frac{\epsilon}{2} \text{dist}^2(AX^*X^T, C) + \left(3j^*\frac{\epsilon^2}{26} (1 + \frac{12}{\epsilon^2})\right) \Delta^*
\]

Further steps are mentioned in the appendix, where we are able to show that for any \(\epsilon < \frac{1}{30}\), the following holds with probability at least \(1 - \delta\)

\[
|\text{dist}^2(A^*, C) - \text{dist}^2(A, C)| \leq \epsilon \text{dist}^2(A, C).
\]
Now, we address the second case of Equation \[15\] which is stated as follows.

**Case 2:** \[
\|A^*X^+\|_F^2 - |AX^+|_F^2 \leq \left(1 + \frac{26}{26}\right)\Delta^*.
\]

We obtain the following due to the above expression along with Equation 17

\[
|\text{dist}^2(A^*, C) - \text{dist}^2(A, C)| \leq \left(1 + \frac{26}{26}\right)\Delta^* + \frac{3\epsilon^2}{2}\text{dist}^2(A, C) + \frac{12}{\epsilon} \left(\frac{\epsilon^2}{26}\|AX^+\|_F^2 + 3j^*\epsilon^2\Delta^*\right).
\]

Further calculations of this case are presented in the appendix, where we are able to show that

\[
|\text{dist}^2(A^*, C) - \text{dist}^2(A, C)| \leq \epsilon\text{dist}^2(A, C).
\]

Case 1 and Case 2 complete a proof of the theorem, and expected running time of Algorithm [2] follows similar to that of Theorem [3].

\[\blacktriangleleft\]

A proof of the following lemma is presented in the appendix.

**Lemma 12.** Let \(X^* \in \mathbb{R}^{d \times j}\) be a matrix with orthonormal columns whose span is \(L^*\), then in Algorithm [2] the following is true with probability at least \(1 - \delta\)

\[
\|A^*X^+X^T - AX^+X^T\|_F^2 \leq \frac{\epsilon^2}{26}\|AX^+\|_F^2 + 3j^*\frac{\epsilon^2}{26}\|A - A(m^*)\|_F^2.
\]

**Remark.** The dimension of \(A^*\) is \(m^*\) which is \(O\left(\frac{k(j+1)}{\epsilon^2}\right)\). Dimensionality reduction step of coreset construction (for projective clustering) of [9] requires computing the SVD of \(A\). While here (similar to discussed in Section [3]), the complexity of dimensionality reduction step of our coreset construction is \(O(nmz(A)(\frac{m^*}{\epsilon} + m^* \log m^*) + (n + d)(\frac{m^*}{\epsilon} + m^* \log m^*)^2)\).

As a corollary of Theorem [3] and using the known techniques from [9, 15, 7] on \(A^*\), we present the cardinality reduction step of coreset construction as follows:

**Corollary 13 (Corollary 9.1 of [9]).** Let \(A \in \{1, 2, \ldots, \Lambda\}^{n \times d}\), with \(\Lambda \in (nd)^{O(1)}\), \(d \in n^{O(1)}\). There is a matrix \(Q \in \mathbb{R}^{n \times d'}\) with \(d' = \text{poly}(2^{j^*}, \frac{1}{\epsilon}, \log n, \log \Lambda)\), \(d' = O\left(\frac{k(j+1)}{\epsilon^2}\right)\), and a weight function associated with the rows of \(Q\), i.e. \(w : Q_{i,:} \rightarrow [0, \infty]\) such that for every closed set \(C\), which is the union of \(k\) affine \(j\)-subspaces of \(\mathbb{R}^d\), the following holds with high probability

\[
(1 - \epsilon)\Sigma_{i=1}^n \text{dist}^2(A_{i,:}, C) \leq \Sigma_{i=1}^n w(Q_{i,:})\text{dist}^2(Q_{i,:}, C) \leq (1 + \epsilon)\Sigma_{i=1}^n \text{dist}^2(A_{i,:}, C).
\]

### 5 Experiments

We accomplished experiments for coreset construction of subspace clustering by implementing Algorithm [1] and similar empirical results can be obtained for the projective clustering as well. An input to our algorithm is a synthetic dataset (with entries normalized between 0 and 1), which is a random matrix of rank (say \(r\))—we obtain it by multiplying two random matrices of sizes \((n \times r)\) and \((r \times d)\). We compare our results with [9] and are able to show that we can achieve a faster coreset construction while preserving a desired accuracy—especially if the input matrix is of small rank.

Suppose we generate a coreset for \(j\)-subspace clustering. Then we can evaluate the quality of our result as follows: we first generate a random matrix \([X^+]_{(d-j) \times d}\) having orthonormal columns which can be viewed as the orthogonal complement of the subspace (say \(L\)) of dimension \(j\). Let \(|AX^+|_F^2\) and \(|AX^+|_F^2\) denote the sum of square distances from the rows of \(A\) and \(\tilde{A}\) to the subspace \(L\), respectively (see preliminaries); and \(\tilde{A}\) be our coreset. Then, the
expression $\|AX - f\|_F^2$ represents the quality of our solution, and we denote it as approximation-ratio. We further compute the running time of our algorithm, and compare it with the coreset construction time of [9], and state it as $\text{Time-ratio} = \frac{\text{Running time of Algorithm 1}}{\text{Coreset construction time of [9]}}$.

We summarized our experimental evaluation in the following figure. Row 1 demonstrates the trade-off between the “approximation-ratio, time-ratio” vs the “rank” of the input matrix (for fixed values of $\epsilon$, dimension, and different values of $j$). This figure demonstrates that if the rank of the matrix is small, then we can achieve a non-trivial time saving while preserving a desired accuracy. As expected, time ratio increases as the value of rank and $j$ increases.

Row 2 demonstrates the trade-off between the “approximation-ratio, time-ratio” vs the “dimension” of the input matrix (for fixed values of $\epsilon$, rank, and different values of $j$). Time ratio increases as the values of dimension and $j$ increases. Row 3, demonstrates the trade-off between the “approximation-ratio, time-ratio” vs the tolerance parameter “$\epsilon$”. The main insight here is that if the matrix is of small rank, then even by choosing a higher value of $\epsilon$, one can achieve substantial time saving while maintaining a desired accuracy.

**Remark.** Algorithm 1 holds for any matrix and for smaller values of $\epsilon$. However, for illustration purpose, we did a simple implementation of the algorithm, and perform experiments for a restricted setting when the matrix is of small rank, and could show substantial time saving. However, we further believe that by a more efficient implementation of the algorithm, especially of line number 3, empirical time saving can be further improved significantly.

6 Conclusion and open problems

We presented randomized coreset constructions for subspace and projective clustering via low-rank approximation. We obtained a coreset of constant size and dimension for subspace...
clustering, and obtained a coreset of constant dimension for projective clustering. As a corollary of projective clustering result, for points on an integer grid with bounded coordinates, we could able to show a sublinear size coreset. Our coreset constructions are significantly faster as compared to the corresponding deterministic construction \[9\] and it also preserve a desired accuracy. Using analysis techniques of \[9\], our result can be easily extended for other data analysis tasks like \(k\)-mean, \(k\)-line clustering etc. Our work leaves several open problems - improving dimensionality reduction bounds for projective clustering is one of the major open problem of this work.

References

1. P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan. Approximating extent measures of points. *J. ACM*, 51(4):606–635, 2004.
2. P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan. geometric approximation via coresets. *Current Trends in Combinatorial and Computational Geometry (E. Welzl, ed.)*, 2007.
3. M. B. Cohen, S. Elder, C. Musco, C. Musco, and M. Persu. Dimensionality reduction for \(k\)-means clustering and low rank approximation. In *Proceedings of the Forty-Seventh Annual ACM on Theory of Computing*, pages 163–172, 2015.
4. A. Deshpande, L. Rademacher, S. Vempala, and G. Wang. Matrix approximation and projective clustering via volume sampling. *Theory of Computing*, 2(12):225–247, 2006.
5. A. Deshpande and S. Vempala. Adaptive sampling and fast low-rank matrix approximation. In *Approximation, Randomization, and Combinatorial Optimization, Algorithms and Techniques, APPROX-RANDOM*, pages 292–303, 2006.
6. D. Feldman, A. Fiat, and M. Sharir. Coresets for weighted facilities and their applications. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006)*, 21-24 October 2006, Berkeley, California, USA, *Proceedings*, pages 315–324, 2006.
7. D. Feldman and M. Langberg. A unified framework for approximating and clustering data. In *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pages 569–578, 2011.
8. D. Feldman, M. Monemizadeh, C. Sohler, and D. P. Woodruff. Coresets and sketches for high dimensional subspace approximation problems. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 630–649, 2010.
9. D. Feldman, M. Schmidt, and C. Sohler. Turning big data into tiny data: Constant-size coresets for \(k\)-means, pca and projective clustering. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1434–1453, 2013.
10. S. Har-Peled. No, coreset, no cry. In *FSTTCS 2005: Foundations of Software Technology and Theoretical Computer Science, 24th International Conference, Chennai, India, December 16-18, 2004, Proceedings*, pages 324–335, 2004.
11. S. Har-Peled and S. Mazumdar. On coresets for \(k\)-means and \(k\)-median clustering. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing, Chicago, IL, USA, June 13-16, 2004*, pages 291–300, 2004.
12. M. W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.
13. J. M. Phillips. Coresets and sketches. *CoRR*, abs/1601.00617, 2016.
14. T. Sarlós. Improved approximation algorithms for large matrices via random projections. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006)*, 21-24 October 2006, Berkeley, California, USA, *Proceedings*, pages 143–152, 2006.
15. K. R. Varadarajan and X. Xiao. A near-linear algorithm for projective clustering integer points. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2012, Kyoto, Japan, January 17-19, 2012*, pages 1329–1342, 2012.
Appendix

A.0.0.1 Proof of Lemma [11]

Here, we present a lower bound on the following expression which was skipped earlier while presenting a proof Lemma [11] in Section 3.

\[ ||\tilde{A}X^\perp||_F^2 - ||AX^\perp||_F^2 \]

= \[ ||\tilde{A}||_F^2 - ||\tilde{A}X||_F^2 - ||A||_F^2 + ||AX||_F^2 \]

= \[ ||AV^\top V^T X||_F^2 - ||\tilde{A}X||_F^2 + ||AX||_F^2 - ||AV^\top V^T X||_F^2 + ||\tilde{A}||_F^2 - ||A||_F^2 \]

\geq \[ ||AV^\top V^T X||_F^2 - ||\tilde{A}X||_F^2 + ||\tilde{A}||_F^2 - ||A||_F^2 \]

\geq \[ -3\epsilon ||A - A^{(m)}||_F^2 + ||A^{(m)}||_F^2 - ||A||_F^2 \]

= \[ -(1 + 3\epsilon)||A - A^{(m)}||_F^2 \]

= \[ -(1 + 2\epsilon)(d - m)\sigma_{m+1} - \epsilon \Sigma_{i=m+1} \sigma_i^2 \]

= \[ -\epsilon(m - j + 1)(1 + 2\epsilon)(d - m)\sigma_{m+1} - \epsilon \Sigma_{i=m+1} \sigma_i^2 \]

(20)

\geq \[ -\epsilon(m - j + 1)\sigma_{m+1} 3(d - m) \frac{1}{j} - \epsilon \Sigma_{i=m+1} \sigma_i^2 \]

(21)

\geq \[ -\epsilon(m - j + 1)\sigma_{m+1} 3(d - j - 1) \frac{1}{j} - \epsilon \Sigma_{i=m+1} \sigma_i^2 \]

(22)

\geq \[ -\epsilon(m - j + 1)\sigma_{m+1} 3(d - j - \frac{3d - 4j + 1}{3j}) \frac{1}{j} - \epsilon \Sigma_{i=m+1} \sigma_i^2 \]

(23)

\geq \[ -\epsilon(m - j + 1)\sigma_{m+1} \left( \frac{3d - 3d + 4j - 1 - j + 1}{3j} \right) - \epsilon \Sigma_{i=m+1} \sigma_i^2 \]

(24)

Equality [18] follows from pythagoras theorem; Inequality [19] holds from Lemma [8] Inequality [20] holds from Equation [2] Inequality [22] holds as \( m = j + \lceil \frac{d}{2} \rceil - 1 \); Inequality [23] holds as \( \epsilon < \frac{3j}{3d - 4j + 1} \).

A.0.0.2 Proof of Theorem [5]

Let \([X^*]_{d \times j}\) be a matrix with orthonormal columns whose span is \(L^*\), and let \(L^\perp\) be the orthogonal complement of \(L^*\) spanned by \([X^\perp]_{d \times (d - j)}\). In our analysis we use the equality, \(\text{dist}^2(A, C) = ||AX^\perp||_F^2 + \text{dist}^2(AX^* X^* T, C)\), which holds true due to the Pythagoras theorem. We have

\[ |\text{dist}^2(A^*, C) - \text{dist}^2(A, C)| \]

\[ = \left( ||A^* X^\perp||_F^2 + \text{dist}^2(A^* X^* X^* T, C) \right) - \left( ||AX^\perp||_F^2 + \text{dist}^2(AX^* X^* T, C) \right) \]
We have two terms to bound in the above expression. The first term can be upper bounded using a similar analysis as of Lemma 11 (see Equations [21,22]) which holds true with probability at least $1 - \delta$. (In Equations [21,22] we replace $j$ by $j^*$, $m$ by $m^*$, $\epsilon$ by $\frac{\epsilon^2}{26}$.) In the remaining proof of the theorem, we denote $\|A - A^{m^*}\|_F^2$ by $\Delta^*$.

$$\left| \left| A^* X^\perp \right|_F^2 - \left| A X^* X^\perp \right|_F^2 \right| \leq \max \left\{ \frac{\epsilon^2}{26} \left| A X^* X^\perp \right|_F^2 + (3j^* \frac{\epsilon^2}{26} - 1) \Delta^*, (1 + \frac{3\epsilon^2}{26}) \Delta^* \right\}$$

(25)

In order to bound the second term $\| \text{dist}(A^* X^* X^T, C) - \text{dist}(A X^* X^T, C) \|_F$, we use a triangle inequality from Lemma 10. For any $\epsilon \in (0, 1)$ and from Lemma 10 we have the following:

$$\left| \text{dist}(A^* X^* X^T, C) - \text{dist}(A X^* X^T, C) \right| \leq \frac{12}{\epsilon} \left( \frac{\epsilon^2}{26} \left| A X^* X^\perp \right|_F^2 + 3j^* \frac{\epsilon^2}{26} \Delta^* \right) + \frac{\epsilon}{2} \text{dist}^2(A, C)$$

(27)

We have two cases in Equation (25), and we consider them one by one.

**Case 1:** $\|A^* X^\perp \|_F^2 - \| A X^* X^\perp \|_F^2 \leq \frac{\epsilon^2}{26} \left| A X^* X^\perp \right|_F^2 + (3j^* \frac{\epsilon^2}{26} - 1) \Delta^*$.

The above expression in conjunction with Equation (27) give us the following:

$$\left| \text{dist}(A^*, C) - \text{dist}(A, C) \right| \leq \frac{\epsilon^2}{26} \left| A X^* X^\perp \right|_F^2 + (3j^* \frac{\epsilon^2}{26} - 1) \Delta^* + \frac{12}{\epsilon} \left( \frac{\epsilon^2}{26} \left| A X^* X^\perp \right|_F^2 + 3j^* \frac{\epsilon^2}{26} \Delta^* \right) + \frac{\epsilon}{2} \text{dist}^2(A, C)$$

$$= \left( 1 + \frac{12}{\epsilon} \right) \frac{\epsilon^2}{26} \left| A X^* X^\perp \right|_F^2 + \frac{\epsilon}{2} \text{dist}^2(A, C) + \left( 3j^* \frac{\epsilon^2}{26} \left( 1 + \frac{12}{\epsilon} \right) - 1 \right) \Delta^*$$

$$\leq \left( 1 + \frac{12}{\epsilon} \right) \frac{\epsilon^2}{26} \text{dist}^2(A, C) + \frac{\epsilon}{2} \text{dist}^2(A, C) + \left( 3j^* \frac{\epsilon^2}{26} \left( 1 + \frac{12}{\epsilon} \right) - 1 \right) \Delta^*$$

(28)
Equality 28 holds by choosing $\varepsilon = \varepsilon$. As we have $\frac{\epsilon^2}{26} + \frac{12\epsilon}{26} < \frac{\epsilon}{2}$, also as we have chosen $\varepsilon < \frac{26}{397\sqrt{2}} = \frac{\epsilon}{2\sqrt{2}}$, we obtain the following:
\[
|\text{dist}^2(A^*, C) - \text{dist}^2(A, C)| \leq \epsilon\text{dist}^2(A, C). \tag{29}
\]

Now, we address the second case of Equation 25 which is stated as follows.

**Case 2:** $||A^* X^{*-1}||_F^2 - ||AX^{*-1}||_F^2 \leq (1 + \frac{26\epsilon}{26})\Delta^*$.

The above inequality in conjunction with Equation 27 gives us the following:
\[
|\text{dist}^2(A^*, C) - \text{dist}^2(A, C)| \leq (1 + \frac{3\epsilon^2}{26})\Delta^* + \frac{\epsilon}{2}\text{dist}^2(A, C) + \frac{12\epsilon}{26}||AX^{*-1}||_F^2 + 3j^* \frac{\epsilon^2}{26}||A - A^{(m^*)}||_F^2
\]
\[
= \frac{12\epsilon^2}{26}||AX^{*-1}||_F^2 + \frac{\epsilon}{2}\text{dist}^2(A, C) + \left(1 + \frac{3\epsilon^2}{26} \left(\frac{12j^*}{\epsilon} + 1\right)\right)\Delta^*
\]
\[
\leq \frac{12\epsilon^2}{26}\text{dist}^2(A, C) + \frac{\epsilon}{2}\text{dist}^2(A, C) + \left(1 + \frac{3\epsilon^2}{26} \left(\frac{12j^*}{\epsilon} + 1\right)\right)\Delta^*
\]
\[
= \left(\frac{12\epsilon}{26} + \frac{\epsilon}{2}\right)\text{dist}^2(A, C) + \left(\frac{3\epsilon^2 + 36\epsilon j^* + 26j^*}{26}\right)\Delta^*
\]
\[
\leq \left(\frac{12\epsilon}{26} + \frac{\epsilon}{2}\right)\text{dist}^2(A, C) + \left(\frac{3\epsilon^2 + 36\epsilon j^* + 26j^*}{26}\right)(d - m^*)\sigma_{m^*+1}^2
\]
\[
\leq \left(\frac{12\epsilon}{26} + \frac{\epsilon}{2}\right)\text{dist}^2(A, C) + \left(\frac{36j^* + 29}{26}\right)\left(d - j^* - \frac{26j^*}{\epsilon^2} + 1\right)\sigma_{m^*+1}^2
\]
\[
\leq \left(\frac{12\epsilon}{26} + \frac{\epsilon}{2}\right)\text{dist}^2(A, C) + \left(\frac{j^* (36j^* + 29)}{26j^*}\right)\left(d - j^* + 1 - \frac{26j^* (d - j - 0.6)}{26j^*}\right)\sigma_{m^*+1}^2
\]
\[
= \left(\frac{12\epsilon}{26} + \frac{\epsilon}{2}\right)\text{dist}^2(A, C) + \frac{j^* (36j^* + 29)}{65j^*}\sigma_{m^*+1}^2
\]
\[
\leq \left(\frac{12\epsilon}{26} + \frac{\epsilon}{2}\right)\text{dist}^2(A, C) + j^* \sigma_{m^*+1}^2
\]
\[
= \left(\frac{12\epsilon}{26} + \frac{\epsilon}{2}\right)\text{dist}^2(A, C) + \frac{\epsilon^2}{26}(m^* - j^* + 1)\sigma_{m^*+1}^2
\]
\[
\leq \left(\frac{12\epsilon}{26} + \frac{\epsilon}{2}\right)\text{dist}^2(A, C) + \frac{\epsilon^2}{26}\prod_{i=j^*+1}^{m^*} \sigma_{i+1}^2
\]
\[
\leq \left(\frac{12\epsilon}{26} + \frac{\epsilon}{2}\right)\text{dist}^2(A, C) + \frac{\epsilon^2}{26}\prod_{i=j^*+1}^{m^*} \sigma_{i+1}^2
\]
\[
\leq \left(\frac{12\epsilon}{26} + \frac{\epsilon}{2}\right)\text{dist}^2(A, C) + \frac{\epsilon^2}{26}||AX^{*-1}||_F^2
\]
\[
\leq \left(\frac{\epsilon^2}{26} + \frac{12\epsilon}{26} + \frac{\epsilon}{2}\right)\text{dist}^2(A, C)
\]
\[
\leq \epsilon\text{dist}^2(A, C) \tag{33}
\]

Equality 30 is true by choosing $\varepsilon = \varepsilon$. Inequality 31 holds true as $\varepsilon < \sqrt{\frac{26j^*}{\sqrt{2} - j^* - 0.6}}$.

Inequality 32 holds true as $\frac{36j^* + 29}{65j^*} < 1$, for all $j^* \geq 1$. From above inequality and Equation 29, we can conclude that for $\varepsilon \in (0, \varepsilon')$ with $\varepsilon' = \min \left\{ \frac{\epsilon}{2\sqrt{2}}, \sqrt{\frac{26j^*}{\sqrt{2} - j^* - 0.6}} \right\}$, the following holds with probability at least $1 - \delta$
\[
|\text{dist}^2(A^*, C) - \text{dist}^2(A, C)| \leq \epsilon\text{dist}^2(A, C).
\]
A.0.0.3 Proof of Lemma 12

We give an upper bound of the following expression

\[ \| A^* X^* X^*^T - A X^* X^*^T \|_F^2 \]

\[ = \| A X^* X^*^T - A^* X^* X^*^T \|_F^2 \]

\[ = \| A X^* X^*^T - A^{(m^*)} X^* X^*^T + A^{(m^*)} X^* X^*^T - A^* X^* X^*^T \|_F^2 \]

\[ \leq \| A X^* X^*^T - A^{(m^*)} X^* X^*^T \|_F^2 + \| A^{(m^*)} X^* X^*^T - A^* X^* X^*^T \|_F^2 \]

\[ \leq \frac{\epsilon^2}{26} \sum_{i=1}^{j^*+1} \sigma_i^2 + \| A V^* V^*^T X^* X^*^T - A^* X^* X^*^T \|_F^2 \]

\[ \leq \frac{\epsilon^2}{26} \sum_{i=1}^{j^*+1} \sigma_i^2 + \| A V^* V^*^T - A^* \|_F^2 \| X^* \|_F^2 \]

\[ \leq \frac{\epsilon^2}{26} \| A X^* \|_F^2 + 3 j^* \frac{\epsilon^2}{26} \| A - A^{(m^*)} \|_F^2 \]

Inequality (34) follows by Lemma 8, where we replace \( \epsilon \) by \( \frac{\epsilon^2}{26} j \); \( j \) by \( j^* \); and \( V^* \) denotes the first \( m^* \) right singular vectors in the SVD of \( A \); and Inequality (35) holds due to Equation (2) and Fact 1.