Coreset for k-means of streaming lines

Marom, Yair
yairmrm@gmail.com

Feldman, Dan
dannyf.post@gmail.com

May 11, 2018

1 Abstract

This motivates the *k-means for lines* problem: Given set of n lines in $\mathbb{R}^d$ compute the set of k centers (points) that minimizes the sum of squared n distances over each line and its nearest point.

We propose the first PTAS for this problem that returns a $(1+\varepsilon)$-approximation in $O(n)$ time for any constant error $\varepsilon \in (0,1)$, and $k, d \geq 1$. Extensions include streaming version that uses $O(\log n)$ points in memory and update time per point, dynamic algorithm that supports deletion (as the sliding window model), sum of (non-squared) distances, robust m-estimators, ignoring m outliers, and more. The algorithms can run on data that is distributed among M machines (e.g. cloud, GPU) where the running time and memory usage reduces linearly with M.

Our main technical results is a constructive proof that *every* such set of n lines has a weighted subset (core-set) of only $O(|CoresetSize|)$ lines whose sum of weighted distances is the same for *every* k centers, up to $(1 \pm \varepsilon)$.

Experimental results on a system for tracking a quadcopter shows that our algorithms significantly outperforms existing heuristics also in practice.

2 Introduction

Clustering is the task that aims to group a subset of objects that are more similar to each other from the other objetcs in the group, that group called a cluster. There is a lot of different clustering techniques, but most noticeable and popular is Lloyd's algorithm or the k-means algorithm [1], [2] that was first published in 1955.

The classical Euclidean k-means problem input is a set of n points in $\mathbb{R}^d$, and the purpose is to compute a set of k-centers (that are also points in $\mathbb{R}^d$) that minimizes the sum of squared distances over each input point to its nearest center. Geometrical objects clustering in $\mathbb{R}^2$ was discussed in Weber Problem [3] and assumes that the dimension of the input is limited. But when handling data in high dimension the task become much more harder. Many variations and restrictions to this standard setting have been addressed. They differ from each other by the type of objects they are clustering, the number of groups, the dimension of the problem and the objective function. Each one of those problems has different motivation and can be related to different fields of research. Since the problem is proven to be NP-Hard [7], [8], a lot of effort was invested in finding approximation algorithms for these problems. Techniques such as PCA/SVD [4], [5] or Johnson-Lindenstraus [6] aims to reduce the the input dimensionality. However, such techniques project the data and turn it into dense data, and the k-means of the projected points are no longer a subset of the original input.

One of the approximation algorithm method for this problem is by using *Coresets*. Given a set of n points in $\mathbb{R}^d$, and an error parameter $\epsilon > 0$, a coreset is a small set of weighted points in $\mathbb{R}^d$, such that the sum of squared distances from the original set of points to any set of k centers in $\mathbb{R}^d$ can be approximated by the sum of weighted squared distances from the points in the coreset. Running an existing clustering algorithm on

the coreset yields approximation to the output of running the same algorithm on the original data. Coresets were first suggested in [9] as a way to improve the theoretical running time of existing algorithms. Moreover, a coreset is a natural tool for handling Big Data using all the computation models that we talked about.

Clustering and shape fitting problems on points have been actively studied in the past decade. For the k-center problem on points in $\mathbb{R}^d$, efficient approximation algorithms with running time polynomially dependent on d are available. A simple greedy algorithm [15, 16], finds a 2-approximation and can be implemented in optimal time $O(nd \log k)$ [17]. Another result using coresets based techniques for k-means achieves a $(1+\epsilon)$-approximation algorithm with running time $k^{O(k/\epsilon^2)}dn$ [18, 10, 11], were based on partitioning the data into cells, and take a representative point from each cell to the coreset, but these algorithms for points in $\mathbb{R}^d$ do not generalize to the case of incomplete data (i.e., lines), and with exponential size of d. Recently, deterministic construction of size $O(1/\epsilon^2)$ was suggested in [25] for the case k = 1, and for $k > 1$ [26] suggest an streaming algorithm for computing a provable approximation to the k-means of sparse Big data.

In facility location problems [13, 14], the input points represent the location of clients, and the centers are called facilities. In this case we might have constraints that some of the centers will be closer to some clients. A major difficulty in such a generalization lies in the lack of a triangle inequality when considering lines. The problem is not that the triangle inequality is slightly violated, but that no relaxation of it holds. No matter how far apart lines a and c are, there is always a line b that intersects both.

2.1 Preliminaries

- Let $d \geq 1$ be an integer. For every $p \in \mathbb{R}^d$, a set of points $X \subseteq \mathbb{R}^d$ and a finite set of such sets $\mathcal{X} = \{X_1, X_2, \ldots\}$ in $\mathbb{R}^d$, let $\text{dist}(X, p) = \min_{x \in X} \|x - p\|$ be the distance from X to p, let $\text{dist}(\mathcal{X}, p) = \min_{X \in \mathcal{X}} \text{dist}(X, p)$ be the distance from $\mathcal{X}$ to p, let $\text{dist}(X', X) = \min_{x' \in X'} \text{dist}(X, x')$ denote the distance from $X' \in \mathcal{X}$ to X, and let $\text{cost}(\mathcal{X}, X) = \sum_{X' \in \mathcal{X}} \text{dist}(X', X)$ be the sum of distances from $\mathcal{X}$ to X.
- For every positive integer $n \in \mathbb{N}$, we denote the set $\{1, \ldots, n\}$ by $[n]$.
- Here and in the rest of the paper, ties are broken arbitrarily

2.2 Problem Statement

In the familiar k-means clustering problem, the input is a set P of n points in $\mathbb{R}^d$, and the goal is to compute a set C of k centers (points) in $\mathbb{R}^d$, that minimizes the sum of squared distances over each point $p \in P$ to its nearest center in C, i.e.

$$C \in \arg \min_{C' \subseteq \mathbb{R}^d, |C'|=k} \sum_{p \in P} \min_{c' \in C'} \|p - c'\|^2 .$$

A natural generalization of the k-means problem is to replace the input set of points P by a set L of n lines in $\mathbb{R}^d$.

Definition 2.1 (k-means for lines) *Let L be a set of n lines in $\mathbb{R}^d$. For an integer $k \geq 1$, a k-means of L is a set $C \subseteq \mathbb{R}^d$ consist of k points that minimizes the sum of n squared distances from C to L, i.e.*

$$C \in \arg \min_{C' \subseteq \mathbb{R}^d, |C'|=k} \sum_{\ell \in L} \min_{c' \in C'} \text{dist}^2(\ell, c').$$

In this paper, for every set L of finite number of lines in $\mathbb{R}^d$, we are interested in seeking a compact representation $S \subseteq L$ that approximates $\text{cost}(L, C)$ for every set $C \subseteq \mathbb{R}^d$ of k points. We will use the following definition.

Definition 2.2 ((k, ε)-coreset) *Let L be a finite set of lines in $\mathbb{R}^d$, $k \geq 1$ be an integer and a small $\varepsilon > 0$. A pair (S, μ) is a (k, ε)-coreset for L if for every set $C \subseteq \mathbb{R}^d$ of k points we have*

$$(1 - \varepsilon)\text{cost}(L, C) \leq \text{cost}'(S, \mu, C) \leq (1 + \varepsilon)\text{cost}(L, C),$$

where $S = \{s_1, \ldots, s_m\}$ *is a set of* m *lines in* $\mathbb{R}^d$, $\mu = \{\mu_1, \ldots, \mu_m\} \subseteq \mathbb{R}$ *and*

$$\text{cost}'(S, \mu, C) = \sum_{i=1}^{m} \mu_i \cdot \text{dist}(s_i, C).$$

We present coreset construction with provable approximations for a family of natural k-clustering optimization problems. The running time is linear in both the number of data lines n, their dimensionality d, and the number k of desired centers. The resulting coreset consists of $O(|CoresetSize|)$ weighted lines that approximate the sum of square distances for any k-centers. In particular, we can use this coreset to compute the k-centers that minimize the sum of squared distances to the input lines (k-centers over time) and for robust m-estimators.

2.3 Related Work

Langebreg and Shculman used Helly's theorem [20] (intersection of convex sets) to introduce a "k-center problem" for lines [21], trying to cover a collection of lines with k balls in $\mathbb{R}^3$. The usual starting point for statistical theory, learning theory, or estimation for control, is an input set consisting of a list of empirically gathered data points in $\mathbb{R}^d$. One of the serious gaps between statistical theory and practice, however, lies with incompletely-specified data. Essentially, the issue is that a high-dimensional data point is not specified by one "measurement" but by many, and that some of those measurements may be missing. They suggests a particular method of imputation – given a "data line", find a ball intersecting it, and choose the point on the line closest to the center of that ball.

Langebreg and Schulman [19] addressed the 1-center problem (i.e., the case $k = 1$ to find a single ball intersecting all input lines). From a computational point of view, the 1-center problem significantly differs from the general k-center problem, the 1-center problem is a convex optimization problem and therefore fundamentally easier than the cases of k-center for $k \geq 2$.

The similarity to our problem is the statistical motivation and specifically the notion that the region of intersection of a line with a ball is a useful imputation of the missing data on that line.

Other works

Also there has been work on "clustering points with k" lines [22], [23], [24], where one finds a set of lines L such that the set of cylinders with radius r about these lines covers all the input points S.

3 Coreset for k-means for lines

We would like to compute a (k, ε)-coreset for our data. A (k, ε)-coreset (S, μ) for a set L of lines, approximates the fitting cost of any query k-means for L up to a small multiplicative error of $1 \pm \varepsilon$. We note that the k-means for L can be computed using the naive algorithm 4-Approximation(L, k), with running time of $n^{O(k)}$; See Algorithm 3. We will do it in a linear time to our input.

The main idea of the coresets is to take a small and smart sample from the input data by its centers of gravity, and show that running the naive algorithm on that small sample brings almost the same results we would get from running it on the original input, and from the size of the sample that running time will be significantly small. However, if we knew the centers of gravity, we could solve the original problem with the optimal clustering. In order to solve this situation of an egg and a chicken, we will calculate a bi-critia for the problem as a starting point (see Algorithm **??**). That is, instead of returning k points that minimize the sum of distances from a set of n lines, we will return $k \cdot log(n)$ points that minimize the sum of distances from these lines up to a constant factor, with a probability of at least $1 - \delta$. We will set it with the following definition.

Definition 3.1 (α, β-approximation) *Let* L *be a set of finite number of lines in* $\mathbb{R}^d$, $k \geq 1$ *be an integer,*

$\alpha, \beta \geq 0$, $B \subseteq \mathbb{R}^d$ *s.t.* $|B| = \beta k$, *and* $\mathcal{B} : L \to B$ *such that*

$$\sum_{\ell \in L} \text{dist}(\ell, \mathcal{B}(\ell)) \leq \alpha \min_{C \subseteq \mathbb{R}^d, |C|=k} \sum_{\ell \in L} \text{dist}(\ell, C).$$

Then $\mathcal{B}$ *is called an* α, β*-approximation for* L*.*

Once we have a starting point as the α, β-approximation, we can perform some analysis on the lines in L relative to these $k\beta$ centers, and get an estimation for the centers of gravity of the lines. The full process and its immediate result will be explained in the following theorem and algorithm.

Theorem 3.2 *Let* L *be a set of* n *lines in* $\mathbb{R}^d$*,* $k \geq 1$ *be an integer and* $\varepsilon, \delta \in (0, 1)$*. Let* $\mathcal{G}$ *be the output set of the call to* α, β-APPROXIMATION$(L, k, \varepsilon, \delta)$ *; See Algorithm* **??***, and let* (S, μ) *be the output pair of the call to* CORESET$(L, k, \mathcal{B}, m)$ *; See Algorithm 1, and* $m = |CoresetSize|$*. Then,* (S, μ) *is a* (k, ε)*-corsets for* L *of size* m*, and can be compute in* $O(???)$ *time, with a probability of* $1 - \delta$*.*

Algorithm 1: CORESET$(L, w, k, \mathcal{B}, m)$

Input: A set L of n lines in $\mathbb{R}^d$, a set $w \subseteq \mathbb{R}$ of n corresponding weights, a positive integer k, α, β-approximation $\mathcal{B} : L \to B$ for L and a positive integer $m < n$.

Output: A pair (S, μ) which is a (k, ε)-coreset for L.

for *every* $b \in B$ **do**
  $L_b \leftarrow \{\ell \in L \mid \mathcal{B}(\ell) = b\}$
$T \leftarrow 0$
for *every* $b \in B$ *and* $\ell \in L_b$ **do**
  $s(\ell) = \frac{\alpha \cdot \text{dist}(\ell, b)}{\text{cost}(L, B)} + 2\alpha \cdot S_{L'}(\ell', k)$ /* See Definition 4.12 and Equation (13) */
  $T \leftarrow T + s(\ell)$
for $\ell \in L$ **do**
  $\text{prob}(\ell) = \frac{s(\ell)}{T}$
Pick a sample S of at least m lines from L where each line is sampled i.i.d. with probability $\text{prob}(\ell)$.
Set $u : L \to [0, \infty)$ s.t. for every $b \in B$ and $\ell \in L_b$,

$$u(\ell) = \begin{cases} \frac{w(\ell)}{|S| \text{prob}(\ell)}, & \text{if } \ell \in S \\ 0, & \text{otherwise} \end{cases}$$

return (S, u)

Set $\delta = \frac{1}{2}$ yields a $1 - \frac{1}{2^x}$ probability of success for running CORESET algorithm for x times, and for efficient k-means we run the $(1 + \varepsilon)$-APPROXIMATION algorithm on our small coreset instead of the original large input - as we summarize as follows.

Theorem 3.3 *Let* L *be a set of finite number of lines in* $\mathbb{R}^d$*. A* $(1 + \varepsilon)$*-approximation to the* k*-means for* L *can be computed in* $O(???)$ *time.*

4 Proof of Correctness

4.1 α, β-Approximation

Let d, k be two positive integers, and let $L = \{\ell_1, \ldots, \ell_n\}$ be a set of n lines in $\mathbb{R}^d$. We denote by $OPT(L, k) \subseteq \mathbb{R}^d$ a set of k points that minimizes the total cost to L, i.e. $OPT(L, k) \in \arg\min_{P \subseteq \mathbb{R}^d, |P|=k} \text{cost}(L, P)$. For convenience, we denote the set of points $OPT(L, k)$ by P^*.

Algorithm 2: α, β-APPROXIMATION(L, k)

Input: A set L of n lines in $\mathbb{R}^d$ and a positive integer k
Output: A set of $k \cdot \log n$ pairs $[\tilde{p}, L']$, where $\tilde{p} \in \mathbb{R}^d$ and $L' \subseteq L$, that satisfy Theorem 4.1

$\mathcal{G} \leftarrow \varnothing$
while $|L| > 1$ **do**
  $S \leftarrow$ i.i.d sample consist of $O\left(\frac{\log\left(\frac{n^{dk\log k}}{\delta}\right)}{2\varepsilon^2}\right)$ lines from the uniform distribution over L
  $\tilde{P} \leftarrow$ 4-APPROXIMATION(S, k)
  $L' \leftarrow \arg\min_{X \subseteq L, |X| = \frac{|L|}{2}} \text{cost}(X, \tilde{P})$
  foreach $\tilde{p} \in \tilde{P}$ **do**
    $L'_{\tilde{p}} = \left\{\ell' \in L' \mid \forall q \in \tilde{P} : \text{dist}(\ell', \tilde{p}) \leq \text{dist}(\ell', q)\right\}$,
    $\mathcal{G} \leftarrow \mathcal{G} \cup \left\{\tilde{p}, L'_{\tilde{p}}\right\}$
  $L \leftarrow L \setminus L'$
return $\mathcal{G}$

Theorem 4.1 *Let L be a set of n lines in $\mathbb{R}^d$ and k be a positive integer, suppose that $\mathcal{G}$ is the output set of the call to α, β-APPROXIMATION(L, k) ; See Algorithm **??**. Let $\tilde{P}$ denote the union of $\tilde{p} \in \left\{\tilde{p}, L'_{\tilde{p}}\right\}$ over every pair $\left\{\tilde{p}, L'_{\tilde{p}}\right\} \in \mathcal{G}$. Then,*

$$\text{cost}(L, \tilde{P}) \leq 4 \cdot \text{cost}(L, P^*).$$

We will prove the correctness of the last theorem in the following two parts.

4.1.1 Robust Approximation

First, we show that for every finite set L of lines in $\mathbb{R}^d$, we can compute a set $P \subseteq \mathbb{R}^d$ of k points that minimizes the sum of distances to L up to a constant factor, and in particular - minimizes the sum of distances to each subset of lines $L' \subseteq L$ up to a constant factor - which will help us a lot later.

Algorithm 3: 4-APPROXIMATION(L, k)

Input: A finite set L of lines in $\mathbb{R}^d$ and a positive integer k
Output: A set $\tilde{P} \subseteq \mathbb{R}^d$ of k points that satisfies Theorem 4.2

$Q \leftarrow \varnothing$
$\tilde{P} \leftarrow \varnothing$
for *every* $\ell \in L$ **do**
  $Q \leftarrow Q \cup Q(L, \ell)$; See Claim 4.4
$\tilde{P} \leftarrow \arg\min_{P' \subseteq Q, |P'| = k} \text{cost}(L, P')$
return $\tilde{P}$

Theorem 4.2 *Let L be a set of n lines in $\mathbb{R}^d$ and k be a positive integer, suppose that $\tilde{P}$ is the output set of the call to 4-APPROXIMATION(L, k) ; See Algorithm 3. Then, for every subset $L' \subseteq L$ of lines,*

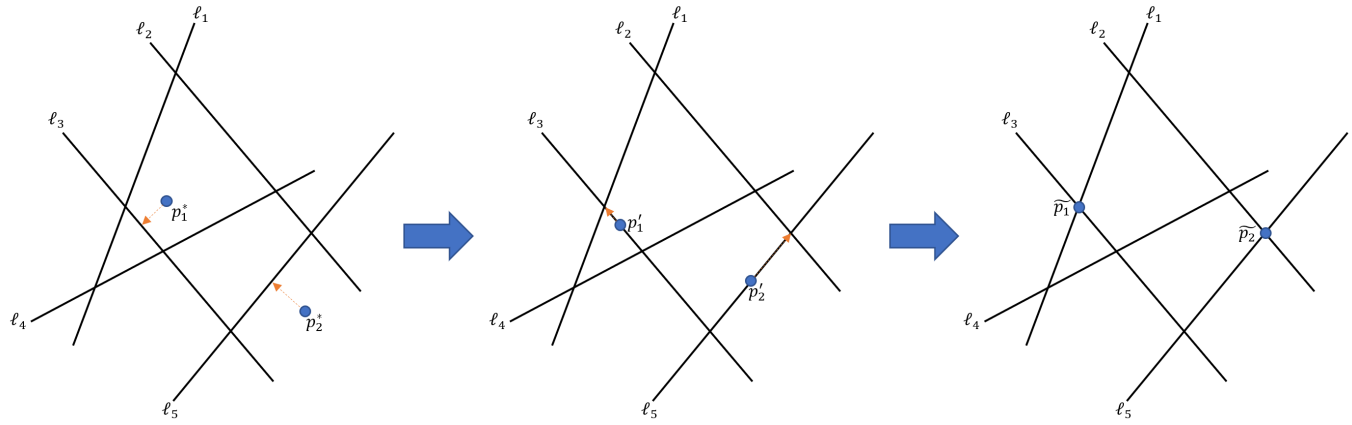
$$\text{cost}(L', \tilde{P}) \leq 4 \cdot \text{cost}(L', P^*).$$

Proof.

Claim 4.3 *For every point $p \in \mathbb{R}^d$, let $L(p) \in L$ denote the closest line to p in L, and let* $\text{proj}(L,p) \in L(p)$ *denote the closest point to p in $L(p)$. For every $p^* \in P^*$, let $p' = \text{proj}(L,p^*)$. That is, the translation of p^* to its nearest line in L (see the first transition in Figure 1). For every $p^* \in P^*$, let the subset of lines in L that are closer to $p^* \in P^*$ than the other points be $L'(p^*) = \{\ell \in L \mid \forall q \in P^* \setminus \{p^*\} : \text{dist}(\ell,p^*) \leq \text{dist}(\ell,q)\}$. Then, for every $p^* \in P^*, \ell \in L'(p^*)$,*

$$\text{dist}(\ell,p') \leq 2 \cdot \text{dist}(\ell,p^*). \tag{1}$$

Figure 1: Example for $k = 2$. The input is a set $L = \{\ell_1,\ldots,\ell_5\}$ of lines, and the translated center points are the sets, from left to right respectively, $P^* = \{p_1^*, p_2^*\}, P' = \{p_1', p_2'\}$ and $\tilde{P} = \{\tilde{p_1}, \tilde{p_2}\}$.



Proof. Let $p^* \in P^*, \ell \in L'(p^*)$. From the right angle inequality and from the fact that for every $\ell' \in L'(p^*)$ we have $\|p^* - p'\| \leq \text{dist}(\ell', p^*)$, we get

$$\text{dist}(\ell,p') \leq \text{dist}(\ell,p^*) + \|p^* - p'\| \leq 2 \cdot \text{dist}(\ell,p^*).$$

□

Claim 4.4 *For every line ℓ in $\mathbb{R}^d$ and a finite set L of lines, let*

$$Q(L,\ell) = \left\{ c \in \ell \mid \exists \ell' \in L : c \in \arg\min_{c' \in \ell} \text{dist}(\ell', c') \right\},$$

i.e. $Q(L,\ell)$ is a set of $|L|-1$ points in ℓ s.t. each point in $Q(L,\ell)$ is the closest point in ℓ to a different $\ell' \in L$. For every $p^* \in P^*$, let $\tilde{p} \in \arg\min_{c \in Q(L,L(p^*))} \|c - p'\|$. That is, moving p' to its closest point in $Q(L,L(p^*))$ (see the second transition in Figure 1). Then, for every $p^* \in P^*, \ell \in L'(p^*)$,

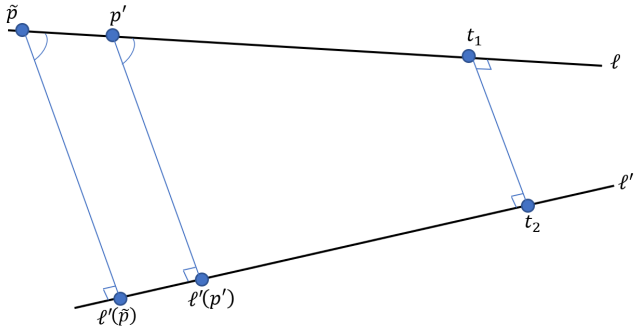
$$\text{dist}(\ell,\tilde{p}) \leq 2 \cdot \text{dist}(\ell,p'). \tag{2}$$

Proof. For every $p \in \mathbb{R}^d$ and a line ℓ in $\mathbb{R}^d$, let $\ell(p) \in \ell$ denote the projection of p onto ℓ. Let $p^* \in P^*$, and let $\ell = L(p^*) \in L$ be the closest line in L to p^*. Let $p', \tilde{p} \in \ell$ be two points in ℓ that were made after the first and the second translations of p^*, respectively, as describe above. For every line $\ell' \in L \setminus \{\ell\}$, let the pair of points $(t_1 \in \ell, t_2 \in \ell') \in \arg\min_{t_1' \in \ell, t_2' \in \ell'} \|t_1' - t_2'\|$, i.e. the closest points in ℓ and ℓ'. ; See Fig. 2.

By the definition of $\tilde{p}$ as the closest point to p' in $Q(L,\ell)$, we obtain

$$\frac{\|\tilde{p} - p'\| + \|p' - t_1\|}{\|p' - t_1\|} \leq \frac{\|p' - t_1\| + \|p' - t_1\|}{\|p' - t_1\|} = 2, \tag{3}$$

Figure 2: Example in 3 dimensional space. Given two lines $\ell, \ell' \in L$, where $\tilde{p}, p' \in \ell$, and $\|t_1 - t_2\|$ is the shortest euclidean distance between ℓ an ℓ'.



from this construction, we have that the two quadrilateral made by the two sets of points $\{t_1, p', \ell'(p'), t_2\}$ and $\{t_1, \tilde{p}, \ell'(\tilde{p}), t_2\}$ are similar (or two similar rectangles if $\|t_1 - t_2\| = 0$ and the two lines are intersect), and from the polygons similarity we get

$$\frac{\|\tilde{p} - \ell'(\tilde{p})\|}{\|p' - \ell(p')\|} = \frac{\|\tilde{p} - p'\| + \|p' - t_1\|}{\|p' - t_1\|}, \tag{4}$$

plugging (3) in (4) yields

$$\frac{\text{dist}(\ell', \tilde{p})}{\text{dist}(\ell', p')} = \frac{\|\tilde{p} - \ell'(\tilde{p})\|}{\|p' - \ell'(p')\|} \leq 2,$$

i.e.

$$\text{dist}(\ell', \tilde{p}) \leq 2 \cdot \text{dist}(\ell', p'). \tag{5}$$

□

Corollary 4.5 *For every line $\ell \in L$, we have*

$$\text{dist}(\ell, \tilde{P}) \leq 4 \cdot \text{dist}(\ell, P^*) \tag{6}$$

Proof. Immediate from the combination of (1) and (2). □

And from the linearity of the cost function, we get that the property in the last corollary is maintained for every subset $L' \subseteq L$, and that proves the theorem. □

4.1.2 Bound on the VC-dimension

In each iteration of the α, β-approximation (see Algorithm 2), we have a set $L' \subseteq L$ consist of x lines, and we compute k pairs $\left\{\tilde{p}, L'_{\tilde{p}}\right\}$, s.t. the union over every $\tilde{p}$ is a set of k centers that minimizes the sum of squared distances to L' up to a constant factor. In order to compute such pair, we need to run the 4-Approximation algorithm at each iteration, which can take a very long time. In this part, we show that running the 4-Approximation algorithm on an i.i.d sample consist of $O\left(\frac{\log\left(\frac{n^{dk\log k}}{\delta}\right)}{2\varepsilon^2}\right)$ lines from the uniform distribution over L' is enough to get the same results up to ε additive error. In order to show that, we use a method due to Warren [ref?] to bound the VC-dimension of following range space.

Definition 4.6 (range) *For every finite set L of lines in $\mathbb{R}^d$, $r > 0$ and a set of k points $P \subseteq \mathbb{R}^d$, let*

$$\text{range}(P, r, L) = \left\{\ell \in L \mid dist^2(\ell, P) \leq r^2\right\}.$$

Lemma 4.7 *For every set L of n lines in $\mathbb{R}^d$, we have*

$$\left|\left\{\text{range}(P, r, L) \mid P \subseteq \mathbb{R}^d, |P| = k, r > 0\right\}\right| \in O\left(n^{dk \log k}\right) \tag{7}$$

Proof. In what follows, $\text{sgn}(x)$ denotes the sign of $x \in \mathbb{R}^d$. More precisely, $\text{sgn}(x) = 1$ if $x > 0$, $\text{sgn}(x) = -1$ if $x < 0$, and $\text{sgn}(x) = 0$ otherwise. We use the following theorem.

Theorem 4.8 (Theorem 3 in (War68)) *Let $\{f1, \ldots, f_m\}$ be real polynomials in $d^* < m$ variables, each of degree at most $\ell \geq 1$. Then the number of sign sequences $(\text{sgn}(f_1(x)), \ldots, \text{sgn}(f_1(x)))$, $x \in \mathbb{R}^d$, that consist of the terms $1, -1$ is at most $\left(\frac{4e\ell m}{d^*}\right)^{d^*}$.*

Corollary 4.9 (Corollary 3.1 in (War68)) *If $\ell \geq 2$ and $m \geq 8d^* \log(\ell)$, then the number of distinct sequences as in the above theorem is less than 2^m.*

We use these results to obtain.

Corollary 4.10 *Let d, k be positive integers. Let $\mathcal{Q}_k$ be the family of all sets which are the union of k points in $\mathbb{R}^d$. Let $L = \{\ell_1, \ldots, \ell_n\}$ be a set of n lines in $\mathbb{R}^d$. Let $F^* = \{f_1, \ldots, f_n\}$, where $f_i(Q) = \text{dist}^2(\ell_i, Q)$ for every $i \in [n], Q \in \mathcal{Q}_k$. Then the dimension of the range space $\mathcal{R}_{\mathcal{Q}_k, F^*}$ that is induced by $\mathcal{Q}_k$ and F^* is $O(dk \log k)$.*

Proof. We first show that in the case $k = 1$ the VC-dimension of the range space $\mathcal{R}_{\mathcal{Q}_k, F^*}$ is $O(dk)$. Then the result follows from the fact that the k-fold intersection of range spaces of VC-dimension $O(dk)$ has VC-dimension $O(dk \log k)$ [BEHW89, EA07 in subspace.pdf].

If $n < d$ then the result is immediate. Thus, we consider the case $n > d$. We will first argue that the distance to a line can be written as a polynomial in $O(d)$ variables. Let $p \in \mathbb{R}^d$, for every $i \in \{1, \ldots, n\}$, we can write $dist^2(\ell_i, Q) = (p - b_i)^t X_i$ where $X_i \in \mathbb{R}^{d \times d-1}$ with $X_i^t X_i = I$ and $b_i \in \mathbb{R}^d$. Therefore, $f_i(p) - r$ is a polynomial of constant degree ℓ with $d^* \in O(d)$ variables.

Consider a subset $G \subset F$ with $|G| = m$, denote the functions in G by $\{f_1, \ldots, f_m\}$. Our next step will be to give an upper bound on the number of different ranges in our range space $\mathcal{R}_{\mathcal{Q}_k, F^*}$ for $k = 1$ that intersect with G. Recall that the ranges are defined as

$$\left\{\ell \in L \mid \text{dist}^2(\ell, P) \leq r^2\right\}$$

for $P \in \mathcal{Q}_k$, and $r \geq 0$. We observe that for every $i \in \{1, \ldots, n\}$ we have $\text{dist}^2(\ell_i, P) \geq r^2$, iff $sgn(f_i(P) - r^2) \geq 0$. Thus, the number of ranges is at most

$$\left|\left\{\left(sgn(f_1(P) - r^2), \ldots, sgn(f_m(P) - r^2)\right) \mid P \in \mathcal{Q}_k, r \geq 0\right\}\right|.$$

We also observe that for every sign sequence that has zeros, there is a sign sequence corresponding to the same range that only contains 1 and -1 (this can be obtain by infinitesimally changing r). Thus, by Theorem 4.8 the number of such sequences is bounded by $\left(\frac{4e\ell m}{d^*}\right)^{d^*}$, where $\ell = O(1)$. By Corollary 4.9 we know that for $\ell \geq 2$ (which we can always assume as ℓ is an upper bound for the degree of the involved polynomials) and $m \geq 8d^* \log \ell$ the number of such ranges is less than 2^m. At the same time, a range space with VC-dimension d must contain a subset G of size d such that any subset of G can be written as $G \cap \text{range}$ for some range $\in$ ranges, which implies that the number of such sets is 2^d. Since this is not possible for G if $m \geq 8d^* \log \ell$, we know that the VC dimension of our range space is bounded by $8d^* \log \ell \in O(d)$ (for the case $k = 1$). Now the he result follows by observing that in the case of k centers every range is obtained by taking the intersection of k ranges of the range space for $k = 1$. □

□

Theorem 4.11 *Let L be a set of n lines in $\mathbb{R}^d$, a positive integer k and $\varepsilon, \delta \in (0,1)$, and let $S \subseteq L$ be an i.i.d sample consist of $O\left(\frac{\log\left(\frac{n^{dk\log k}}{\delta}\right)}{2\varepsilon^2}\right)$ lines from the uniform distribution over L. Then, with a probability of $1-\delta$, for every $r > 0$ and a set $P \subseteq \mathbb{R}^d$ of k points, the following inequality is satisfied,*

$$\frac{|\text{range}(P,r,L)|}{|L|} - \frac{|\text{range}(P,r,S)|}{|S|} < \varepsilon.$$

Proof. Let $Er(P,r) = \left|\frac{|\text{range}(P,r,L)|}{|L|} - \frac{|\text{range}(P,r,S)|}{|S|}\right|$ denote the size of the error between the ranges ratio from the original data and the ranges ratio from the sample. By Hoeffding's bound, for every set $P \subseteq \mathbb{R}^d$ of k points and $r \geq 0$, we bound that probability as follows

$$Pr\left(Er(P,r) > \varepsilon\right) \leq \frac{1}{e^{2|S|\varepsilon^2}},$$

from Union Bound, and from Lemma 4.7 that claims we have only $O\left(n^{dk\log k}\right)$ distinct ranges, we get that the probability of the failure is bounded by

$$\bigcup_{P\subseteq\mathbb{R}^d, |P|=k, r\geq 0} Pr\left(Er(P,r) > \varepsilon\right) \leq \frac{n^{dk\log k}}{e^{2|S|\varepsilon^2}}.$$

Since the probability of the failure should be at most δ, we get

$$\frac{n^{dk\log k}}{e^{2|S|\varepsilon^2}} \leq \delta,$$

hence,

$$|S| \geq \frac{\log\left(\frac{n^{dk\log k}}{\delta}\right)}{2\varepsilon^2}.$$

□

Combining Theorem 4.2 with Corollary 4.10 proves Theorem 4.1.

4.2 Sensitivity

Our coreset is basically a sample consist of small amount of lines from a distribution over our input set of lines - a distribution that reflect the optimal clustering which we get from a pre-process in a linear time over the data. In order to get such a good and reflecting distribution, we need to estimate how much a line influences on the sum of distances in the optimal solution, and give to lines with higher such influence a higher probability to be chosen. We measure that influence by the line sensitivity.

Let $L = \{\ell_1, \ldots, \ell_n\}$ be a set of n lines in $\mathbb{R}^d$, let $P^* = \{p_1^*, \ldots, p_k^*\} \subseteq \mathbb{R}^d$ denote a set of k points that minimizes the total sum of distances to L over every k points in $\mathbb{R}^d$, i.e. $P^* = OPT(L,k)$, and let $\hat{P} \subseteq \mathbb{R}^d$ be a set of $k\beta$ points that satisfies $\text{cost}(L,\hat{P}) \leq \alpha \cdot \text{cost}(L,P^*)$, for any $\alpha > 0, \beta > 0$. For every $\ell \in L$, let ℓ' be the parallel line to ℓ that passes through the closest point to ℓ in $\hat{P}$, and let $L' = \{\ell_1', \ldots, \ell_n'\}$ denote their union. ; See Figure (3).

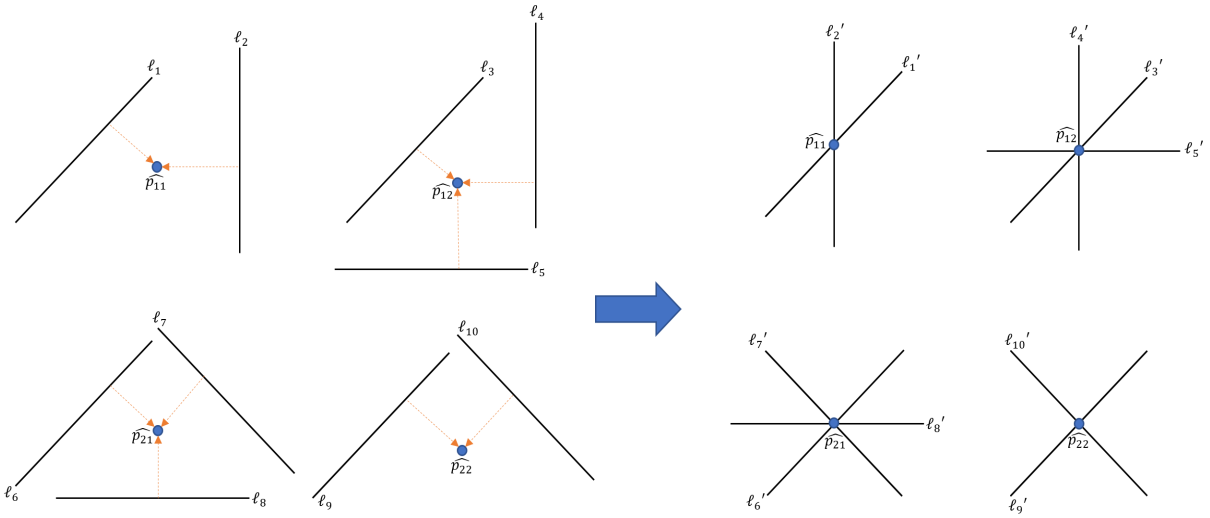
Definition 4.12 (Sensitivity) *For every integer $k \geq 1$, the sensitivity of a line $\ell \in L$ is defined by*

$$S_L(\ell,k) = \max_{P\subseteq\mathbb{R}^d, |P|=k} \frac{\text{dist}(\ell,P)}{\text{cost}(L,P)}, \tag{8}$$

and the total sensitivity is then defined to be

$$S(L,k) = \sum_{\ell\in L} S_L(\ell,k). \tag{9}$$

Figure 3: Example of a set $L = \{\ell_1, \ldots, \ell_{10}\}$ and its corresponding sets $L' = \{\ell'_1, \ldots, \ell'_{10}\}$ and $\hat{P} = \{\hat{p_{11}}, \hat{p_{12}}, \hat{p_{21}}, \hat{p_{22}}\}$, when $k = \beta = 2$. Each cluster of lines on the left is projected onto its center, as we can see on the right side.



Theorem 4.13 *For every integer $k \geq 1$ and a set L of finite number of lines in $\mathbb{R}^d$ we have*

$$S(L, k) \in O(\alpha\beta k^{O(k)} \log n).$$

Proof.

The proof will be divided into two parts, where in the first part we will bound $S(L, k)$ in terms of $S(L', k)$, and in the second part we will bound the size of $S(L', k)$; See Figure (3) for the first part.

Bound on $S(L, k)$. Let $k \geq 1$ be an integer, $P \subseteq \mathbb{R}^d$ be a set of k points, let $\ell \in L$, and let $\ell' \in L'$ denote the corresponding parallel line to ℓ. Let $x \in \ell$ and $x' \in \ell'$ be the closest points to P in ℓ and ℓ', respectively, and let $p' \in P$ denote the closest point to ℓ' in P. Hence,

$$\text{dist}(\ell, P) = \text{dist}(\ell, p') \leq \|x - p'\| \leq \|x - x'\| + \|x' - p'\| = \text{dist}(\ell, \ell') + \text{dist}(\ell', P). \tag{10}$$

From the definition of the sensitivity, we get

$$\text{dist}(\ell', P) = \frac{\text{dist}(\ell', P)}{\text{cost}(L', P)} \cdot \text{cost}(L', P) \leq S_{L'}(\ell', k) \cdot \text{cost}(L', P), \tag{11}$$

since $\text{cost}(L, \hat{P}) \leq \alpha \cdot \text{cost}(L, P)$, we have

$$\begin{aligned} \text{cost}(L', P) &= \sum_{\ell' \in L'} \text{dist}(\ell', P(\ell')) \leq \sum_{\ell' \in L'} \text{dist}(\ell', P(\ell)) \leq \sum_{\ell' \in L'} [\text{dist}(\ell, P(\ell)) + \text{dist}(\ell, \ell')] \\ &= [\text{cost}(L, P) + \text{cost}(L, \hat{P})] \leq 2\alpha \cdot \text{cost}(L, P). \end{aligned} \tag{12}$$

where $P(\ell)$ denote the closest point to ℓ in P. Combining (11) and (12) yields a bound on the right term in the right hand side of (10),

$$\text{dist}(\ell', P) = \frac{\text{dist}(\ell', P)}{\text{cost}(L', P)} \cdot \text{cost}(L', P) \leq S_{L'}(\ell', k) \cdot \text{cost}(L', P) \leq 2\alpha \cdot S_{L'}(\ell', k) \cdot \text{cost}(L, P). \tag{13}$$

We bound the left term in the right hand side of (10) by

$$\operatorname{dist}(\ell,\ell') = \frac{\operatorname{dist}(\ell,\ell')}{\operatorname{cost}(L,P)} \cdot \operatorname{cost}(L,P) \leq \frac{\alpha \cdot \operatorname{dist}(\ell,\ell')}{\operatorname{cost}(L,\hat{P})} \cdot \operatorname{cost}(L,P) = \frac{\alpha \cdot \operatorname{dist}(\ell,\hat{P})}{\operatorname{cost}(L,\hat{P})} \cdot \operatorname{cost}(L,P). \tag{14}$$

Plugging (13) and (14) in (10) and we thus obtain,

$$\operatorname{dist}(\ell,P) \leq \cdot\operatorname{dist}(\ell,\ell') + \cdot\operatorname{dist}(\ell',P) = \operatorname{cost}(L,P) \cdot \left(\frac{\alpha \cdot \operatorname{dist}(\ell,\hat{P})}{\operatorname{cost}(L,\hat{P})} + 2\alpha \cdot S_{L'}(\ell',k) \right). \tag{15}$$

From the last equation and from the definition of the sensitivity of $\ell \in L$ yields

$$S_L(\ell,k) \leq \frac{\operatorname{cost}(L,P) \cdot \left(\frac{\alpha \cdot \operatorname{dist}(\ell,\hat{P})}{\operatorname{cost}(L,\hat{P})} + 2\alpha \cdot S_{L'}(\ell',k) \right)}{\operatorname{cost}(L,P)} = \frac{\alpha \cdot \operatorname{dist}(\ell,\hat{P})}{\operatorname{cost}(L,\hat{P})} + 2\alpha \cdot S_{L'}(\ell',k).$$

Summing over every $\ell \in L$ yields

$$S(L,k) = \sum_{\ell \in L} S_L(\ell,k) \leq \sum_{\ell \in L} \left(\frac{\alpha \cdot \operatorname{dist}(\ell,\hat{P})}{\operatorname{cost}(L,\hat{P})} + 2\alpha \cdot S_{L'}(\ell',k) \right) = \alpha + 2\alpha \cdot S(L',k). \tag{16}$$

Bound on $S(L',k)$. For every $i \in \{1,\ldots,k\beta\}$, let $L'_i \subseteq L'$ denote the subset of lines in L' that intersect at $\hat{p}_i$. That yields a redefinition for the cost function for every $P \subseteq \mathbb{R}^d$ as follows

$$\operatorname{cost}(L',P) = \sum_{i=1}^{k\beta} \operatorname{cost}(L'_i,P), \tag{17}$$

and the total sensitivity is then bounded by

$$\begin{aligned} S(L',k) &= \sum_{\ell' \in L'} \max_{P \subseteq \mathbb{R}^d, |P|=k} \frac{\operatorname{dist}(\ell',P)}{\operatorname{cost}(L',P)} = \sum_{i=1}^{k\beta} \sum_{\ell' \in L'_i} \max_{P \subseteq \mathbb{R}^d, |P|=k} \frac{\operatorname{dist}(\ell',P)}{\operatorname{cost}(L',P)} \\ &\leq \sum_{i=1}^{k\beta} \sum_{\ell' \in L'_i} \max_{P \subseteq \mathbb{R}^d, |P|=k} \frac{\operatorname{dist}(\ell',P)}{\operatorname{cost}(L'_i,P)} = \sum_{i=1}^{k\beta} S(L'_i,k). \end{aligned} \tag{18}$$

Hence, we will bound $S(L'_i,k)$ for some subset of lines $L'_i \subseteq L'$, and thus we will bound the total sensitivity. Fix a subset $L'_i \subseteq L'$ that are intersected at $\hat{p}_i \in \hat{P}$. For convenience, assume that all the lines in L'_i were shifted by $\hat{p}_i$ and now are intersected at the origin. Let S denote the unit-sphere. For every $P \subseteq \mathbb{R}^d$, let $\operatorname{proj}(p,S) \in S$ denote the closest point in S to $p \in P$, and let $\operatorname{proj}(P,S) \subseteq S$ denote their union over every $p \in P$. By Thales' theorem, we have

$$\operatorname{dist}(\ell',p) = \|p\| \cdot \operatorname{dist}(\ell', \operatorname{proj}(p,S)), \tag{19}$$

for every $\ell' \in L'_i, p \in \mathbb{R}^d$. By denoting with $P(\ell) \in P$ the closest point in $P \subseteq \mathbb{R}^d$ to a line ℓ' in $\mathbb{R}^d$, we can use (19) and obtain

$$S(L'_i,k) = \sum_{\ell' \in L'_i} \max_{P \subseteq \mathbb{R}^d, |P|=k} \frac{\operatorname{dist}(\ell',P)}{\sum_{\ell \in L'_i} \operatorname{dist}(\ell,P)} = \sum_{\ell' \in L'_i} \max_{P \subseteq \mathbb{R}^d, |P|=k} \frac{\|P(\ell')\| \cdot \operatorname{dist}(\ell', \operatorname{proj}(S,P))}{\sum_{\ell \in L'_i} \|P(\ell)\| \cdot \operatorname{dist}(\ell, \operatorname{proj}(S,P))},$$

that is, a set of k weighted points on the unit-sphere that maximizes the sensitivity of each line. For simplicity, we write it as follows,

$$S(L'_i,k)=\sum_{\ell'\in L'_i}\max_{P\subseteq S,|P|=k}\frac{w(P(\ell'))\cdot \mathrm{dist}(\ell',P)}{\sum_{\ell\in L'_i} w(P(\ell))\cdot \mathrm{dist}(\ell,P)}. \tag{20}$$

For every $\ell'\in L'_i$, let $Q(\ell')\in \ell'\cap S$ denote an arbitrary point from the two possible intersection points of ℓ' and the unit-sphere, and let $Q(L'_i)\subseteq S$ denote their union over every $\ell'\in L'_i$. From the right-angle triangle, for every $\ell'\in L'_i, p\in S$, exists a positive $\alpha(\ell',p)\leq 1$ that satisfies the following

- $\frac{2\alpha(\ell',p)}{\pi}<\sin(\alpha(\ell',p))<\frac{\pi\alpha(\ell',p)}{2}$
- $\mathrm{dist}(\ell',p)=\sin(\ell',p)\cdot\|p-Q(\ell')\|$

Combining the last two with (20) yields

$$\begin{aligned} S(L'_i,k)&=\sum_{\ell'\in L'_i}\max_{P\subseteq S,|P|=k}\frac{w(P(\ell'))\cdot \mathrm{dist}(\ell',P)}{\sum_{\ell\in L'_i} w(P(\ell))\cdot \mathrm{dist}(\ell,P)}\\ &\in O(1)\cdot\sum_{\ell'\in L'_i}\max_{P\subseteq S,|P|=k}\frac{w(P(\ell'))\cdot\|\min\{Q(\ell')-P(\ell'),Q(\ell')+P(\ell')\}\|}{\sum_{\ell\in L'_i} w(P(\ell))\cdot\|\min\{Q(\ell)-P(\ell),Q(\ell)+P(\ell)\}\|}. \end{aligned} \tag{21}$$

Since every k points and their neg are just the optimal $2k$ points, we have

$$\begin{aligned} S(L'_i,k)&=\sum_{\ell'\in L'_i}\max_{P\subseteq S,|P|=k}\frac{w(P(\ell'))\cdot\|\min\{Q(\ell')-P(\ell'),Q(\ell')+P(\ell')\}\|}{\sum_{\ell\in L'_i} w(P(\ell))\cdot\|\min\{Q(\ell)-P(\ell),Q(\ell)+P(\ell)\}\|}\\ &=\sum_{\ell'\in L'_i}\max_{P\subseteq S,|P|=2k}\frac{w(P(\ell'))\cdot\|Q(\ell')-P(\ell')\|}{\sum_{\ell\in L'_i} w(P(\ell))\cdot\|Q(\ell)-P(\ell)\|}. \end{aligned} \tag{22}$$

In order to bound the last expression, which is the total sensitivity of $2k$ weighted clustering points, we use a method due to Feldman and Schulman theorem [**?**].

Theorem 4.14 *Let $P\subseteq$ be as set of n points in $\mathbb{R}^d$. For every $p\in P$ and a positive integer $k\geq 1$, let*

$$s(p,k):=\max_{\substack{\{c_1,\ldots,c_k\}\subseteq\mathbb{R}^d\\ \{w_1,\ldots,w_k\}\in[0,\infty)^k}}\frac{\min_{i\in\{1,\ldots,k\}} w_i\cdot\|p-c_i\|}{\sum_{p'\in P}\min_{i\in\{1,\ldots,k\}} w_i\cdot\|p'-c_i\|},$$

then,

$$\sum_{p\in P} s(p)\leq k^{O(k)}\log n. \tag{23}$$

Plugging (23) in (22) yields

$$S(L'_i,k)\leq k^{O(k)}\log n,$$

and substitute the last upper bound in (18) proves the theorem.

□

References

[1] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms.

[2] R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy. The effectiveness of lloyd-type methods for the k-means problem.

[3] Weber, Alfred, 1909, The Theory of the Location of Industries, Chicago, Chicago University Press, 1929, 256 pages.

[4] J. A. Lee and M. Verleysen. Nonlinear dimensionality reduction. Springer Science & Business Media, 2007.

[5] D. Feldman, M. Schmidt, and C. Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms.

[6] W. B. Johnson and J. Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. Contemporary mathematics, 26(189-206):1, 1984.

[7] N. Megiddo and K. J. Supowit. On the complexity of some common geometric location problems. SIAM J. Comput., 13:182196, 1984.

[8] C. H. Papadimitriou. Worst-case and probabilistic analysis of a geometric location problem. SIAM Journal of Computing, 10:542 557, 1981.

[9] P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan. Approximating extent measures of points. Journal of the ACM, 51(4):606-635, 2004.

[10] S. Har-Peled and A. Kushal. Smaller coresets for k-median and k-means clustering. In SCG 05: Proceedings of the twenty-first annual symposium on Computational geometry, pages 126134, New York, NY, USA, 2005. ACM.

[11] S. Har-Peled and S. Mazumdar. On coresets for k-means and k-median clustering. In STOC 04: Proceedings of the thirty-sixth annual ACM symposium on Theory of computing, pages 291300, New York, NY, USA, 2004. ACM.

[12] D. Feldman, M. Monemizadeh, and C. Sohler. A ptas for k-means clustering based on weak coresets. In SCG 07: Proceedings of the twenty-third annual symposium on Computational geometry, pages 1118, New York, NY, USA, 2007. ACM.

[13] D. Feldman, A. Fiat, and M. Sharir. Coresets forweighted facilities and their applications. In Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on, pages 315-324. IEEE, 2006.

[14] H. W. Hamacher and Z. Drezner. Facility location: applications and theory. Springer Science & Business Media, 2002.

[15] Gonzlez, T. 1985. Clustering to minimize the maximum intercluster distance. Theoret. Comput. Sci. 38, 293–306.

[16] Dorit S. Hochbaum , David B. Shmoys, A unified approach to approximation algorithms for bottleneck problems, Journal of the ACM (JACM), v.33 n.3, p.533-550, July 1986.

[17] Toms Feder, Daniel Greene, Optimal algorithms for approximate clustering, Proceedings of the twentieth annual ACM symposium on Theory of computing, p.434-444, May 02-04, 1988, Chicago, Illinois, USA .

[18] Mihai Bādoiu , Sariel Har-Peled , Piotr Indyk, Approximate clustering via core-sets, Proceedings of the thiry-fourth annual ACM symposium on Theory of computing, May 19-21, 2002, Montreal, Quebec, Canada .

[19] Jie Gao , Michael Langberg , Leonard J. Schulman, Analysis of Incomplete Data and an Intrinsic-Dimension Helly Theorem, Discrete & Computational Geometry, v.40 n.4, p.537-560, October 2008

[20] Danzer, L., Grnbaum, B., Klee, V. (1963), Helly's theorem and its relatives, Convexity, Proc. Symp. Pure Math., 7, American Mathematical Society, pp. 101179.

[21] Clustering lines in high dimensional space: classification of incomplete data, with J. Gao and M. Langberg. ACM Trans.

[22] Pankaj K. Agarwal , Cecilia M. Procopiuc, Approximation algorithms for projective clustering, Proceedings of the eleventh annual ACM-SIAM symposium on Discrete algorithms, p.538-547, January 09-11, 2000, San Francisco, California, USA.

[23] Pankaj K. Agarwal , Cecilia M. Procopiuc , Kasturi R. Varadarajan, A $(1+\epsilon)$-approximation algorithm for 2-line-center, Computational Geometry: Theory and Applications, v.26 n.2, p.119-128, October 2003.

[24] Sariel Har-Peled , Kasturi Varadarajan, Projective clustering in high dimensions using core-sets, Proceedings of the eighteenth annual symposium on Computational geometry, p.312-318, June 05-07, 2002, Barcelona, Spain.

[25] Dan Feldman, Mikhail Volkov, Daniela Rus, Dimensionality Reduction of Massive Sparse Datasets Using Coresets, CoRR abs/1503.01663 (2015).

[26] k-Clustering of Big Data via Coresets of Size $O(k)$, Artem Barger, Dan Feldman, TODO: ???.

$\frac{d \log n + \log\left(\frac{1}{\delta}\right)}{\varepsilon^2}$