Coresets and Sketches for High Dimensional Subspace Approximation Problems *

Dan Feldman† Morteza Monemizadeh‡ Christian Sohler§ David P. Woodruff¶

Abstract

We consider the problem of approximating a set P of n points in $\mathbb{R}^d$ by a j-dimensional subspace under the ℓ_p measure, in which we wish to minimize the sum of ℓ_p distances from each point of P to this subspace. More generally, the $F_q(\ell_p)$-subspace approximation problem asks for a j-subspace that minimizes the sum of qth powers of ℓ_p-distances to this subspace, up to a multiplicative factor of $(1+\epsilon)$.

We develop techniques for subspace approximation, regression, and matrix approximation that can be used to deal with massive data sets in high dimensional spaces. In particular, we develop coresets and sketches, i.e. small space representations that approximate the input point set P with respect to the subspace approximation problem. Our results are:

- A dimensionality reduction method that can be applied to $F_q(\ell_p)$-clustering and shape fitting problems, such as those in [8, 15].
- The first strong coreset for $F_1(\ell_2)$-subspace approximation in high-dimensional spaces, i.e. of size polynomial in the dimension of the space. This coreset approximates the distances to any j-subspace (not just the optimal one).
- A $(1+\epsilon)$-approximation algorithm for the j-dimensional $F_1(\ell_2)$-subspace approximation problem with running time $nd(j/\epsilon)^{O(1)} + (n+d)2^{\text{poly}(j/\epsilon)}$.
- A streaming algorithm that maintains a coreset for the $F_1(\ell_2)$-subspace approximation problem and uses a space of $d\left(\frac{2\sqrt{\log n}}{\epsilon^2}\right)^{\text{poly}(j)}$ (weighted) points.
- Streaming algorithms for the above problems with bounded precision in the turnstile model, i.e, when coordinates appear in an arbitrary order and undergo multiple updates. We show that bounded precision can lead to further improvements. We extend results of [7] for approximate linear regression, distances to subspace approximation, and optimal rank-j approximation, to error measures other than the Frobenius norm.

1 Introduction

The analysis of high-dimensional massive data sets is an important task in data mining, machine learning, statistics and clustering. Typical applications include: pattern recognition in computer vision and image processing, bioinformatics, internet traffic analysis, web spam detection, and classification of text documents. In these applications, we often have to process huge data sets that do not fit into main memory. In order to process these very large data sets, we require streaming algorithms that read the data in a single pass and use only a small amount of memory. In other situations, data is collected in a distributed way, and shall be analyzed centrally. In this case, we need to find a way to send a small summary of the data that contains enough information to solve the problem at hand.

The second problem one has to overcome is the dimensionality of the data. High-dimensional data sets are often hard to analyze, and at the same time many data sets have low intrinsic dimension. Therefore, a basic task in data analysis is to find a low dimensional space, such that most input points are close to it. A well-known approach to this problem is principle component analysis (PCA) which, for a given set of n points in d-dimensional space, computes a linear j-dimensional subspace, such that the sum of squared distances to this subspace is minimized. Since this subspace is given by certain eigenvectors of the corresponding covariance matrix, one can compute it in $O(\min\{nd^2, dn^2\})$ time.

However, for massive data sets this computation may already be too slow. Therefore, the problem of approximating this problem in linear time has been studied in the standard and in the streaming model of computation. Depending on the problem, it is also interesting to study other error measures, like the sum of ℓ_p-distances to the subspace or, more generally, the sum of qth powers of ℓ_p-distances, and other related problems like linear regression [8] or low-rank matrix approximation [12].

For example, an advantage of the sum of distances measure is its robustness in the presence of outliers when compared to sum of non squared distances measure. However, unlike for the sum of squared errors, no closed formula exists for the optimal solution, even for the case of $j = 1$ (a line) in three-dimensional space [20].

In this extended abstract we mainly focus on the problem of approximating the optimal j-dimensional sub-

*Supported in parts by DFG project So 514/4-2.

†School of Computer Science; Tel Aviv University; Tel Aviv 69978, Israel; dannyf@post.tau.ac.il

‡ Department of Computer Science, University of Dortmund, Germany; morteza.monemizadeh@tu-dortmund.de

§ Department of Computer Science, University of Dortmund, Germany; christian.sohler@tu-dortmund.de

¶ IBM Almaden Research Center, San Jose, CA; dpwoodru@us.ibm.com

space with respect to sum of distances. That is, we are given a set P of n points with the objective to find a j-space C that minimizes $\text{cost}(P, C) = \sum_{p \in P} \min_{c \in C} \|p - c\|_2$. We call this problem the $F_1(\ell_2)$-*subspace approximation problem*. Most of our results generalize (in a non-trivial way) to $F_q(\ell_p)$-subspace approximations. Details will be given in the full version of this paper. As discussed above, we are interested in developing algorithms for huge high-dimensional point sets. In this case we need to find small representations of the data that approximate the original data, which allows us to solve the problem in a distributed setting or for a data stream.

In this paper, we develop such representations and apply them to develop new approximation and streaming algorithms for subspace approximation. In particular, we develop *strong coresets* and *sketches*, where the coresets apply to the case of unbounded and the sketches to bounded precision arithmetic.

A *strong coreset* [1, 16] is a small weighted set of points such that for *every* j-subspace of $\mathbb{R}^d$, the cost of the coreset is approximately the same as the cost of the original point set. In contrary, weak coresets [14, 12, 9] are useful only for approximating the optimal solution. One of the benefits of strong coresets is that they are closed under the union operation, which is, for example, desirable in a distributed scenario as sketched below.

Application scenarios. For an application of coresets and/or sketches, consider the following scenario. We are aggregating data at a set of clients and we would like to analyze it at a central server by first reducing its dimensionality via subspace approximation, projecting the data on the subspace and then clustering the projected points. Using such a client-server architecture, it is typically not feasible to send all data to the central server. Instead, we can compute coresets of the data at the clients and collect them centrally. Then we solve the subspace approximation problem on the union of the coresets and send the computed subspace to all clients. Each client projects the points on the subspace and computes a coreset for the clustering problem. Again, this coreset is sent to the server and the clustering problem is solved.

Specific applications for the j-subspace approximation problems, include the well known "Latent Semantic Analysis" technique for text mining applications, the PageRank algorithm in the context of web search, or the Eigenvector centrality measure in the field of social network analysis (see [12, 9] and the references therein).

These problems also motivate the turnstile streaming model that is defined below, which is useful when new words (in latent semantic analysis) or new connections between nodes (in social network analysis) are updated over time, rather than just the insertion or deletion of entire documents and nodes.

Results and relation to previous work.

- We develop a dimensionality reduction method that can be applied to $F_q(\ell_p)$-clustering and shape fitting problems [15]. For example, the cluster centers can be point sets, subspaces, or circles.

- We obtain the first strong coreset for $F_1(\ell_p)$-subspace approximation in high-dimensional spaces, i.e. of size $dj^{O(j^2)} \cdot \epsilon^{-2} \cdot \log n$ (weighted) points. Previously, only a strong coreset construction with an exponential dependence on the dimension of the input space was known [11]. Other previous research [9, 25, 10, 15] in this area constructed so-called weak coresets. A weak coreset is a small set $\mathcal{A}$ of points such that the span of $\mathcal{A}$ contains a $(1+\epsilon)$-approximation to the optimal j-subspace. The authors [9, 10, 15] show how to find a weak coreset for sum of squared distances, and in [10] Deshpande and Varadarajan obtain a weak coreset for sum of qth power of distances. All of these algorithms are in fact $\text{poly}(j, \epsilon^{-1})$-pass streaming algorithms.

- Our next result is an improved $(1+\epsilon)$-approximation algorithm for the j-dimensional subspace approximation problem under the measure of sum of distances. The running time of our algorithm is $nd(j/\epsilon)^{O(1)} + (n+d)2^{(j/\epsilon)^{O(1)}}$. This improves upon the previously best result of $nd2^{(j/\epsilon)^{O(1)}}$ [25, 10].

- We then show that one can maintain a coreset in a data stream storing $\tilde{O}(d(\frac{j2^{O(\sqrt{\log n})}}{\epsilon^2})^{\text{poly}(j)})$ (weighted) points. From this coreset we can extract a $(1+\epsilon)$-approximation to the optimal subspace approximation problem. We remark that we do not have a bound on the time required to extract the subspace from the data points. Previously, no 1-pass streaming algorithm for this problem was known, except for the case of the $F_2(\ell_2)$ objective function [7].

- We also study bounded precision in the turnstile model, i.e., when coordinates are represented with $O(\log(nd))$ bits and encode, w.l.o.g., integers from $-(nd)^{O(1)}$ to $(nd)^{O(1)}$. The coordinates appear in an arbitrary order and undergo multiple updates. Bounded precision is a practical assumption, and using it now we can extract a $(1+\epsilon)$-approximation to the optimal j-space in a data stream in polynomial time for fixed j/ϵ. Along the way we extend the results of [7] for linear regression, distance to subspace approximation, and best rank-j approximation, to error measures other than the Frobenius norm.

Techniques. In order to obtain the strong coreset, we first develop a dimensionality reduction technique for subspace approximation. The main idea of the dimensionality reduction is to project the points onto a low-dimensional subspace and approximate the *difference* between the projected points and the original point set. In order to do so, we need to introduce points with negative weights in the coreset. While this technique only gives an additive error, we remark that for many applications this additive error can be directly translated into a multiplicative error with respect to $\text{cost}(P, C)$. The non-uniform sampling technique we are using to estimate the difference between the projected points and the original point set is similar to that in previous work [9, 10, 14, 25].

Although we apply the dimensionality reduction here in the context of subspaces, we remark that this technique can be easily generalized. In fact, we can replace the subspace by any closed set on which we can efficiently project points. For example, we can easily extend the dimensionality reduction method to geometric clustering problems where the centers are low dimensional objects. We can also use the technique for any problem where we are trying to minimize the sum of distances to manifolds [3, 4] (if the projection on the manifold is well-defined), which might, for example, occur in the context of kernel methods [3, 4, 21].

In order to obtain a strong coreset, we apply our dimensionality reduction recursively using the fact that, for a set P of points that are contained in an $(i+1)$-subspace of $\mathbb{R}^d$, a coreset for i-subspaces is also a coreset for j-subspaces, $1 \leq j \leq d$. This recursion is applied until $i = 0$ and the points project to the origin. This way, we obtain a small weighted sample set S of size $O(\log(1/\delta) \cdot j^{O(j^2)}/\epsilon^2)$ in $O(nd j^2 + |S|)$ time, such that for an arbitrary query j-subspace C, we have $(1-\epsilon) \cdot \text{cost}(P, C) \leq \text{cost}(S, C) \leq (1+\epsilon) \cdot \text{cost}(P, C)$. This result is then used to construct a strong coreset by showing that the approximation guarantee holds simultaneously for all solutions from a certain grid near the input points.

Using the (standard) merge-and-reduce technique [1, 16] we use our coresets to obtain a streaming algorithm. However, we remark that the use of negatively weighted points leads to some technical complications that do not allow us to get down to $\log^{O(1)} n$ space.

With bounded precision, we use space-efficient sketches of ℓ_p-distances, a search over grid points, and a lower bound on the singular values of A to approximate the ℓ_p-regression problem in low dimensions in a data stream . A consequence is that we can efficiently approximate the sum of q-th powers of ℓ_p-distances, denoted $F_q(\ell_p)$, to any fixed j-subspace. We then approximate the optimal j-subspace for $F_q(\ell_2)$ distances, $1 \leq q \leq 2$, using a structural result of Shyamalkumar and Varadarajan [24] that proves that a $(1+\epsilon)$-approximate solution is spanned by $r = O(j/\epsilon)$ rows (points) of A. That is, there are two matrices B and C of size $j \times r$ and $r \times n$, respectively, such that the columns of $B \cdot C \cdot A$ span a $(1+\epsilon)$-approximate solution. It is not clear how to find these matrices in the streaming model, but we can use algebraic methods together with bounded precision to limit the number of candidates. We can test candidates offline using the linearity of our sketch.

1.1 Preliminaries

A *weighted point* is a point $r \in \mathbb{R}^d$ that is associated with a weight $w(r) \in \mathbb{R}$. We consider an (unweighted) point $r \in \mathbb{R}^d$ as having a weight of one. The (Euclidean) distance of a point $r \in \mathbb{R}^d$ to a set (usually, subspace) $C \subseteq \mathbb{R}^d$ is $\text{dist}(r, C) := \inf_{c \in C} \|r - c\|_2$. The set C is called a *center*. For a closed set C, we define $\text{proj}(r, C)$ to be the closest point to r in C, if it exists, where ties are broken arbitrarily. We further define the weight of $\text{proj}(r, C)$ as $w(\text{proj}(r, C)) = w(r)$. Similarly, $\text{proj}(P, C) = \{\text{proj}(r, C) \mid r \in P\}$. We let $\text{cost}(P, C) = \sum_{r \in P} w(r) \cdot \text{dist}(r, C)$ be the weighted sum of distances from the points of P to C. Note that points with negative weights are also assigned to their closest (and not farthest) point $r \in C$.

For a specific class of centers $\mathcal{C}$ we can now define the $F_q(\ell_p)$-clustering problem as the problem to minimize $\sum_{r \in P} \inf_{c \in C} \|r - c\|_p^q$. For example, if $\mathcal{C}$ is the collection of sets of k points from $\mathbb{R}^d$, then the $F_1(\ell_2)$-clustering problem is the standard k-median problem with Euclidean distances, and the $F_2(\ell_2)$-clustering problem is the k-means problem.

One specific variant of clustering that we are focusing on is the j-*subspace approximation problem* with $F_1(\ell_p)$ objective function. The term j-*subspace* is used to abbreviate j-dimensional linear subspace of $\mathbb{R}^d$.

Given a set P of n points in $\mathbb{R}^d$, the j-subspace approximation problem with $F_1(\ell_p)$ objective function is to find a j-dimensional subspace C of $\mathbb{R}^d$ that minimizes $\text{cost}(P, C)$.

DEFINITION 1.1. (CORESET [1, 16]) *Let P be a weighted point set in $\mathbb{R}^d$, and $\epsilon > 0$. A weighted set of points Q is called a* strong ϵ-coreset *for the j-dimensional subspace approximation problem, if for every linear j-dimensional subspace C of $\mathbb{R}^d$, we have*

$$(1-\epsilon) \cdot \text{cost}(P, C) \leq \text{cost}(Q, C) \leq (1+\epsilon) \cdot \text{cost}(P, C) \ .$$

2 Dimensionality Reduction for Clustering Problems

In this section we present our first result, a general dimensionality reduction technique for problems that involve sums of distances as a quality measure. Our result is that

for an arbitrary fixed subset $C \subseteq \mathbb{R}^d$, $\text{cost}(P, C)$ can be approximated by a small weighted sample and the projection of P onto a low dimensional subspace. This result can be immediately applied to obtain a dimensionality reduction method for a large class of clustering problems, where the cluster centers are objects contained in low-dimensional spaces. Examples include: k-median clustering, subspace approximation under ℓ_1-error, variants of projective clustering and more specialized problems where cluster centers are, for example, discs or curved surfaces.

For these type of problems, we suggest an algorithm that computes a low dimensional weighted point set Q such that, with probability at least $1 - \delta$, for any fixed query center C, $\text{cost}(Q, C)$ approximates $\text{cost}(P, C)$ to within a factor of $1 \pm \epsilon$. The algorithm is a generalization of a technique developed in [14] to compute coresets for the k-means clustering problem.

The main new idea that allows us to handle any type of low dimensional center is the use of points that are associated with negative weights. To obtain this result, we first define a randomized algorithm DIMREDUCTION; see the figure below. For a given (low-dimensional) subspace C^* and a parameter $\epsilon > 0$, the algorithm DIMREDUCTION computes a weighted point set Q, such that most of the points of Q lie on C^*, and for any fixed query center C we have $\mathbf{E}[\text{cost}(Q, C)] = \text{cost}(P, C)$, i.e., $\text{cost}(Q, C)$ is an unbiased estimator of the cost of C with respect to P. Then we show that, with probability at least $1 - \delta$, the estimator has an additive error of at most $\epsilon \cdot \text{cost}(P, C^*)$.

DIMREDUCTION $(P, C^*, \delta, \epsilon)$

1. Pick $r = \left\lceil \frac{2 \lg(2/\delta)}{\epsilon^2} \right\rceil$ points $s_1, \ldots, s_r$ i.i.d. from P, s.t. each $p \in P$ is chosen with probability
$$\mathbf{Pr}[p] = \frac{\text{dist}(p, C^*)}{\text{cost}(P, C^*)}.$$
2. For $i \leftarrow 1$ **to** r **do**
$$w(s_i) \leftarrow \frac{1}{r \cdot \mathbf{Pr}[s_i]}$$
3. Return the multiset $Q = \text{proj}(P, C^*) \cup \{s_1, \ldots, s_r\} \cup \{\text{proj}(s_1^-, C^*), \ldots, \text{proj}(s_r^-, C^*)\}$, where s_i^- is the point s_i with weight $-w(s_i)$.

We can then apply this result to low dimensional clustering problems in two steps. First, we observe that, if each center is a low dimensional object, i.e. is contained in a low dimensional j-subspace, then k centers are contained in a (kj)-subspace and so clustering them is at least as expensive as $\text{cost}(P, C')$, where C' is a (kj)-subspace that minimizes $\text{cost}(P, C')$. Thus, if we compute an α-approximation C^* for the (kj)-dimensional subspace approximation problem, and replace ϵ by ϵ/α, we obtain the result outlined above.

Analysis of Algorithm DIMREDUCTION. Let us fix an arbitrary set C. Our first step will be the following technical lemma that shows that $\text{cost}(Q, C)$ is an unbiased estimator for $\text{cost}(P, C)$. Let X_i denote the random variable for the sum of contributions of the sample points s_i and $\text{proj}(s_i^-, C)$ to C, i.e.

$$\begin{aligned} X_i &= w(s_i) \cdot \text{dist}(s_i, C) + w(s_i^-) \cdot \text{dist}(\text{proj}(s_i^-, C)) \\ &= w(s_i) \cdot \big(\text{dist}(s_i, C) - \text{dist}(\text{proj}(s_i, C^*), C)\big). \end{aligned}$$

LEMMA 2.1. *Let P be a set of points in $\mathbb{R}^d$. Let $\epsilon > 0$, $0 < \delta \leq 1$, and Q be the weighted set that is returned by the randomized algorithm* DIMREDUCTION$(P, C^*, \delta, \epsilon)$. *Then* $\mathbf{E}[\text{cost}(Q, C)] = \text{cost}(P, C)$.

Proof. We have

$$\begin{aligned} &\mathbf{E}[X_i] \\ &= \sum_{p \in P} \mathbf{Pr}[p] \cdot w(p)\big(\text{dist}(p, C) - \text{dist}(\text{proj}(p, C^*), C)\big) \\ &= \sum_{p \in P} \frac{1}{r} \frac{1}{\mathbf{Pr}[p]} \cdot \mathbf{Pr}[p]\big(\text{dist}(p, C) - \text{dist}(\text{proj}(p, C^*), C)\big) \\ &= \frac{1}{r} \cdot \big(\text{cost}(P, C) - \text{cost}(\text{proj}(P, C^*), C)\big) \ . \end{aligned}$$

By linearity of expectation we have

$$\mathbf{E}[\sum_{i=1}^{r} X_i] = \text{cost}(P, C) - \text{cost}(\text{proj}(P, C^*), C) \ .$$

Since algorithm DIMREDUCTIONcomputes the union of $\text{proj}(P, C^*)$ and the points s_i and s_i^-, we obtain

$$\begin{aligned} \mathbf{E}[\text{cost}(Q, C)] &= \text{cost}(\text{proj}(P, C^*), C) + \mathbf{E}[\sum_{i=1}^{r} X_i] \\ &= \text{cost}(P, C). \end{aligned}$$

The lemma follows. □

Our next step is to show that $\text{cost}(Q, C)$ is sharply concentrated about its mean.

THEOREM 2.1. *Let P be a set of n points in $\mathbb{R}^d$, and let C^* be a j-subspace. Let $0 < \delta, \epsilon \leq 1$, and Q be the weighted point set that is returned by the algorithm* DIMREDUCTION$(P, C^*, \delta, \epsilon)$. *Then for a fixed query set $C \subseteq \mathbb{R}^d$ we have*

$$|\mathrm{cost}(P, C) - \mathrm{cost}(Q, C)| \leq \epsilon \cdot \mathrm{cost}(P, C^*),$$

with probability at least $1 - \delta$*. Moreover, only*

$$r = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$$

points of Q *are not contained in* $\mathrm{proj}(P, C^*)$*. This algorithm runs in* $O(ndj + r)$ *time.*

Proof. Let $P = \{p_1, \ldots, p_n\}$ be a set of n points in $\mathbb{R}^d$. We first prove the concentration bound and then discuss the running time.

In order to apply Chernoff-Hoeffding bounds [2] we need to determine the range of values X_i can attain. By the triangle inequality we have

$$\mathrm{dist}(s_i, C) \leq \mathrm{dist}(s_i, C^*) + \mathrm{dist}(\mathrm{proj}(s_i, C^*), C)$$

and

$$\mathrm{dist}(\mathrm{proj}(s_i, C^*), C) \leq \mathrm{dist}(s_i, C) + \mathrm{dist}(s_i, C^*).$$

This implies

$$|\mathrm{dist}(s_i, C) - \mathrm{dist}(\mathrm{proj}(s_i, C^*), C)| \leq \mathrm{dist}(s_i, C^*).$$

We then have

$$\begin{aligned} |X_i| &= \left|w(s_i) \cdot \left(\mathrm{dist}(s_i, C) - \mathrm{dist}(\mathrm{proj}(s_i, C^*), C)\right)\right| \\ &\leq w(s_i) \cdot \mathrm{dist}(s_i, C^*) = \frac{\mathrm{cost}(P, C^*)}{r}. \end{aligned}$$

Thus, $-\mathrm{cost}(P, C^*)/r \leq X_i \leq \mathrm{cost}(P, C^*)/r$. Using additive Chernoff-Hoeffding bounds [2] the result follows.

In order to achieve the stated running time, we proceed as follows. We first compute in $O(ndj)$ time for each point $p \in P$ its distance $\mathrm{dist}(p, C^*)$ to C^* and store it. This can easily be done by first computing an orthonormal basis of C^*. We sum these distances in order to obtain $\mathrm{cost}(P, C^*)$ in $O(n)$ time. From this we can also compute $\mathbf{Pr}[p]$ and $w(p)$ for each $p \in P$, in $O(n)$ overall time. We let P be the array of probabilities $p_1, \ldots, p_n$. It is well known that one can obtain a set of r samples according to a distribution given as a length-n array in $O(n + r)$ time, see [26]. □

3 From Dimensionality Reduction to Adaptive Sampling

In this section we show how to use Theorem 2.1 to obtain a small weighted set S that, with probability at least $1 - \delta$, approximates the cost to an arbitrary fixed j-subspace. The first step of the algorithm is to apply our dimensionality reduction procedure with a j-subspace C_j^* that is, with probability at least $2/3$ an $O(j^{j+1})$-approximation to the optimal j-dimensional linear subspace with respect to the ℓ_1-error. The success probability can be amplified to $1 - \delta$ in time $O(ndj \log(1/\delta))$. Such an approximation can be computed in $O(ndj)$ time using the algorithm APPROXIMATEVOLUMESAMPLING by Deshpande and Varadarajan [10]. Once we have projected all the points on C_j^*, we apply the same procedure using a $(j-1)$-dimensional linear subspace C_{j-1}^*. We continue this process until all the points are projected onto a 0-dimensional linear subspace, i.e. the origin. As we will see, this procedure can be used to approximate the cost of a fixed j-subspace C.

ADAPTIVESAMPLING (P, j, δ, ϵ)

1. $P_{j+1} \leftarrow P$.
2. **For** $i = j$ **Downto** 0
   (a) $C_i^* \leftarrow$ APPROXIMATEVOLUMESAMPLING(P_{i+1}, i).
   (b) $Q_i \leftarrow$ DIMREDUCTION$(P_{i+1}, C_i^*, \delta, \epsilon)$.
   (c) $P_i \leftarrow \mathrm{proj}(P_{i+1}, C_i^*)$.
   (d) $S_i \leftarrow Q_i \setminus P_i$, where S_i consists of the positively and negatively weighted sample points.
3. Return $S = \bigcup_{i=0}^{j} S_i$.

Note that P_0 is the origin, and so $\mathrm{cost}(P_0, C) = 0$ for any j-subspace C. Let C_i^* be an arbitrary but fixed sequence of linear subspaces as used in the algorithm.

THEOREM 3.1. *Let* P *be a set of* n *points in* $\mathbb{R}^d$*, and* $\epsilon', \delta' > 0$*. Let* C *be an arbitrary* j*-dimensional linear subspace. If we call algorithm* ADAPTIVESAMPLING *with the parameters* $\delta = O(\delta'/(j+1))$ *and* $\epsilon = \epsilon'/j^{c \cdot j^2}$ *for a large enough constant* c*, then we get*

$$\begin{aligned} (1 - \epsilon') \cdot \mathrm{cost}(P, C) &\leq \mathrm{cost}(S, C) \\ &\leq (1 + \epsilon') \cdot \mathrm{cost}(P, C), \end{aligned}$$

with probability at least $1 - \delta'$*. The running time of the algorithm is*

$$O(ndj(j + \log(1/\delta)) + \frac{j^{O(j^2)} \log(1/\delta')}{\epsilon'^2}.$$

Proof. Let C be an arbitrary j-subspace. We split the proof of Theorem 3.1 into two parts. The first and easy part is to show that $\mathrm{cost}(S, C)$ is an unbiased estimator of $\mathrm{cost}(P, C)$. The hard part is to prove that $\mathrm{cost}(S, C)$ is sharply concentrated.

We can apply Lemma 2.1 with $C^* = C_i^*$ to obtain that for any $1 \le i \le j$ we have $\mathbf{E}[\mathrm{cost}(Q_i, C)] = \mathrm{cost}(P_{i+1}, C)$ and hence

$$\mathbf{E}[\mathrm{cost}(S_i, C)] = \mathrm{cost}(P_{i+1}, C) - \mathrm{cost}(P_i, C) \ .$$

Therefore,

$$\begin{aligned}\mathbf{E}[\mathrm{cost}(S, C)] &= \sum_{i=0}^{j} \mathbf{E}[\mathrm{cost}(S_i, C)] \\ &= \mathrm{cost}(P_{j+1}, C) - \mathrm{cost}(P_0, C) \\ &= \mathrm{cost}(P, C) \ ,\end{aligned}$$

where the last equality follows from $P_{j+1} = P$ and P_0 being a set of n points at the origin.

Now we show that $\mathrm{cost}(S, C)$ is sharply concentrated. We have

$$\begin{aligned}&\left|\mathbf{E}[\mathrm{cost}(S, C)] - \mathrm{cost}(S, C)\right| \\ &\le \sum_{i=0}^{j} \left|\mathbf{E}[\mathrm{cost}(S_i, C)] - \mathrm{cost}(S_i, C)\right| \ .\end{aligned}$$

The following observation was used in [11] for $j = 1$, and generalized later in [13].

LEMMA 3.1. *Let C be a j-subspace, and L be an $(i+1)$-subspace, such that $i+1 \le j$. Then there exists an i-subspace C_i, and a constant $0 < \nu_L \le 1$, such that for any $p \in L$ we have $\mathrm{dist}(p, C) = \nu_L \cdot \mathrm{dist}(p, C_i)$.*

Let $0 \le i \le j$. By substituting $L = \mathrm{Span}\{P_{i+1}\}$ in Lemma 3.1, there is an i-subspace C_i and a constant ν_L, such that

$$\begin{aligned}&\left|\mathbf{E}[\mathrm{cost}(S_i, C)] - \mathrm{cost}(S_i, C)\right| \\ &= \left|\mathrm{cost}(P_{i+1}, C) - \mathrm{cost}(P_i, C) - \mathrm{cost}(S_i, C)\right| \\ &= \nu_L \cdot |\mathrm{cost}(P_{i+1}, C_i) - \mathrm{cost}(P_i, C_i) - \mathrm{cost}(S_i, C_i)| \\ &= \nu_L \cdot |\mathrm{cost}(P_{i+1}, C_i) - \mathrm{cost}(Q_i, C_i)| \ .\end{aligned}$$

Here, the second equality follows from the fact that the solution computed by approximate volume sampling is spanned by input points and so $P_i \subseteq \mathrm{Span}\{P_{i+1}\}$. We apply Theorem 2.1 with $C = C_i$ and $C^* = C_i^*$ to obtain

$$|\mathrm{cost}(P_{i+1}, C_i) - \mathrm{cost}(Q_i, C_i)| \le \epsilon \cdot \mathrm{cost}(P_{i+1}, C_i^*),$$

with probability at least $1 - \delta$. By our choice of C_i^*, we also have

$$\mathrm{cost}(P_{i+1}, C_i^*) \le O(i^{i+1}) \cdot \mathrm{cost}(P_{i+1}, C_i).$$

Combining the last three inequalities yields

$$\begin{aligned}&|\mathbf{E}[\mathrm{cost}(S_i, C)] - \mathrm{cost}(S_i, C)| \\ &\le \nu_L \cdot \epsilon \cdot \mathrm{cost}(P_{i+1}, C_i^*) \\ &\le O(\nu_L \cdot \epsilon \cdot i^{i+1}) \cdot \mathrm{cost}(P_{i+1}, C_i) \\ &= O(\epsilon \cdot i^{i+1}) \cdot \mathrm{cost}(P_{i+1}, C) \ ,\end{aligned}$$

with probability at least $1 - \delta$. Hence,

$$\begin{aligned}&\left|\mathbf{E}[\mathrm{cost}(S, C)] - \mathrm{cost}(S, C)\right| \\ &\le O\left(\sum_{i=0}^{j-1} \epsilon \cdot i^{i+1}\right) \cdot \mathrm{cost}(P_{i+1}, C),\end{aligned}$$

with probability at least $1 - j \cdot \delta$. Therefore, for our choice of δ and ϵ, a simple induction gives

$$\left|\mathbf{E}[\mathrm{cost}(S, C)] - \mathrm{cost}(S, C)\right| \le \epsilon \cdot j^{O(j^2)} \cdot \mathrm{cost}(P, C)$$

with probability at least $1 - j \cdot \delta$. Further, the running time is proven as in the proof of Theorem 2.1. □

4 Coresets

In order to construct a coreset, we only have to run algorithm ADAPTIVESAMPLING using small enough δ. One can compute δ by discretizing the space near the input points using a sufficiently fine grid. Then snapping a given subspace to the nearest grid points will not change the cost of the subspace significantly. If a subspace does not intersect the space near the input points, its cost will be high and the overall error can be easily charged.

THEOREM 4.1. *Let P denote a set of n points in $\mathbb{R}^d$, $j \ge 0$, and $1 > \epsilon', \delta' > 0$, $d \le n$. Let Q be the weighted set that is returned by the algorithm* ADAPTIVESAMPLING *with the parameters $\delta = \frac{1}{j} \cdot \delta'/(10nd)^{10dj}$ and $\epsilon = \epsilon'/(j+1)^{c' \cdot j^2}$ for a large enough constant c. Then, with probability at least $1 - \delta' - 1/n^2$, Q is a strong ϵ-coreset. The size of the coreset in terms of the number of (weighted) points saved is*

$$O(dj^{O(j^2)} \cdot \epsilon'^{-2} \log n).$$

First we prove some auxiliary lemmata.

LEMMA 4.1. *Let P be a set of points in a subspace A of $\mathbb{R}^d$. Let $M, \epsilon > 0$, $M > \epsilon$, and let $G \subseteq A$ be such that for every $c \in A$, if $\mathrm{dist}(c, P) \le 2M$ then $\mathrm{dist}(c, G) \le \epsilon/2$. Let $C \subseteq A$ be a 1-subspace (i.e, a line that intersects the origin of $\mathbb{R}^d$), such that $\mathrm{dist}(p, C) \le M$ for every $p \in P$. Then there is a 1-subspace D that is spanned by a point in G, such that,*

$$|\mathrm{dist}(p, C) - \mathrm{dist}(p, D)| \le \epsilon \text{ for every } p \in P.$$

Proof. Let g be a point such that the angle between the lines C and $\mathrm{Span}\{g\}$ is minimized over $g \in G$. Let $D = \mathrm{Span}\{g\}$, and $p \in P$. We prove the lemma using the following case analysis: **(i)** $\mathrm{dist}(p, D) \ge \mathrm{dist}(p, C)$, and **(ii)** $\mathrm{dist}(p, D) < \mathrm{dist}(p, C)$.

(i) $\mathrm{dist}(p, D) \ge \mathrm{dist}(p, C)$: Let $c = \mathrm{proj}(p, C)$. We have $\mathrm{dist}(c, P) \le \|c - p\| = \mathrm{dist}(p, C) \le M$. By the

assumption of the lemma, we thus have $\text{dist}(c, G) \leq \epsilon$. By the construction of D, we also have $\text{dist}(c, D) \leq \text{dist}(c, G)$. Combining the last two inequalities yields $\text{dist}(c, D) \leq \epsilon$. Hence

$$\text{dist}(p, D) \leq \|p - c\| + \text{dist}(c, D) \leq \text{dist}(p, C) + \epsilon.$$

(ii) $\text{dist}(p, D) < \text{dist}(p, C)$: Let $q = \text{proj}(p, D)$, and $q' = \text{proj}(q, C)$. We can assume that $\text{dist}(q, q') > \epsilon$ since otherwise by the triangle inequality, $\text{dist}(p, C) \leq \text{dist}(p, q) + \text{dist}(q, q') \leq \text{dist}(p, D) + \epsilon$, and we are done.

Define $\ell = \frac{q+q'}{2}$ and $\ell' = \ell / \|\ell\|_2$. Now consider the point $r = \ell + \frac{\epsilon}{2}\ell'$. We claim that r has distance from C and D more than $\epsilon/2$. Assume, this is not the case. Then C (the proof for D is identical) intersects a ball B with center r and radius $\epsilon/2$. Let r' be an intersection point of C with B. Let r'' be the projection of r' on the span of r. Since, B has radius $\epsilon/2$, we have that $\text{dist}(r'', r') \leq \epsilon/2$. However, the intercept theorem implies that $\text{dist}(r'', r') > \epsilon/2$, a contradiction. To finish the proof, we observe that $\text{dist}(p, r) \leq \text{dist}(p, q) + \text{dist}(q, \ell) + \text{dist}(\ell, r) \leq \text{dist}(p, C) + \epsilon \leq M + \epsilon$. Using $M > \epsilon$ the assumption of lemma implies $\text{dist}(r, G) < \epsilon/2$, but $\text{dist}(r, C) > \epsilon/2$ and $\text{dist}(r, D) > \epsilon/2$, which means there is a grid point g' for which $\angle(\text{Span}\{g'\}, C) < \angle(\text{Span}\{g\}, C)$, contradicting the choice of g. □

LEMMA 4.2. *Let P be a set of n points in $\mathbb{R}^d$, and $M, \epsilon > 0$, $M > \epsilon$. Let $G \subseteq \mathbb{R}^d$ be such that for every $c \in \mathbb{R}^d$, if $\text{dist}(c, P) \leq 2M$ then $\text{dist}(c, G) \leq \epsilon/2$. Let C be a j-subspace, such that $\text{dist}(p, C) \leq M - (j-1)\epsilon$ for every $p \in P$. Then there is a j-subspace D that is spanned by j points from G, such that*

$$|\text{dist}(p, C) - \text{dist}(p, D)| \leq j\epsilon \qquad \textit{for every } p \in P.$$

Proof. The proof is by induction on j. The base case of $j = 1$ is furnished by substituting $A = \mathbb{R}^d$ in Lemma 4.1. We now give a proof for the case $j \geq 2$. Let $e_1, \cdots, e_j$ denote a set of orthogonal unit vectors on C. Let $C^\perp$ be the orthogonal complement of the subspace that is spanned by $e_1, \cdots, e_{j-1}$. Finally, fix $p \in P$. The key observation is that for any j-subspace T in $\mathbb{R}^d$ that contains $e_1, \cdots, e_{j-1}$, we have

$$\text{dist}(p, T) = \text{dist}(\text{proj}(p, C^\perp), \text{proj}(T, C^\perp)).$$

Note that for such a j-subspace T, $\text{proj}(T, C^\perp)$ is a 1-subspace.

Let $P' = \text{proj}(P, C^\perp)$, and let $c' \in C^\perp$ be such that $\text{dist}(c', P') \leq 2M$. Hence, there is a point $q' \in P'$ such that

$$\|q' - c'\| = \text{dist}(c', P') \leq 2M. \tag{4.1}$$

Let $q \in P$ be such that $\text{proj}(q, C^\perp) = q'$. Let $c \in \mathbb{R}^d$ be such that $\text{proj}(c, C^\perp) = c'$ and $\|c - c'\| = \|q - q'\|$. Hence,

$$\begin{aligned}\|q - c\| &= \sqrt{\|q - c'\|^2 - \|c' - c\|^2} \\ &= \sqrt{\|q - c'\|^2 - \|q' - q\|^2} = \|q' - c'\|.\end{aligned}$$

By (4.1) and the last equation, $\|q - c\| \leq 2M$, i.e, $\text{dist}(c, P) \leq \|q - c\| \leq 2M$. Using the assumption of this lemma, we thus have $\text{dist}(c, G) \leq \epsilon/2$, so, clearly $\text{dist}\big(c', \text{proj}(G, C^\perp)\big) \leq \epsilon/2$.

From the previous paragraph, we conclude that for every $c' \in C^\perp$, if $\text{dist}(c', P') \leq 2M$ then $\text{dist}\big(c', \text{proj}(G, C^\perp)\big) \leq \epsilon/2$. Clearly, we also have $\text{dist}(\text{proj}(p, C^\perp), \text{proj}(C, C^\perp)) = \text{dist}(p, C) \leq M$. Using this, we apply Lemma 4.1 while replacing A with $C^\perp$, P with P', C with $\text{proj}(C, C^\perp)$ and G with $\text{proj}(G, C^\perp)$.

We obtain that there is a 1-subspace $D \subseteq C^\perp$ that is spanned by a point from $\text{proj}(G, C^\perp)$, such that

$$|\text{dist}\big(\text{proj}(p, C^\perp), \text{proj}(C, C^\perp)\big) - \text{dist}\big(\text{proj}(p, C^\perp), D\big)| \leq \epsilon.$$

Since $\text{dist}\big(\text{proj}(p, C^\perp), \text{proj}(C, C^\perp)\big) = \text{dist}(p, C)$ by the definition of $C^\perp$, the last two inequalities imply

$$|\text{dist}(p, C) - \text{dist}\big(\text{proj}(p, C^\perp), D\big)| \leq \epsilon. \tag{4.2}$$

Let E be the j-subspace of $\mathbb{R}^d$ that is spanned by D and $e_1, \cdots, e_{j-1}$. Let $D^\perp$ be the $(d-1)$-subspace that is the orthogonal complement of D in $\mathbb{R}^d$. Since $D \subseteq E$, we have that $\text{proj}(E, D^\perp)$ is a $(j-1)$-subspace of $\mathbb{R}^d$. We thus have

$$\begin{aligned}\text{dist}(\text{proj}(p, C^\perp), D) &= \text{dist}(\text{proj}(p, D^\perp), \text{proj}(E, D^\perp)) \\ &= \text{dist}(p, E).\end{aligned} \tag{4.3}$$

Using (4.2), with the assumption of this lemma that $\text{dist}(p, C) \leq M - (j-1)\epsilon$, yields

$$\begin{aligned}\text{dist}(\text{proj}(p, C^\perp), D) &\leq \text{dist}(p, C) + \epsilon \\ &\leq M - (j-2)\epsilon.\end{aligned}$$

By the last inequality and (4.3), we get

$$\text{dist}(\text{proj}(p, D^\perp), \text{proj}(E, D^\perp)) \leq M - (j-2)\epsilon. \tag{4.4}$$

For $P' = \text{proj}(P, D^\perp)$ and $c' \in D^\perp$, we have that if $\text{dist}(c', P') \leq 2M$ then $\text{dist}(c', \text{proj}(G, D^\perp) \leq \epsilon/2$. This can be proved similarly to the case $P' = \text{proj}(P, C^\perp)$ that was already proven. Using this and (4.4), we apply this lemma inductively with C as $\text{proj}(E, D^\perp)$, G as $\text{proj}(G, D^\perp)$ and P as $\text{proj}(P, D^\perp)$, to obtain a $(j -$

1)-subspace F that is spanned by $j-1$ points from $\mathrm{proj}(G, D^\perp)$, such that $|\mathrm{dist}(\mathrm{proj}(p, D^\perp), \mathrm{proj}(E, D^\perp)) - \mathrm{dist}(\mathrm{proj}(P, D^\perp), F)| \le (j-1)\epsilon$. Hence,

(4.5)
$$\begin{aligned}&|\mathrm{dist}(p, E) - \mathrm{dist}(\mathrm{proj}(p, D^\perp), F)| = \\ &|\mathrm{dist}(\mathrm{proj}(p, D^\perp), \mathrm{proj}(E, D^\perp)) - \mathrm{dist}(\mathrm{proj}(p, D^\perp), F)| \\ &\le (j-1)\epsilon.\end{aligned}$$

Let R be the j-subspace of $\mathbb{R}^d$ that is spanned by D and F. Hence, R is spanned by j points of G. We have

$$\begin{aligned}&|\mathrm{dist}(p, C) - \mathrm{dist}(p, R)| \\ &= |\mathrm{dist}(p, C) - \mathrm{dist}(\mathrm{proj}(p, D^\perp), F)| \\ &\le |\mathrm{dist}(p, C) - \mathrm{dist}(p, E)| \\ &\quad + |\mathrm{dist}(p, E) - \mathrm{dist}(\mathrm{proj}(p, D^\perp), F)|.\end{aligned}$$

By (4.3), we have $\mathrm{dist}(p, E) = \mathrm{dist}(\mathrm{proj}(p, C^\perp), D)$. Together with the previous inequality, we obtain

$$\begin{aligned}&|\mathrm{dist}(p, C) - \mathrm{dist}(p, R)| \\ &\le |\mathrm{dist}(p, C) - \mathrm{dist}(\mathrm{proj}(p, C^\perp), D)| \\ &\quad + |\mathrm{dist}(p, E) - \mathrm{dist}(\mathrm{proj}(p, D^\perp), F)|.\end{aligned}$$

Combining (4.2) and (4.5) in the last inequality proves the lemma. □

NET (P, M, ϵ)

1. $G \leftarrow \emptyset$.
2. **For** each $p \in P$ **Do**
   (a) $G_p \leftarrow$ vertex set of a d-dimensional grid that is centered at p. The side length of the grid is $2M$, and the side length of each cell is $\epsilon/(2\sqrt{d})$.
   (b) $G \leftarrow G \cup G_p$.
3. **Return** G.

LEMMA 4.3. *Let $0 < \epsilon, \delta' < 1$, and P be a set of n points in $\mathbb{R}^d$ with $d \le n$. Let C^* be a j-subspace, and Q be the weighted set that is returned by the algorithm* DIMREDUCTION *with the parameter $\delta = \delta'/(10nd)^{10jd}$. Then, with probability at least $1 - \delta' - 1/n^2$, for every j-subspace $C \subseteq \mathbb{R}^d$ we have (simultaneously)*

$$|\mathrm{cost}(P, C) - \mathrm{cost}(Q, C)| \le \epsilon \cdot \mathrm{cost}(P, C^*) + \epsilon \cdot \mathrm{cost}(P, C).$$

The following two propositions prove the lemma.

PROPOSITION 4.1. *For every j-subspace C of $\mathbb{R}^d$ such that*

$$\mathrm{cost}(P, C) > 2\mathrm{cost}(P, C^*)/\epsilon,$$

we have

$$|\mathrm{cost}(P, C) - \mathrm{cost}(Q, C)| \le \epsilon \cdot \mathrm{cost}(P, C).$$

Proof. Let C be a j-subspace such that

$$\mathrm{cost}(P, C) > 2\mathrm{cost}(P, C^*)/\epsilon.$$

Let $S = Q \setminus \mathrm{proj}(P, C^*)$. Hence,

(4.6)
$$\begin{aligned}&|\mathrm{cost}(P, C) - \mathrm{cost}(Q, C)| \\ &= |\mathrm{cost}(P, C) - \mathrm{cost}(\mathrm{proj}(P, C^*), C) - \mathrm{cost}(S, C)| \\ &\le |\mathrm{cost}(P, C) - \mathrm{cost}(\mathrm{proj}(P, C^*), C)| + |\mathrm{cost}(S, C)|.\end{aligned}$$

We now bound each term in the right hand side of (4.6).

Let s_i denote the ith point of S, $1 \le i \le |S|$. By the triangle inequality,

$$|\mathrm{dist}(s_i, C) - \mathrm{dist}(\mathrm{proj}(s_i, C^*), C)| \le \mathrm{dist}(s_i, C^*),$$

for every $1 \le i \le |S|$. Hence,

$$\begin{aligned}&|\mathrm{cost}(S, C)| \\ &= \left| \sum_{1 \le i \le |S|} w(s_i)(\mathrm{dist}(s_i, C) - \mathrm{dist}(\mathrm{proj}(s_i, C^*), C)) \right| \\ &\le \sum_{1 \le i \le |S|} w(s_i)|\mathrm{dist}(s_i, C^*)| = \mathrm{cost}(P, C^*).\end{aligned}$$

Similarly,

$$\begin{aligned}&|\mathrm{cost}(P, C) - \mathrm{cost}(\mathrm{proj}(P, C^*), C)| \\ &= \left| \sum_{p \in P} \mathrm{dist}(p, C) - \sum_{p \in P} \mathrm{dist}(\mathrm{proj}(p, C^*), C) \right| \\ &\le \sum_{p \in P} \mathrm{dist}(p, C^*) \\ &= \mathrm{cost}(P, C^*).\end{aligned}$$

Combining the last two inequalities in (4.6) yields

$$\begin{aligned}&|\mathrm{cost}(P, C) - \mathrm{cost}(Q, C)| \\ &\le |\mathrm{cost}(P, C) - \mathrm{cost}(\mathrm{proj}(P, C^*), C)| + |\mathrm{cost}(S, C)| \\ &\le 2\mathrm{cost}(P, C^*) \le \epsilon \cdot \mathrm{cost}(P, C).\end{aligned}$$

□

PROPOSITION 4.2. *Let* $0 < \epsilon < 1$ *and* $d \leq n$. *With probability at least*

$$1 - \delta' - 1/n^2,$$

for every j*-subspace* C *such that*

$$\mathrm{cost}(P, C) \leq 2\mathrm{cost}(P, C^*)/\epsilon,$$

we have (simultaneously)

$$|\mathrm{cost}(P, C) - \mathrm{cost}(Q, C)| \leq \epsilon \cdot \mathrm{cost}(P, C) + \epsilon \mathrm{cost}(P, C^*).$$

Proof. Let G denote the set that is returned by the algorithm $\textsc{Net}(P \cup \mathrm{proj}(P, C^*), M, \epsilon')$, where $M = 10\mathrm{cost}(P, C^*)/\epsilon$, and $\epsilon' = \epsilon \mathrm{cost}(P, C^*)/n^{10}$. Note that G is used only for the proof of this proposition.

By Theorem 2.1, for a fixed center $D \in G$ we have

$$\begin{aligned} &|\mathrm{cost}(P, D) - \mathrm{cost}(Q, D)| \\ &\leq \epsilon \cdot \mathrm{cost}(P, D) \\ &\leq \epsilon \cdot \mathrm{cost}(P, C) + \epsilon \cdot |\mathrm{cost}(P, C) - \mathrm{cost}(P, D)|, \end{aligned} \tag{4.7}$$

with probability at least

$$1 - \delta \geq 1 - \frac{\delta'}{(10nd)^{10jd}} \geq 1 - \frac{\delta'}{|G|^j}.$$

Using the union bound, (4.7) holds simultaneously for every j-subspace D that is spanned by j points from G, with probability at least $1 - \delta'$.

Let $p \in P$. By the assumption of this claim, we have

$$\mathrm{dist}(p, C) \leq \mathrm{cost}(P, C) \leq 2\mathrm{cost}(P, C^*)/\epsilon,$$

and also

$$\begin{aligned} &\mathrm{dist}(\mathrm{proj}(p, C^*), C) \\ &\leq \|\mathrm{proj}(p, C^*) - p\| + \mathrm{dist}(p, C) \\ &\leq \mathrm{dist}(p, C^*) + \frac{2\mathrm{cost}(P, C^*)}{\epsilon} \\ &\leq \frac{3\mathrm{cost}(P, C^*)}{\epsilon}. \end{aligned}$$

By the last two inequalities, for every $p \in P \cup \mathrm{proj}(P, C^*)$ we have

$$\begin{aligned} \mathrm{dist}(p, C) &\leq \frac{3\mathrm{cost}(P, C^*)}{\epsilon} \leq \frac{10\mathrm{cost}(P, C^*)}{\epsilon} - \frac{\mathrm{cost}(P, C^*)}{\epsilon} \\ &\leq M - (j-1)\epsilon', \end{aligned}$$

where in the last derivation we used the assumption $j \leq d \leq n$ and $0 \leq \epsilon \leq 1$. By the construction of G, for every $c \in \mathbb{R}^d$, if $\mathrm{dist}(c, P) \leq 2M$, then $\mathrm{dist}(c, G) \leq \epsilon'/2$. Using this, applying Lemma 4.2 with $P \cup \mathrm{proj}(P, C^*)$ yields that there is a j-subspace D that is spanned by j points from G, such that

$$|\mathrm{dist}(p, C) - \mathrm{dist}(p, D)| \leq j \cdot \epsilon',$$

for every $p \in P \cup \mathrm{proj}(P, C^*)$. Using the last equation with (4.7) yields

(4.8)

$$\begin{aligned} &|\mathrm{cost}(P, C) - \mathrm{cost}(Q, C)| \\ &\leq |\mathrm{cost}(P, C) - \mathrm{cost}(P, D)| + |\mathrm{cost}(P, D) - \mathrm{cost}(Q, D)| \\ &+ |\mathrm{cost}(Q, D) - \mathrm{cost}(Q, C)| \\ &\leq (1+\epsilon)|\mathrm{cost}(P, C) - \mathrm{cost}(P, D)| + \epsilon \mathrm{cost}(P, C) \\ &+ |\mathrm{cost}(Q, D) - \mathrm{cost}(Q, C)| \\ &\leq \epsilon \mathrm{cost}(P, C) \\ &+ 3 \sum_{p \in P \cup Q} |w(p)| \cdot |\mathrm{dist}(p, C) - \mathrm{dist}(p, D)| \\ &\leq \epsilon \mathrm{cost}(P, C) + 3j\epsilon' \sum_{p \in P \cup Q} |w(p)|, \end{aligned}$$

with probability at least $1 - \delta'$.

Let $s \in S$ be such that $w(s) > 0$. By the construction of S, we have

$$\mathrm{dist}(s, C^*) \geq \mathrm{cost}(P, C^*)/(n^2|S|)$$

with probability at least $1 - 1/(n^2|S|)$. Hence, with probability at least $1 - 1/n^2$, for every $s \in S$ we have

$$|w(s)| = \frac{\mathrm{cost}(P, C^*)}{|S|\mathrm{dist}(s, C^*)} \leq n^2.$$

Combining the last two equations with (4.8) yields

$$\begin{aligned} &|\mathrm{cost}(P, C) - \mathrm{cost}(Q, C)| \\ &\leq \epsilon \mathrm{cost}(P, C) + 3j\epsilon' \sum_{p \in P \cup Q} |w(p)| \\ &\leq \epsilon \mathrm{cost}(P, C) + \epsilon \mathrm{cost}(P, C^*), \end{aligned}$$

with probability at least $1 - 1/n^2 - \delta'$, as desired. □

Proof. [of Theorem 4.1] Let P_i, S_i, Q_i and C_i^* denote the set that are defined in the ith iteration of $\textsc{AdaptiveSampling}$, for every $0 \leq i \leq j$. For every i, $0 \leq i \leq j$, we have $|S_i| = O(\log(1/\delta)/\epsilon^2)$. Hence,

$$\begin{aligned} |Q| &= \bigcup_{0 \leq i \leq j} S_i = O\left(\frac{j \log(1/\delta)}{\epsilon^2}\right) \\ &\leq j^{O(j^2)} \cdot \frac{\log(1/\delta')}{\epsilon'^2}. \end{aligned}$$

This bounds the size of Q. For the correctness, let $0 \leq i \leq j$.

Fix $0 \leq i \leq j$. By the previous lemma and our choice of δ, we conclude that, with probability at least $1 - \delta'/j - 1/n^2$, for any j-subspace C we have for our choice of ϵ (assuming c' large enough)

$$\begin{aligned}
&|\mathrm{cost}(P_{i+1}, C) - \mathrm{cost}(Q_i, C)| \\
&\leq \epsilon \mathrm{cost}(P_{i+1}, C) + \epsilon \mathrm{cost}(P_{i+1}, C_i^*) \\
&\leq O(\frac{\epsilon'}{j^{j+1}})\mathrm{cost}(P_{i+1}, C) + O(\frac{\epsilon'}{j^{j+1}})\mathrm{cost}(P_{i+1}, C_i^*).
\end{aligned}$$

By construction of C_i^*, we have

$$\begin{aligned}
\mathrm{cost}(P_{i+1}, C_i^*) &\leq O(j^{j+1}) \min_{C'} \mathrm{cost}(P_{i+1}, C') \\
&\leq O(j^{j+1})\mathrm{cost}(P_{i+1}, C).
\end{aligned}$$

Combining the last two inequalities yields

$$|\mathrm{cost}(P_{i+1}, C) - \mathrm{cost}(Q_i, C)| \leq O(\frac{\epsilon'}{j^{j+1}}) \cdot \mathrm{cost}(P_{i+1}, C),$$

with probability at least $1 - \delta'/j - 1/n^2$.

Summing the last equation over all the j iterations of ADAPTIVESAMPLING yields

$$\begin{aligned}
&|\mathrm{cost}(P, C) - \mathrm{cost}(Q, C)| \\
&= |\mathrm{cost}(P, C) - \bigcup_{0 \leq i \leq j} \mathrm{cost}(S_i, C)| \\
&\leq |\sum_{0 \leq i \leq j} \big(\mathrm{cost}(P_{i+1}, C) - \mathrm{cost}(P_i, C) - \mathrm{cost}(S_i, C)\big)| \\
&= |\sum_{0 \leq i \leq j} \big(\mathrm{cost}(P_{i+1}, C) - \mathrm{cost}(Q_i, C)\big)| \\
&\leq \sum_{0 \leq i \leq j} |\mathrm{cost}(P_{i+1}, C) - \mathrm{cost}(Q_i, C)| \\
&\leq O(\frac{\epsilon'}{j^{j+1}}) \sum_{0 \leq i \leq j} \mathrm{cost}(P_{i+1}, C),
\end{aligned}$$

with probability at least $1 - \delta' - 1/n^2$.

By Lemma 3.1, there is an i-subspace C_i and a constant $0 < \nu_L \leq 1$, such that for any $p \in L$ we have $\mathrm{dist}(p, C) = \nu_L \cdot \mathrm{dist}(p, C_i)$. Hence, $|\mathrm{cost}(P_i, C) - \mathrm{cost}(P_{i+1}, C)| = \nu_L \cdot |\mathrm{cost}(P_i, C_i) - \mathrm{cost}(P_{i+1}, C_i)|$. We thus have

$$\begin{aligned}
&\mathrm{cost}(P_i, C) \\
&\leq \mathrm{cost}(P_{i+1}, C) + \mathrm{cost}(P_i, C) - \mathrm{cost}(P_{i+1}, C) \\
&\leq \mathrm{cost}(P_{i+1}, C) + |\mathrm{cost}(P_i, C) - \mathrm{cost}(P_{i+1}, C)| \\
&= \mathrm{cost}(P_{i+1}, C) + \nu_L \cdot |\mathrm{cost}(P_i, C_i) - \mathrm{cost}(P_{i+1}, C_i)| \\
&\leq \mathrm{cost}(P_{i+1}, C) + \nu_L \cdot \mathrm{cost}(P_{i+1}, C_i^*) \\
&\leq \mathrm{cost}(P_{i+1}, C) + \nu_L \cdot O(i^{i+1}) \cdot \mathrm{cost}(P_{i+1}, C_i) \\
&= \mathrm{cost}(P_{i+1}, C) + O(i^{i+1}) \cdot \mathrm{cost}(P_{i+1}, C) \\
&= O(i^{i+1}) \cdot \mathrm{cost}(P_{i+1}, C)
\end{aligned}$$

Hence,

$$\mathrm{cost}(P_{i+1}, C) \leq O(j^{j+1})\mathrm{cost}(P, C)$$

for every $0 \leq i \leq j$.
Combining the last inequalities together yields,

$$\begin{aligned}
&\mathbf{Pr}[|\mathrm{cost}(P, C) - \mathrm{cost}(Q, C)| \leq \epsilon' \mathrm{cost}(P, C)] \\
&\geq 1 - \delta' - 1/n^2.
\end{aligned}$$

□

5 Subspace Approximation

In this section we show how to construct in

$$O(nd \cdot \mathrm{poly}(j/\epsilon) + (n + d) \cdot 2^{\mathrm{poly}(j/\epsilon)})$$

time, a small set $\mathfrak{C}$ of candidate solutions (i.e., j-subspaces) such that $\mathfrak{C}$ contains a $(1+\epsilon/3)$-approximation to the subspace approximation problem, i.e., for the point set P, one of the j-subspaces in $\mathfrak{C}$ is a $(1 + \epsilon/3)$-approximation to the optimal j-subspace. Given such a candidate set $\mathfrak{C}$, we run the algorithm ADAPTIVESAMPLING with parameters $\delta/|\mathfrak{C}|$ and $\epsilon/6$. By the union bound it follows that every $C \in \mathfrak{C}$ is approximated by a factor of $(1 \pm \epsilon/6)$ with probability at least $1 - \delta$. It follows that the cost of the optimal candidate solution in $\mathfrak{C}$ is a $1 + O(\epsilon)$-approximation to the cost of the optimal j-subspace of the original set of points P.

The intuition behind the algorithm and the analysis. The first step of the algorithm is to invoke approximate volume sampling due to Deshpande and Varadarajan [10] to obtain in $O(nd \cdot \mathrm{poly}(j/\epsilon))$ time, an $\widetilde{O}(j^4 + (j/\epsilon)^3)$-dimensional subspace A that contains a $(1+\epsilon/6)$-approximation j-subspace. We use C_0 to denote a linear j-dimensional subspace of A with

$$\mathrm{cost}(P, C_0) \leq (1 + \epsilon/6) \cdot \mathrm{Opt}.$$

Our candidate set $\mathfrak{C}$ will consist of subspaces of A. Then the algorithm proceeds in j phases. In phase i, the algorithm computes a set G_i of points in A. We define $G_{\leq i} = \bigcup_{1 \leq l \leq i} G_l$. The algorithm maintains, with probability at least $1 - \frac{i \cdot \delta}{j}$, the invariant that i points from $G_{\leq i}$ span an i-subspace H_i such that there exists another j-subspace C_i, $H_i \subseteq C_i \subseteq A$, with

$$\begin{aligned}
\mathrm{cost}(P, C_i) &\leq (1 + \epsilon/6) \cdot (1 + \gamma)^i \cdot \mathrm{Opt} \\
&\leq (1 + \epsilon/3) \cdot \mathrm{Opt},
\end{aligned}$$

where Opt is the cost of an optimal subspace (not necessarily contained in A) and $\gamma = \epsilon/(12j)$ is an approximation parameter. The candidate set $\mathfrak{C}$ would be the spans of every j points from $G_{\leq j}$.

Notation	Meaning
$\mathfrak{C}$	The set of candidate solutions (i.e., j-subspaces)
A	The poly(j/ϵ)-subspace that contains a $(1+\epsilon/6)$-approximation
Opt	The cost of an optimal subspace (not necessarily contained in A)
C_i	A j-subspace which is a $(1+\gamma)^i$-approximation to the optimal j-subspace of P
H_i	An i-subspace which is the span of i points from $G_{\le i}$ where $H_i \subseteq C_i$
$H_i^\perp$	The orthogonal complement of the linear subspace H_i in $\mathbb{R}^d$
C_i^*	The projection of C_i on $H_i^\perp$
N_i	$\{p \in P_{i+1} : \text{dist}(p, C_i) \le 2 \cdot \text{cost}(P, C_i) \cdot \mathbf{Pr}[p]\}$
r_l	A point in $N_i \subseteq H_i^\perp$ that has $\mathbf{Pr}[r_l] > 0$
q in case 1	$\text{proj}(r_l, C_i^*)$
q in case 2	A point in $C_i^* \cap B(\text{proj}(r_l, C_i^*), 5 \cdot \text{Opt} \cdot \mathbf{Pr}[r_l], A \cap H_i^\perp)$ s.t. $\text{dist}(q, 0) \ge 5 \cdot \text{Opt} \cdot \mathbf{Pr}[r_l]$
q'	A point in $\mathcal{N}(r_l, 10 \cdot \text{Opt} \cdot \mathbf{Pr}[r_l], A \cap H_i^\perp, \gamma/20)$ s.t $\text{dist}(q, q') \le \frac{\gamma}{2} \cdot \text{Opt} \cdot \mathbf{Pr}[r_l]$
ℓ	$\text{Span}\{q\}$
ℓ'	$\text{Span}\{q'\}$
$C_i^\perp$	The orthogonal complement of ℓ in C_i
L_i	The orthogonal complement of $C_i^\perp$ in $\mathbb{R}^d$
C_{i+1}	A j-subspace which is the span of $C_i^\perp$ and ℓ'
$\mathcal{N}(p, R, A, \gamma)$	A γ-net of a ball $B(p, R, A)$ in the subspace A with radius R centered at p

Table 1: Notation in Section 5.

5.1 The algorithm. In the following, we present our algorithm to compute the candidate set. We use $H_i^\perp$ to denote the orthogonal complement of a linear subspace H_i in $\mathbb{R}^d$. We use $\mathcal{N}(p, R, A, \gamma)$ to denote a γ-net of a ball $B(p, R, A)$ in the subspace A with radius R centered at p, i.e. a set of points such that for every point $t \in B(p, R, A)$ there exists a point q in $\mathcal{N}(p, R, A, \gamma)$ with $\text{dist}(t, q) \le \gamma R$. It is easy to see that a γ-net of a ball $B(p, R, A)$ of size $O(\sqrt{d'}/\gamma^{d'})$ (See [2]) exists, where d' is the dimension of A. The input to the algorithm is the point set $P' = \text{proj}(P, A)$ in the space A, an i-dimensional linear subspace H_i and the parameters i and γ. The algorithm is invoked with $i = 0$ and $H_i = 0$ and j being the dimension of the subspace that is sought. Notice that the algorithm can be carried out in the space A since $H_i \subseteq A$ and so the projection of P' to $H_i^\perp$ will be inside A. Note, that although the algorithm doesn't know the cost Opt of an optimal solution, it is easy to compute the cost of an $O(j^{j+1})$-approximation using approximate volume sampling. From this approximation we can generate $O(j \log j)$ guesses for Opt, one of which includes a constant factor approximation.

CANDIDATESET (P', H_i, i, j, γ)

1. **if** $i = j$ **then return** H_i.
2. $P_{i+1} \leftarrow \text{proj}(P', H_i^\perp)$.
3. Sample $s = \lceil \log(j/\delta) \rceil$ points $r_1, \ldots, r_s$ i.i.d. from P_{i+1} s.t. each $p \in P_{i+1}$ is chosen with probability $\mathbf{Pr}[p] = \text{dist}(p, 0) / \sum_{q \in P_{i+1}} \text{dist}(q, 0)$
4. $G_{i+1} \leftarrow \bigcup_{l=1}^{s} \mathcal{N}(r_l, 10 \cdot \text{Opt} \cdot \mathbf{Pr}[r_l], A \cap H_i^\perp, \gamma/20)$.
5. **return** $\bigcup_{q \in G_{i+1}}$ CANDIDATESET$(P', \text{Span}\{H_i \cup q\}, i+1, j, \gamma)$.

5.2 Invariant of algorithm CANDIDATESET**.** We will prove that the algorithm satisfies the following lemma.

LEMMA 5.1. *Let $C_i \subseteq A$ be a subspace that contains H_i. Assume that C_i is a $(1+\gamma)^i$-approximation to the optimal j-subspace of P. Then, with probability at least $1 - \delta/j$, there is a j-subspace $C_{i+1} \subseteq A$ containing H_i and a point from G_{i+1}, such that C_{i+1} is a $(1+\gamma)^{i+1}$ approximation to the optimal j-subspace of P.*

Once the lemma is proved, we can apply it inductively to show that with probability at least $1 - \delta$ we have a subspace C_j that is spanned by j points from $G_{\le j}$ and

that has

$$\begin{aligned}\text{cost}(P, C_j) &\leq (1+\gamma)^j \cdot \text{cost}(P, C_0)\\ &\leq (1+\epsilon/6) \cdot (1+\gamma)^j \cdot \text{Opt}\\ &\leq (1+\epsilon/6) \cdot (1+\epsilon/12) \cdot \text{Opt}\\ &\leq (1+\epsilon/3) \cdot \text{Opt}.\end{aligned}$$

The running time of the algorithm is dominated by the projections in line 2, j of which are carried out for each element of the candidate set. Since the input P' to the algorithm is in the subspace A, its running time is $n \cdot 2^{\text{poly}(j/\epsilon)}$. To initialize the algorithm, we have to compute space A and project all points on A. This can be done in $O(nd \cdot \text{poly}(j/\epsilon))$ time [10].

Finally, we run algorithm ADAPTIVESAMPLING to approximate the cost for every candidate solution generated by algorithm CANDIDATESET. For each candidate solution, we have to project all points on its span. This can be done in $O(d \cdot 2^{\text{poly}(j/\epsilon)})$ time, since the number of candidate solutions is $2^{\text{poly}(j/\epsilon)}$ and the size of the sample is $\text{poly}(j/\epsilon)$. Thus we can summarize the result in the following theorem setting $\delta = 1/6$ in the approximate volume sampling and in our algorithm.

THEOREM 5.1. *Let P be a set of n points in $\mathbb{R}^d$, $0 < \epsilon < 1$ and $1 \leq j \leq d$. A $(1+\epsilon)$-approximation for the j-subspace approximation problem can be computed, with probability at least $2/3$, in time*

$$O(nd \cdot \text{poly}(j/\epsilon) + (n+d) \cdot 2^{\text{poly}(j/\epsilon)}).$$

5.3 Overview of the proof of Lemma 5.1. The basic idea of the proof follows earlier results of [25]. We show that by sampling with probability proportional to the distance from the origin, we can find a point p whose distance to the optimal solution is only a constant factor more than the weighted average distance (where the weighting is done according to the distance from the origin). If we then consider a ball with radius a constant times the average weighted distance and that is centered at p, then this ball must intersect the projection of the current space C_i solution on $H_i^\perp$. If we now place a fine enough net on this ball, then there must be a point q of this net that is close to the projection. We can then define a certain rotation of the current subspace to contain q to obtain the new subspace C_{i+1}. This rotation increases the cost only slightly and C_{i+1} contains $\text{Span}\{H_i \cup \{q\}\}$.

5.4 The complete proof of Lemma 5.1. We assume that there is a j-subspace C_i, $H_i \subseteq C_i \subseteq A$, with

$$\begin{aligned}\text{cost}(P, C_i) &\leq (1+\gamma)^i \cdot \text{cost}(P, C_0)\\ &\leq (1+\epsilon/3) \cdot \text{Opt}.\end{aligned}$$

We use C_i^* to denote the projection of C_i on $H_i^\perp$. Note that C_i^* has $j-i$ dimensions as $H_i \subseteq C_i$. The idea is to find a point q from $G_{i+1} \subseteq H_i^\perp \cap A$ such that we can rotate C_i^* in a certain way to contain q and this rotation will not change the cost with respect to P significantly. Let

$$N_i = \{p \in P_{i+1} : \text{dist}(p, C_i) \leq 2 \cdot \text{cost}(P, C_i) \cdot \mathbf{Pr}[p]\}.$$

N_i contains all points that are close to the subspace C_i, where closeness is defined relative to the distance from the origin. We will first show that by sampling points with probability proportional to their distance from the origin, we are likely to find a point from N_i.

PROPOSITION 5.1.

$$\mathbf{Pr}[\exists r_l, 1 \leq l \leq s : r_l \in N_i] \geq 1 - \delta/j \ .$$

Proof. We first prove by contradiction that the probability to sample a point from N_i is at least $1/2$. Assume that

$$\sum_{p \in P_{i+1} \setminus N_i} \mathbf{Pr}[p] > 1/2.$$

Observe that $\text{cost}(P, C_i) \geq \text{cost}(P', C_i)$ since $C_i \subseteq A$ and $P' = \text{proj}(P, A)$. Further, $\text{cost}(P', C_i) = \text{cost}(P_{i+1}, C_i)$ since $P_{i+1} = \text{proj}(P', H_i^\perp)$ and $H_i \subseteq C_i$. It follows that

$$\begin{aligned}\text{cost}(P, C_i) \geq \text{cost}(P', C_i) &= \text{cost}(P_{i+1}, C_i)\\ &\geq \sum_{p \in P_{i+1} \setminus N_i} \text{dist}(p, C_i)\\ &> 2 \cdot \text{cost}(P, C_i) \cdot \sum_{p \in P_{i+1} \setminus N_i} \mathbf{Pr}[p]\\ &> \text{cost}(P, C_i),\end{aligned}$$

which is a contradiction. Hence,

$$\mathbf{Pr}[r_l \in N_i] = \sum_{p \in N_i} \mathbf{Pr}[p] \geq 1/2.$$

It follows that

$$\begin{aligned}\mathbf{Pr}[\exists l, 1 \leq l \leq s : r_l \in N_i] &\geq 1 - (1-1/2)^s\\ &\geq 1 - \delta/j.\end{aligned}$$

□

We now make a case distinction in order to prove Lemma 5.1.

Case 1: Points are on average much closer to C_i than to the origin.
We first consider the case that

$$\sum_{p\in P_{i+1}} \text{dist}(p,0) \geq 4 \sum_{p\in P_{i+1}} \text{dist}(p,C_i).$$

In this case, the points in N_i are much closer to C_i than to the origin.

Now let r_l be a point from $N_i \subseteq H_i^\perp$ that has $\mathbf{Pr}[r_l] > 0$. Since $C_i \subseteq A$ and

$$\text{dist}(r_l, C_i^*) = \text{dist}(r_l, C_i) \leq 2 \cdot \text{cost}(P, C_i) \cdot \mathbf{Pr}[r_l]$$

we know that $B(r_l, 10 \cdot \text{Opt} \cdot \mathbf{Pr}[r_l], A \cap H_i^\perp)$ intersects C_i^*. This also implies that $q := \text{proj}(r_l, C_i^*)$ lies in $B(r_l, 10 \cdot \text{Opt} \cdot \mathbf{Pr}[r_l], A \cap H_i^\perp)$. Hence, there is a point

$$q' \in \mathcal{N}(r_l, 10 \cdot \text{Opt} \cdot \mathbf{Pr}[r_l], A \cap H_i^\perp, \gamma/20)$$

with $\text{dist}(q,q') \leq \frac{\gamma}{2} \cdot \text{Opt} \cdot \mathbf{Pr}[r_l]$.

Let ℓ be the line through q and let ℓ' be the line through q'. Let $C_i^\perp$ denote the orthogonal complement of ℓ in C_i. Define the subspace C_{i+1} as the span of $C_i^\perp$ and ℓ'. Since q lies in C_i^* (and hence in $H_i^\perp$) we have that $C_i^\perp$ contains H_i. Hence, C_{i+1} also contains H_i. It remains to show that

$$\text{cost}(P, C_{i+1}) \leq (1+\gamma) \cdot \text{cost}(P, C_i).$$

We have

$$\begin{aligned}
&\text{cost}(P, C_{i+1}) - \text{cost}(P, C_i) && (5.9)\\
&\leq \sum_{p\in P} \text{dist}(\text{proj}(p, C_i), C_{i+1}) && (5.10)\\
&= \sum_{p\in P} \text{dist}(\text{proj}(\text{proj}(p, A), C_i), C_{i+1}) && (5.11)\\
&= \sum_{p\in P'} \text{dist}(\text{proj}(p, C_i), C_{i+1}) && (5.12)\\
&= \sum_{p\in P_{i+1}} \text{dist}(\text{proj}(p, C_i), C_{i+1}) && (5.13)
\end{aligned}$$

where Step 5.11 follows from the fact that $C_i \subseteq A$ and so $\text{proj}(\text{proj}(p, A), C_i) = \text{proj}(p, C_i)$ for all $p \in \mathbb{R}^d$ and Step 5.13 follows from $H_i \subseteq C_i, C_{i+1}$.

Now define L_i to be the orthogonal complement of $C_i^\perp$ in $\mathbb{R}^d$. Note that for any $p \in \mathbb{R}^d$ and its projection $p' = \text{proj}(p, L_i)$ we have $\text{dist}(p, C_i) = \text{dist}(p', C_i)$ and $\text{dist}(p, C_{i+1}) = \text{dist}(p', C_{i+1})$. Further observe that C_i corresponds to the line ℓ in L_i and C_{i+1} corresponds to a line $\ell'' = \text{proj}(\ell', L_i)$. Define α to be the angle between ℓ and ℓ' and β the angle between ℓ and ℓ''. Note that $\alpha \geq \beta$. Then

$$\begin{aligned}
\text{dist}(\text{proj}(p, C_i), C_{i+1}) &= \text{dist}(\text{proj}(\text{proj}(p, C_i), L_i), \ell'')\\
&= \text{dist}(\text{proj}(p, \ell), \ell'').
\end{aligned}$$

This implies

$$\begin{aligned}
\text{dist}(\text{proj}(p, \ell), \ell'') &= \text{dist}(\text{proj}(p, \ell), 0) \cdot \sin\beta\\
&\leq \text{dist}(p, 0) \cdot \sin\alpha.
\end{aligned}$$

We need the following claim that the distance of q to the origin is not much smaller than the distance of r_l to the origin.

PROPOSITION 5.2. *If*

$$\sum_{p\in P_{i+1}} \text{dist}(p,0) \geq 4 \sum_{p\in P_{i+1}} \text{dist}(p, C_i)$$

then

$$\text{dist}(q,0) \geq \frac{1}{2}\text{dist}(r_l, 0).$$

Proof. Since $r_l \in N_i$ we have

$$\text{dist}(r_l, C_i) \leq 2\text{Opt}\frac{\text{dist}(r_l,0)}{\sum_{p\in P_{i+1}} \text{dist}(p,0)}.$$

By our assumption we have

$$\sum_{p\in P_{i+1}} \text{dist}(p,0) \geq 4 \sum_{p\in P_{i+1}} \text{dist}(p, C_i),$$

which implies $\text{dist}(r_l, C_i) \leq \frac{1}{2}\text{dist}(r_l, 0)$ by plugging in into the previous inequality. We further have $\text{dist}(r_l, C_i) = \text{dist}(r_l, C_i^*)$ and so

$$\text{dist}(q,0) \geq \text{dist}(r_l, 0) - \text{dist}(r_l, C_i^*) \geq \frac{1}{2}\text{dist}(r_l, 0)$$

by the triangle inequality. □

We get

$$\begin{aligned}
\sin\alpha &\leq \frac{\text{dist}(q,q')}{\text{dist}(q,0)}\\
&\leq \frac{1/2 \cdot \gamma \cdot \text{Opt} \cdot \mathbf{Pr}[r_l]}{1/2 \cdot \text{dist}(r_l, 0)}\\
&= \frac{\gamma \cdot \text{Opt} \cdot \text{dist}(r_l, 0)}{\text{dist}(r_l, 0) \cdot \sum_{p\in P_{i+1}} \text{dist}(p,0)}\\
&= \frac{\gamma \cdot \text{Opt}}{\sum_{p\in P_{i+1}} \text{dist}(p,0)}.
\end{aligned}$$

The latter implies

$$\begin{aligned}
\text{cost}(P, C_{i+1}) - \text{cost}(P, C_i) &\leq \sum_{p\in P_{i+1}} \text{dist}(p,0) \cdot \sin\alpha\\
&\leq \gamma \cdot \text{Opt}\\
&\leq \gamma \cdot \text{cost}(P, C_i)
\end{aligned}$$

which implies the lemma in Case 1.

Case 2: Points are on average much closer to the origin than to C_i.
Now we consider the case that

$$\sum_{p\in P_{i+1}} \mathrm{dist}(p,0) < 4 \sum_{p\in P_{i+1}} \mathrm{dist}(p,C_i).$$

Let r_l be a point from $P_{i+1} \subseteq H_i^{\perp}$ that is in N_i and that has $\mathbf{Pr}[r_l] > 0$. Since $C_i \subseteq A$ and

$$\mathrm{dist}(r_l, C_i^*) = \mathrm{dist}(r_l, C_i) \le 2 \cdot \mathrm{cost}(P, C_i) \cdot \mathbf{Pr}[r_l],$$

we know that $B(r_l, 10 \cdot \mathrm{Opt} \cdot \mathbf{Pr}[r_l], A \cap H_i^{\perp})$ intersects C_i^*. This implies that $\mathrm{proj}(r_l, C_i^*)$ lies also in $B(r_l, 10 \cdot \mathrm{Opt} \cdot \mathbf{Pr}[r_l], A \cap H_i^{\perp})$.

In fact,

$$2 \cdot \mathrm{cost}(P, C_i) \cdot \mathbf{Pr}[r_l] \le 5 \cdot \mathrm{Opt} \cdot \mathbf{Pr}[r_l]$$

implies that

$$B(\mathrm{proj}(r_l, C_i^*), 5 \cdot \mathrm{Opt} \cdot \mathbf{Pr}[r_l], A \cap H_i^{\perp}) \subseteq B(r_l, 10 \cdot \mathrm{Opt} \cdot \mathbf{Pr}[r_l], A \cap H_i^{\perp}).$$

Since $C_i^* \subseteq A \cap H_i^{\perp}$ we also have that there is a point

$$q \in C_i^* \cap B(\mathrm{proj}(r_l, C_i^*), 5 \cdot \mathrm{Opt} \cdot \mathbf{Pr}[r_l], A \cap H_i^{\perp})$$

with $\mathrm{dist}(q,0) \ge 5 \cdot \mathrm{Opt} \cdot \mathbf{Pr}[r_l]$.

Now consider the set which is the intersection of

$$\mathcal{N}(r_l, 10 \cdot \mathrm{Opt} \cdot \mathbf{Pr}[r_l], A \cap H_i^{\perp}, \gamma/20)$$

with

$$B(\mathrm{proj}(r_l, C_i), 5 \cdot \mathrm{Opt} \cdot \mathbf{Pr}[r_l], A \cap H_i^{\perp}),$$

which is a $(\gamma/10)$-net of

$$B(\mathrm{proj}(r_l, C_i), 5 \cdot \mathrm{Opt} \cdot \mathbf{Pr}[r_l], A \cap H_i^{\perp}).$$

Hence, there is a point

$$q' \in \mathcal{N}(r_l, 10 \cdot \mathrm{Opt} \cdot \mathbf{Pr}[r_l], A \cap H_i^{\perp}, \gamma/20)$$

with $\mathrm{dist}(q,q') \le \frac{\gamma}{10} \cdot 5 \cdot \mathrm{Opt} \cdot \mathbf{Pr}[r_l] \le \gamma \cdot \mathrm{Opt} \cdot \mathbf{Pr}[r_l]$.

Let ℓ be the line through q and let ℓ' be the line through q'. Let $C_i^{\perp}$ denote the orthogonal complement of ℓ in C_i. Define the subspace C_{i+1} as the span of $C_i^{\perp}$ and ℓ'. Since q lies in C_i^* we have that $C_i^{\perp}$ contains H_i. Hence, C_{i+1} also contains H_i.

It remains to show that

$$\mathrm{cost}(P, C_{i+1}) \le (1+\gamma) \cdot \mathrm{cost}(P, C_i).$$

Now define L_i to be the orthogonal complement of $C_i^{\perp}$. Note that for any $p \in \mathbb{R}^d$ and its projection $p' = \mathrm{proj}(p, L_i)$ we have $\mathrm{dist}(p, C_i) = \mathrm{dist}(p', C_i)$ and $\mathrm{dist}(p, C_{i+1}) = \mathrm{dist}(p', C_{i+1})$. Further observe that C_i corresponds to the line ℓ in L_i and C_{i+1} corresponds to a line $\ell'' = \mathrm{proj}(\ell', L_i)$.

Define α to be the angle between ℓ and ℓ' and β the angle between ℓ and ℓ''. Note that $\alpha \ge \beta$. Then

$$\begin{aligned}\mathrm{dist}(\mathrm{proj}(p, C_i), C_{i+1}) &= \mathrm{dist}(\mathrm{proj}(\mathrm{proj}(p, C_i), L_i), \ell'') \\ &= \mathrm{dist}(\mathrm{proj}(p, \ell), \ell'').\end{aligned}$$

This implies

$$\begin{aligned}\mathrm{dist}(\mathrm{proj}(p, \ell), \ell'') &= \mathrm{dist}(\mathrm{proj}(p, \ell), 0) \cdot \sin\beta \\ &\le \mathrm{dist}(p, 0) \cdot \sin\alpha.\end{aligned}$$

We have

$$\sin\alpha \le \frac{\gamma \cdot \mathrm{Opt} \cdot \mathbf{Pr}[r_l]}{5 \cdot \mathrm{Opt} \cdot \mathbf{Pr}[r_l]} \le \frac{\gamma}{5}.$$

Similar to the first case it follows that

$$\begin{aligned}\mathrm{cost}(P, C_{i+1}) - \mathrm{cost}(P, C_i) &\le \sum_{p\in P_{i+1}} \mathrm{dist}(p,0) \cdot \sin\alpha \\ &\le \frac{\gamma}{5} \cdot \sum_{p\in P_{i+1}} \mathrm{dist}(p,0).\end{aligned}$$

Since we are in Case 2 we have

$$\sum_{p\in P_{i+1}} \mathrm{dist}(p,0) < 4 \cdot \mathrm{cost}(P_{i+1}, C_i),$$

which implies

$$\begin{aligned}\mathrm{cost}(P, C_{i+1}) - \mathrm{cost}(P, C_i) &\le \frac{\gamma}{5} \cdot \sum_{p\in P_{i+1}} \mathrm{dist}(p,0) \\ &\le \gamma \cdot \mathrm{cost}(P_{i+1}, C_i) \\ &\le \gamma \cdot \mathrm{cost}(P, C_i).\end{aligned}$$

This concludes the proof of Lemma 5.1.

6 Streaming Algorithms in the Read-Only Model

We can maintain our coreset with

$$\tilde{O}\left(d\left(\frac{j 2^{O(\sqrt{\log n})}}{\epsilon^2}\right)^{\mathrm{poly}(j)}\right)$$

(weighted) points via known merge and reduce technique [1, 16] in the read-only streaming model where only insertion of a point is allowed. The presence of negative points makes the process of maintaining a coreset harder. The problem is that the sum of the absolute weights of the coreset is about three times the size of the input point set.

If we now apply our coreset construction several times (as is required during merge and reduce), we blow up the sum of absolute weights with each application by a constant factor. This blow-up, together with the fact that we have to estimate the difference between positively and negatively weighted points, cannot be controlled as well as in the case of a standard merge and reduce approach, and requires taking larger sample sizes with every merge step. The proof of the following theorem will appear in the full version of this paper.

THEOREM 6.1. *Let C be a fixed j-subspace of $\mathbb{R}^d$. Let P be a set of n points in $\mathbb{R}^d$, $j \geq 0$, and $\epsilon, \delta > 0$. In the read-only streaming model we can maintain two sets S'' and Q using*

$$\tilde{O}\left(d\left(\frac{j \cdot 2^{\sqrt{\log n}}}{\epsilon^2}\right)^{\mathrm{poly}(j)}\right)$$

weighted points such that, with probability at least $1-\delta$,

$$|\mathrm{cost}(P,C) - \mathrm{cost}(S'',C) - \mathrm{cost}(Q,C)| \leq \epsilon \cdot \mathrm{cost}(P,C).$$

Moreover, $\tilde{O}()$ notation hides $\mathrm{poly}(\log n)$ factors.

7 Streaming Algorithms with Bounded Precision in the Turnstile Model

In this section, we consider the problems of previous sections in the 1-pass turnstile streaming model. In this model, coordinates of points may arrive in an arbitrary order and undergo multiple updates. We shall assume that matrices and vectors are represented with *bounded precision*, that is, their entries are integers between $-\Delta$ and Δ, where $\Delta \geq (nd)^B$, and $B > 1$ is a constant. We also assume that $n \geq d$.

The rank of a matrix A is denoted by $\mathrm{rank}(A)$. The best rank-j approximation of A is the matrix A_j such that $\|A - A_j\|_F$ is minimized over every matrix of rank at most j, and $\|\cdot\|_F$ is the Frobenius norm (sum of squares of the entries). Recall that using the singular value decomposition every matrix A can be expressed as $U\Sigma V^T$, where the columns of U and V are orthonormal, and Σ is a diagonal matrix with the singular values along the diagonal (which are all positive).

LEMMA 7.1. ([7]) *Let A be an $n \times d$ integer matrix represented with bounded precision. Then for every w, $1 \leq w \leq \mathrm{rank}(A)$, the w-th largest singular value of A is at least $1/\Delta^{5(w-1)/2}$.*

COROLLARY 7.1. *Let A be an $n \times d$ integer matrix represented with bounded precision. Let A_j be the best rank-j approximation of A. If $\mathrm{rank}(A) \geq j+1$ then $\|A - A_j\|_F \geq 1/\Delta^{5j/2}$.*

Proof. Let σ_{j+1} denote the $(j+1)$-th singular value of A. By Lemma 7.1, we have $\|A - A_j\|_2 \geq \sigma_{j+1} \geq 1/\Delta^{5j/2}$, where the first inequality follows by standard properties of the singular values. □

For an $n \times d$ matrix A, let $F_q(\ell_p)(A) = \sum_{i=1}^{n} \|A^{(i)}\|_p^q$, where A^i is the i-th row of A. Let $A_{j,p}^q$ be the matrix which minimizes $F_q(\ell_p)(B-A)$ over every $n \times d$ matrix B of rank j. The next corollary follows from relations between norms and singular values.

COROLLARY 7.2. *Let A be an $n \times d$ matrix, and $p, q = O(1)$. If $\mathrm{rank}(A) \geq j+1$ then*

$$F_q(\ell_p)(A - A_{j,p}^q) \geq 1/\Delta^{O(j)}.$$

Proof. By Corollary 7.1,

$$F_2(\ell_2)(A_j - A) = \sum_{i=1}^{n} \|(A_j)^i - A^i\|_2^2 \geq 1/\Delta^{5j}.$$

We use the following relations between norms. Let x be a d-dimensional vector. For any $a \geq b$,

$$\frac{\|x\|_b}{d^{(a-b)/ab}} \leq \|x\|_a \leq \|x\|_b. \tag{7.14}$$

It follows that for any $p \leq 2$,

$$\begin{aligned} F_2(\ell_p)(A_{j,p}^q - A) &= \sum_{i=1}^{n} \|(A_{j,p}^q)^i - A^i\|_p^2 \\ &\geq \sum_{i=1}^{n} \|(A_{j,p}^q)^i - A^i\|_2^2 \\ &\geq \sum_{i=1}^{n} \|(A_j)^i - A^i\|_2^2 \\ &\geq 1/\Delta^{5j}, \end{aligned}$$

On the other hand, if $p > 2$,

$$\begin{aligned} F_2(\ell_p)(A_{j,p}^q - A) &= \sum_{i=1}^{n} \|(A_{j,p}^q)^i - A^i\|_p^2 \\ &\geq \sum_{i=1}^{n} \left(\frac{\|(A_{j,p}^q)^i - A^i\|_2}{d^{(p-2)/(2p)}}\right)^2 \\ &\geq \sum_{i=1}^{n} \frac{\|(A_j)^i - A^i\|_2^2}{d^{(p-2)/p}} \\ &\geq \frac{1}{\Delta^{5j+1}}, \end{aligned}$$

where we use $\Delta \geq nd$. Hence, in either case,

$$\left(F_2(\ell_p)(A_{j,p}^q - A)\right)^{1/2} \geq 1/\Delta^{5j/2+1/2}.$$

Again, appealing to the right part of (7.14), if $q \leq 2$, then

$$\left(F_q(\ell_p)(A_{j,p}^q - A)\right)^{1/q} \geq \left(F_2(\ell_p)(A_{j,p}^q - A)\right)^{1/2} \geq \frac{1}{\Delta^{5j/2+1/2}}.$$

If instead $q > 2$,

$$\left(F_q(\ell_p)(A_{j,p}^q - A)\right)^{1/q} \geq \frac{\left(F_2(\ell_p)(A_{j,p}^q - A)\right)^{1/2}}{n^{(q-2)/(2q)}} \geq \frac{\left(F_2(\ell_p)(A_{j,p}^q - A)\right)^{1/2}}{n^{1/2}} \geq \frac{1}{\Delta^{5j/2+1}},$$

using that $\Delta \geq nd$. Hence, in all cases,

$$F_q(\ell_p)(A_{j,p}^q - A) \geq 1/\Delta^{5jq/2+q} = 1/\Delta^{\Theta(j)}.$$

□

In the remainder of the section, we shall assume that p and q are in the interval $[1, 2]$. It is known [5, 6] that estimating $\|x\|_r$ for any $r > 2$ requires polynomial space in the turnstile model, so this assumption is needed. Also, $p, q \geq 1$ in order to be norms.

We start by solving approximate linear regression. We use this to efficiently solve distance to subspace approximation. Finally, we show how to efficiently $(1+\epsilon)$-approximate the best rank-j approximation via a certain discretization of subspaces. Our space is significantly less than the input matrix description $O(nd \log(nd))$.

7.1 Approximate Linear Regression

DEFINITION 7.1. (APPROXIMATE LINEAR REGRESSION) *Let A be an $n \times d$ matrix, and b be an $n \times 1$ vector. Assume that A and b are represented with bounded precision, and given in a stream. The approximate linear regression problem is to output a vector $x' \in \mathbb{R}^d$ so that with probability at least $2/3$,*

$$\left| \|Ax' - b\|_p - \min_{x \in \mathbb{R}^d} \|Ax - b\|_p \right| \leq \epsilon \min_{x \in \mathbb{R}^d} \|Ax - b\|_p.$$

Let $G_{p,\gamma,d,Z}$ be the d-dimensional grid in $\mathbb{R}^d$ of all points whose entries are integer multiples of $\gamma/(d^{1+1/p}\Delta)$, and bounded in absolute value by Z. We show that if we restrict the solution space to the grid $G_{p,\gamma,d,\Delta^{\Theta(d)}}$, then we can minimize $\|Ax - b\|_p$ up to a small additive error γ. The proof follows by bounding the entries of an optimal solution $x^* \in \mathbb{R}^d$, where

$$x^* = \text{argmin}_{x \in \mathbb{R}^d} \|Ax - b\|_p.$$

LEMMA 7.2. *Suppose A is an $n \times d$ matrix, and b is an $n \times 1$ column vector with bounded precision. Then,*

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|_p \leq \min_{x \in G_{p,\gamma,d,\Delta^{\Theta(d)}}} \|Ax - b\|_p \leq \min_{x \in \mathbb{R}^d} \|Ax - b\|_p + \gamma.$$

Proof. Let $x^* = \text{argmin}_{x \in \mathbb{R}^d} \|Ax - b\|_p$.

We first argue that the entries of x^* cannot be too large. We can suppose $x^* \neq 0^d$, as otherwise the entries are all bounded in absolute value by 0. By the triangle inequality,

$$\|Ax^* - b\|_p + \|b\|_p \geq \|Ax^*\|_p.$$

Now,

$$\|Ax^* - b\|_p + \|b\|_p \leq 2\|b\|_p \leq 2n\Delta.$$

Also, $\|Ax^*\|_p \geq \|Ax^*\|_2$. Since $x^* \neq 0^d$, it holds that $\|Ax^*\|_2 \geq \sigma_r \|x^*\|_2$, where $r = \text{rank}(A)$ and σ_r is the smallest singular value of A. Hence,

$$\|x^*\|_2 \leq 2n\Delta/\sigma_r.$$

By Lemma 7.1, $\sigma_r \geq \Delta^{-5d/2}$, and so

$$\|x^*\|_\infty \leq \|x^*\|_2 \leq 2n\Delta \cdot \Delta^{5d/2} \leq \Delta^{6d}.$$

Put $G = G_{p,\gamma,d,\Delta^{6d}}$. Then

$$\begin{aligned} \min_{x \in G} \|Ax - b\|_p &\leq \min_{x \in \mathbb{R}^d} \|Ax - b\|_p \\ &\leq \max_{y \in \{0, \gamma/(d^{1+1/p}\Delta)\}^d} \min_{x \in G} \|Ax + Ay - b\|_p \\ &\leq \min_{x \in G} \|Ax - b\|_p + \max_{y \in \{0, \gamma/(d^{1+1/p}\Delta)\}^d} \|Ay\|_p \\ &\leq \min_{x \in G} \|Ax - b\|_p + (d(d\Delta\gamma/(d^{1+1/p}\Delta)^p)^{1/p} \\ &\leq \min_{x \in G} \|Ax - b\|_p + \gamma. \end{aligned}$$

□

We use the following sketching result (see also [17, 23]).

THEOREM 7.1. ([19]) *For $1 \leq p \leq 2$, if one chooses the entries of an $(\log 1/\delta)/\epsilon^2 \times n$ matrix S with entries that are p-stable $O(\log 1/\epsilon)$-wise independent random variables rounded to the nearest integer multiple of $\Delta^{-2} = (nd)^{-2B}$ and bounded in absolute value by $\Delta^2 = (nd)^{2B}$, then for any fixed $x \in \mathbb{R}^n$, with integer entries bounded in absolute value by Δ, there is an efficient algorithm $\mathcal{A}$ which, given Sx, outputs a $(1 \pm \epsilon/3)$-approximation to $\|x\|_p$ with probability at least $1 - \delta$. The algorithm can be implemented in $O(\log(nd) \log 1/\delta)/\epsilon^2$ bits of space. Moreover, it can be assumed to output an implicit representation of S.*

THEOREM 7.2. *There is a 1-pass algorithm, which given the entries of A and b in a stream, solves the Approximate Linear Regression Problem with $O(d^3 \log^2(nd))/\epsilon^2$ bits of space and $\Delta^{\Theta(d^2)}$ time (we will only invoke this later as a subroutine for small values of d).*

Proof. [of Theorem 7.2] We first consider the case that b is in the columnspace of A. In this case, we have

$$0 = \min_{x \in \mathbb{R}^d} \|Ax - b\|_p = \min_{x \in \mathbb{R}^d} \|Ax - b\|_2.$$

Let y be such that $Ay = b$. In [24], it is shown how to recover y with $O(d^2 \log(nd))$ bits of space with probability at least $11/12$, and to simultaneously report that b is in the columnspace of A.

We now consider the case that b is not in the columnspace of A. We seek to lower bound $\min_{x \in \mathbb{R}^d} \|Ax - b\|_p$. Consider the $n \times (d+1)$ matrix A' whose columns are the columns $a_1, \ldots, a_d$ of A adjoined to b. Also consider any $n \times (d+1)$ matrix T whose columns are in the columnspace of A. Then

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|_p = \min_{T} \|T - A'\|_p.$$

Since b is not in the columnspace of T, $\text{rank}(A') = \text{rank}(T) + 1$. By Corollary 7.2, it follows that

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|_p \geq 1/\Delta^{\Theta(d)}.$$

Put $\gamma = \epsilon/(3\Delta^{\Theta(d)})$, and let $G = G_{p,\gamma,d,\Delta^{\Theta(d)}}$ be as defined above, so

$$|G| \leq (3d^{1+1/p}\Delta^{\Theta(d)}/\epsilon)^d.$$

For $1 \leq p \leq 2$, let S be a random $(\log 1/\delta')/\epsilon^2 \times n$ matrix as in Theorem 7.1, where $\delta' = \Theta(1/|G|)$. The algorithm maintains $S \cdot A$ in the data stream, which can be done with $O(d^3 \cdot \log^2(nd))/\epsilon^2$ bits of space. Let $\mathcal{A}$ be the efficient algorithm in Theorem 7.1. Then for a sufficiently small δ', with probability at least $3/4$, for every $x \in G$,

$$\begin{aligned} &|\mathcal{A}(SAx - Sb) - \|Ax - b\|_p| \\ &\leq \frac{\epsilon}{3}\|Ax - b\|_p. \end{aligned} \tag{7.15}$$

By Lemma 7.2, there is an $x' \in G$ for which

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \|Ax - b\|_p &\leq \|Ax' - b\|_p \\ &\leq \min_{x \in \mathbb{R}^d} \|Ax - b\|_p + \frac{\epsilon}{3\Delta^{\Theta(d)}}. \end{aligned}$$

Moreover,

$$\begin{aligned} \left(1 - \frac{\epsilon}{3}\right)\|Ax' - b\|_p &\leq \mathcal{A}(SAx' - Sb) \\ &\leq \left(1 + \frac{\epsilon}{3}\right)\|Ax' - b\|_p. \end{aligned}$$

Moreover, for $\epsilon \leq 1$, $(1 + \epsilon/3)\epsilon/3 \leq 2\epsilon/3$. Hence,

$$\begin{aligned} \left(1 - \frac{\epsilon}{3}\right) \min_{x \in \mathbb{R}^d} \|Ax - b\|_p &\leq \mathcal{A}(SAx' - Sb) \\ &\leq (1 + \epsilon) \min_{x \in \mathbb{R}^d} \|Ax - b\|_p. \end{aligned}$$

By a union bound, the algorithms succeed with probability $\geq 3/4 - 1/12 = 2/3$. The time complexity is dominated by enumerating grid points, which can be done in $\Delta^{\Theta(d^2)} = (nd)^{\Theta(j^2)}$ time. This assumes that $\epsilon > \Delta^{-\Theta(d)}$, since when $\epsilon = \Delta^{-\Theta(d)}$ the problem reduces to exact computation. The theorem follows. □

7.2 Distance to Subspace Approximation Given an $n \times d$ integer matrix A in a stream with bounded precision, we consider the problem of maintaining a sketch of A so that from the sketch, for any subspace F in $\mathbb{R}^d$ of dimension j, represented by a $j \times d$ matrix of bounded precision, with probability at least $2/3$, one can output a $(1 \pm \epsilon)$-approximation to

$$F_q(\ell_p)(\text{proj}_F(A) - A),$$

where $\text{proj}_F(A)$ is the projection of A onto F.

THEOREM 7.3. *For $p, q \in [1, 2]$, there is a 1-pass algorithm, which solves the Distance to Subspace Approximation Problem with $O(nj^3 \log^3(nd)/\epsilon^2)$ bits of space and $\Delta^{O(j^2)}$ time. If $p = 2$ this can be improved to $O(nj^2 \log(nd))/\epsilon$ space and $\text{poly}(j \log(nd)/\epsilon)$ time.*

Proof. Set $\delta = 1/(3n)$. We sketch each of the n rows of A as in the algorithm of Theorem 7.2. That algorithm also outputs a representation of its sketching matrix S. In the offline phase, we are given F, and we compute $F \cdot S$. We independently solve the ℓ_p-regression problem with matrix F and each of the n rows of A. For each row A^i, we approximate $\min_{x \in \mathbb{R}^j} \|xF - A^i\|_p$. By a union bound, from the estimated costs for the rows, we get a $(1 \pm \epsilon)$-approximation to $F_q(\ell_p)(\text{proj}_F(A) - A)$. The value of d in the invocation of Theorem 7.2 is j. Using results in [7], for $p = 2$ this can be somewhat improved. □

7.3 Best Rank-j Approximation Given an $n \times d$ matrix A with bounded precision in a stream, we consider the problem of maintaining a sketch of A so that one can $(1 + \epsilon)$-approximate the value $F_q(\ell_p)(A^q_{j,p} - A)$ with probability at least $2/3$. We shall only consider $p = 2$. We first give a 1-pass algorithm near-optimal in space, but with poor running time, using sketches of [18]. We then improve this to achieve polynomial running time. Note that the case $(q, p) = (2, 2)$ was solved in [7].

For the time-inefficient solution, the work of [18] gives a sketch SA of the $n \times d$ input matrix A so that

$F_q(\ell_2)(A)$ (recall $q \leq 2$) is estimable to within $(1+\epsilon)$ with probability $1-\delta$ using $\text{poly}(\log(nd)/\epsilon)\log 1/\delta$ bits of space, where all entries are integer multiples of $1/\Delta$ and bounded in magnitude by Δ. In the offline phase, we enumerate all rank-j matrices F with a certain precision. In Lemma 7.3 we show we can consider only $\Delta^{O(j^2(d+n))}$ different F. We compute $SF - SA$ for each F by linearity, and we choose the F minimizing the estimate. Setting $\delta = \Theta(\Delta^{-O(j^2(d+n))})$, we get a 1-pass $(n+d)\text{poly}((\log nd)/\epsilon)$-space algorithm, though the time complexity is poor.

LEMMA 7.3. *There is a set of $\Delta^{O(j^2(d+n))}$ different matrices F containing a $(1+\epsilon)$-approximation to the best rank-j approximation to A.*

Proof. By Corollary 7.2, we can restrict to F with entries that are integer multiples of $\Delta^{-\Theta(j)}$ and bounded in magnitude by $\text{poly}(\Delta)$. Choose a subset of j rows of F to be linearly independent. There are $\binom{n}{j}$ choices and $\Delta^{O(dj^2)}$ assignments to these rows. They contain j linearly independent columns, and once we fix the values of other rows on these columns, this fixes the other rows. In total, the number of F is $\Delta^{O(j^2(d+n))}$. □

We now show how to achieve polynomial time complexity. We need a theorem of Shyamalkumar and Varadarajan (stated for subspaces instead of flats).

THEOREM 7.4. *([25]) Let A be an $n \times d$ matrix. There is a subset Q of $O(\frac{j}{\epsilon}\log\frac{1}{\epsilon})$ rows of A so that span(Q) contains a j-dimensional subspace F with*

$$F_q(\ell_2)(\textit{proj}_F(A) - A) \leq (1+\epsilon)F_q(\ell_2)(A^q_{j,2} - A).$$

Let $r \stackrel{\text{def}}{=} O(\frac{j}{\epsilon}\log\frac{1}{\epsilon})$. Theorem 7.4 says that given a matrix A, there is a $j \times r$ matrix B and an $r \times n$ subset-matrix C (i.e., a binary matrix with one 1 in each row and at most one 1 in each column), with

$$F_q(\ell_2)(\text{proj}_{B\cdot C\cdot A}(A) - A) \leq (1+\epsilon)F_q(\ell_2)(A^q_{j,2} - A).$$

We enumerate all possible C in n^r time, which is polynomial for $j/\epsilon = O(1)$, but we need a technical lemma to discretize the possible B.

LEMMA 7.4. *Suppose rank$(A) > j$. Then there is a discrete set of $\Delta^{O(j^3\log^2 1/\epsilon)/\epsilon^2}$ different B, each with entries that are integer multiples of $\Delta^{-\Theta(j)}$ and bounded in magnitude by $\Delta^{O(j\log 1/\epsilon)/\epsilon}$, so that for every A, there is a B in the set and a subset-matrix C with*

$$F_q(\ell_2)(\textit{proj}_{B\cdot C\cdot A}(A) - A) \leq (1+\epsilon)F_q(\ell_2)(A^q_{j,2} - A).$$

Proof. We can assume rank$(A) \geq r > j$. If rank$(A) < r$ but rank$(A) > j$, we can just repeat this process for each value of ℓ between j and r, replacing r in the analysis below with the value ℓ. We then take the union of the sets of matrices that are found. This multiplies the number of sets by a negligible factor of r.

From Theorem 7.4, there is a $j \times r$ matrix B for which BCA has orthonormal rows and for which we have

$$F_q(\ell_2)(\text{proj}_{B\cdot C\cdot A}(A) - A) \leq (1+\epsilon)F_q(\ell_2)(A^q_{j,2} - A).$$

There may be multiple such B; for a fixed C, A we let B be the matrix that minimizes $F_q(\ell_2)(\text{proj}_{B\cdot C\cdot A}(A) - A)$.

Note that one can find such a B, w.l.o.g., since rank$(A) > r$. Furthermore, we can assume CA has full row rank since rank$(A) \geq r$. Note that if CA does not have this property, there is some $C'A$ with this property whose rowspace contains the rowspace of CA, so this is without loss of generality.

Fix any row A^i of A, and consider the y that minimizes

$$\min_{y\in\mathbb{R}^j} \|yBCA - A^i\|_2.$$

It is well-known that

$$y = A^i(BCA)^T[(BCA)(BCA)^T]^{-1},$$

but since BCA has orthonormal rows, $y = A^i(BCA)^T$. For conforming matrices we have $\|y\|_2 \leq \|A^i\|_2\|BCA\|_F$. Since BCA has orthonormal rows, $\|BCA\|_F = \sqrt{j}$. Moreover, $\|A^i\|_2 \leq \sqrt{d}\Delta$. It follows that

$$\|y\|_\infty \leq \|y\|_2 \leq \sqrt{dj}\Delta \leq \Delta^2.$$

Consider the expression $\|yBH - a\|_2$, where $a = A^i$ for some i, and $H = CA$. The entries of both a and C are integers bounded in magnitude by Δ.

Let the r rows of H be $H_1, \ldots, H_r$, and the j rows of BH be

$$\sum_{\ell=1}^{r} B_{1,\ell}H_\ell,\ \sum_{\ell=1}^{r} B_{2,\ell}H_\ell,\ \ldots,\ \sum_{\ell=1}^{r} B_{j,\ell}H_\ell.$$

Then the expression $\|yBH - a\|_2^2$ has the form

$$\sum_{\nu=1}^{d}\left(a_\nu - \sum_{u=1}^{j}\sum_{\ell=1}^{r} y_u B_{u,\ell}H_{\ell,\nu}\right)^2. \tag{7.16}$$

Notice that $|y_u H_{\ell,\nu}| \leq \Delta^3$ for every u, ℓ, and ν. It follows that if we replace B with the matrix B', in which entries are rounded down to the nearest multiple of Δ^{-cj} for a constant $c > 0$, then a routine calculation shows that expression 7.16 changes by at most $\Delta^{-\Theta(cj)}$, where the

constant in the $\Theta(\cdot)$ does not depend on c. As this was for one particular row $a = A^i$, it follows by another routine calculation that

$$\begin{aligned} &F_q(\ell_2)(\mathrm{proj}_{B\cdot C\cdot A}(A) - A) \\ \overset{(7.17)}{\leq} &F_q(\ell_2)(\mathrm{proj}_{B'\cdot C\cdot A}(A) - A) \\ &\leq F_q(\ell_2)(\mathrm{proj}_{B\cdot C\cdot A}(A) - A) + \Delta^{-\Theta(cj)}. \end{aligned}$$

We would like to argue that the RHS of inequality 7.17 can be turned into a relative error. For this, we appeal to Corollary 7.2, which shows that if $\mathrm{rank}(A) \geq j+1$, the error incurred must be at least $\Delta^{-\Theta(j)}$. Since ϵ can be assumed to be at least $\Delta^{-\Theta(j)}$, as otherwise the problem reduces to exact computation, it follows that if $\mathrm{rank}(A) \geq j+1$, then for a large constant $c > 0$,

$$\begin{aligned} &F_q(\ell_2)(\mathrm{proj}_{B\cdot C\cdot A}(A) - A) \\ &\leq F_q(\ell_2)(\mathrm{proj}_{B'\cdot C\cdot A}(A) - A) \\ &\leq (1+\epsilon)F_q(\ell_2)(\mathrm{proj}_{B\cdot C\cdot A}(A) - A). \end{aligned}$$

It remains to bound the number of different B'. Observe that $\|B'\|_F \leq \|B\|_F$. Now,

$$B = B(CA)(CA)^-,$$

where M^- denotes the Moore-Penrose inverse of a matrix M (that is, if we write $M = U\Sigma V^T$ using the singular value decomposition, then $M^- = V\Sigma^{-1}U^T$). Here we use the fact that CA has full row rank. Hence,

$$\|B\|_F \leq \|BCA\|_F\|(CA)^-\|_F = \sqrt{j}\|(CA)^-\|_F,$$

since the rows of BCA are orthonormal. Let $e = \mathrm{rank}(CA)$. Now,

$$\|(CA)^-\|_F^2 = \sum_{i=1}^{e} \sigma_i^{-2},$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_e > 0$ are the non-zero singular values of CA. Since CA is an integer matrix with entries bounded in magnitude by Δ, by Lemma 7.1 all singular values of CA are at least $1/\Delta^{\Theta(e)}$, and thus

$$\|(CA)^-\|_F^2 \leq e\Delta^{\Theta(e)} \leq \Delta^{O(j/\epsilon \log 1/\epsilon)}.$$

In summary,

$$\|B'\|_F \leq \|B\|_F \leq \Delta^{O(j/\epsilon \log 1/\epsilon)}.$$

As B' contains entries that are integer multiples of Δ^{-cj}, the number of different values of an entry in B' is $\Delta^{O(j/\epsilon \log 1/\epsilon)}$. Since B' is a $j \times r$ matrix, where $r = O(j/\epsilon \log 1/\epsilon)$, it follows that the number of different B' is $\Delta^{O(j^3/\epsilon^2 \log^2 1/\epsilon)}$, which completes the proof. □

We sketch each row A^i of A independently, treating it as the vector b in Theorem 7.2 with the d there equaling the j here, thereby obtaining A^iS for sketching matrix S and each $i \in [n]$. Offline, we guess each of $n^r \cdot \Delta^{O(j^3 \log^2 1/\epsilon)/\epsilon^2}$ matrix products BC, and by linearity compute $BCAS$. We can $(1 \pm \epsilon)$-approximate

$$\min_{x \in \mathbb{R}^j} \|xBCA - A^i\|_p$$

for each i, B, C provided S is a $d \times O(j^3 \log^2 1/\epsilon)/\epsilon^2$ matrix. Finally by Theorem 7.1,

THEOREM 7.5. *There is a 1-pass algorithm for Best Rank-j Approximation with*

$$O(nj^4 \log(nd) \log^3 1/\epsilon)/\epsilon^5$$

bits of space and $\Delta^{\mathrm{poly}(j/\epsilon)}$ time. The algorithm also obtains the $n \times j$ basis representation of rows of A in BC for the choice of B and C resulting in a $(1+\epsilon)$-approximation. In another pass we can obtain the subspace BCA in $O(jd \log(nd))$ additional bits of space.

References

[1] P. K. Agarwal, S. Har-Peled, and K.R. Varadarajan. Approximating extent measures of points, *J. ACM*, volume 51(4), pp. 606–635, 2004.

[2] N. Alon and J. Spencer, *The probabilistic method*, J. Wiley & Sons, New York, 2nd edition, 2000.

[3] M. Belkin and P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, *Neural Computation*, volume 15(6), pp. 1373–1396, 2003.

[4] M. Belkin and P. Niyogi, Semi-supervised learning on riemannian manifolds, *Machine Learning Journal*, 56, 2004.

[5] Z. Bar-Yossef, T. S. Jayram, R. Kumar, and D. Sivakumar, An information statistics approach to data stream and communication complexity., In *Proc. 43th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, pp. 209–218, 2002.

[6] A. Chakrabarti, S. Khot, and X. Sun, Near-optimal lower bounds on the multi-party communication complexity of set disjointness, In *Proc. 18th Ann. IEEE Conf. on Computational Complexity (CCC)*, pp. 107–117, 2003.

[7] K. Clarkson and D. Woodruff, Numerical linear algebra in the streaming model, *Proc. 41th Annu. ACM Symp. Theory Comput. (STOC)*, 2009.

[8] A. Dasgupta, P. Drineas, B. Harb, R. Kumar, and M. W. Mahoney. Sampling algorithms and coresets for l_p regression. *Siam J. Comput.*, 38(5), pp. 2060-2078, 2009.

[9] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang, Matrix approximation and projective clustering via volume sampling, *Theory of Computing*, volume 2(12), pp. 225–247, 2006.

[10] A. Deshpande and K. Varadarajan, Sampling-based dimension reduction for subspace approximation, *Proc. 39th Annu. ACM Symp. Theory Comput. (STOC)*, 2007.

[11] D. Feldman, A. Fiat, and M. Sharir, Coresets for weighted facilities and their applications, *Proc. 47th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, 2006.

[12] A. Frieze, R. Kannan, and S. Vempala, *Fast monte-carlo algorithms for finding low-rank approximations*, *JACM*, 32(2), pp. 269–288, 2004.

[13] D. Feldman and M. Landberg, *Algorithms for approximating subspaces by subspaces*, 2007.

[14] D. Feldman, M. Monemizadeh, and C. Sohler, *A ptas for k-means clustering based on weak coresets*, *Proc. 23st Annu. ACM Sympos. Comput. Geom. (SOCG)*, pp. 11–18, 2007.

[15] S. Har-Peled, *How to get close to the median shape*, *CGTA*, 36(1), pp. 39–51, 2007.

[16] S. Har-Peled and S. Mazumdar, *Coresets for* k*-means and* k*-median clustering and their applications*, *Proc. 36th Annu. ACM Sympos. Theory Comput. (STOC)*, pp. 291–300, 2004.

[17] P. Indyk, *Stable distributions, pseudorandom generators, embeddings, and data stream computation*, *J. ACM*, 53(3), pp. 307–323, 2006.

[18] T.S. Jayram and D. Woodruff, *The data stream space complexity of cascaded norms*, In *Proc. 50th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, *FOCS*, 2009.

[19] D. Kane, J. Nelson, and D. Woodruff. On the Exact Space Complexity of Sketching and Streaming Small Norms. In *Proc. 21th ACM-SIAM Symp. Discrete Algorithms (SODA)*, 2010.

[20] J. Kleinberg and M. Sandler, Using mixture models for collaborative filtering, In *Proc. 36th Annu. ACM Sympos. Theory Comput. (STOC)*, pp. 569– 578, 2004.

[21] D. Kuzmin and M. K. Warmuth, Online kernel PCA with entropic matrix updates, In *Proc. 24th Intl. Conf. for Machine Learning* , pp. 465–472, 2007.

[22] A. Lakhina, M. Crovella, and C. Diot, Characterization of network-wide anomalies in traffic flows. In *Proc. 4th ACM SIGCOMM Conf. on Internet measurement*, pp. 201–206, 2004.

[23] P. Li, Estimators and tail bounds for dimension reduction in ℓ_α $(0 < \alpha \leq 2)$ using stable random projections. In *SODA*, pp. 10–19, 2008.

[24] T. Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proc. 47th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, pp. 143–152, 2006.

[25] N.D. Shyamalkumar and K. Varadrajan. Efficient subspace approximation algorithms. In *Proc. 18th ACM-SIAM Symp. Discrete Algorithms (SODA)*, pp. 532–540, 2007.

[26] M. Vose. A linear algorithm for generating random numbers with a given distribution. In *IEEE Trans. Software Eng.*, 17:9, pp. 972–975, 1991.