

Coresets for Discrete Integration and Clustering*

Sariel Har-Peled[†]

October 5, 2006

“The problem received the title of ‘Buridan’s sheep.’ The biological code was taken from a young merino sheep, by the Casparo-Karpov method, at a moment when the sheep was between two feeding troughs full of mixed fodder. This code, along with additional data about sheep in general, was fed into CODD. The machine was required: a) to predict which trough the merino would choose, and b) to give the psychophysiological basis for this choice.”

– The mystery of the hind leg, Arkady and Boris Strugatsky

Abstract

Given a set P of n points on the real line and a (potentially infinite) family of functions, we investigate the problem of finding a small (weighted) subset $\mathcal{S} \subseteq P$, such that for any $f \in \mathcal{F}$, we have that $f(P)$ is a $(1 \pm \varepsilon)$ -approximation to $f(\mathcal{S})$. Here, $f(Q) = \sum_{q \in Q} w(q)f(q)$ denotes the weighted discrete integral of f over the point set Q , where $w(q)$ is the weight assigned to the point q .

We study this problem, and provide tight bounds on the size \mathcal{S} for several families of functions. As an application, we present some coreset constructions for clustering.

1 Introduction

Motivated by recent work on clustering, we investigate the following natural problem.

Problem 1.1 Let P be a set of points on the real line, and let \mathcal{F} be a (potentially infinite) family of functions. A ε -coreset for P is a weighted subset $\mathcal{S} \subseteq P$, such that

$$\forall f \in \mathcal{F}, \quad f(P) \approx_{\varepsilon} f(\mathcal{S}),$$

where $f(P) = \sum_{p \in P} f(p)$, $f(\mathcal{S}) = \sum_{p \in \mathcal{S}} f(p)w(p)$ and $w(p)$ is the associated weight of p in \mathcal{S} . (The notation $x \approx_{\varepsilon} y$ denotes the fact that $(1-\varepsilon)x \leq y \leq (1+\varepsilon)x$ and $(1-\varepsilon)y \leq x \leq (1+\varepsilon)y$.)

The problem is to find the smallest ε -coreset for P and \mathcal{F} . Note that such a coreset always exists as we can just take $\mathcal{S} = P$.

*The latest version of this paper is available online [Har06a].

[†]Department of Computer Science; University of Illinois; 201 N. Goodwin Avenue; Urbana, IL, 61801, USA; sariel@uiuc.edu; <http://www.uiuc.edu/~sariel/>. Work on this paper was partially supported by a NSF CAREER award CCR-0132901.

If P is uniformly distributed on an interval on the real line, this is (very similar to) the standard problem of numerical integration on the real line studied in numerical-analysis. However, as our investigation demonstrates, this is fundamentally a different problem once the point set is not uniform. Similarly, there seems to be some indirect connection to discrepancy [Mat99]. However, the author is unaware of any direct previous work on this problem.

To see how this problem naturally arises from clustering, consider the problem of computing *k-median clustering* (say, in the plane). For any set of centers C in the plane, every point of P is assigned the cost of being clustered using C ; namely, the k -median clustering cost of P by C is the sum of distances of the points of P to their closest neighbor in C . Let $f_C(\cdot)$ the cost function induced by C , and let \mathcal{F}_{kwc} be the set of all such functions. If one can find a small ε -coreset for P (for \mathcal{F}_{kwc}) then one can compute a good clustering of P directly on the (considerably smaller) coreset. Har-Peled and Kushal [HK05] showed that this problem, for a point set in \mathbb{R}^d , can be reduced to (a variant) of this one dimensional problem.

Coresets for k -median clustering are now relatively well understood, see [HM04, HK05, Che06]. See [AHV05] for more details on the usage of coresets for clustering. However, if we are interested in handling more general clustering problems, like the centers being lines instead of points, we need to better understand the aforementioned more general problem. See the work by Feldman *et al.* [FFS06] for preliminary results on this problem for the line clustering problem. For more information about clustering, see [ADPR00, AP00, BC03, Epp98, FG88, Gon85, Ind99, MOP01, OR00, KMN⁺04, IKI94]. In particular, our work yields better coreset constructions for k -line median clustering (see Theorem 6.4) than what was known before [FFS06].

Our approach is to systematically classify which families of functions have coresets and of what sizes, starting from (the trivial) family of linear functions and ending in the clustering functions mentioned above. The basic approach is quite natural, and has long history: We will partition the points into groups, and from each group pick one representative point (with weight equal to the group size). This is a classical technique used in computing estimates to summations and integrals (for example, bounding $\sum_{i=1}^n 1/i$ by partitioning the range $1, \dots, n$ into the blocks $2^i, \dots, 2^{i+1} - 1$, for $i = 1, \dots, \lfloor \lg n \rfloor$). What makes our study (maybe) interesting, is that our partitions do not work just for a single function but for a family of functions, and require (especially towards the end) delicate and not completely trivial constructions. In particular, this gives rise to new partition schemes of point sets on the line (i.e., safe and secure partitions) which might be of independent interest.

The paper is organized as follows. In Section 2 we define some preliminary definitions. In Section 3, we study the problem for some simple families of functions. In Section 4, we introduce some partition schemes of point sets that are useful in constructing coresets for clustering functions. In Section 5, we use these partition scheme to construct coreset for the weighted k -median clustering problem on the line, and in Section 6, we extend this to handle the k -median function induced by k lines. We conclude in Section 7 with conclusions and some open problems.

2 Preliminaries

Two non-negative numbers x and y are $(1 \pm \varepsilon)$ -approximation of each other if $(1 - \varepsilon)x \leq y \leq (1 + \varepsilon)x$ and $(1 - \varepsilon)y \leq x \leq (1 + \varepsilon)y$. We denote this fact by $x \approx_{\varepsilon} y$.

Observation 2.1 *Let x and y be two positive numbers and $\varepsilon < 1/4$. We have: (i) If $x \approx_{\varepsilon} y$ and $y \approx_{\varepsilon} z$ then $x \approx_{3\varepsilon} z$. (ii) If $|x - y| \leq \varepsilon x$ then $x \approx_{2\varepsilon} y$. (iii) If $x \leq (1 + \varepsilon)y$ and $y \leq (1 + \varepsilon)x$ then $x \approx_{\varepsilon} y$.*

Two non-negative functions $f(\cdot)$ and $g(\cdot)$ are $(1 \pm \varepsilon)$ -approximation of each other, denoted by $f \approx_{\varepsilon} g$, if $f(x) \approx_{\varepsilon} g(x)$, for all x .

3 Basic Coresets for integration

In this section, we present coresets for several simple families of functions.

3.1 Linear functions

Let $\mathcal{F}_{\text{linear}}$ be the set of affine functions of the form $f(x) = ax + b$. Then, clearly for the set P , the centroid point of P (i.e., the average value of P) with assigned weight $|P|$ is a coreset.

Lemma 3.1 *The family $\mathcal{F}_{\text{linear}}$ of linear functions have a coreset of size 1.*

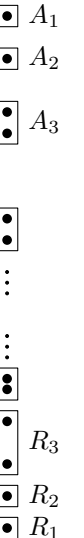
3.2 Monotone functions

Let \mathcal{F}_{dec} (resp., \mathcal{F}_{inc}) be the family of monotone decreasing non-negative functions (resp., family of monotone increasing non-negative functions) from \mathbb{R} to \mathbb{R}^+ .

Lemma 3.2 *Given a set $P \subseteq \mathbb{R}$ of n numbers, one can compute an ε -coreset of P of size $O(\varepsilon^{-1} \log n)$ for the family of functions \mathcal{F}_{dec} .*

Proof: The following construction is somewhat of an overkill, but it would be useful later for other purposes.

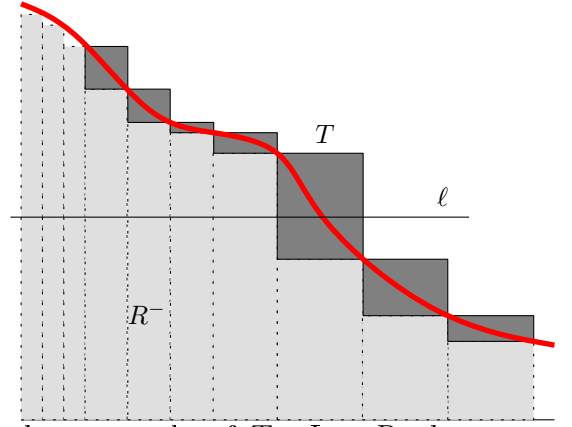
The construction follows the exponential construction used in the 1-median coreset in [Har06b]. Indeed, let z_i denote the i th point in the sorted order of the points of P . We partition P symmetrically into subsets, such that the size of the subsets increase in size as one comes toward the middle of the set. Formally, the set $A_i = \{z_i\}$ contains the i th point on the line, for $i = 1, \dots, m$, where $m \geq 10/\varepsilon$ is a parameter to be determined shortly. Similarly, $R_i = \{z_{n-i+1}\}$, for $i = 1, \dots, m$. Set $\alpha_m = m$, and let $\alpha_{i+1} = \min(\lceil (1 + \varepsilon/10)\alpha_i \rceil, n/2)$, for $i = m, \dots, m'$, where $\alpha_{m'}$ is the first number in this sequence equal to $n/2$. Now, let $A_i = \{z_{\alpha_{i-1}+1}, \dots, z_{\alpha_i}\}$ and $R_i = \{z_{n-\alpha_{i-1}}, \dots, z_{n-\alpha_i+1}\}$, for $i = m+1, \dots, m'$. See figure on the right. We will refer to a set A_i or R_i as a *chunk*. Consider the partition of P formed by the chunks $A_1, A_2, \dots, A_{m'}, B_{m'}, \dots, B_2, B_1$. To simplify the exposition, let C_1, \dots, C_M denote the resulting sequence of sets, where $M = 2m'$. This is a partition of P into “exponential sets”. The first/last m sets on the boundary are singletons, and all the other sets grow exponentially in cardinality, till they cover the whole set P .



Next, we pick an arbitrary point $l_i \in C_i$ and assign it weight $w_i = |C_i|$, for $i = 1, \dots, M$. Let \mathcal{S} be the resulting weighted set of points. We claim that this is a coresset for any function of \mathcal{F}_{dec} .

For the sake of analysis we can assume that $P = \{1, 2, \dots, n\}$. Indeed, this can be realized by stretching and translating the real-axis appropriately and observing that the resulting function is still monotone decreasing. The claim now follows by a simple integration argument.

Indeed, let $f \in \mathcal{F}_{\text{dec}}$ be an arbitrary function. For a chunk $C_i = \{j_i, \dots, j_{i+1} - 1\}$, we observe that its contribution to $f(\mathcal{S})$ can be interpreted as a bar (i.e., rectangle) in a histogram based at the interval $\mathcal{I}_i = [j_i - 1, \dots, j_{i+1} - 1]$ and having height $f(l_i)$, for $i = 1, \dots, M$. Consider the error zone formed by this rectangle, as it gets its maximum height (by setting $l_i = j_i$) and its minimum height (by setting $l_i = j_{i+1} - 1$). Clearly, this zone of error is a rectangle $r_i = \mathcal{I}_i \times [f(j_{i+1} - 1), f(j_i)]$. Let $T = r_1 \cup \dots \cup r_M$ be the resulting set formed by these of rectangles. Since $f(\cdot)$ is monotone decreasing, a horizontal line crosses the interior of at most one of the rectangle of T . Let R^- be the histogram $\bigcup_i \mathcal{I}_i \times f(j_{i+1} - 1)$, and let R^+ be the histogram $\bigcup_i \mathcal{I}_i \times f(j_i)$. Clearly, $R^- \subseteq R^+$, $R^+ = R^- \cup T$, $\text{area}(R^+) = \text{area}(R^-) + \text{area}(T)$, $\text{area}(R^-) \leq f(\mathcal{S}) \leq \text{area}(R^+)$, and $\text{area}(R^-) \leq f(P) \leq \text{area}(R^+)$.



Furthermore, by construction, we have that

$$|\mathcal{I}_i| \leq \frac{\varepsilon}{4} \sum_{j < i} |\mathcal{I}_j|,$$

for $i = m + 1, \dots, M$, where $|\mathcal{I}_i|$ denotes the length of this interval. In particular, this implies that for any horizontal line ℓ we have that $|T \cap \ell| \leq (\varepsilon/4) |R^- \cap \ell|$. Now, imagine computing the area of T by integrating the length of the intersection of a horizontal line with T . We have that $\text{area}(T) \leq (\varepsilon/4) \text{area}(R^-)$. This implies that

$$\begin{aligned} \mathcal{E} &= |f(\mathcal{S}) - f(P)| \leq \text{area}(R^+) - \text{area}(R^-) = \text{area}(T) \leq (\varepsilon/4) \text{area}(R^-) \\ &\leq (\varepsilon/4) \min(f(P), f(\mathcal{S})), \end{aligned}$$

as required. ■

Clearly, the same construction works for monotonically increasing function. In fact, the construction also works for a function which is decreasing and then increasing, as can be verified. In particular, let $\mathcal{F}_{\text{dec} \rightarrow \text{inc}}$ denote the set of non-negative functions which are first monotonically decreasing, and then they become monotonically increasing. We summarize:

Theorem 3.3 *Let P be a set of n points on the real line. One can construction a ε -coreset, of size $O(\varepsilon^{-1} \log n)$, that works for any function that belongs to $\mathcal{F}_{\text{inc}} \cup \mathcal{F}_{\text{dec}} \cup \mathcal{F}_{\text{dec} \rightarrow \text{inc}}$.*

Since non-negative convex functions are also in $\mathcal{F}_{\text{dec} \rightarrow \text{inc}}$, it follows that this also holds for convex functions.

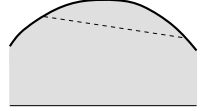
Let $\mathcal{F}_{\text{inc} \rightarrow \text{dec}}$ denote the set of non-negative functions which are monotonically increasing, and then they become monotonically decreasing afterwards.

Lemma 3.4 Any ε -coreset for $\mathcal{F}_{\text{inc} \rightarrow \text{dec}}$ for a set P of n points on the real line, must include all points of P .

Proof: Consider the function f_i that assigns 1 to the i th point p_i of P and zero everywhere else. Clearly, $f_i \in \mathcal{F}_{\text{inc} \rightarrow \text{dec}}$ and $f_i(P) = 1$, and as such, it must be that p_i must be in the coreset, for all i . ■

3.3 Concave Functions

Definition 3.5 A function $f : \mathbb{R} \rightarrow \mathbb{R}^+$ is *concave* on the interval \mathcal{I} , if for all $x, y \in \mathcal{I}$ and $\alpha \in [0, 1]$ we have that $\alpha f(x) + (1 - \alpha)f(y) \leq f(\alpha x + (1 - \alpha)y)$.



Let $\mathcal{F}_{\text{concave}}(\mathcal{I})$ denote the family of concave non-negative functions defined over the interval \mathcal{I} .

Definition 3.6 An interval \mathcal{J} is ε -oblivious for $\mathcal{F}_{\text{concave}}(\mathcal{I})$, if $\mathcal{J} \subseteq \mathcal{I}$ and for any $f \in \mathcal{F}_{\text{concave}}(\mathcal{I})$ we have that $f(x) \underset{\varepsilon/10}{\approx} f(y)$, for all $x, y \in \mathcal{J}$.

Lemma 3.7 Let $\mathcal{I} = [a, b]$, and consider $x \in [a, b]$. Then, there is an ε -oblivious interval, for $\mathcal{F}_{\text{concave}}(\mathcal{I})$, centered at x of length $(\varepsilon/40) \min((x - a), (b - x))$.

Proof: This follows by easy convexity arguments and the proof is included here only for the sake of completeness. Assume $a = 0$ and $b = 1$. Let f be any function of $\mathcal{F}_{\text{concave}}(\mathcal{I})$. By convexity, we have that

$$\frac{x}{y}f(y) \leq \frac{x}{y}f(y) + \frac{y-x}{y}f(0) \leq f\left(\frac{x}{y}y + \frac{y-x}{y}0\right) = f(x).$$

By symmetry, we have that $\frac{1-y}{1-x}f(x) \leq f(y)$. In particular, let $\alpha = \min((\varepsilon/40)x, (\varepsilon/40)(1 - x))$. If $y \in [x, x + \alpha]$ then

$$\left(1 - \frac{\varepsilon}{40}\right) f(x) \leq \frac{1-y}{1-x}f(x) \leq f(y) \leq \frac{y}{x}f(x) \leq \left(1 + \frac{\varepsilon}{40}\right) f(x).$$

Similarly, if $y \in [x - \alpha, x]$ then $\frac{y}{x}f(x) \leq f(y)$ and $\frac{1-x}{1-y}f(y) \leq f(x)$; namely,

$$\left(1 - \frac{\varepsilon}{40}\right) f(x) \leq \frac{y}{x}f(x) \leq f(y) \leq \frac{1-y}{1-x}f(y) \leq \left(1 + \frac{\varepsilon}{40}\right) f(x).$$

By Observation 2.1 (ii) it now follows that $f(x) \underset{\varepsilon/20}{\approx} f(y)$, for any $y \in [x - \alpha, x + \alpha]$, implying the claim. ■

Clearly, if \mathcal{I} is a ε -oblivious interval, then we can pick an arbitrary point of $Q = \mathcal{I} \cap P$ and assign it weight equal to $|Q|$ as a representative of Q in the resulting coreset. Lemma 3.7 now implies that $f(Q)$ would be well approximated by this single point. The problem is that one can not cover a given interval by oblivious intervals, since an infinite number of such intervals is required.

Theorem 3.8 *Let P be a set of n points on the real line, let \mathcal{I} be an interval containing P . Then one can compute a ε -coreset for $\mathcal{F}_{\text{concave}}(\mathcal{I})$ of size $O(\varepsilon^{-1} \log n)$.*

Proof: Assume $\mathcal{I} = [0, 1]$. Tile the interval $[\varepsilon/100, 1 - \varepsilon/100]$ by ε -oblivious intervals. The number of oblivious intervals required is $O(\varepsilon^{-1} \log(1/\varepsilon))$. Indeed, we start tiling from the middle, and let $x_1 = 1/2$. The i th ε -oblivious interval is $[x_{i+1}, x_i]$, where by Lemma 3.7 we can set $x_{i+1} = (1 - \varepsilon/40)x_i$. Thus, for $j > 2 \lceil (40/\varepsilon) \rceil \ln(100/\varepsilon)$ intervals, we have that $x_j < \varepsilon/100$, as can be easily verified. We use the symmetric tiling for the interval $[1/2, 1 - \varepsilon/100]$.

Compute for each such oblivious interval its coreset representative. As such, we are left with handling the margin intervals. Let $\mathcal{J} = [0, \varepsilon/100]$, and consider a function $f \in \mathcal{F}_{\text{concave}}(\mathcal{I})$.

The coreset we use for $Q = \mathcal{J} \cap P$ is the $(\varepsilon/30)$ -coreset for monotone increasing functions of Theorem 3.3. Let \mathcal{S} denote the resulting coreset for Q . We use, up to symmetry, the same coreset construction for $[1 - \varepsilon/100, 1] \cap P$. Every oblivious interval contributes one point to the coreset. Overall, we have a coreset of size $O(\varepsilon^{-1} \log n)$. We remain with the task of proving that this subset is indeed the required ε -coreset.

Let $g(x) = \max_{0 \leq y \leq x} f(y)$. Note that $g(x) - f(x)$ is maximized in \mathcal{J} , if the maximum of $f(\cdot)$ lies at point β that is inside \mathcal{J} and $x = \varepsilon/100$ is at the right endpoint of \mathcal{J} . But then, we have by convexity, that $f(x) \geq (1 - \varepsilon/100)f(\beta)$. This implies that for any $x \in \mathcal{J}$, we have that $f(x) \leq g(x) \leq f(\beta) \leq f(x)/(1 - \varepsilon/100) \leq (1 + \varepsilon/50)f(x)$.

Thus, we can use $g(x)$ instead of $f(x)$ in \mathcal{J} , introducing a small $(\varepsilon/50)$ -error in the process. Now, since the $(\varepsilon/10)$ -coreset \mathcal{S} for Q can handle monotone functions, it follows that it is a $(\varepsilon/30)$ -coreset for $f(\cdot)$ on this interval. Formally,

$$f(Q) \leq g(Q) \leq (1 + \varepsilon/30)g(\mathcal{S}) \leq (1 + \varepsilon/50)(1 + \varepsilon/30)f(\mathcal{S}) \leq (1 + \varepsilon/20)f(\mathcal{S}),$$

and similarly,

$$f(Q) \geq \frac{1}{1 + \varepsilon/50}g(Q) \geq \frac{1 - \varepsilon/30}{1 + \varepsilon/50}g(\mathcal{S}) \geq \frac{1 - \varepsilon/30}{(1 + \varepsilon/50)^2}f(\mathcal{S}) \geq (1 - \varepsilon/10)f(\mathcal{S}).$$

This implies that \mathcal{S} is an ε -coreset for Q for the function f . Thus, collecting the two coresets of the margin intervals, and all the coresets for the oblivious intervals results in a ε -coreset for P . \blacksquare

Somewhat surprisingly, the coreset constructed in Theorem 3.8 is of optimal size, up to a multiplicative constant.

Lemma 3.9 *In the worst case, any ε -coreset for $\mathcal{F}_{\text{concave}}(\mathcal{I})$ is of size $\Omega(\varepsilon^{-1} \log n)$.*

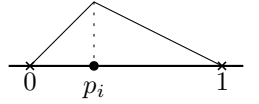
Proof: Consider the point set $P = \{p_1, \dots, p_n\}$, where $p_i = 1/(100n/\varepsilon)^{10i}$. Consider the function $f_i(x) = \min(x/p_i, 1)$, for $i = 1, \dots, n$. Clearly,

$$f_i(P) = \sum_k f_i(p_k) = \sum_{k=1}^i f_i(p_k) + \sum_{k=i+1}^n f_i(p_k) = i + \sum_{k=i+1}^n \left(\frac{\varepsilon}{100n}\right)^{10(k-i)} \leq i + \frac{\varepsilon}{50n^{10}},$$

which also implies that $i \leq f_i(P)$. Now, consider a ε -coreset $\mathcal{S} \subseteq P$. Clearly, for $f_i(\mathcal{S})$ is just a prefix sum of all the weights of the points in the range $[p_i, 1]$ (the contribution of the points of \mathcal{S} in the range $[0, p_i]$ is so small that it can be ignored, arguing as above). This prefix sum has to be in the range $\mathcal{K}(i) = [(1 - \varepsilon)i, (1 + \varepsilon)i]$. Pick a set, as large as possible, of such intervals which are disjoint and contained in the range $[1, n]$. To this end, let $i_0 = \lceil 10/\varepsilon \rceil$, and $i_k = \lceil (1 + \varepsilon)^4 i_{k-1} \rceil$, for $k = 1, \dots, m$, where $m = \lfloor (\log n) / \log((1 + \varepsilon)^4) \rfloor = \Theta(\varepsilon^{-1} \log n)$.

Let $\mathcal{K}_k = \mathcal{K}(i_k)$, for $k = 1, \dots, m$. These intervals are interior disjoint. Now, $f_{i_k}(P) = k + o(1)$, and as such $f_{i_k}(\mathcal{S}) \approx_\varepsilon f_{i_k}(P)$, implying that $f_{i_k}(\mathcal{S}) \in \mathcal{K}_k$. Now, arguing as above $f_{i_k}(\mathcal{S})$ is (up to a tiny noise) a sum of some prefix of the weights of the sorted points of the coreset. Note, that there must be one prefix sum of the coreset points with the resulting value lying inside \mathcal{K}_k , for $k = 1, \dots, m$. As such, it follows that the coreset size is at least m , as claimed. \blacksquare

Remark 3.10 Somewhat surprisingly the lower bound of Lemma 3.9 works even when the function $f_i(x)$ is a “triangle” function, namely, $f_i(x) = \min(x/p_i, (1 - x)/(1 - p_i))$. The same estimates as above hold with minor “noise” which do not effect the conclusion.



4 Partitioning schemes

Let P be a set of n points on the real line. In this section, we investigate different ways of partitioning P into subsets, such that each subset has some kind of separation property from the rest of the point set. Intuitively, all the coreset constructions so far were based on partition schemes, and we need to have better understanding of such partitions to have more general coresets.

Definition 4.1 (Partition.) A partition \mathcal{P} of a point-set $P \subseteq \mathbb{R}$ is a set of disjoint subsets S_1, \dots, S_m of P , such that $\cup_i S_i = P$ and $\mathcal{B}(S_i) \cap \mathcal{B}(S_j) = \emptyset$, for $i \neq j$, where $\mathcal{B}(S)$ denotes is the smallest interval containing the set $S \subseteq \mathbb{R}$.

In the following, for an interval \mathcal{I} and a positive real number c , we denote by $c\mathcal{I}$ the interval resulting from scaling \mathcal{I} up by a factor of c around the middle point of \mathcal{I} .

For a partition \mathcal{P} of P into disjoint subsets $\langle S_1, \dots, S_m \rangle$ and a set $I \subseteq \mathbb{R}$, let

$$\mathcal{P} \setminus I = \left\{ S_i \mid S_i \cap I = \emptyset \right\}$$

denote the family of sets of \mathcal{P} that are completely outside I . Similarly, let

$$\mathcal{P} \cap I = \left\{ S_i \mid S_i \subseteq I \right\}$$

denote the family of sets of \mathcal{P} that are completely inside I . Note that $(\mathcal{P} \cap I) \cup (\mathcal{P} \setminus I)$ is not necessarily equal to \mathcal{P} , as some sets in \mathcal{P} might intersect both I and $\mathbb{R} \setminus I$. We remind the reader that $\cup(\mathcal{P} \setminus I) = \cup_{S \in \mathcal{P} \setminus I} S$.

4.1 Safe Partitions

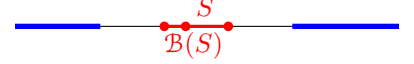


Figure 1: The interval $\mathcal{B}(S)$ and its associated two safe regions. The safe regions must contain at least $|S|/\varepsilon$ points.

Definition 4.2 (ε -safe partition.) A set $S \subseteq \mathbb{R}$ is ε -safe in relation to a set $P \subseteq \mathbb{R}$ if either $|S| = 1$ or alternatively $|P \setminus 3\mathcal{B}(S)| \geq |S|/\varepsilon$. Namely, there are a “lot” of points of P that are “faraway” from S .

As such, a partition \mathcal{P} is ε -safe if either $|S_i| = 1$ or we have that $|\cup(\mathcal{P} \setminus 3\mathcal{B}(S_i))| \geq |S_i|/\varepsilon$, for $i = 1, \dots, m$.

Lemma 4.3 For a set P of n points on the real line, there exists an ε -safe partition of size $O(\varepsilon^{-1} \log n)$. And in the worst case, any ε -safe partition of P must be of size $\Omega(\varepsilon^{-1} \log n)$.

Proof: The construction is recursive. During the recursive construction, if the point-set handled has less than $O(\varepsilon^{-1} \log n)$ points, we just add every point of it as a singleton to the partition.

Otherwise, let p_i denote the i th point of P when we sort the numbers of P from smallest to largest. Consider the interval $I = [p_\alpha, p_\beta]$, where $\alpha = \lceil ((3 - \varepsilon)/6)n \rceil$ and $\beta = \lfloor ((3 + \varepsilon)/6)n \rfloor$. Split I in the middle into two equal length intervals I_L, I_R . We next recursively compute a partition \mathcal{R} of $P \setminus I$. Let $\mathcal{Q} = \mathcal{R} \cup \{S_L\} \cup \{S_R\}$, where $S_L = I_L \cap P$ and $S_R = I_R \cap P$. We claim that \mathcal{Q} is ε -safe.

Clearly, the base of the recursion always generates a ε -safe partition. Thus, we now prove by induction that the resulting partition is safe. Indeed, assume that \mathcal{R} is safe, and observe that $|S_R|, |S_L| \leq (\varepsilon/3)n$. Furthermore, $3\mathcal{B}(S_L)$ does not contain any of the points $p_{\beta+1}, \dots, p_n$, and as such $|P \setminus \mathcal{B}(S_L)| \geq n/3 = (\varepsilon n/3)(1/\varepsilon) \geq |S_L|/\varepsilon$. Namely, S_L is ε -safe. Similarly, it follows that S_R is safe, and as such \mathcal{Q} is safe.

As for the size of the partition, we observe that $T(n) = T((1 - \varepsilon/3)n) + 2 = O(\varepsilon^{-1} \log n)$.

As for the lower bound, consider the point-set $P = \{p_1, \dots, p_n\}$, where $p_i = -4^{-i}$, for $i = 1, \dots, n$. Let $\mathcal{P} = \langle S_1, \dots, S_m \rangle$ be an ε -safe partition, sorted so that $\mathcal{B}(S_i)$ is to the left of $\mathcal{B}(S_j)$ for $i < j$. Let l_i be the left most point in S_i , and observe that if S_i contains more than one point, then all the points of P to the right of l_i are contained in $3\mathcal{B}(S_i)$. Thus, for S_i to be safe, we must have the property that

$$|S_i| \leq \varepsilon \sum_{k < i} |S_k|.$$

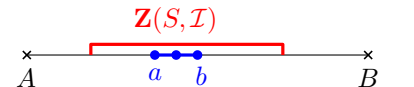
Thus, for $n_i = \sum_{k=1}^i |S_k|$, we have that $n_i \leq (1 + \varepsilon)n_{i-1} = (1 + \varepsilon)^i$. Thus, for $n_m = m$, we must have that $(1 + \varepsilon)^m \geq n$, which implies that $m = \Omega(\varepsilon^{-1} \log n)$. ■

4.2 Secure Partitions

As reality had demonstrated repeatedly, there is no real safety without security (the reader mystified by this comment, should see Remark 5.3).

Definition 4.4 (ε -secure partition.)

For an interval $\mathcal{I} = [A, B]$ and a set $S \subseteq \mathcal{I}$, the *security zone* for S , denoted by $\mathbf{Z}(S, \mathcal{I})$ is the interval $[(a + A)/2, (b + B)/2]$, where $\mathcal{B}(S) = [a, b]$.



Let P be a set of n points on the real line, and let $\mathcal{I} = [a, b]$ be a given interval that contains P . Informally, a partition $\mathcal{P} = \{S_1, \dots, S_m\}$ of P is ε -secure in \mathcal{I} , if we have that S_i is ε -safe in relation to the set $P \cap \mathbf{Z}(S_i, \mathcal{I})$, for all i . Formally, let $\mathcal{P}_i = \mathcal{P} \cap \mathbf{Z}(S_i, \mathcal{I})$ be the partition inside $\mathbf{Z}(S_i, \mathcal{I})$ induced by \mathcal{P} , and we require that S_i is ε -safe in relation to the partition \mathcal{P}_i . See Figure 2 for an alternative equivalent definition.

Intuitively, \mathcal{P} is ε -secure if every subset in the partition has enough sets with sufficient total mass in the partition “faraway” from it, that are not too close to the endpoints of the host interval \mathcal{I} .

Lemma 4.5 *Let P be a set of n points on the real line contained inside the interval \mathcal{I} . Then, there exists a ε -secure partition of P in relation to \mathcal{I} of size $O(\varepsilon^{-2} \log^2 n)$.*

Proof: For the sake of simplicity of exposition, assume that $\mathcal{I} = [0, 1]$. We break this interval into two equal length intervals, and first handle the interval $[1/2, 1]$. The other part is handled in a similar fashion. Consider the interval $\mathcal{I}_i = [1 - 2^{-i}, 1 - 2^{-i-1}]$, let $P[i] = \mathcal{I}_i \cap P$, and let $P[j, k] = \cup_{i=j}^k P[i]$ (in particular, $P[j, k]$ is empty if $k < j$).

Construction. Let k_1 be the first index such that $P[k_1]$ is not empty, and break $P[k_1]$ into sets using Lemma 4.3, and add these sets to the resulting partition. Next, assume that we had handled all the points in the set $P[1, k]$, and let Δ be a maximal integer, such that

$$|P[k+1, k+\Delta-1]| > \varepsilon |P[1, k-1]| \quad \text{and} \quad |P[k+1, k+\Delta-2]| \leq \varepsilon |P[1, k-1]|. \quad (1)$$

Note, that $\Delta \geq 2$ (if $\Delta = 2$ then the set $P[k+1, k+\Delta-2] = P[k+1, k]$ is empty). Next, we add the set $X = P[k+1, k+\Delta-2]$ to the resulting partition and compute a ε -safe partition of $P[k+\Delta-1]$, using Lemma 4.3, and add the sets of this partition to the output partition. We thus handled all points in the set $P[1, k+\Delta-1]$. We continue in this fashion till we handled all the points of P in the interval $[1/2, 1]$. We apply the symmetric construction to the interval $[0, 1/2]$.

Correctness. Consider an ε -safe partition \mathcal{R} of $P[i] = \mathcal{I}_i \cap P$, for any i . It is easy to verify that a set $S \in \mathcal{R}$ is ε -secure in $[0, 1]$. Indeed, observe that for any $S \subseteq \mathcal{I}_i$ we have that its security zone contains the interval \mathcal{I}_i .¹ So, consider an iteration of our construction, with the appropriate values of k and Δ , see Eq. (1). The set $X = P[k+1, k+\Delta-2]$ is ε -secure (in relation to the interval $[0, 1]$), since

$$|\mathcal{B}(X)| \leq \sum_{i=k+1}^{\infty} |\mathcal{I}_i| = \sum_{i=k+1}^{\infty} 1/2^{-i-1} = 1/2^{k-1} = |\mathcal{I}_k|.$$

In particular, $3\mathcal{B}(X) \cap P[1, k-1] = \emptyset$, implying the required security property. By the above discussion, the sets forming the ε -safe partition of $P[k+\Delta-1]$ are ε -secure in the resulting partition.

¹Thus, the temptation is to construct a ε -secure partition, for $[0, 1]$, by taking all such sets, and partitioning the point set lying inside each one of them into a ε -safe partition, which overall would result in an ε -secure partition. The problem with this scheme is that the number of non-empty intervals might be prohibitly large, and as such the resulting partition would be too large.

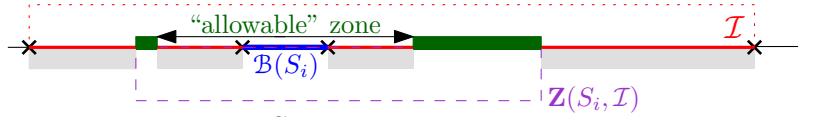


Figure 2: The set S_i and its associated “allowable” zone $\mathbf{Z}(S_i, \mathcal{I}) \setminus 3\mathcal{B}(S_i)$. For \mathcal{P} to be secure, the “allowable” zone of S_i must include sets of \mathcal{P} with total mass exceeding $|S_i|/\varepsilon$. And this has to hold for all sets S_i in \mathcal{P} .

Size. Let n_i be the number of points of P in the output partition in the end of the i th iteration. Clearly, we have that $n_i \geq (1 + \varepsilon)n_{i-2}$, which implies that the algorithm performs at most $O(\varepsilon^{-1} \log n)$ iterations. At every iteration, we add $O(\varepsilon^{-1} \log n)$ sets to the partition by Lemma 4.3. Thus, the claim follows. \blacksquare

4.2.1 An improved construction

The better bound follows by a careful inspection of the above construction. Assume, that we had already constructed a partition for $P[1, k]$. In particular, let $k_1 < k_2$ be the two smallest indices for which $P[k_1] \neq \emptyset$ and $P[k_2] \neq \emptyset$. Partition $P[k_1]$ and $P[k_2]$ by constructing a ε -safe partition for these two sets, using Lemma 4.3. Let $k = k_2$ and $m = |P[1, k - 1]|$. In somewhat similar fashion to the scheme above, we find a Δ such that

$$\beta = |P[k + 1, k + \Delta - 1]| \geq 2m \text{ and } |P[k + 1, k + \Delta - 2]| < 2m.$$

If $\beta \leq 10m$ then we can just split the set $P[k + 1, k + \Delta - 1]$ into sets of size $\leq (\varepsilon/10)m$. Clearly, the resulting partition is secure, and at each such step we added $O(1/\varepsilon)$ sets to the partition.

Thus, assume that $\beta > 10m$. We first partition the set $P[k + 1, k + \Delta - 2]$ into $O(1/\varepsilon)$ sets as above, this works since this set has less than $2m$ points. Thus, we only left with the task of partitioning the set $Q = P[k + \Delta - 1]$.

The set Q has m points of $P[1, k - 1]$ it can use ε -securely, without any problem since \mathcal{I}_k acts as a buffer zone. Thus, we have these m points that we can use to bottom out the recursion when constructing an ε -safe set for Q . Formally, we repeat the recursive partition of Lemma 4.3 for the set Q , with the twist that when the set becomes smaller than m , we stop, and just partition it into consecutive sets of cardinality at most $(\varepsilon/10)m$. Clearly, the resulting partition is ε -secure. The number of sets used is

$$O\left(\frac{1}{\varepsilon} \left(1 + \log \frac{\beta}{m}\right)\right).$$

Observe, that we can have at most $\mu = O(\log n)$ iteration of this process. In particular, let $\gamma_i = n_{i+1}/n_{i-1}$, where n_i is the number of points stored in the partition in the beginning of the i th iteration. Observe, that the number of sets generated by the algorithm in the i th iteration is bounded by $O(\frac{1}{\varepsilon}(1 + \log \gamma_i))$. Thus, the number of sets in the resulting partition is

$$M = O\left(\frac{\log n}{\varepsilon} + \sum_i^\mu \frac{1}{\varepsilon}(1 + \log \gamma_i)\right) = O\left(\frac{\log n}{\varepsilon} + \frac{1}{\varepsilon} \log \left(\prod_i \gamma_i\right)\right).$$

Observe that by cancellation $\prod_{i=1}^\mu \gamma_i = \prod_i \frac{n_{i+1}}{n_{i-1}} \leq n_{\mu+1} n_\mu \leq n^2$. As such, the number of sets in the resulting partition is bounded by

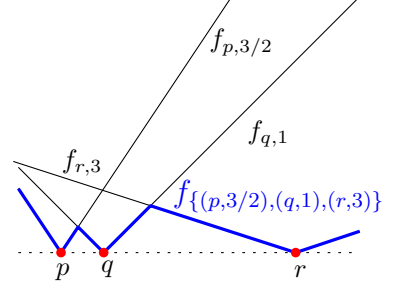
$$M = O\left(\frac{\log n}{\varepsilon} + \frac{1}{\varepsilon} \log(n^2)\right) = O\left(\frac{\log n}{\varepsilon}\right).$$

We conclude:

Theorem 4.6 *Let P be a set of n points on the real line contained inside the interval \mathcal{I} . Then, there exists a ε -secure partition of P in relation to \mathcal{I} of size $O(\varepsilon^{-1} \log n)$.*

5 Coreset for weighted centers

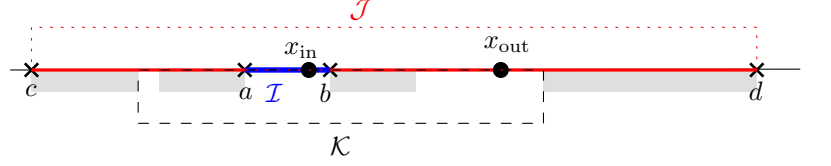
Let \mathcal{F}_{kwc} be the family of functions induced by k weighted centers on the real line. Formally, for a point $p \in \mathbb{R}$ and weight $w \in \mathbb{R}^+$, let $f_{p,w}(x) = w * |x - p|$. Given a set W of k weighted points $(p_1, w_1), \dots, (p_k, w_k)$, let $f_W(x) = \min_i f_{p_i, w_i}(x)$ denote the cost function of clustering x using its nearest weighted center in W , for all $x \in \mathbb{R}$. Let \widehat{W} denote the set of all possible k weighted points on the real line. Then, we have that



$$\mathcal{F}_{kwc} = \left\{ f_W(\cdot) \mid W \in \widehat{W} \right\}.$$

Thus, for a point set $P \subseteq \mathbb{R}$ and $W \in \widehat{W}$, the quantity $f_W(P)$ is the price of k -median clustering of P by the weighted centers of W . We are now interested in computing a coreset for this family. We need the following technical lemma.

Lemma 5.1 *Let $\mathcal{I} = [a, b]$ and $\mathcal{J} = [c, d]$ be two intervals on the real line, such that $\mathcal{I} \subseteq \mathcal{J}$, and let $\mathcal{K} = [(a + c)/2, (b + d)/2]$. Let $f_W \in \mathcal{F}_{kwc}$ be a function, such that none of its centers are in the set $\mathcal{J} \setminus \mathcal{I}$ (i.e., $W \subseteq (\mathbb{R} \setminus \mathcal{J}) \cup \mathcal{I}$). Finally, let $x_{\text{out}} \in \mathcal{K} \setminus 3\mathcal{I}$ and $x_{\text{in}} \in \mathcal{I}$ be any two points. Then $f_W(x_{\text{in}}) \leq 3f_W(x_{\text{out}})$.*



Proof: Let \mathbf{I} be the weighted centers of f_W that are located inside \mathcal{I} and let $W \setminus \mathbf{I}$ be the set of weighted centers of f_W that are located outside \mathcal{J} . Let $f_{\mathbf{I}}$ and $f_{W \setminus \mathbf{I}}$ be the associated functions induced by these two sets of weighted centers. Observe that, for any $\alpha \in \mathbb{R}$, we have that $f_W(\alpha) = \min(f_{W \setminus \mathbf{I}}(\alpha), f_{\mathbf{I}}(\alpha))$. As such, it is enough to prove this claim for those two functions.

Consider the center in \mathbf{I} with lowest weight w , and observe that $f_{\mathbf{I}}(x_{\text{in}}) \leq w|\mathcal{I}|$. On the other hand, the distance of x_{out} from \mathcal{I} is at least $|\mathcal{I}|$, which implies that $f_{\mathbf{I}}(x_{\text{out}}) \geq w|\mathcal{I}|$. Thus, $f_{\mathbf{I}}(x_{\text{in}}) \leq f_{\mathbf{I}}(x_{\text{out}})$.

As for the other function, observe that $f_{W \setminus \mathbf{I}}(\cdot)$ is concave over \mathcal{J} , as all its weighted centers lies outside \mathcal{J} . Assume that x_{out} is to the right \mathcal{I} . Observe that since $x_{\text{out}} \in \mathcal{K}$ we have that $|x_{\text{out}} - b| < |x_{\text{out}} - d|$. Since $x_{\text{out}} \notin 3\mathcal{I}$, it follows that $|x_{\text{out}} - b| > |\mathcal{I}|$. Thus, by concavity of $f_{W \setminus \mathbf{I}}$, we have that

$$\begin{aligned} f_{W \setminus \mathbf{I}}(x_{\text{out}}) &\geq \frac{x_{\text{out}} - x_{\text{in}}}{d - x_{\text{in}}} f_{W \setminus \mathbf{I}}(d) + \frac{d - x_{\text{out}}}{d - x_{\text{in}}} f_{W \setminus \mathbf{I}}(x_{\text{in}}) \\ &\geq 0 + \frac{d - x_{\text{out}}}{d - x_{\text{in}}} f_{W \setminus \mathbf{I}}(x_{\text{in}}) \geq \frac{d - x_{\text{out}}}{|\mathcal{I}| + |b - x_{\text{out}}| + |x_{\text{out}} - d|} f_{W \setminus \mathbf{I}}(x_{\text{in}}) \\ &\geq \frac{d - x_{\text{out}}}{3|x_{\text{out}} - d|} f_{W \setminus \mathbf{I}}(x_{\text{in}}) = \frac{f_{W \setminus \mathbf{I}}(x_{\text{in}})}{3}, \end{aligned}$$

as claimed. ■

Theorem 5.2 *Let P be a set of n points on the real line. There exists a ε -coreset for \mathcal{F}_{kwc} of size $O\left((\varepsilon^{-1} \log n)^{k+1}\right)$.*

Proof: The construction is done recursively. We construct a ε -coreset for P that can handle k centers inside the interval $\mathcal{B} = \mathcal{B}(P)$ and arbitrary number of centers outside.

Thus, for $k = 0$, the function induced by the set of centers (outside \mathcal{B}) is a concave function, and we use the $(\varepsilon/4)$ -coreset construction of Theorem 3.8. It would be however more convenient to consider the partition induced by this construction.

Otherwise, for $k > 0$, we compute a (ε/\bar{c}) -secure partition \mathcal{P} of P in relation to \mathcal{B} using Theorem 4.6, where \bar{c} is a constant to be specified shortly. For every set $S \in \mathcal{P}$, we recursively compute a partition for $(k - 1)$ centers.

This recursively induce a partition of P into subsets, and let $T(n, k)$ denote the number of subsets in the partition. We have that $T(n, 0) = O(\varepsilon^{-1} \log n)$, and $T(n, k) = T(n, k - 1)O(\varepsilon^{-1} \log n) = O(\varepsilon^{-k-1} \log^{k+1} n)$. Next, for every subset S in the new partition \mathcal{P} , we pick an arbitrary representative point and assign it weight equal to the number of points of S . We claim that the resulting weighted set \mathcal{S} is the required coreset.

Consider a function $f_C \in \mathcal{F}_{kwc}$. If $k = 0$ the coreset constructed is ε -coreset by Theorem 3.8. If f_C has at most $k - 1$ centers at each interval $\mathcal{B}(S)$, for all $S \in \mathcal{P}$, then the corresponding coresets for each such sets are ε -coresets, by induction, and it follows that \mathcal{S} is ε -coreset in this case. (Note, that here in the inductive step we used the fact that the coreset can handle an arbitrary number of additional centers outside the interval $\mathcal{B}(S)$.)

Thus, the only possible bad case is when all the k centers of f_C falls into a single interval $\mathcal{K} = \mathcal{B}(S) = [a, b]$, for a set $S \in \mathcal{P}$. Observe, that f is being $(1 \pm \varepsilon/4)$ -approximated correctly for all the other sets of \mathcal{P} , since they have no centers of C inside them. (Indeed, for each such set of \mathcal{P} the recursive construction further breaks it down into subintervals, and in the bottom of the recursion for each subset inside a subinterval we construct a $(\varepsilon/4)$ -coreset for concave functions.)

Now, that there are at least $(\bar{c}/\varepsilon) |S|$ points of P in $\mathbf{Z}(S, \mathcal{B}) \setminus 3\mathcal{K}$, since \mathcal{P} is (ε/\bar{c}) -secure. In particular, let \mathcal{T} be the coreset constructed for the points of S , and \mathcal{R} be the coreset constructed for the point-set $Q = \cup(\mathcal{P} \cap (\mathbf{Z}(S, \mathcal{B}) \setminus 3\mathcal{K}))$. We have, by Lemma 5.1, that

$$\begin{aligned} \max(f(\mathcal{T}), f(S)) &\leq |S| \max_{x \in \mathcal{I}} f(x) \leq |S| \cdot 3 \min_{x \in Q} f(x) \leq \frac{\varepsilon}{\bar{c}} |Q| 3 \min_{x \in Q} f(x) \leq \frac{3\varepsilon}{\bar{c}} f(Q) \\ &\leq (1 + \varepsilon/4) \frac{3\varepsilon}{\bar{c}} f(\mathcal{R}) \leq \frac{4\varepsilon}{\bar{c}} f(\mathcal{R}). \end{aligned}$$

It follows that

$$\mathcal{E}' = |f(\mathcal{T}) - f(S)| \leq \frac{4\varepsilon}{\bar{c}} f(\mathcal{R}) \leq \frac{4\varepsilon}{\bar{c}} f(P).$$

Thus,

$$\begin{aligned} \mathcal{E} = |f(\mathcal{S}) - f(P)| &\leq |f(\mathcal{T}) - f(S)| + |f(\mathcal{S} \setminus \mathcal{T}) - f(P \setminus S)| \leq \frac{4\varepsilon}{\bar{c}} f(P) + \frac{\varepsilon}{4} f(P \setminus S) \\ &\leq \frac{\varepsilon}{2} f(P), \end{aligned}$$

as required. ■

Remark 5.3 The reader might wonder why we need secure partitions for proving Theorem 5.2, and maybe safe partitions are enough. However, a careful inspection of the proof reveals that the inductive hypothesis in the proof fails if we use only safe partitions. Thus, we need the more involved and painful construction of secure partitions for our purposes.

5.1 A lower bound

By Remark 3.10, we know that the family of functions describable as a “triangle” requires a coreset of size $\Omega(\varepsilon^{-1} \log n)$. Now, clearly, such a triangle function can be realized by two weighted centers. Thus, we get a lower bound on the size of ε -coreset for k weighted centers.

We can push this even further, by partitioning our set of n points into $k/2$ groups, each group has the structure of the point set of Lemma 3.9. We place these $k/2$ sets very far away from each other. Clearly, now we can realize a function which is a triangle function on one of the sets and (arbitrarily close to) zero on all other sets using k weighted centers. It follows, that every set must include $\Omega(\varepsilon^{-1} \log(n/k))$ points in the ε -coreset, and overall the size of the ε -coreset is $\Omega((k/\varepsilon) \log(n/k))$.

5.1.1 Improved lower bound

We construct a point-set P such that any coreset of P must include all the points of P . Let k be the parameter which is the number of weighted centers in a function of \mathcal{F}_{kwc} . The construction would be recursive and would be done on the interval, say, $[0, 1]$. Let Δ be an arbitrary constant which is very large.

The recursive construction receives as input a parameter a number k and a construction interval \mathcal{I} .

For $k = 1$, we just place two points at the two endpoints of \mathcal{I} . Otherwise, for $k > 1$, let \mathcal{J} and \mathcal{K} be the two intervals of length $|\mathcal{I}|/\Delta$ contained inside \mathcal{I} , such that \mathcal{J} shares its left endpoint with \mathcal{I} , and \mathcal{K} shares its right endpoint of \mathcal{I} . Let P_L be the point set generated by this recursive construction for the interval \mathcal{J} for $k - 1$, and similarly, let P_R be the point set generated by this construction for the interval \mathcal{K} . Let $P = P_L \cup P_R$. Clearly, $n = |P| = 2^k$. We claim that any $1/2$ -coreset for P must contain all the points of P .

Indeed, assume for the sake of contradiction that this is false, and let $\mathcal{S} \subsetneq P$ be a $(1/2)$ -coreset for P and let q be a point of $P \setminus \mathcal{S}$. We demonstrate that there exists a function $f \in \mathcal{F}_{kwc}$ such that it assigns value 1 to q and value smaller than $2/\Delta$ to all the other points of P . In particular, we have that $f(P) \geq 1$ and $f(\mathcal{S}) \leq 2n/\Delta$, which is a contradiction to the fact that \mathcal{S} is a $(1/2)$ -coreset for, say, $\Delta > 32n$.

In particular, for a points x on the real line, let $g_x(\cdot)$ denote the weighted cone that has its apex at x , and $g_x(q) = 1$. We will pick k points, each one of them will induce a cone function $g_{x_i}(\cdot)$. Together, their lower envelope would form the required function f .

To this end, we track the recursive construction of P . At every point in our process, we will have a parameter k , a subset $S \subseteq P$ which was not “served” yet containing q , and an interval \mathcal{I} . For $k = 1$, the set S would be the two endpoints of \mathcal{I} and one of them would be q . Let p be the other point of S , and set $x_1 = p$.

Otherwise, for $k > 1$, the set S is made out of 2^k points. The set S is the union of the left subset P_L and the right subset P_R . Assume that $q \in P_L$. Let u be the right endpoint of

\mathcal{I} , and set $x_k = u$.

Observe that for any $x \in P_R$ we have that $g_u(x) \leq 2/\Delta$ since the distance between x and u is at most $|\mathcal{I}|/\Delta$. Also, the slope of $g_u(x)$ is at most $2/|\mathcal{I}|$ since $g_u(u) = 0$ and $g_u(q) = 1$, where $u \in P_R$ and $x \in P_L$ and the distance between P_R and P_L is at least $|\mathcal{I}|/2$. Now, all the points of P_R are contained in an interval of length $|\mathcal{I}|/\Delta$, by construction. Thus, for any $x \in P_R$ we have that $g_u(x) \leq (2/|\mathcal{I}|) \cdot (|\mathcal{I}|/\Delta) = 2/\Delta$, as claimed.

We now continue the choice of apexes with $k-1$, recursing over the set P_R . Let x_1, \dots, x_k be the apexes picked, and let $f(z) = \min_{i=1}^k g_{x_i}(z)$. Clearly, $f \in \mathcal{F}_{kwc}$, $f(q) = 1$ and for all other points of $p \in P \setminus \{q\}$ we have that $f(p) \leq 2/\Delta$. The claim now easily follows.

Theorem 5.4 *A ε -coreset for a set of n points for the set of functions \mathcal{F}_{kwc} , has to be of size $\Omega\left(\max\left[(k/\varepsilon) \log(n/k), 2^k\right]\right)$ in the worst case.*

6 Coreset for weighted lines

Consider a line ℓ in \mathbb{R}^d , for $d > 1$, and another line σ . For any point $p \in \sigma$, we are interested in its distance to ℓ . It is easy to verify that if we parameterize σ uniformly by a number $t \in \mathbb{R}$, then the *line distance function*

$$f(t) = \min_{x \in \ell} \|\sigma(t) - x\|,$$

has the form $f(t) = \sqrt{\gamma^2 + \beta^2(t - \alpha)^2}$, where $\beta \leq 1$ and $\gamma \geq 0$. Note, that $f(t)$ is non-negative, symmetric function around α , realizing its minimum at $t = \alpha$. Let $\mathcal{F}_{\text{dline}}$ denote the family of all such functions.

Lemma 6.1 *Let $f(t) = \sqrt{\gamma^2 + \beta^2(t - \alpha)^2}$ be a line distance function, and let x and y be two real numbers, such that $y < x < \alpha$ and $\alpha - y \leq \eta(\alpha - x)$, where $\eta \geq 1$ is a constant. Then, $f(y) \leq \eta * f(x)$.*

Proof: Let $\Delta = \alpha - x$. Since $\alpha - y \leq \eta\Delta$, We have that

$$f(y) = \sqrt{\gamma^2 + \beta^2(y - \alpha)^2} \leq \sqrt{\gamma^2 + \beta^2(\eta\Delta)^2} \leq \eta\sqrt{\gamma^2 + \beta^2\Delta^2} \leq \eta * f(x). \quad \blacksquare$$

Let $\mathcal{F}_{k\text{-lines}}$ be the function formed by the minimization diagram of k functions of $\mathcal{F}_{\text{dline}}$. A function of such family has a natural interpretation as the distance of a point on a line, to the closest line in a set of k lines. Next, we prove the analogue of Lemma 5.1 for this family of functions.

Lemma 6.2 *Let $\mathcal{I} = [a, b]$ and $\mathcal{J} = [c, d]$ be two intervals on the real line, such that $\mathcal{I} \subseteq \mathcal{J}$, and let $\mathcal{K} = [(a+c)/2, (b+d)/2]$. Let $f \in \mathcal{F}_{k\text{-lines}}$ be a function, such that all its minimums are either in \mathcal{I} or outside \mathcal{J} . Finally, let $x_{\text{out}} \in \mathcal{K} \setminus 3\mathcal{I}$ and $x_{\text{in}} \in \mathcal{I}$ be any two points. Then $f(x_{\text{in}}) \leq 3f(x_{\text{out}})$.*

Proof: The function f is the minimization function of k functions f_1, \dots, f_k . In particular, let F_O be the set of these functions with their minimum outside \mathcal{J} , and F_I be the set of these functions with minimums inside \mathcal{I} .

For a function $g \in F_I$, let $u \in \mathcal{I}$ be the location of its minimum. Consider the two points $x_1 = u - |x_{\text{out}} - u|$ and $x_2 = u + |x_{\text{out}} - u|$. Clearly, $\mathcal{I} \subseteq [x_1, x_2]$, and by the symmetry of g around u , we have that outside $[x_1, x_2]$ it is larger than it is inside this interval. As such, $g(x_{\text{in}}) \leq g(x_1) = g(x_2) = g(x_{\text{out}})$.

For a function $h \in F_O$, assume that x_{out} is to the right of \mathcal{I} . If the minimum of h is to the left of \mathcal{J} , then we are done as h is increasing on the interval \mathcal{J} , and as such $h(x_{\text{out}}) \geq h(x_{\text{in}})$. If the minimum of h is to the right of \mathcal{J} , as depicted on the right, then arguing as in Lemma 5.1 we have that $|x_{\text{out}} - b| > |\mathcal{I}|$. Also, we have that $|x_{\text{out}} - d| \geq |x_{\text{out}} - b|$. Thus, by Lemma 6.1, we have that

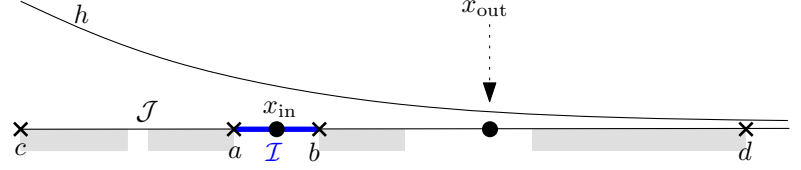


Figure 3: The gray areas represent the “forbidden/insecure” areas for x_{out} .

$$\begin{aligned} h(x_{\text{out}}) &\geq \frac{d - x_{\text{out}}}{d - x_{\text{in}}} f_M(x_{\text{in}}) \geq \frac{d - x_{\text{out}}}{|\mathcal{I}| + |b - x_{\text{out}}| + |x_{\text{out}} - d|} f_M(x_{\text{in}}) \\ &\geq \frac{d - x_{\text{out}}}{3|x_{\text{out}} - d|} f_M(x_{\text{in}}) = \frac{f_M(x_{\text{in}})}{3}, \end{aligned}$$

as claimed. ■

To construct a coreset for $\mathcal{F}_{k\text{-lines}}$ using a construction similar to Theorem 5.2, we need to be able to handle the base case $k = 0$. Fortunately, the functions induced in such a case are “almost” concave.

Lemma 6.3 *Let $\mathcal{I} = [a, b]$ be an interval, and $P \subseteq \mathcal{I}$ be a set of n points. Then one can compute a ε -coreset \mathcal{S} , of size $O(\varepsilon^{-1} \log n)$, for the set of functions of $\mathcal{F}_{k\text{-lines}}$ with minimums outside \mathcal{I} .*

Formally, the coreset works for functions f , such that $f \in \mathcal{F}_{k\text{-lines}}$ is the minimization diagram of k line distance functions $f_1, \dots, f_k \in \mathcal{F}_{\text{line}}$, and all of them having minimums outside \mathcal{I} .

Proof: We only sketch the proof since it is similar to previous proofs. First, observe that one can define the notion of ε -oblivious intervals, as done in Definition 3.6. Clearly, such intervals get shorter as you get closer to the endpoints of \mathcal{I} , but in fact similar bounds hold as in Lemma 3.7. Thus, we pave the middle of \mathcal{I} by ε -oblivious intervals. As in the proof of Theorem 3.8, we have to handle only two “tiny” margin intervals of length $\varepsilon^2 |\mathcal{I}|$ adjacent to the endpoints of \mathcal{I} (i.e., the rest of \mathcal{I} is covered by $O(\varepsilon^{-1} \log(1/\varepsilon))$ ε -oblivious intervals).

So, consider the left side such interval \mathcal{K} . We use the monotone increasing $(\varepsilon/10)$ -coreset construction for the points of P lying inside \mathcal{K} , see Lemma 3.2. The symmetric construction is applied to the other interval.

Every oblivious interval contribute one coreset point, and these two margin intervals contribute $O(\varepsilon^{-1} \log n)$ points. As such, the overall coreset size is as claimed.

As for showing that it is indeed the required coreset. Observe that for the oblivious intervals the claim trivially holds. Thus, consider \mathcal{K} , and consider a function f as above,

and let $L = \{f_1, \dots, f_m\}$ (resp. $R = \{f_{m+1}, \dots, f_k\}$) be the functions that are inducing f and have centers to the left (resp., right) of \mathcal{I} .

Observe, that for functions of R , the distance to \mathcal{K} is so large compared to the length of $|\mathcal{K}|$, that we can essentially consider them to be a constant on this interval (this is not precise and some minor noise is introduced doing this). On the other hand, the functions of L induce a function $g(x) = \min_{h \in L} h(x)$ which is monotone increasing on \mathcal{K} . Thus, up to small noise, the function f can be treated as a monotone increasing function on \mathcal{I} . Thus, the coresets provide the required approximation. The exact details of the above argumentation are very similar to the proof of Theorem 3.8 and are thus omitted.

We conclude that the resulting coreset is the required ε -coreset. \blacksquare

Theorem 6.4 *Let P be a set of n points on the real line. There exists a ε -coreset for \mathcal{F}_k -lines of size $O\left((\varepsilon^{-1} \log n)^{k+1}\right)$.*

Proof: The construction is similar to the construction of Theorem 5.2, where in the bottom of the construction we use the construction of Lemma 6.3. The proof of correctness now proceeds identically to the proof of Theorem 5.2, with the modification that we use Lemma 6.2 instead of Lemma 5.1. \blacksquare

Remark 6.5 Theorem 6.4 is a substantial improvement over the result of Fiat *et al.* [FFS06] that had coreset of size

$$\frac{2^{O(k^2)}}{\varepsilon^{2k+1}} \log^{4k-3} n,$$

for this problem. Also, our construction is arguably simpler and more intuitive (well, at least for the author).

6.1 Applications

The results above immediately imply that there exists a k -line median coreset for clustering of small size. Namely, given a set P of n points in \mathbb{R}^d one can find a small coreset of small size such that finding the k -lines of minimum median price (i.e., every point pays its distance to the closest line in the set k lines that serves as centers).

Theorem 6.6 *Given a set P of n points in \mathbb{R}^d , and a $k > 0$, there exists a ε -coreset for P for the problem of k -line median clustering. The size of the coreset is $O(k\varepsilon^{-k-d} \log^{k+2} n)$.*

Proof: Consider the k lines realizing the centers in the optimal solution. Next, spread copies of this lines in an exponential grid fashion [HM04]. Every line gets $O(\varepsilon^{-(d-1)} \log n)$ copies. Next, snap every point of P to its nearest line. For each line now, and the points lying on it, we apply the coreset construction of Theorem 6.4. The resulting coreset has size

$$O\left(\left(\frac{\log n}{\varepsilon}\right)^{k+1} \cdot \frac{k}{\varepsilon^{d-1}} \log n\right),$$

as claimed. \blacksquare

Plugging our construction into the standard machinery of random sampling leads to an efficient clustering algorithm, which computes $(1 + \varepsilon)$ -approximate k -median line clustering in near linear time. We omit any further details, see [FFS06]

7 Conclusions

We had introduced the problem of computing coresets for discrete integration, and showed some tight coreset constructions for this problem, for various families of functions. We also used it to improve the coreset size for the problem of clustering points in \mathbb{R}^d for the k -median line clustering.

In particular, we showed a coreset of size (roughly) $O(\log^{k+2} n)$ and a lower bound of 2^k , see Theorem 5.4. Previously, no non-trivial lower bound was known for this problem. Although there is still a gap between the upper and lower bound it is relatively “small”, and we leave the improvement of both bounds as an open problem for further research. The lower bound also leaves open the question of how to cluster efficiently a set of n points with $k = \Omega(\log n)$ line centers. The lower bound implies that coresets can not help us in solving this problem efficiently, but maybe other techniques might work here. We leave this as an open problem for further research.

Another open problem is to extend the study of the discrete integration problem to dimensions higher than one.

Acknowledgments

The author would like to thank Danny Feldman for useful and insightful discussions on the problems studied in this paper.

References

- [ADPR00] N. Alon, S. Dar, M. Parnas, and D. Ron. Testing of clustering. In *Proc. 41th Annu. IEEE Sympos. Found. Comput. Sci.*, pages 240–250, 2000.
- [AHV05] P. K. Agarwal, S. Har-Peled, and K. Varadarajan. Geometric approximation via coresets. In J. E. Goodman, J. Pach, and E. Welzl, editors, *Combinatorial and Computational Geometry*, Math. Sci. Research Inst. Pub. Cambridge, 2005.
- [AP00] P. K. Agarwal and C. M. Procopiuc. Approximation algorithms for projective clustering. In *Proc. 11th ACM-SIAM Sympos. Discrete Algorithms*, pages 538–547, 2000.
- [BC03] M. Bădoiu and K. Clarkson. Smaller coresets for balls. In *Proc. 14th ACM-SIAM Sympos. Discrete Algorithms*, pages 801–802, 2003.
- [Che06] K. Chen. On k -median clustering in high dimensions. In *Proc. 17th ACM-SIAM Sympos. Discrete Algorithms*, pages 1177–1185, 2006.
- [Epp98] D. Eppstein. Fast hierarchical clustering and other applications of dynamic closest pairs. In *Proc. 9th ACM-SIAM Sympos. Discrete Algorithms*, pages 619–628, 1998.

- [FFS06] D. Feldman, A. Fiat, and M. Sharir. Coresets for weighted facilities and their applications. manuscript, 2006.
- [FG88] T. Feder and D. H. Greene. Optimal algorithms for approximate clustering. In *Proc. 20th Annu. ACM Sympos. Theory Comput.*, pages 434–444, 1988.
- [Gon85] T. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoret. Comput. Sci.*, 38:293–306, 1985.
- [Har06a] S. Har-Peled. Coresets for discrete integration and clustering. Available from <http://www.uiuc.edu/~sariel/papers/06/integrate>, 2006.
- [Har06b] S. Har-Peled. How to get close to the median shape. In *Proc. 22nd Annu. ACM Sympos. Comput. Geom.*, 2006. To appear. Available from http://www.uiuc.edu/~sariel/papers/05/11_fitting/.
- [HK05] S. Har-Peled and A. Kushal. Smaller coresets for k -median and k -means clustering. In *Proc. 21st Annu. ACM Sympos. Comput. Geom.*, pages 126–134, 2005.
- [HM04] S. Har-Peled and S. Mazumdar. Coresets for k -means and k -median clustering and their applications. In *Proc. 36th Annu. ACM Sympos. Theory Comput.*, pages 291–300, 2004.
- [IKI94] M. Inaba, N. Katoh, and H. Imai. Applications of weighted voronoi diagrams and randomization to variance-based k -clustering. In *Proc. 10th Annu. ACM Sympos. Comput. Geom.*, pages 332–339, 1994.
- [Ind99] P. Indyk. A sublinear time approximation scheme for clustering in metric spaces. In *Proc. 40th Annu. IEEE Sympos. Found. Comput. Sci.*, 1999. 100–110.
- [KMN⁺04] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. A local search approximation algorithm for k -means clustering. *Comput. Geom. Theory Appl.*, 28:89–112, 2004.
- [Mat99] J. Matoušek. *Geometric Discrepancy*. Springer, 1999.
- [MOP01] N. Mishra, D. Oblinger, and L. Pitt. Sublinear time approximate clustering. In *Proc. 12th ACM-SIAM Sympos. Discrete Algorithms*, pages 439–447, 2001.
- [OR00] R. Ostrovsky and Y. Rabani. Polynomial time approximation schemes for geometric k -clustering. In *Proc. 41st Symp. Foundations of Computer Science*, pages 349–358. IEEE, Nov 2000.