Coresets for Weighted Facilities and Their Applications

Dan Feldman[*]
School of Computer Science
Tel Aviv University
Tel Avivs 69978, Israel
dannyf@post.tau.ac.il

Amos Fiat[*]
School of Computer Science
Tel Aviv University
Tel Aviv 69978, Israel
fiat@post.tau.ac.il

Micha Sharir[†]
School of Computer Science
Tel Aviv University
Tel Aviv 69978, Israel
michas@post.tau.ac.il

Abstract

We develop efficient $(1+\varepsilon)$-approximation algorithms for generalized facility location problems. Such facilities are not restricted to being points in $\mathbb{R}^d$, and can represent more complex structures such as linear facilities (lines in $\mathbb{R}^d$, j-dimensional flats), etc. We introduce coresets for weighted (point) facilities. These prove to be useful for such generalized facility location problems, and provide efficient algorithms for their construction. Applications include: k-mean and k-median generalizations, i.e., find k lines that minimize the sum (or sum of squares) of the distances from each input point to its nearest line. Other applications are generalizations of linear regression problems to multiple regression lines, new SVD/PCA generalizations, and many more. The results significantly improve on previous work, which deals efficiently only with special cases. Open source code for the algorithms in this paper is also available.

1 Introduction

An avalanche of recent work has been generated by the seminal work of Agarwal, Har-Peled, and Varadarajan [1] that formally defined the notion of a *coreset*. Intuitively, given some property for a set of points $P \subset \mathbb{R}^d$ (such as its width, diameter, smallest bounding box, etc.), a coreset for this property is a small (possibly weighted) subset of P, that approximately preserves this property [2]. Small coresets often imply efficient approximation algorithms for related optimization problems.

As a motivating example, let P be a set of points in $\mathbb{R}^d$. Har-Peled and Mazumdar [16] describe how to construct a *coreset for k-median*: This is a weighted subset $\mathcal{S}$ of P, so that for any set of k points in $\mathbb{R}^d$ (called *facilities*), the weighted sum of distances from points in $\mathcal{S}$ to their nearest facilities is approximately the same as (i.e., differs by a factor of $1 \pm \varepsilon$ from) the sum of distances from the points of P to their nearest facilities. The sum of distances is used in "median" problems; in "mean" problems we replace it by the sum of squared distances.

Such a coreset implies an efficient approximation algorithm for the k-median (or k-mean) problem: An optimal set of k facilities for $\mathcal{S}$ is a good approximation to the optimal set for P, and, if $\mathcal{S}$ is sufficiently small, the former set can be found efficiently via brute force.

Generalized facilities. We seek to study generalizations of k-median/mean like problems where facilities are not restricted to being points in $\mathbb{R}^d$, but possibly more general structures. In particular, we are interested in linear facilities (lines or j-dimensional flats, $j \geq 2$, in $\mathbb{R}^d$).

Finding a small set of low-dimensional flats that approximately matches the input points is a problem that appears in a great many areas. For example, "one of the most fundamental problems in computer vision is to find straight lines in an image" [4]. Other examples include: matrix approximation [9], image processing [24], data compression [21], graphics [18], socioeconomics [19], and many more.

In some cases such problems are amenable to algebraic techniques, in particular finding one flat that approximately minimizes the sum of squared distances can be done using SVD or PCA techniques. This does not hold for the sum of distances, and does not hold for $k > 1$ generalized facilities. The class of problems we consider include many such variants that are unlikely to admit a polynomial-time solution when k is part of the input [20] (unless $P = NP$). For such problems no non-trivial approximation algorithms were previously

[*] Supported by the German-Israel Foundation for Scientific Research and Development

[†] Supported by NSF Grant CCF-05-14079, by a grant from the U.S.-Israeli Binational Science Foundation, by grant 155/05 from the Israel Science Fund, and by the Hermann Minkowski–MINERVA Center for Geometry at Tel Aviv University.

known. E.g., given a set of points in 3-dimensional space, find a 1-dimensional flat (line) in $\mathbb{R}^3$ that approximately minimizes the sum of distances from to the points. We give a linear time PTAS for this problem.

Coresets for weighted facilities. To tackle optimization problems of this kind that deal with linear facilities, we introduce a novel tool, called *coresets for weighted facilities*. Specifically, let P be a set of weighted points *on a line* ℓ, and C be a set of weighted facilities (points) in $\mathbb{R}^d$, where each $c \in C$ has some positive weight $W(c)$. We define $\nu'_C(P)$ $\big($resp. $\mu'_C(P)\big)$ as the overall sum of minimal weighted distances (resp. squared weighted distances) from points to facilities. That is

$$\nu'_C(P) = \sum_{p \in P} \left(w(p) \cdot \min_{c \in C} \{W(c)\,\|p-c\|\} \right), \text{ and}$$

$$\mu'_C(P) = \sum_{p \in P} \left(w(p) \cdot \min_{c \in C} \left\{ (W(c)\,\|p-c\|)^2 \right\} \right).$$

Fix k and $\varepsilon > 0$. A (possibly differently) weighted set $\mathcal{S} \subseteq P$ is called a (k, ε)-*coreset for weighted facilities*, if *for any* weighted set of k facilities (points) $C \subset \mathbb{R}^d$, (*i*) and (*ii*) hold:

$$(i) \quad (1-\varepsilon)\nu'_C(P) \le \nu'_C(\mathcal{S}) \le (1+\varepsilon)\nu'_C(P), \quad (1.1)$$

$$(ii) \quad (1-\varepsilon)\mu'_C(P) \le \mu'_C(\mathcal{S}) \le (1+\varepsilon)\mu'_C(P).$$

In other words, a coreset for the weighted facilities problem is a (weighted) subset of the input set, so that *for any k facilities*, with *any associated weights*, the sum of minimum weighted (squared) distances to the facilities is about the same for the original set and for the weighted subset.

This problem is interesting in its own right, and arises naturally in facility location (see [10]). However, we only know how to construct (k, ε)-coresets for weighted facilities when the points of P all lie on a *line* (but the facilities can be anywhere in $\mathbb{R}^d$), and it is open at the moment whether the construction can be extended to arbitrary input sets in $\mathbb{R}^d$, $d \ge 2$.

Nevertheless, (k, ε)-coresets for weighted facilities for point sets on a line, are sufficient for solving optimization problems for generalized facilities of the kinds mentioned above, for aribtrary point sets. Specifically, they lead to construction of new coresets for generalized facilities, with no restriction on the input set P in $\mathbb{R}^d$.

For a collection of generalized facilities Y, let $\operatorname{dist}(p, Y)$, $p \in \mathbb{R}^d$, denote distance from point p to the closest generalized facility $y \in Y$. We obtain *linear and point facilities* coresets for arbitrary $P \subset \mathbb{R}^d$. *I.e.,* given k and ε, the coreset $\mathcal{S}$ computed from P has the property that for any (mixed) set Y that contains $0 \le j \le k$ lines and at most $k - j$ points in $\mathbb{R}^d$, (*i*) and (*ii*) hold:

$$(i) \quad (1-\varepsilon)\nu_Y(P) \le \nu_Y(\mathcal{S}) \le (1+\varepsilon)\nu_Y(P)$$

$$(ii) \quad (1-\varepsilon)\mu_Y(P) \le \mu_Y(\mathcal{S}) \le (1+\varepsilon)\mu_Y(P)$$

where $\nu_Y(P) = \sum_{p \in P} \big(w(p) \cdot \operatorname{dist}(p, Y) \big)$,
and $\mu_Y(P) = \sum_{p \in P} \Big(w(p) \cdot (\operatorname{dist}(p, Y))^2 \Big)$.

Thus, this coreset is a generalization of coresets for k-median, and simultaneously, a generalization of coresets for k-mean. Additionally, this coreset approximately preserves distances to both *point* facilities and *line* facilities. For Y restricted to point sets, [16] give coresets for k-median and k-mean. It is intresting that, unlike prior constructions, we get the *same* coreset for both k-mean and k-median. However, the significance of our construction mainly lies in its applications to generalized linear facilities.

In addition, for arbitrary input point sets P in $\mathbb{R}^d$, our coreset $\mathcal{S}$ has the property that for any single j-dimensional flat f, with $0 \le j \le d-1$, (*i*) and (*ii*) hold:

$$(i) \quad (1-\varepsilon)\nu_{\{f\}}(P) \le \nu_{\{f\}}(\mathcal{S}) \le (1+\varepsilon)\nu_{\{f\}}(P)$$

$$(ii) \quad (1-\varepsilon)\mu_{\{f\}}(P) \le \mu_{\{f\}}(\mathcal{S}) \le (1+\varepsilon)\mu_{\{f\}}(P)$$

where $\nu_{\{f\}}(\cdot)$, and $\mu_{\{f\}}(\cdot)$ are defined in an analogous manner to the preceding definitions.

Further Results. As mentioned, we define the notion of a weighted facilities (k, ε)-coreset $\mathcal{S}$ for a point set P on a line. We give an algorithm to construct such coresets, in $O(nk)$ time, where the coreset is of size $2^{O(k)}\varepsilon^{-2k-1}\log^{4k-3} n$. Given any set of points $P \subset \mathbb{R}^d$, for fixed $d \ge 1$, we construct these weighted facility coresets for projections of P onto certain lines, and then combine them to form the desired coreset for P itself. Recently, Har-Peled [15] proposed a set of size $2^{O(k)}\varepsilon^{-k-1}\log^{k+1} n$) that satisfies only (1.1)(*i*).

Using these coresets we obtain LTAS's ($O(n)$-time $(1+\varepsilon)$-approximation algorithms) for the following problems.

C1 *Coreset for linear and point facilities*: Find a small weighted subset that well approximates the sum of distances, or of squared distances, from the points of P to *any* given set of $0 \le i \le k$ lines and at most $k - i$ points in $\mathbb{R}^d$, up to a factor of $(1+\varepsilon)$. The same coreset also approximates sum of (squared) regression distances (i.e, distances measured in the x_d-dimension). We construct such coresets of size $\varepsilon^{-d-k}(\log n)^{O(1)}$ in $O(n)$ time, for any fixed $k, d \ge 1$.

C2 *Coreset for a flat*: Find a small weighted subset that well approximates the sum of distances, or of squared distances, from P to *any* (single) j-dimensional flat,

$0 \le j \le d-1$. We construct such coresets of size $\varepsilon^{-d-1}(\log n)^{O(j^2)}$ in $O(n)$ time, for any fixed $d \ge 1$.

P1 *Approximate k-line median/mean*: Find a set of k lines in $\mathbb{R}^d$ such that the sum of the distances, or the squared distances, from the points of P to their closest lines is minimized, up to a factor of $(1+\varepsilon)$.

P2 *Approximate j-flat median/mean*: Find a j-dimensional flat f such that the sum of the distances, or the squared distances, from the points of P to f is at most $(1+\varepsilon)$ times the optimum value of such a sum. The solution uses a single coreset that is good for *any* dimension j.

P3 *Restricted Facility Location*: Approximate the k-line median/mean or j-flat median/mean with additional constraints on the allowed location of the lines/flat, by forbidding them, or alternatively forcing them, to pass through certain locations.

P4 *Approximate k-regression lines and M-estimators*: Solve problems P1–P3, now with vertical (regression) distances (in the direction of the x_d-axis), squared or non-squared.

P5 *Data Fitting with outliers*: For a fixed k and k', or for a fixed value of $k+k'$, find a set of k lines and k' points that minimizes the sum of distances, or of squared distances, from each point to its nearest facility (with or without location constraints). Note that k' represents the number of outlier *clusters* and not the number of outliers. This may suggest a way to deal with outliers when their exact number is not known.

We remark that for $\varepsilon = \Theta(1)$ we can also generalize all the above results for high-dimensional spaces, *i.e.,* where d is not constant. In this case our construction runs in time linear in d and yields a coreset of size independent on d. See Remark 2.3.

Related Work.

C1-C2 Although coresets for linear facilities are discussed in several places, no constructions have yet been suggested [9, 2].

P1 For $k = 1$ and $d = 2$, Yamamoto et al. [25] give an $O(n^{1.5}\log^2 n)$ time algorithm that computes a 1-line median for a set of input points P. Using Dey's improved bound on the number of halving lines [8], the algorithm can be improved to $O(n^{4/3}\log^2 n)$.

The 1-line mean can be computed in $O(n)$ time using the SVD technique, for any fixed d. In previous work [12] we gave an exact (optimal) solution for the k-line-mean in the plane that takes $O(n^3)$ time for $k = 2$, and $n^{O(k^2)}$ for $k \ge 3$. Recently, [9] give an $(n/\varepsilon)^{O(k/\varepsilon)}$ PTAS for computing the k-line mean. Many heuristics for this problem, such as the Hough transform and Independent Component Analysis (ICA), have been proposed (see references in [17]).

P2 Although the j-flat mean can be computed in $O(n)$ time for any fixed d and j using SVD, no analogous efficient algorithms are known for the j-flat median or its approximations for $1 \le j < d-1$.

The $(d-1)$-dimensional flat that minimizes the sum of distances to P can be computed in $O(n^d)$ time [3]. Prior to our work, no polynomial time approximation was known for a j-dimensional flat ($j < d-1$) that minimizes the sum of distances to P (even for $j = 1$ and $d = 3$); these are cited as "interesting open problems" [22, 10, 3]. Our PTAS runs in linear time for fixed d and ε.

P3 Polynomial-time algorithms for a good approximate $(d-1)$-flat with respect to the sum of distances or distances squared, and subject to additional restrictions, are given in [10, 22].

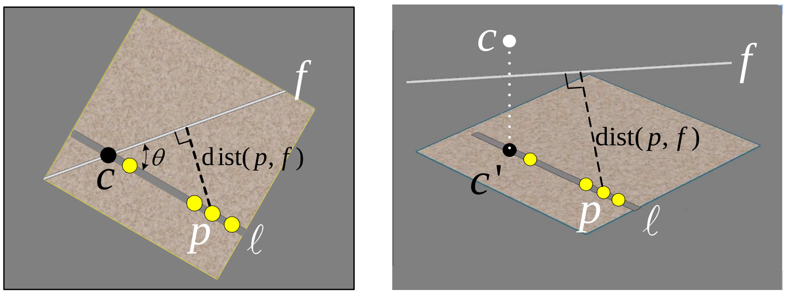
For fixed ε and d, we give linear time PTAS algorithms for computing approximate k lines or a single j-dimensional flat ($2 \le j \le d-1$), subject to various constraints. Note that even in the case of one flat, or even one line in the plane ($j = 1$, $d = 2$), algebraic methods, such as the SVD/PCA, cannot handle constraints.

P4 A $(1+\varepsilon)$-approximation for the j-flat mean, for squared regression distances with no constraints, can also be computed in $O(n)$ time using SVD. For the median (regression) line in the plane ($d = 2$), the 1-line median can be computed in $O(n)$ time [25]. For $d > 2$ and $j = d-1$, a PTAS that takes $O(n\log n)d^{O(1)} + O(n)(1/\varepsilon)^{O(1)}$ time was recently suggested by [7] for the $(d-1)$-flat median with vertical (regression) distances. No results are known for the case $1 < j < d-1$, or where there are constraints on the location of the flat/lines.

P5 Outliers were investigated for the k-(point) mean and median problems [6, 5]. However, we do not know of any generalization for linear facilities, even for a single line in the plane.

Why coresets for weighted facilities? To motivate the relationship between weighted facilities and linear facilities, consider the following (restrictive) scenario: The (unweighted) input point set P resides on some line $\ell \subset \mathbb{R}^d$, $f \subset \mathbb{R}^d$ is another line, and $\ell \cap f \neq \emptyset$. It follows from elementary trigonometry, that the distance between a point $p \in \ell$ and f is equal to $\|c-p\|\sin\theta$, where c is the point of intersection between ℓ and f, and θ is the angle formed at c by these two lines. See Fig. 1(left).

This simple observation lies at the heart of our work. It extends to arbitrary (skew) lines ℓ and f $\big($see Fig. 1(right)$\big)$. I.e., for any lines ℓ and f, such that f is



$f \cap \ell \neq \emptyset$ $f \cap \ell = \emptyset$

Fig. 1: ***(left)*** $\mathrm{dist}(p, f) = W(c) \cdot \mathrm{dist}(p, c)$, with $W(c) = \sin\theta$. Hence, c, weighted by $\sin\theta$, replaces f for points on ℓ. ***(right)*** $\mathrm{dist}(p, f) = W(c) \cdot \mathrm{dist}(p, c)$ with $W(c) = \sin\theta$, for any pair (ℓ, f) of lines in $\mathbb{R}^d$, where c is a point on the line that spans the shortest distance between ℓ and f, at distance $\mathrm{dist}(\ell, f)/\sin\theta$ from the point $c' \in \ell$, nearest to f, and θ is the angle between the lines ℓ and f (a routine exercise in stereometry).

not a translation of ℓ, there exist some weighted point facility $c \in \mathbb{R}^d$ such that the (weighted) distance from any point $p \in \ell$ to c is equal to the distance between p and f. This claim can be further generalized to the case where f is a j-dimensional flat, of arbitrary dimension $j \leq d - 1$, and also for vertical (regression) distances.

This seemingly suggests a very general transformation. Subject to the restriction that the input point set P be contained in some line, there is a general reduction from any optimization problem that involves distances between points of P and arbitrary j-dimensional flats, to another optimization problem that involves the points of P and weighted (point) facilities.

Unfortunately, for general sets of points $P \subset \mathbb{R}^d$, there is no point $c \in \mathbb{R}^d$ such that the distance between a linear facility f and a point $p \in P$ is proportional to the distance between p and c. We show how to overcome this setback by reducing the general case to several subproblems involving points on a line.

The paper is organized as follows. In Section 2 we present the construction of coresets for weighted facilities for point sets on a line. Then, in Section 3, we use this construction to obtain coresets for k-line median/mean for arbitrary point sets in $\mathbb{R}^d$ (problem C1), and also for flat median/mean (problem C2).

For lack of space, this abstract omits most of the proofs. They can be found in the full version of the paper, available on-line, together a companion source code, in C++ and Mathlab, at `www.cs.tau.ac.il/~dannyf`. The applications of these coresets for solving problems P1–P5 are also omitted. These applications, however, are trivial: Since the coreset $\mathcal{S}$ has small size, we can use any (possibly inefficient) algorithm for computing, say, the (exact or approximate) k-line median for $\mathcal{S}$, and then report it as an approximate k-line median for the whole input set. We present such algorithms in [14] and [12].

2 Coresets for Weighted Facilities

2.1 ε-Coresets for a single facility

Let P be a weighted set of points in $\mathbb{R}^d$ and $0 < \varepsilon \leq 1$. A weighted set $\mathcal{S}$, where $\mathcal{S} \subseteq P$, is called an *ε-coreset for a single facility* if, for every facility (point) $c \in \mathbb{R}^d$, (*i*) and (*ii*) hold:

$$(i) \quad (1-\varepsilon)\nu_{\{c\}}(P) \leq \nu_{\{c\}}(\mathcal{S}) \leq (1-\varepsilon)\nu_{\{c\}}(P) \tag{2.1}$$
$$(ii) \quad (1-\varepsilon)\mu_{\{c\}}(P) \leq \mu_{\{c\}}(\mathcal{S}) \leq (1-\varepsilon)\mu_{\{c\}}(P).$$

The algorithm SINGLE-FACILITY-CORESET given below is very similar to the one in [16], but, unlike [16], produces a single coreset that satisfies both (2.1)(i) and (ii). We use this algorithm later in this section, and in Section 3.

Algorithm SINGLE-FACILITY-CORESET(P, ε)
Input: Weighted point set $P \subset \mathbb{R}^d$, $0 < \varepsilon \leq 1$
Output: A single-facility 9ε-coreset for P of size $O(\varepsilon^{-d} \log W)$, where $W = \sum_{p \in P} w(p)$

```
$\overline{P} \leftarrow \sum_{p \in P} \big(w(p) \cdot p\big)/W$,
  $R \leftarrow \sum_{p \in P} \|p - \overline{P}\|/W$
for $j \leftarrow 1$ to $\lceil \log W \rceil$
  do $B_j \subset \mathbb{R}^d$ := the closed ball with radius
       $2^j R$ centered at $\overline{P}$    (* See Figure 2 *)
       (* Note: $P \subset B_{\lceil \log W \rceil}$, since
       $\|p - \overline{P}\| \leq WR \quad \forall p \in P$ *)
     $G_j \leftarrow$ vertex set of an infinite grid of
           cell size $2^j \varepsilon R/\sqrt{d}$ centered at $\overline{P}$
     if $j = 1$
        then $V_1 \leftarrow G_1 \cap B_1$
```

```
7           else  $V_j \leftarrow G_j \cap (B_j \setminus B_{j-1})$
8       for each nonempty cell $\Delta \in V_j$
9         do choose arbitrary point $p'$ in $P \cap \Delta$
            (* Note: $\|p - p'\| \leq \varepsilon \left\|p - \overline{P}\right\|$ if $j > 1$,
            i.e., $\|p - p'\| \leq \varepsilon \cdot \max\{R, \left\|p - \overline{P}\right\|\}$. *)
10          $w(p') \leftarrow \sum_{p \in P \cap \Delta} w(p)$
11          $\mathcal{S} \leftarrow \mathcal{S} \cup \{p'\}$
12    return $\mathcal{S}$
```

With careful implementation, SINGLE-FACILITY-CORESET takes $O(n)$ time, using the log and floor functions. The proof that SINGLE-FACILITY-CORESET indeed returns an ε-coreset for P is a consequence of the following lemma; its somewhat cumbersome notation is needed for further applications of proving the correctness of constructions of other coresets of interest, in Section 3.

Lemma 2.1. *Let P and $\mathcal{S}$ be two weighted sets in $\mathbb{R}^d$, and let g be a mapping from P to $\mathcal{S}$ such that the weight $w(p')$ of $p' \in \mathcal{S}$ is equal to the sum of the weights of all points $p \in P$ with $g(p) = p'$.*

Let $\{P_1, P_2, \ldots, P_m\}$ be some partition of P, and let $C_1, C_2, \ldots, C_m$, $C_i \subset \mathbb{R}^d$, be a collection of m "objects" (sets) in $\mathbb{R}^d$. Define $R = \sum_{i=1}^{m} \nu_{C_i}(P_i)/w(P)$, where $w(P) = \sum_{p \in P} w(p)$. Assume that for some $\varepsilon > 0$ we have

$$\|p - p'\| \leq \varepsilon \cdot \max\{R, \operatorname{dist}(p, C_i)\},$$

for every $p \in P_i$, $1 \leq i \leq m$, and its image $p' = g(p)$. Then, for any $Q \subset \mathbb{R}^d$,

(i) $\sum_{i=1}^{m} \nu_{C_i}(P_i) \leq \alpha \nu_Q(P)$ for some $\alpha \geq 1$, implies that $|\nu_Q(P) - \nu_Q(\mathcal{S})| \leq 2\alpha\varepsilon \cdot \nu_Q(P)$.

(ii) $\sum_{i=1}^{m} \mu_{C_i}(P_i) \leq \beta \mu_Q(P)$ for some $\beta \geq 1$, implies that $|\mu_Q(P) - \mu_Q(\mathcal{S})| \leq 9\beta\varepsilon \cdot \mu_Q(P)$.

Corollary 2.2. *Let P be a set of n points in $\mathbb{R}^d$ for constant d, and $0 < \varepsilon \leq 1$. Then SINGLE-FACILITY-CORESET$(P, \varepsilon/9)$ returns, in $O(n)$ time, a single-facility ε-coreset for P, of size $O\big(\varepsilon^{-d} \log n\big)$.*

Proof. Define $\overline{P} = \sum_{p \in P} p/w(P)$, and $R = \nu_{\overline{P}}(P)/n$. As noted in line 2.1 of SINGLE-FACILITY-CORESET, we have $\|p - p'\| \leq \varepsilon \max\{R, \left\|p - \overline{P}\right\|\}$ for every $p \in P$ and its representative p' in the coreset. It is also well known that for any $q \in \mathbb{R}^d$ we have $\nu_{\overline{P}}(P) \leq 2\nu_q(P)$, and also $\mu_{\overline{P}}(P) \leq \mu_q(P)$ (see [11]). Thus, substituting $m = 1$, $C_1 = \{\overline{P}\}$, $Q = \{q\}$, $\alpha = 2$, $\beta = 1$ in Lemma 2.1, yields the corollary. □

Remark 2.3. The algorithm SINGLE-FACILITY-CORESET is the only algorithm in this paper which explicitly uses space and running time that are exponential in d. However, for $\varepsilon = \Theta(1)$, choosing an arbitrary single point p' in $P \cap (B_j \setminus B_{j-1})$ at line 9 of the algorithm SINGLE-FACILITY-CORESET, makes the construction time linear in d. This yields constant approximation for the coresets constructed in this paper, also for high-dimensional spaces (non-constant d).

2.2 (k, ε)-Coresets for weighted facilities

In this section we assume that P is a set of points on a line ℓ in $\mathbb{R}^d$.

Voronoi region. Given a weighted set of facilities $C \subset \mathbb{R}^d$, with an associated weight function $W : C \mapsto \mathbb{R}$, we define the *Voronoi region* $V(c)$ associated with $c \in C$ to be the set of points $x \in \mathbb{R}^d$ such that $W(c)\,\|x - c\| \leq W(c')\,\|x - c'\|$ for all $c' \in C$. See Fig. 2.

Voronoi intervals and boundaries. Given a line ℓ, a set of facilities $C \subset \mathbb{R}^d$, and an associated weight function $W : C \mapsto \mathbb{R}^+$, a *Voronoi interval* for a facility $c \in C$ is a connected component of $V(c) \cap \ell$. Endpoints of Voronoi intervals are called *Voronoi boundaries*. Two Voronoi intervals are called *adjacent* if they share a Voronoi boundary.

Remark: Note that if all the facilities have the same weight, then each facility has a single connected Voronoi interval; see Fig. 2(left). However, if their weights are unequal, then a single facility may serve multiple intervals; see Fig. 2(right).

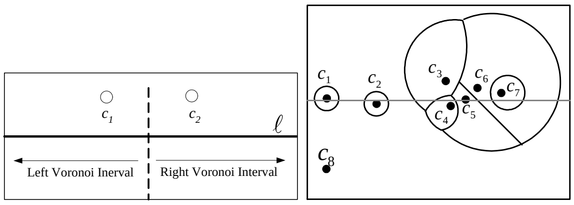


Fig. 2: **(left)** Two facilities of equal weight induce two Voronoi intervals. **(right)** Eight weighted facilities in the plane, and the resulting partition of their eight Voronoi regions into 12 Voronoi intervals.

Lemma 2.5. *Let $C \subset \mathbb{R}^d$ be a weighted set of k facilities. The total number of their Voronoi intervals on a fixed line ℓ is at most $2k - 1$.*

Proof. Follows from the easy observation that the sequence of facilities that own the Voronoi intervals is a Davenport-Schinzel sequence of order 2 on k symbols [23]. □

In what follows we assume, without loss of generality, that ℓ is the x-axis.

(k,ε)-V-coreset for P. A weighted set $\mathcal{S} \subseteq P$ is called a (k,ε)-*V-coreset*, if, for any weighted set of facilities $C \subset \mathbb{R}^d$, such that P is contained in at most k adjacent Voronoi intervals of C, (*i*) and (*ii*) hold:

$$(i) \quad (1-\varepsilon)\nu'_C(P) \le \nu'_C(\mathcal{S}) \le (1+\varepsilon)\nu'_C(P),$$

$$(ii) \quad (1-\varepsilon)\mu'_C(P) \le \mu'_C(\mathcal{S}) \le (1+\varepsilon)\mu'_C(P).$$

Note that k here differs from the number of facilities (but at most by a factor of 2).

2.3 The construction of (k,ε)-V-coresets

Let P be a set of n points on the x-axis, $k \ge 1$ an integer, and $0 < \varepsilon \le 1$. The algorithm V-CORESET returns a weighted set $\mathcal{S}$ of size $|\mathcal{S}| = O\left(\log^{2k-1} n\right)$, that is a (k,ε)-V-coreset for P.

Without loss of generality we may assume that $|P| > \lceil \delta/\varepsilon \rceil$. Otherwise, we take P itself as the coreset; see Line 2 of V-CORESET. The algorithm is recursive and makes use of $(k-1,\varepsilon)$-V-coresets for various subsets of P, where the base case for the recursion is the case $k = 1$. In this case (Line 3) the weight of the single facility is irrelevant for the property that we seek. Thus, a 3ε-coreset for P, as constructed above, is also a $(1,3\varepsilon)$-V-coreset for P. For $k > 1$ the algorithm is given below.

Algorithm V-CORESET (P,k,ε)
Input: P : set of n points on a line, $k \ge 1$ an integer, and $0 < \varepsilon \le 1$.
Output: Weighted-facilities $(k, 3\varepsilon)$-V-Coreset for P.

```
1   if $|P| \le \lceil \delta/\varepsilon \rceil$ (* $\delta$ is a constant that defined in the
    full version of this paper *)
2     then return $P$
3   if $k = 1$
4     then return SINGLE-FACILITY-CORESET ($P$, $3\varepsilon$)
5   $p_1 \leftarrow$ leftmost point of $P$,
        $p_{\lfloor n/2 \rfloor} \leftarrow \lfloor n/2 \rfloor$-leftmost point of $P$
6   $end \leftarrow p_{\lfloor n/2 \rfloor}$
7   for $i \leftarrow 1$ to $\lceil \log n \rceil$
8     do $begin \leftarrow end - 2^{i-1}\left|p_{\lfloor n/2 \rfloor} - p_1\right|/n^2$
9        $B_i \leftarrow P \cap (begin, end]$
10       $end \leftarrow begin$
11    $\mathcal{Z} \leftarrow \emptyset$    (* $\mathcal{Z}$ is a collection of sets *)
12    for $i \leftarrow 1$ to $\lceil \log n \rceil$
13      do $B_{i1} \leftarrow \emptyset$;   $size \leftarrow 1$;   $j \leftarrow 1$;
14        for $m \leftarrow 1$ to $|B_i|$
15          do if i=1
16                then add to $B_{ij}$ the $m$th leftmost point
                   of $B_i$
18                 else add to $B_{ij}$ the $m$th rightmost point
                   of $B_i$
19              if $|B_{ij}| = size$
20                then $\mathcal{Z} \leftarrow \mathcal{Z} \cup \{B_{ij}\}$
21                     $j \leftarrow j + 1$
22                     $B_{ij} \leftarrow \emptyset$
23                     if $(j \mod \lceil \delta/\varepsilon \rceil) = 0$
23                       then $size \leftarrow 2size$
24    $\mathcal{S}_\ell \leftarrow \emptyset$
25    for each $B \in \mathcal{Z}$
26      do $\mathcal{S}_\ell \leftarrow \mathcal{S}_\ell \cup \left(\text{V-CORESET}(B, k-1, \varepsilon)\right)$
27    Repeat lines 5–26 for the $\lceil n/2 \rceil$ rightmost
        points of $P$, resulting in a set $\mathcal{S}_r$
      (* Use a mirror-image construction *)
28    return $\mathcal{S}_\ell \cup \mathcal{S}_r$
```

Lemma 2.6. *The number of sets $B_{ij} \in \mathcal{Z}$ is $O(\varepsilon^{-1}\log^2 n)$.*

Proof. No $B_{ij} \in \mathcal{Z}$ can have more than $\lceil 2\varepsilon n/\delta \rceil$ points. Indeed, for $j \le \lfloor \delta/\varepsilon \rfloor$, $|B_{ij}| = 1$ by construction, and for $j > \lfloor \delta/\varepsilon \rfloor$, the construction implies that each of the $\lfloor \delta/\varepsilon \rfloor$ sets $B_{ij'}$ that precedes B_{ij} satisfies $|B_{i,j'}| \ge {}^1\!/\!{}_2\,|B_{ij}| \ge \lceil \varepsilon n/\delta \rceil$, and this would imply that there are more than n points overall in P. The size of the largest subset of each B_i is thus at most $\lceil 2\varepsilon n/\delta \rceil$, which is easily seen to imply that the number of subsets of B_i is $O(\varepsilon^{-1}\log n)$. Since there are $O(\log n)$ sets B_i, it follows that $|\mathcal{Z}| = O(\varepsilon^{-1}\log^2 n)$. □

Lemma 2.7. *The number of points in the set $\mathcal{S}$ is at most $2^{O(k)}\varepsilon^{-k}\log^{2k-1} n$.*

Proof. The proof is by induction on k. For the general induction step, we only consider $\mathcal{S}_\ell$; the proof is similar for $\mathcal{S}_r$. Define $T(k,\varepsilon)$ to be the maximum size of $\mathcal{S}_\ell$ for a given k and ε. From the construction for $k = 1$, it follows that $T(1,\varepsilon) = O({}^1\!/\!{}_\varepsilon \log n)$, which satisfies the bound. Therefore, by Lemma 2.6, for an appropriate absolute constant b we have

$$T(k,\varepsilon) = |\mathcal{Z}| \cdot T(k-1,\varepsilon) \le b^k \varepsilon^{-k} \log^{2k-1} n. \quad \square$$

To prove that $\mathcal{S}$ is a (k,ε)-V-coreset, we will frequently use the following simple observation. We denote by $\mathcal{I}(X)$ the smallest interval containing a set X.

Observation 2.8. **(i)** *The size of the interval $\mathcal{I}(B_i)$ for $i > 1$ is less than twice the size of all the intervals to its right, i.e (cf. Fig. 3(left)), $|\mathcal{I}(B_i)| \le 2\sum_{m<i}|\mathcal{I}(B_m)|$.*
(ii) *For any $B_{ij} \in \mathcal{Z}$ that contains at least two points, we have (cf. Fig. 3(right)) $|B_{ij}| \le 2\varepsilon/\delta) \sum_{m<j}|B_{im}|$.*

Theorem 2.9. *Let P be a set of points on a line, $k \ge 1$ an integer, and $0 < \varepsilon \le 1$. The algorithm* V-CORESET(P,k,ε) *returns, in $O(kn)$ time, a $(k,3\varepsilon)$-V-coreset $\mathcal{S}$ for P of size $2^{O(k)}\varepsilon^{-k}\log^{2k-1} n$.*

Proof. The bound on the size of $\mathcal{S}$ is given in Lemma 2.7. Each run of V-CORESET, excluding the

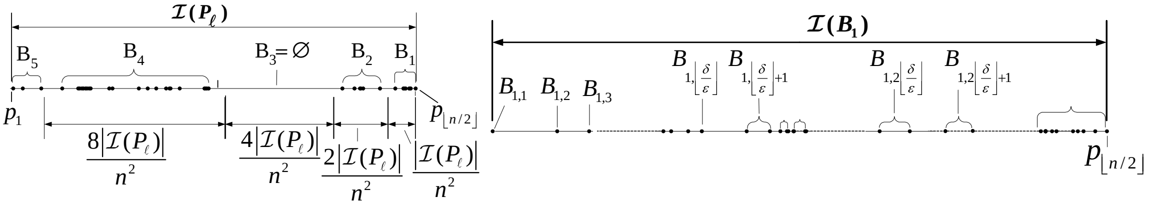

$\mathcal{I}(P_\ell)$
B_5
B_4
$B_3 = \emptyset$
B_2
B_1
p_1
$p_{\lfloor n/2 \rfloor}$
$\frac{8|\mathcal{I}(P_\ell)|}{n^2}$
$\frac{4|\mathcal{I}(P_\ell)|}{n^2}$
$\frac{2|\mathcal{I}(P_\ell)|}{n^2}$
$\frac{|\mathcal{I}(P_\ell)|}{n^2}$
$\mathcal{I}(B_1)$
$B_{1,1}$
$B_{1,2}$
$B_{1,3}$
$B_{1,\lfloor \frac{\delta}{\varepsilon} \rfloor}$
$B_{1,\lfloor \frac{\delta}{\varepsilon} \rfloor+1}$
$B_{1,2\lfloor \frac{\delta}{\varepsilon} \rfloor}$
$B_{1,2\lfloor \frac{\delta}{\varepsilon} \rfloor+1}$
$p_{\lfloor n/2 \rfloor}$

Fig. 3: **(left)** The high-level partition of the $\lfloor n/2 \rfloor$ set P_ℓ into intervals and sets. **(right)** Partition of B_1 into subsets. The other subsets B_i of $\mathcal{Z}$, for $i > 1$, are similarly partitioned, but from right to left rather than from left to right.

recursive call, can be implemented in $O(n)$ time, for $O(kn)$ time overall. Lines 5–10 of V-CORESET can be easily implemented, in $O(n)$ time (without sorting P) using the log and floor functions, similarly to the implementation of SINGLE-FACILITY-CORESET. Lines 11–23 can be implemented in $O(n)$ time by finding the points in the $\lfloor \delta/\varepsilon \rfloor$ last (largest) subsets $B_{ij} \subseteq B_i$ in $O(|B_i|)$ time, using a linear-time algorithm for order statistics, and then by continuing recursively on the remaining points.

Interestingly enough, the following proof of correctness for non-squared distances remains true for squared distances, if we use everywhere the cost function μ' instead of ν', and replace $\|\cdot\|$ by $\|\cdot\|^2$. For the case $k = 1$ the weight of the single facility is irrelevant for the property that we seek. Thus, by Corollary 2.2, the 3ε-coreset that is returned by SINGLE-FACILITY-CORESET is also a $(1, 3\varepsilon)$-V-coreset for P; see Line 4. It is left to prove the case $k > 1$.

Let $C \subset \mathbb{R}^d$ be any weighted set of facilities, such that P falls into no more than k adjacent Voronoi intervals of C. We denote by P_ℓ the set of $\lfloor n/2 \rfloor$ leftmost points of P. By line 2.3, $\mathcal{S} = \mathcal{S}_\ell \cup \mathcal{S}_r$, where $\mathcal{S}_\ell$ is the coreset for P_ℓ and $\mathcal{S}_r$ is the coreset for $P_r = P \setminus P_\ell$. We have

$$\begin{aligned} &|\nu'_C(P) - \nu'_C(\mathcal{S})| = \\ &\left|\big(\nu'_C(P_\ell) + \nu'_C(P_r)\big) - \big(\nu'_C(\mathcal{S}_\ell) + \nu'_C(\mathcal{S}_r)\big)\right| \quad (2.2)\\ &\leq |\nu'_C(P_\ell) - \nu'_C(\mathcal{S}_\ell)| + |\nu'_C(P_r) - \nu'_C(\mathcal{S}_r)|\,. \end{aligned}$$

We will prove that $|\nu'_C(P_\ell) - \nu'_C(\mathcal{S}_\ell)| \leq (3/2)\varepsilon\nu'_C(P)$. A symmetric proof will then imply $|\nu'_C(P_r) - \nu'_C(\mathcal{S}_r)| \leq (3/2)\varepsilon\nu'_C(P)$. The coreset property of $\mathcal{S}$ then follows from (2.2). If it so happens that for every $B \in \mathcal{Z}$, the interval $\mathcal{I}(B)$ intersects no more than $k-1$ Voronoi intervals, we are done, because then, by the recursive construction,

$$|\nu'_C(B) - \nu'_C(\mathcal{S}_B)| \leq \varepsilon\nu'_C(B)$$

for each $B \in \mathcal{Z}$. The coreset $\mathcal{S}_\ell$ is the union of the coresets for all $B \in \mathcal{Z}$, and thus

$$\begin{aligned} |\nu'_C(P_\ell) - \nu'_C(\mathcal{S}_\ell)| &= \left|\sum_{B\in\mathcal{Z}} \big(\nu'_C(B) - \nu'_C(\mathcal{S}_B)\big)\right| \\ \leq \sum_{B\in\mathcal{Z}} |\nu'_C(B) - \nu'_C(\mathcal{S}_B)| &\leq \sum_{B\in\mathcal{Z}} \varepsilon\nu'_C(B) \\ = \varepsilon\nu'_C(P_\ell) &\leq \varepsilon\nu'_C(P) < (3/2)\varepsilon\nu'_C(P). \end{aligned}$$

We are left to handle the case where there is some set $B \in \mathcal{Z}$ such that $\mathcal{I}(B)$ intersects all k Voronoi intervals (and thus contains $k-1$ Voronoi boundaries — see Fig. 4). In this case the sum of errors contributed by the rest of the $(k-1, \varepsilon)$ V-coresets is then

$$\begin{aligned} \sum_{X\in\mathcal{Z}\setminus\{B\}} |\nu'_C(X) - \nu'_C(\mathcal{S}_X)| &\leq \sum_{X\in\mathcal{Z}\setminus\{B\}} \varepsilon\nu'_C(X) \\ \leq \sum_{X\in\mathcal{Z}} \varepsilon\nu'_C(X) &= \varepsilon\nu'_C(P_\ell) \leq \varepsilon\nu'_C(P). \end{aligned}$$

We will show that in this case

$$|\nu'_C(B) - \nu'_C(\mathcal{S}_B)| \leq \frac{\varepsilon}{2}\nu'_C(P), \quad \text{and thus} \quad (2.3)$$

$$|\nu'_C(P_\ell) - \nu'_C(\mathcal{S}_\ell)| \leq \varepsilon\nu'_C(P) + \frac{\varepsilon}{2}\nu'_C(P) = \frac{3\varepsilon}{2}\nu'_C(P).$$

It is left to prove (2.3). Let c be any facility in C. For simplicity, we abuse notation, and write $\nu'_c(P)$ instead of $\nu'_{\{c\}}(P)$. By construction, $\mathcal{S}_B$ is a $(k-1, \varepsilon)$-V-coreset, so, by definition, $\mathcal{S}_B$ is also a $(1, \varepsilon)$-V-Coreset. Hence, $|\nu'_c(\mathcal{S}_B) - \nu'_c(B)| \leq \varepsilon\nu'_c(B) \leq \nu'_c(B)$, so, $\nu'_c(\mathcal{S}_B) \leq 2\nu'_c(B)$. Thus, for any facility $c \in C$, the left hand side of (2.3) can be bounded by

$$\begin{aligned} |\nu'_C(B) - \nu'_C(\mathcal{S}_B)| &\leq \nu'_C(B) + \nu'_C(\mathcal{S}_B) \quad (2.4)\\ &\leq \nu'_c(B) + 2\nu'_c(B) = 3\nu'_c(B). \end{aligned}$$

Let the facility c' be the projection of c on the x-axis, with weight $W(c') = W(c)$. See Fig. 4. Using the triangle inequality, $\nu'_c(B)$ can be bounded by

$$\nu'_c(B) = W(c) \sum_{p\in B} \|p - c\| \quad (2.5)$$

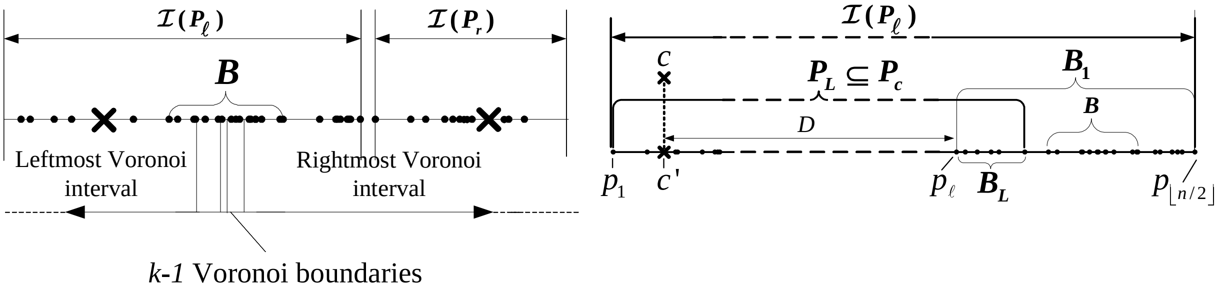


Fig. 4: **(left)** All the k Voronoi intervals fall into $\mathcal{I}(B)$ for some $B \in \mathcal{Z}$. The two 'x' facilities in this figure serve the leftmost and rightmost Voronoi intervals. **(right)** B intersects k Voronoi intervals, and is also contained in B_1. The facility $c \in C$ serves the leftmost Voronoi interval, and c' denotes its projection on the line. Since c' can be anywhere on the line, its nearest point in B_1 can be any point of B_1.

$$\leq W(c) \sum_{p \in B} (\|c - c'\| + \|p - c'\|)$$

$$= W(c)\,|B| \cdot \|c - c'\| + \nu'_{c'}(B).$$

We now bound each of the two terms in the right hand side of (2.5) by $(\varepsilon/12)\nu'_C(P)$, which, using (2.4), will prove (2.3) and conclude the proof of this theorem. Let B_i be the set that contains $B = B_{ij}$. We distinguish between the two following cases. ***(i)*** $B_i = B_1$: Let $c \in C$ be the facility that serves the leftmost Voronoi interval, and denote by P_c the points of P that are served by c. Also, let B_L denotes the set of points of B_1 that lie to the left of B, and note that $B_L \subseteq P_c$ (see Fig. 4(right)). By Observation 2.8(*ii*) we have, $|B| \leq (\varepsilon/12)\,|B_L|$, and thus $|B| \leq (\varepsilon/12)\,|P_c|$. Clearly, c' is the nearest point on the x-axis to c, and therefore $\|c - c'\| \leq \|p - c\|$ for any $p \in P$. Hence, $|P_c| \cdot \|c - c'\| \leq \nu_c(P_c)$. Altogether we have $W(c)\,|B| \cdot \|c - c'\| \leq W(c) \cdot \dfrac{\varepsilon}{12}\,|P_c| \cdot \|c - c'\|$ $\leq \dfrac{\varepsilon}{12}\nu'_c(P_c) \leq \dfrac{\varepsilon}{12}\nu'_C(P)$.

To bound the second term of 2.5, let P_L denote the points of P to the left of B, and note that $P_L \subseteq P_c$; see Fig. 4(right). Clearly, $\|p - c'\| \leq \|p - c\|$ for every p on the x-axis, and we get $\nu_{c'}(P_L) \leq \nu_c(P_L) \leq \nu_c(P_c)$. Using Lemma 2.10(*i*) below, we have $\nu_{c'}(B) \leq (\varepsilon/12)\nu_{c'}(P_L)$ (note that $|B| > 1$, since a single point cannot intersect $k > 1$ Voronoi intervals). After multiplying by $W(c)$, this yields $\nu'_{c'}(B) \leq (\varepsilon/12)\nu'_c(P_c)$.

(ii) $B \subseteq B_i \neq B_1$: The proof is symmetric, taking c to be the facility that serves the *rightmost* Voronoi interval. The sets B_R, P_R and Lemma 2.10(*ii*), should then replace B_L, P_L and Lemma 2.10(*i*), respectively. □

To conclude the proof of Theorem 2.9, we still need to show that $\nu'_c(B) \leq (\varepsilon/12)\nu'_c(P_L)$ (for $B \subseteq B_1$) or $\nu'_c(B) \leq (\varepsilon/12)\nu'_c(P_R)$ (for $B \subseteq B_i \neq B_1$), for a facility c' on the x-axis, and the same for squared distances. This is proven in the full version as stated in the following Lemma.

Lemma 2.10. *Let $P \subset \mathbb{R}$ be a set of points. Let $\mathcal{Z} = \{B_{ij}\}$ be the partition of P given in Lines 11–23 of the algorithm* V-CORESET, *for the specified $0 < \varepsilon \leq 1$ and k. Consider a set $B = B_{ij} \in \mathcal{Z}$, where $|B| > 1$ (i.e., $j > \lfloor \delta/\varepsilon \rfloor$, see Fig. 3(right)), and let P_L, P_R denotes the set of points of P that lie to the left and right of B, respectively (see Fig. 4(right)). Then, for any facility $c' \in \mathbb{R}$, we have (i) for $i = 1$, $\nu_{c'}(B) \leq (\varepsilon/12)\nu'_c(P_L)$, and $\mu_{c'}(B) \leq (\varepsilon/12)\mu'_c(P_L)$; (ii) for $i > 1$, $\nu_{c'}(B) \leq (\varepsilon/12)\nu'_c(P_R)$, and $\mu_{c'}(B) \leq (\varepsilon/12)\mu'_c(P_R)$.*

Lemma 2.5 implies the following trivial modification of Theorem 2.9.

Theorem 2.11. *Let P be a set of n points on a line, $k \geq 1$ an integer, and $\varepsilon > 0$. The algorithm* V-CORESET$(P, 2k-1, \varepsilon/3)$ *returns, in $O(nk)$ time, a weighted-facilities (k, ε)-coreset for P, of size $|\mathcal{S}| = 2^{O(k)}\varepsilon^{-2k-1}\log^{4k-3} n$.*

3 Coresets for $P \subseteq \mathbb{R}^d$

So far we have constructed (k, ε)-coresets for a set of points on a fixed line. In this section we use these coresets to construct the following coreset for a set of points in $\mathbb{R}^d$.

(k, j, ε)-Coreset. Let P be a set of n points in $\mathbb{R}^d$, $k \geq 1$ and $1 \leq j \leq d-1$ integers. A weighted set $\mathcal{S}$, where $\mathcal{S} \subset P$ is called a (k, j, ε)-*coreset* for P, if for any set L of $0 \leq k' \leq k$ lines and at most $k - k'$ points we have

$$(i) \qquad (1-\varepsilon)\nu_L(P) \leq \nu_L(\mathcal{S}) \leq (1+\varepsilon)\nu_L(P),$$

and if, for any flat f of dimension at most j, we have

$$(ii) \qquad (1-\varepsilon)\nu_{\{f\}}(P) \leq \nu_{\{f\}}(\mathcal{S}) \leq (1+\varepsilon)\nu_{\{f\}}(P).$$

These properties also holds for squared distances, or regression distances (squared or non-squared).

Our construction of this coreset crucially relies on a randomized bicriteria constant-factor approximation algorithm BICRITERIA-APPROXIMATION(P, k, j), which is described in a companion paper [13]. It receives as input a point set $P \subset \mathbb{R}^d$, and integers k, j, such that $k \geq 1$ and $1 \leq j \leq d - 1$. It outputs a set $F = \{f_1, f_2, \ldots, f_m\}$ of $m = 2^{O(j)}(kj)^{j+1}\log^{j+2} n$ j-dimensional flats and a partition $\Pi = \{P_1, P_2, \ldots, P_m\}$ of P, such that, with probability $1/2$, for any set Y of at most k flats, all of dimension no greater than j, $\sum_{i=1}^{m} \nu_{f_i}(P_i) \leq 2^{j+2} \cdot \nu_Y(P)$ and $\sum_{i=1}^{m} \mu_{f_i}(P_i) \leq 2^{j+2} \cdot \mu_Y(P)$. The algorithm takes $dn(kj\log n)^{j+1}2^{O(j)}$ time; see [13] for details.

In the following algorithm LINEAR-FACILITIES-CORESET, we denote by $\mathrm{project}(q, X)$ the projection of a point q on a set of points X (i.e., $\mathrm{project}(q, X)$ is the nearest point to q among all points of X). For a set Q, we define $\mathrm{project}(Q, X) = \bigcup_{q \in Q} \mathrm{project}(q, X)$.

Algorithm LINEAR-FACILITIES-CORESET(P, k, j, ε)

Input: A set of points $P \subset \mathbb{R}^d$, k, j, and $0 < \varepsilon \leq 1$, where $k \geq 1$ and $1 \leq j \leq d-1$ are integers.
Output: A set $\mathcal{S} \subseteq P$ such that, with probability at least $1/2$, $\mathcal{S}$ is $(k, j, 5\varepsilon)$-coreset for P.

```
$(F, \Pi)$ ←BI-CRITERIA-APPROXIMATION$(P, k, j)$
$\mathcal{S} \leftarrow \emptyset$
for $i \leftarrow 1$ to $|F|$
do $f_i \leftarrow$ the $i$th $j$-dimensional flat in $F$.
   $P_i \leftarrow$ the $i$th set of points in $\Pi$.
   $f_i^{\perp} \leftarrow$ a $(d-j)$-dimensional flat that is
         orthogonal to $f_i$.     (* See Fig. 5(up) *)
   $P_i^* \leftarrow \mathrm{project}(P_i, f_i^{\perp})$.
   for each $p^* \in P_i^*$ do $w(p^*) \leftarrow |P| / |P_i|$
   $\mathcal{S}_i$ ←SINGLE-FACILITY-CORESET$(P_i^*, \varepsilon)$
   for each $p' \in \mathcal{S}_i$
   do $f \leftarrow$ the $j$-dimensional flat that passes
            through $p'$, and parallel to $f_i$.
            (* See Fig. 5(down) *)
      $P_f \leftarrow$ the set of those $p \in P_i$, such that $p'$ is
            the representative of
            $p^* = \mathrm{project}(p, f_i^{\perp})$ in $\mathcal{S}_i$. (* See
            Line 9 of SINGLE-FACILITY-
            CORESET *)
      $\tilde{P}_f \leftarrow \mathrm{project}(P_f, f)$
      if $j = 1$
         then $\tilde{\mathcal{S}}_f \leftarrow$ V-CORESET$(\tilde{P}_f, 2k-1, \varepsilon)$
         else $\tilde{\mathcal{S}}_f$ ←LINEAR-FACILITIES
                  -CORESET $(\tilde{P}_f, k, j-1, \varepsilon)$
      $\mathcal{S} \leftarrow \mathcal{S} \cup \{p \in P_f \mid \mathrm{project}(p, f) \in \tilde{\mathcal{S}}_f\}$
      (* each point in $\mathcal{S}$ is also assigned the weight
         of its correspondence point in $\tilde{\mathcal{S}}_f$ *)
return $\mathcal{S}$
```

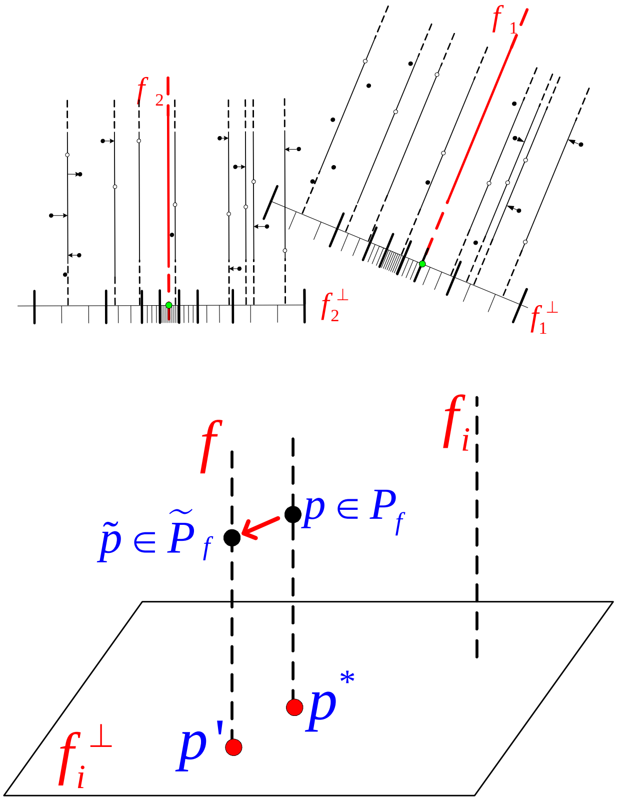


Fig. 5: **(up)** For $j = 1$ and $P \subset \mathbb{R}^2$, each $f_i \in F$ is a line, as its orthogonal $f_i^{\perp}$. **(down)** For $j = 1$ and $P \subset \mathbb{R}^3$, each $f_i \in F$ is a line, and its orthogonal $f_i^{\perp}$ is a plane.

Omitting all further details, the main result of the paper is:

Theorem 3.1. *Let P be a set of n points in $\mathbb{R}^d$, where $d \geq 1$ is constant, $1 \leq j \leq d-1$, and $0 < \varepsilon \leq 1$. Also, let $k \geq 1$ be a constant. Then, for sufficiently large constant b, the algorithm* LINEAR-FACILITIES-CORESET$(k, \varepsilon/b, j)$ *computes, in $O(n)$ time, with probability at least 1/2, a (k, j, ε)-coreset for P of size $\varepsilon^{-d-k}(\log n)^{O(j^2)+2k-1}$. The running time is $O(n)$.*

Acknowledgements. We would like to thank Sariel Har-Peled for helpful discussions and comments concerning the topics and presentation of this paper.

References

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

[2] P. K. Agarwal, S. Har-Peled and K. R. Varadarajan, Geometric approximation via coresets, in *Current Trends in Combinatorial and Computational Geometry*, (J.E. Goodman, J. Pach and E. Welzl, eds.), Cambridge University Press, New York, 2006, pp. 1–30.

[3] V. Boltyanski, H. Martini and V. Soltan, *Geometric Methods and Optimization Problems*, Kluwer Academic Publishers, The Netherlands, 1999.

[4] T.M. Breuel, Finding lines under bounded error, *Pattern Recognition* 29 (1996), 167–178.

[5] M. Charikar, S. Khuller, D. M. Mount and G. Narasimhan, Algorithms for facility location problems with outliers, *Proc. 12th Annu. ACM-SIAM Sympos. Discrete Algorithms*, 2001, 642–651.

[6] M. Charikar, L. O'Callaghan and R. Panigrahy, Better streaming algorithms for clustering problems, *Proc. 35th Annu. ACM Sympos. Theory Comput.*, 2003, 30–39.

[7] K. L. Clarkson, Subgradient and sampling algorithms for L_1-regression, *Proc. 16th Annu. ACM-SIAM Sympos. Discrete algorithms*, 2005, 257–266.

[8] T. K. Dey, Improved bounds for planar k-sets and related problems, *Discrete Comput. Geom.* 19 (1998), 373–382.

[9] A. Deshpande, L. Rademacher, S. Vempala and G. Wang, Matrix approximation and projective clustering via volume sampling, *Proc. 17th Annu. ACM-SIAM Sympos. Discrete Algorithms*, 2006, 1117–1126.

[10] Z. Drezner and H. W. Hamacher, (eds.), *Facility Location: Applications and Theory*, Springer Verlag, Heidelberg, 2002, pp. 20–22.

[11] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, Wiley-Interscience, 2000, p. 115.

[12] D. Feldman, *Algorithms for Fitting Points by k Lines*, M.Sc. Thesis, School of Computer Science, Tel-Aviv university, 2004

[13] D. Feldman, A. Fiat and M. Sharir, Bicriteria approximations for k-line mean and k-line median, Manuscript, 2006; `www.cs.tau.ac.il/~dannyf`

[14] D. Feldman, A. Fiat, and M.Sharir, PTAS Algorithms for Approximating Points by Flats, Manuscript, 2006.

[15] S. Har-Peled, Coresets for discrete integration and clustering, Manuscript, 2006.

[16] S. Har-Peled and S. Mazumdar, On coresets for k-means and k-median clustering, *Proc. 36th Annu. ACM Sympos. Theory computing*, 2004, pp. 291–300.

[17] A. Hyvärinen, E. Oja and J. Karhunen, *Independent Component Analysis*, Wiley-Interscience, 2001.

[18] M. Kirby and L. Sirovich, Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces., *IEEE Trans. Pattern Anal. Mach. Intell.*, 1990, 103–108

[19] S. Kolenikov and G. Angeles, *The Use of Discrete Data in PCA: Theory, Simulations, and Applications to Socioeconomic Indices*, Carolina PopulationCenter, University of North Carolina, Chapel Hill, 2004

[20] N. Meggido and A. Tamir, Finding least-distance lines, *SIAM J. Algebric Discrete Methods* 4 (1983), 207–211.

[21] J.A. Richards, *Remote Sensing Digital Image Analysis*, Springer-Verlag, 1986

[22] A. Schöbel, *Locating Lines and Hyperplanes: Theory and Algorithms*, Springer-Verlag, New York, 1999.

[23] M. Sharir and P. K. Agarwal, *Davenport-Schinzel Sequences and Their Geometric Applications*, Cambridge University Press, New York, 1995.

[24] J.S. Taur and C.W. Tao, Medical Image Compression using Principal Component Anlyasis, *International Conference on Image Processing, IEEE International Conference on Image Processing (ICIP'96)*, 1996, 903–906

[25] P. Yamamoto, K. Kato, K. Imai and H. Imai, Algorithms for vertical and orthogonal L_1-linear approximation of points, *Proc. 4th Annu. ACM Sympos. Comput. Geom.*, 1988, 352–361.