Smaller Coresets for k-Median and k-Means Clustering*

Sariel Har-Peled† Akash Kushal‡

November 29, 2004

Abstract

In this paper, we show that there exists a (k, ε)-coreset for k-median and k-means clustering of n points in $\mathbb{R}^d$, which is of size independent of n. In particular, we construct a (k, ε)-coreset of size $O(k^2/\varepsilon^d)$ for k-median clustering, and of size $O(k^3/\varepsilon^{d+1})$ for k-means clustering.

1 Introduction

Clustering is a widely used technique in Computer Science with applications to unsupervised learning, classification, data mining and other fields. We study two variants of the clustering problem in the geometric setting. The *geometric k-median clustering* problem is the following: Given a set P of n points in $\mathbb{R}^d$, compute a set of k points (i.e., medians) such that the sum of the distances of the points in P to their respective nearest median is minimized. The k-means differs from the above in that instead of the sum of distances, we minimize the sum of squares of distances. Interestingly the 1-mean is the center of mass of the points, while the 1-median problem, also known as the Fermat-Weber problem, has no such closed form. As such the problems have usually been studied separately from each other even in the approximate setting.

The basic question underlying approximation algorithms, is what portion of the data is necessary to compute (approximately) a certain quantity. The smaller this portion is, the more efficient the resulting algorithm would be. A coreset is a small portion of the data, such that running a clustering algorithm on it, generates a clustering for the whole data, which is approximately optimal. In particular, a small coreset indicates that a problem is easy to approximate. Furthermore, it implies that one can summarize and sketch the data efficiently. This is useful for database applications, where one can store such sketches efficiently, and perform efficient clustering on a database, or portions of it using the sketches.

*See http://www.uiuc.edu/~sariel/papers/04/small_coreset/ for the most recent version of this paper.

†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.

‡Department of Computer Science; University of Illinois; 201 N. Goodwin Avenue; Urbana, IL, 61801, USA; kushal@uiuc.edu; http://www-cvr.ai.uiuc.edu/~kushal/.

In particular, the size of the smallest coreset needed is a fundamental combinatorial property of the clustering problem at hand. Among other things, coresets of size independent of n imply "strong" fixed parameter algorithms [DF95] (i.e., algorithms with running time $O(n+\mathrm{poly}(k, \log n, 1/\varepsilon)+\mathrm{func}(k, \varepsilon))$, where poly denotes a polynomial, and $\mathrm{func}(k, \varepsilon)$ denotes a function that depends only on k and ε (and the dimension d).

k-median clustering. The k-median problem is nontrivial even in low dimensions and achieving a good approximation proved to be a challenge. Motivated by the work of Arora [Aro98], which proposed a new technique for geometric approximation algorithms, Arora, Raghavan and Rao [ARR98] presented a $O\left(n^{O(1/\varepsilon)+1}\right)$ time $(1+\varepsilon)$-approximation algorithm for points in the plane. This was significantly improved by Kolliopoulos and Rao [KR99] who proposed an algorithm with a running time of $O(\varrho n \log n \log k)$ for the discrete version of the problem, where the medians must belong to the input set and $\varrho = \exp\left[O((1+\log 1/\varepsilon)/\varepsilon)^{d-1}\right]$. The k-median problem has been studied extensively for arbitrary metric spaces and is closely related to the un-capacitated facility location problem. See [CGTS99, GNMO00, MP02] for more information.

The running time of Kolliopoulos and Rao [KR99] was further improved to $O\left(n+\varrho k^{O(1)} \log^{O(1)} n\right)$ by Har-Peled and Mazumdar [HM04] by using coresets. Formally, a weighted subset $S \subseteq P$ is a (k, ε)-coreset for the k-median problem, if for any set C of k centers in the $\mathbb{R}^d$, the price of clustering P using C, and the price of clustering S using C, is the same up to $1 \pm \varepsilon$. Har-Peled and Mazumdar [HM04] showed that there exists a coreset of P of size $O(k\varepsilon^{-d} \log n)$, and by computing such a coreset quickly and running the algorithm on this coreset, one gets the aforementioned fast approximation algorithm.

k-means clustering. Inaba *et al.* [IKI94] observe that the number of Voronoi partitions of k points in $\mathbb{R}^d$ is n^{kd} and can be done exactly in time $O(n^{kd+1})$. They also propose approximation algorithms for the 2-means clustering problem with time complexity $O(n^{O(d)})$. de la Vega *et al.* [dlVKKR03] proposed a $(1+\varepsilon)$-approximation algorithm, for high dimensions, with running time $O(g(k, \varepsilon) dn \log^k n)$, where $g(k, \varepsilon) = \exp[(k^3/\varepsilon^8)(\ln(k/\varepsilon)) \ln k]$ (they refer to it as ℓ_2^2 k-median clustering). This was improved to $O(h(k, \varepsilon)dn)$ time algorithm, by Kumar *et al.* [KSS04], where $h(k, \varepsilon) = 2^{(k/\varepsilon)^{O(1)}}$ (as such, this algorithm is only appropriate when the data is high dimensional). Matoušek [Mat00] proposed a $(1+\varepsilon)$-approximation algorithm for the geometric k-means problem with running time $O\left(n\varepsilon^{-2k^2 d} \log^k n\right)$. Again, using coresets, Har-Peled and Mazumdar [HM04], presented an algorithm with running time $O(n + k^{k+2}\varepsilon^{-(2d+1)k} \log^{k+1} n \log^k \frac{1}{\varepsilon})$, which is linear for fixed k and ε. Effros and Schulman [ES03] showed that there exists a centroid set of size independent of n. A centroid set is a set that contains at least one k-tuple, which forms (approximately) optimal centers for k-means clustering. While the resulting algorithm is slower than the algorithm of Har-Peled and Mazumdar it does hint to the possibility that a coreset of size independent of n should exist for the k-means problem.

Our Results. In light of the aforementioned results, it is natural to ask what is the smallest coreset one can extract, and compute approximate clustering using it. In particular, can one compute a coreset of size independent of n?

In this paper, we answer this question positively, by showing a coreset of size $O(k^2/\varepsilon^d)$

for k-median and $O(k^3/\varepsilon^{d+1})$ for k-means. Interestingly, unlike the previous results, while the intuition for the two cases is similar, the proof and construction are fundamentally different. In particular, the coreset construction for the k-means case is slightly easier than the k-median case.

The previous construction of coresets for clustering relied on first computing a set of k centers which were a constant factor approximation to the optimal clustering. Next, using an exponential grid of $O(\log n)$ levels around each center, and snapping the points to this grid resulted in the required coreset. The correctness of the above coreset follows since the price of snapping the points to the exponential grid is smaller than $\varepsilon\nu_{\mathrm{opt}}(P,k)$, where $\nu_{\mathrm{opt}}(P,k)$ is the price of the optimal k-median clustering of P. In Appendix A, we show that any such approach of computing a small set C of points such that snapping the points of P to C is "cheap" (i.e., $\leq \varepsilon\nu_{\mathrm{opt}}(P,k)$) is doomed, as such a set must have size $\Omega(\log n)$. To overcome this, we need to be considerably more careful in picking C, such that the errors introduced by the snapping *cancel each other out*.

To this end, we replace each exponential grid around a center point, by a set of $O(1/\varepsilon^{d-1})$ lines. We now snap the points to the lines. We end up with $O(k/\varepsilon^{d-1})$ point sets, each one of them is one dimensional (although the centers are not necessarily on the line). We compute a coreset for each such line separately, and we take the union of those coresets, to form the resulting coreset of the whole set.

To figure out how to pick our coreset on each such line, we first solve the toy problem, of computing a coreset for a set of points on a line, where the centers are also on the line. This is done by breaking the line into chunks of small size (this idea is somewhat similar to the analysis of Effros and Schulman [ES03], although our analysis is considerably simpler). We then extend it to the case where the centers are not necessarily on the line. We do this analysis for the k-median and k-means cases separately, since the analysis is substantially different.

Note, that we reduced the question of computing a d-dimensional coreset to a one and two dimensional problem (since the Voronoi diagram on a line of k points in $\mathbb{R}^d$, can always be simulated by k points in two dimensions). This reduction considerably simplifies our analysis.

The paper is organized as follows. In Section 2, we present the coreset construction for the k-median case. In Section 3, we handle the k-means case. We conclude in Section 4.

2 Coreset for k-median

2.1 Preliminaries

For a point set X, and a point p, both in $\mathbb{R}^d$, let $\mathbf{d}(p,X) = \min_{x\in X}\|xp\|$ denote the *distance of p from X*. For a set B of points on a line in $\mathbb{R}^d$, let $\mathcal{I}(B)$ denote the smallest closed segment containing all the points of B.

Weighted set. A *weighted* point set P is a set of points, where every point $p \in P$ is assigned a *weight* w_p, which is a real positive number. We denote by $w(P) = \sum_{p\in P} w_p$ the total weight of the set P.

k-median clustering. For a weighted point set P with points from $\mathbb{R}^d$, with an associated weight function $w : P \to \mathbb{R}^+$ and any point set C, we define $\nu_C(P) = \sum_{p \in P} w_p \cdot \mathbf{d}(p, C)$ as the *price* of the k-median clustering provided by C. Further let $\nu_{\mathrm{opt}}(P, k) = \min_{C \subseteq \mathbb{R}^d, |C|=k} \nu_C(P)$ denote the price of the *optimal k-median* clustering for P.

(k, ε)-coreset for k-median. For a weighted point set $P \subseteq \mathbb{R}^d$, a weighted set $\mathcal{S} \subseteq \mathbb{R}^d$, is a (k, ε)-*coreset* of P for the k-median clustering, if for any set C of k points in $\mathbb{R}^d$, we have $(1 - \varepsilon)\nu_C(P) \le \nu_C(\mathcal{S}) \le (1 + \varepsilon)\nu_C(P)$.

mean/center of mass. For a weighted point set P in $\mathbb{R}^d$, let $\overline{\mathrm{m}}(P) = \sum_{p \in P}(w_p/w(P))p$ denote the *mean* of P (this is also known as the center of mass of P). We define the *cumulative error* (or just the *error*) for a point set P in $\mathbb{R}$ as $\mathcal{E}_\nu(P) = \nu_{\overline{\mathrm{m}}}(P) = \sum_{p \in P} \|p\overline{\mathrm{m}}\|$, where $\overline{\mathrm{m}} = \overline{\mathrm{m}}(P)$.

2.2 One Dimension

The basic idea for the coreset construction in one dimension (here, both the points and the centers lie in one dimension), is to break the point set into smaller sets, and use the mean point of every subset, as the representative for the coreset. We first prove, in Lemma 2.1, that the cumulative error of a point set bounds the error that it might contribute if we use the median point as the coreset. In Lemma 2.2, we show that cumulative error is a 2-approximation to the optimal 1-median clustering of a point set. As such, we can use the mean point of a point set for a coreset representative. In Lemma 2.3, we extend this observation to several point sets. Then, in Theorem 2.4, we describe the construction and prove that it works.

Lemma 2.1 *Let P be a set of n points on an oriented line ℓ in $\mathbb{R}^d$, and let $\overline{\mathrm{m}} = \overline{\mathrm{m}}(P)$. We have:*

(i) $\sum_L \|\overline{\mathrm{m}}p\| = \sum_R \|\overline{\mathrm{m}}p\|$, *where L (resp. R) are the points of P left (resp. right) of $\overline{\mathrm{m}}$ on ℓ.*

(ii) For a point $q \in \ell$ such that $q \notin \mathcal{I}(P)$, we have that $\nu_q(P) = n\|q\overline{\mathrm{m}}\|$.

(iii) For any set of points $C \subseteq \mathbb{R}^d$, we have $|\nu_C(P) - \nu_C(\mathcal{S})| \le \mathcal{E}_\nu(P)$, where $\mathcal{S}$ is a coreset made out of the single point $\overline{\mathrm{m}}$ with weight $|P|$.

Proof: (i) Rotate space, such that ℓ becomes the x-axis. Then we have $\sum_{p \in P, x_p < x_{\overline{\mathrm{m}}}}(x_{\overline{\mathrm{m}}} - x_p) = \sum_{p \in P, x_p \ge x_{\overline{\mathrm{m}}}}(x_p - x_{\overline{\mathrm{m}}})$, since $\overline{\mathrm{m}}$ is the mean point of $P \subseteq \ell$, where x_p denotes the x-coordinate of a point $p \in \mathbb{R}^d$. Now, $\sum_{p \in L}\|\overline{\mathrm{m}}p\| = \sum_{p \in L}(x_{\overline{\mathrm{m}}} - x_p) = \sum_{p \in R}(x_p - x_{\overline{\mathrm{m}}}) = \sum_{p \in R}\|\overline{\mathrm{m}}p\|$.

(ii) Assume that $x_q < x_{\overline{\mathrm{m}}}$, and then we have

$$
\begin{aligned}
\nu_q(P) &= \sum_{p \in P} \|pq\| = \sum_{p \in P, x_p < x_{\overline{\mathrm{m}}}} (\|q\overline{\mathrm{m}}\| - \|p\overline{\mathrm{m}}\|) + \sum_{p \in P, x_p > x_{\overline{\mathrm{m}}}} (\|q\overline{\mathrm{m}}\| + \|p\overline{\mathrm{m}}\|) \\
&= n\|q\overline{\mathrm{m}}\| + \left(\sum_{p \in P, x_p < \overline{\mathrm{m}}} -\|p\overline{\mathrm{m}}\| + \sum_{p \in P, x_p > \overline{\mathrm{m}}} \|p\overline{\mathrm{m}}\| \right) = n\|q\overline{\mathrm{m}}\|\,,
\end{aligned}
$$

by the first claim. The case $x_q > x_{\overline{m}}$ follows by symmetry.

(iii) We have $|\nu_C(P) - \nu_C(\mathcal{S})| = \left|\sum_{p \in P} (\mathbf{d}(p, C) - \mathbf{d}(\overline{m}, C))\right| \leq \sum_{p \in P} \|p\overline{m}\| = \mathcal{E}_\nu(P)$, since $\mathbf{d}(q, C) - \|pq\| \leq \mathbf{d}(p, C) \leq \mathbf{d}(q, C) + \|pq\|$, for any $p, q \in \mathbb{R}^d$. ■

Lemma 2.2 *Let $P \subseteq \mathbb{R}$ be a set of n points. Then $\mathcal{E}_\nu(P) \leq 2\nu_{\mathrm{opt}}(P, 1)$.*

Proof: Let ξ be a median point of P and $\overline{m} = \overline{m}(P)$. The optimal clustering $\nu_{\mathrm{opt}}(P, 1)$ is achieved at a median point $\xi = \mathrm{Median}(P)$. Thus,

$$\begin{aligned}
\mathcal{E}_\nu(P) = \nu_{\overline{m}}(P) &= \sum_{p \in P} \|p\overline{m}\| = \sum_{p \in P} \left| p - \frac{1}{n}\sum_{q \in P} q \right| = \sum_{p \in P} \left| \frac{1}{n}\sum_{q \in P} p - \frac{1}{n}\sum_{q \in P} q \right| \\
&\leq \sum_{p,q \in P} \frac{1}{n}\|pq\| \leq \sum_{p,q \in P} \frac{1}{n}(\|p\xi\| + \|q\xi\|) = 2\sum_{p \in P}\|p\xi\| = 2\nu_{\mathrm{opt}}(P, 1).
\end{aligned}$$

■

Lemma 2.3 *Let P be a set of points in $\mathbb{R}$. And let $P_1, \cdots, P_k$ be a partition of P into k sets. Then $\nu_{\mathrm{opt}}(P, 1) \geq (\mathcal{E}_\nu(P_1) + \mathcal{E}_\nu(P_2) + \cdots + \mathcal{E}_\nu(P_k))/2$, where $\mathcal{E}_\nu(P_i) = \nu_{\overline{m}(P_i)}(P_i)$.*

Proof: Let $\xi = \mathrm{Median}(P)$ and

$$\begin{aligned}
\nu_{\mathrm{opt}}(P, 1) &= \sum_{p \in P}\|p\xi\| = \sum_{p \in P_1}\|p\xi\| + \sum_{p \in P_2}\|p\xi\| + \cdots + \sum_{p \in P_k}\|p\xi\| \\
&\geq \nu_{\mathrm{opt}}(P_1, 1) + \nu_{\mathrm{opt}}(P_2, 1) + \cdots \nu(P_k, 1) \\
&\geq \frac{1}{2}(\mathcal{E}_\nu(P_1) + \mathcal{E}_\nu(P_2) + \cdots + \mathcal{E}_\nu(P_k)),
\end{aligned}$$

by Lemma 2.2. ■

Theorem 2.4 *Let P be a point set in $\mathbb{R}$, k and $\varepsilon > 0$ parameters. Then, there exists a (k, ε)-coreset $\mathcal{S}$ of P of size $O(k/\varepsilon)$.*

Proof: Assume that we have an approximation V, such that $\nu_{\mathrm{opt}}(P, k) \leq V \leq c\nu_{\mathrm{opt}}(P, k)$, where c is a constant (this can be done efficiently in linear time for small k [HM04]). We scan the points from left to right and group them into batches with cumulative error equal to $\phi = \frac{\varepsilon}{10ck}V$. This is done by allowing the first and last point in the batch to be a fraction of a point of P (i.e., a point of P might appear in two consecutive batches, as two points with total weight one). The last batch is of weight $\leq \phi$. Observe that $\phi \geq \frac{\varepsilon}{10ck}\nu_{\mathrm{opt}}(P, k)$. Let $\mathcal{B} = \{B_1, \ldots, B_u\}$ denote the resulting batches.

It is now straightforward to verify that $|\mathcal{B}| = O(k/\varepsilon)$. Indeed, let C_{opt} be the set of k medians realizing $\nu_{\mathrm{opt}}(P, k)$. Since P is a one dimensional point set, there are at most $k - 1$ batches that are being served by more than one center in C_{opt}. For any other batches $B \in \mathcal{B}$, it is being served by a single center of C_{opt}. Thus, we have $\nu_{C_{\mathrm{opt}}}(B) \geq \mathcal{E}_\nu(B)/2 = \phi/2$, by Lemma 2.2. Thus, the number of batches is bounded by $O(k + \nu_{\mathrm{opt}}(P, k)/\phi) = O(k/\varepsilon)$.

Next, for the coreset construction, we set $\overline{m}(B_i)$ to be the representative point for B_i with weight $|B_i|$. Let $\mathcal{S}$ be the resulting coreset. We claim that this is a (k, ε)-coreset. Indeed,

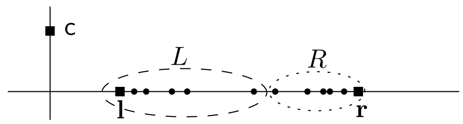

Figure 1: Case (i) of Lemma 2.5

consider any point set $C = \{x_1, x_2, \dots x_k\}$. For a point $x_i \in C$, let I_i denote the interval on the real line that it serves. For a batch B, let $\mathcal{I}(B)$ denote the smallest interval containing B. If a batch $B \subseteq \mathcal{I}_i$, and $x_i \notin \mathcal{I}(B)$, then by Lemma 2.1, we have $\nu_{x_i}(B) = |B| \cdot \|\overline{\mathrm{m}}(B)x_i\|$. Namely, the contribution of the points of B to $\nu_C(P)$ and $\nu_C(\mathcal{S})$ are identical.

Thus, the only batches that might contribute to the error, are the ones that contain an endpoint of $I_1, \dots, I_k$ (there are most $k-1$ such batches), and batches that contain a point of C in their interior (there are at most k such batches). By Lemma 2.1 (iii), every such batch B contributes at most $\mathcal{E}_\nu(B)$ to the overall error. Let $B'_1, \dots, B'_{2k-1}$ be those "problematic" batches. We have that

$$|\nu_C(P) - \nu_C(\mathcal{S})| \leq \sum_{i=1}^{2k-1} \mathcal{E}_\nu(B'_i) \leq (2k-1) \cdot \phi \leq \varepsilon \nu_{\mathrm{opt}}(P, k).$$

■

2.3 Extending to higher Dimension

We need the following technical lemma.

Lemma 2.5 *Let $\mathsf{c} = (0, \alpha)$ be a point in the plane, let L and R be two weighted sets of points on the positive portion of the x-axis such that all the points of L have smaller x-axis value than the points of R, and let $\mathbf{l}$ and $\mathbf{r}$ be two points on the x-axis such that $\nu_{\mathbf{l}}(L) = \nu_{\mathbf{r}}(R)$. Furthermore, let $\mathcal{S}_L = \{(\mathbf{l}, w(L))\}$ and $\mathcal{S}_R = \{(\mathbf{r}, w(R))\}$, be the respective coresets. Also, let $\mathcal{E} = \nu_{\mathsf{c}}(L) + \nu_{\mathsf{c}}(R) - \nu_{\mathsf{c}}(\mathcal{S}_L) - \nu_{\mathsf{c}}(\mathcal{S}_R)$ be the error caused by using the coresets $\mathcal{S}_L$ and $\mathcal{S}_R$ instead of the sets L and R, respectively, in relation to the center c. Then*

(i) If $x_{\mathbf{l}} \leq x_{l'} \leq x_{r'} \leq x_{\mathbf{r}}$, for all $l' \in L$ and $r' \in R$, then $\mathcal{E} \leq 0$. See Figure 1.

(ii) If $x_{l'} \leq x_{\mathbf{l}} \leq x_{\mathbf{r}} \leq x_{r'}$, for all $l' \in L$ and $r' \in R$, then $\mathcal{E} \geq 0$.

Proof: (i) For two points, p, q on the x-axis, such that $x_p \leq x_q$, let $e(p, q) = \|q\mathsf{c}\| - \|p\mathsf{c}\|$. In particular, for any four points a, b, c, d on the x-axis, such that $x_a \leq x_b \leq x_c \leq x_d$, we have $e(a, b)/\|ab\| \leq e(c, d)/\|cd\|$. This follows since the function $f(x) = \|\mathsf{c} - (x, 0)\|$ is a convex function with positive second derivative, as can be easily verified. In particular, for any $a \leq b$ we have $f'(a) \leq e(a, b)/\|ab\| \leq f'(b)$. Thus, for a point z on the real line between R and L, we have

$$\begin{aligned}
\mathcal{E} &= \nu_{\mathsf{c}}(L) + \nu_{\mathsf{c}}(R) - \nu_{\mathsf{c}}(\mathcal{S}_L) - \nu_{\mathsf{c}}(\mathcal{S}_R) = \sum_{p \in L} w_p (\|\mathsf{c}p\| - \|\mathsf{c}\,\mathbf{l}\|) + \sum_{p \in R} w_p (\|\mathsf{c}p\| - \|\mathsf{c}\,\mathbf{r}\|) \\
&= \sum_{p \in L} w_p e(\mathbf{l}, p) - \sum_{p \in R} w_p e(p, \mathbf{r}) = \sum_{p \in L} w_p \|p\mathbf{l}\| \frac{e(\mathbf{l}, p)}{\|p\mathbf{l}\|} - \sum_{p \in R} w_p \|\mathbf{r}p\| \frac{e(p, \mathbf{r})}{\|\mathbf{r}p\|} \\
&\leq \sum_{p \in L} w_p \|p\mathbf{l}\|\, f'(z) - \sum_{p \in R} w_p \|\mathbf{r}p\|\, f'(z) = f'(z)(\nu_{\mathbf{l}}(L) - \nu_{\mathbf{r}}(R)) = 0.
\end{aligned}$$

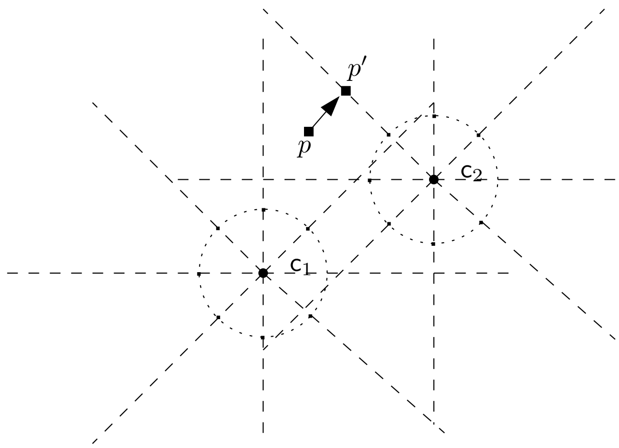

Figure 2: Computing the coreset.

since $e(\mathbf{l}, p)/\|p\mathbf{l}\| \leq f'(z) \leq e(q, \mathbf{r})/\|q\mathbf{r}\|$ for any $p \in L$ and $q \in R$.

The second claim follows by similar argumentation. ■

2.3.1 Construction

The following lemma states the existence of a ε-net for the sphere. See [Mat02] for details.

Lemma 2.6 *There exists a set of points Q on a sphere of unit radius in d-dimensions centered at the origin with the following properties: (i) Q has $O(\varepsilon^{-(d-1)})$ points, and (ii) $\forall p$ that lie on the unit radius ball, $\exists q \in Q$ such that $\|pq\| \leq \varepsilon$. Furthermore, Q can be computed in $O(\varepsilon^{-(d-1)})$ time.*

We compute a set $\mathsf{C} = \{\mathsf{c}_1, \ldots, \mathsf{c}_k\}$ of centers which is a c-approximation to $\nu_{\mathrm{opt}}(P, k)$; namely, $\nu_{\mathsf{C}}(P) \leq c\nu_{\mathrm{opt}}(P, k)$, where c is a constant. Now, we divide the set of points P into k sets based on which point in C is nearest to them. This gives us a partition of P into k subsets $P_1, P_2, \ldots P_k$ where P_i is closest to $\mathsf{c}_i \in \mathsf{C}$. Around each of the points $\mathsf{c}_i \in \mathsf{C}$ we place a fan $\mathcal{L}_i$ of lines passing through it. This is done by taking a unit sphere centered at c_i, and placing an $\varepsilon/(3c)$-net N_{c_i} on this sphere, using Lemma 2.6. For every $p \in N_{\mathsf{c}_i}$, we generate the line spanning the segment $\mathsf{c}_i p$.

For each point of $p \in P_i$, let ℓ_p be its closest line in $\mathcal{L}_i$, and let p' be the projection of p into ℓ_p. Let the new set of snapped points be P'. Next, we compute a coreset for each of the lines using the one dimensional method. Namely, we scan every line ℓ, and break the point set, P_ℓ, along it into batches, such that for each batch B (except the last one), we have $\mathcal{E}_\nu(B) = A_\ell/(20c\varepsilon k)$, where A_ℓ is a c-approximation to $\nu_{\mathrm{opt}}(P_\ell, k)$ (again, allowing a boundary point to appear in two batches with a fractional weight). See Figure 2. (In the following, for the sake of simplicity of exposition, we ignore the fact that a batch contains weighted points. This is a minor technicality, and it can be easily handled.)

Hence, we get $O(k/\varepsilon)$ points selected in the coreset on each of the lines through c_i, and $O(k/\varepsilon^d)$ coreset points for P_i. Thus, the total number of points in the coreset $\mathcal{S}$ is $O(k^2/\varepsilon^d)$.

2.3.2 Correctness

Observation 2.7 *Let p be a point of P, and let x_i be its nearest point in C, let p' be the corresponding point in P'. We have $\|pp'\| \leq \|px_i\| \varepsilon/(3c)$.*

Lemma 2.8 *Let P be a set of points on a line ℓ, and let $\mathcal{S}_\ell$ be the coreset constructed for it. Also, let C be a set of k points in $\mathbb{R}^d$. Then $|\nu_{\mathsf{C}}(P) - \nu_{\mathsf{C}}(\mathcal{S}_\ell)| \leq (\varepsilon/3)\nu_{\mathrm{opt}}(P, k)$.*

Proof: The proof follows the one dimensional case (i.e., Theorem 2.4), although the analysis is somewhat more involved. We rotate and translate all the points so that the line ℓ coincides with the x-axis. Let $\mathsf{C} = \{\mathsf{c}_1, \ldots, \mathsf{c}_k\}$, and let $\mathsf{c}'_1, \ldots, \mathsf{c}'_k$ denote the projection of $\mathsf{c}_1, \ldots, \mathsf{c}_k$ into ℓ, respectively.

Next, we partition the line into k intervals $\mathcal{I}_1, \ldots, \mathcal{I}_k$, such that $\mathcal{I}_i$ is the portion of ℓ closer to c_i than any other point of C (note that the points of C are not necessarily on ℓ).

The every point in the coreset $\mathcal{S}_\ell$ corresponds to a subset (i.e., batch) of P. By construction, all the batches have the same cumulative error (except the last batch, which might have smaller cumulative error). In particular, for any batch B we have that the cumulative error $\mathcal{E}_\nu(B) \leq (\varepsilon/20k)\nu_{\mathrm{opt}}(P, k)$.

Let $\widehat{B}$ be the set of all batches B, which are served by more than one center of C, or alternatively, the interval $\mathcal{I}(B)$ contains the projection of a center point of C to ℓ. We also add the last batch on ℓ to $\widehat{B}$. Clearly, $\left|\widehat{B}\right| \leq 2k$. Let $U = \cup_{B \in \widehat{B}} B$ be the points of P in $\widehat{B}$, and let $\mathcal{S}_U \subseteq \mathcal{S}_\ell$ be the corresponding coreset. It follows that the total error contributed by the points of U is

$$E^* = |\nu_C(U) - \nu_C(\mathcal{S}_U)| \leq \sum_{B \in \widehat{B}} \mathcal{E}_\nu(B) \leq 2k \frac{\varepsilon}{20k} \nu_{\mathrm{opt}}(P, k) \leq \frac{\varepsilon}{10} \nu_{\mathrm{opt}}(P, k),$$

by Lemma 2.1 (iii).

Let us fix a center $\mathsf{c} \in C$, and let $\mathcal{I}$ be its Voronoi cell on ℓ. Next, consider the set R (resp. L) of the batches to the right (resp. left) of c' that lie in its interval $\mathcal{I}$. Let $B^1, B^2, \ldots, B^t$ denote the batches of R sorted from left to right. Furthermore, let L^i and R^i be the set of points of B^i, to the left and right of the mean $\overline{\mathrm{m}}(B^i)$, respectively, for $i = 1, \ldots, t$. Finally, let $\mathcal{S}_l^i$ and $\mathcal{S}_r^i$ denote the coresets formed by placing a point at $\overline{\mathrm{m}}(B^i)$ with weight $w(L^i)$ and $w(R^i)$, respectively. Let $\mathcal{S}^i = \mathcal{S}_l^i \cup \mathcal{S}_r^i$ be the one point coreset placed at $\overline{\mathrm{m}}(B^i)$ with weight $w(L^i) + w(R^i)$. Let $\mathcal{S}_R$ denote the resulting coreset for all the batches in R. Let P_R denote the points in R. By Lemma 2.5 (ii), we have that

$$\nu_{\mathsf{c}}(B^i) - \nu_{\mathsf{c}}(\mathcal{S}^i) = \nu_{\mathsf{c}}(L^i) - \nu_{\mathsf{c}}(\mathcal{S}_l^i) + \nu_{\mathsf{c}}(R^i) - \nu_{\mathsf{c}}(\mathcal{S}_r^i) \geq 0.$$

Thus, the error contributed by the coreset of R is $E = \nu_{\mathsf{c}}(P_R) - \nu_{\mathsf{c}}(\mathcal{S}_R) = \sum_{i=1}^{t} (\nu_{\mathsf{c}}(B^i) - \nu_{\mathsf{c}}(\mathcal{S}^i)) \geq 0$. On the other hand, by Lemma 2.5 (i), we have

$$E' = \sum_{i=1}^{t-1} \bigl(\nu_{\mathsf{c}}(R^i) - \nu_{\mathsf{c}}(\mathcal{S}_r^i) + \nu_{\mathsf{c}}(L^{i+1}) - \nu_{\mathsf{c}}(\mathcal{S}_l^{i+1})\bigr) \leq 0.$$

Thus,

$$0 \leq E = E' + \nu_{\mathsf{c}}(L^1) - \nu_{\mathsf{c}}(\mathcal{S}_l^1) + \nu_{\mathsf{c}}(R^t) - \nu_{\mathsf{c}}(\mathcal{S}_r^t) \leq \mathcal{E}_\nu(B^1) + \mathcal{E}_\nu(B^t) = 2(\varepsilon/20k)\nu_{\mathrm{opt}}(P, k).$$

Namely, the total error induced by using the coreset for batches in R is bounded by $2(\varepsilon/20k)\nu_{\text{opt}}(P,k)$. By symmetry, the same hold of the batches in L. Thus, the total error induced by such batches is $k \cdot 2 \cdot 2(\varepsilon/20k)\nu_{\text{opt}}(P,k) \leq (\varepsilon/5)\nu_{\text{opt}}(P,k)$. Thus, we have

$$|\nu_{\mathsf{C}}(P) - \nu_{\mathsf{C}}(\mathcal{S}_\ell)| \leq (\varepsilon/5)\nu_{\text{opt}}(P,k) + (\varepsilon/10)\nu_{\text{opt}}(P,k) \leq (\varepsilon/3)\nu_{\text{opt}}(P,k)$$

as desired. ■

Theorem 2.9 *Let P be a point set of n points in $\mathbb{R}^d$, and let $\mathcal{S}$ be the coreset constructed for it in Section 2.3.1. Then $\mathcal{S}$ is a weighted set of size $O(k^2/\varepsilon^d)$ and it is a (k,ε)-coreset of P for the k-median clustering.*

Proof: By Lemma 2.8 we know that the error between the distance of any set C of size k and the snapped points P' on the fans can be well approximated using the coreset. Furthermore, the error introduced by the snapping is bounded by

$$E = \sum_{p \in P} \|pp'\| \leq \sum_{i=1}^{k} \left(\sum_{p \in P_i} (\varepsilon/3c)\|px_i\| \right) \leq \frac{\varepsilon}{3}\nu_{\text{opt}}(P,k),$$

by Observation 2.7.

So, $|\nu_{\mathsf{C}}(P) - \nu_{\mathsf{C}}(P')| \leq \sum_{p \in P}\|pp'\| \leq \frac{\varepsilon}{3}\nu_{\text{opt}}(P,k)$. Thus, for any set C of k points, we have

$$\begin{aligned} |\nu_{\mathsf{C}}(P) - \nu_{\mathsf{C}}(\mathcal{S})| &= |\nu_{\mathsf{C}}(P) - \nu_{\mathsf{C}}(P')| + |\nu_{\mathsf{C}}(P') - \nu_{\mathsf{C}}(\mathcal{S})| \\ &\leq (\varepsilon/3)\nu_{\text{opt}}(P,k) + (\varepsilon/3)\nu_{\text{opt}}(P',k) \\ &\leq (\varepsilon/3)\nu_{\text{opt}}(P,k) + (\varepsilon/3)\nu_{\text{opt}}(P,k)(1+\varepsilon/3) \leq \varepsilon\nu_{\text{opt}}(P,k), \end{aligned}$$

by Lemma 2.8. ■

3 Coreset for k-means

3.1 Preliminaries

k-mean clustering. Let $\mu_C(P) = \sum_{p \in P} w_p \cdot (\mathbf{d}(p,C))^2$ denote the price of the k-means clustering of P as provided by the set of centers C. Let $\mu_{\text{opt}}(P,k) = \min_{C \subseteq \mathbb{R}^d, |C|=k} \mu_C(P)$ denote the price of the *optimal k-means clustering* of P. In the following, we will abuse notation, and for $x \in \mathbb{R}^d$, we will denote by $\nu_x(P)$ and $\mu_x(P)$ the quantities $\nu_{\{x\}}(P)$ and $\mu_{\{x\}}(P)$, respectively.

(k,ε)-coreset for k-mean. Similarly, $\mathcal{S}$ is a (k,ε)-*coreset* of P for the k-means clustering, if for any set C of k points in $\mathbb{R}^d$, we have $(1-\varepsilon)\mu_C(P) \leq \mu_C(\mathcal{S}) \leq (1+\varepsilon)\mu_C(P)$.

Centroid Set. Given a set P of n points in $\mathbb{R}^d$, a set $\mathcal{D} \subseteq \mathbb{R}^d$ is an (k,ε)-*approximate centroid set* for P, if there exists a subset $C \subseteq \mathcal{D}$ of size k, such that $\nu_C(P) \leq (1+\varepsilon)\nu_{\text{opt}}(P,k)$.

Definition 3.1 For a point set P, the *error* of P is $\widehat{\mathcal{E}}(P) = \sum_{p \in P}\|p\overline{\mathrm{m}}\|^2$, where $\overline{\mathrm{m}} = \overline{\mathrm{m}}(P)$.

3.2 The 1D case

3.2.1 Construction

Let P be a given set of n points on the real line. We consider the points from left to right and group them into batches, such that a batch B has $\widehat{\mathcal{E}}(B) \le \xi$, and for two consecutive batches B and B' we have $\widehat{\mathcal{E}}(B \cup B') \ge \xi$, where $\xi \le \frac{\varepsilon^2}{100k^2}\mu_{\mathrm{opt}}(P,k)$. As in the k-median case, the number of batches is $O(k^2/\varepsilon^2)$. Let $\mathcal{B}(P)$ denote the resulting set of batches.

Lemma 3.2 *Let B be a set points on a line. There exist two weighted points (q_1, w_1) and (q_2, w_2) both lying completely within $\mathcal{I}(B)$, such that*

- $w_1 + w_2 = |B|$,
- $\frac{q_1 w_1 + q_2 w_2}{w_1 + w_2} = \overline{\mathrm{m}}$, *where* $\overline{\mathrm{m}} = \overline{\mathrm{m}}(P)$,
- *and* $w_1 \|q_1\overline{\mathrm{m}}\|^2 + w_2\|q_2\overline{\mathrm{m}}\|^2 = \sum_{p\in B}\|p\overline{\mathrm{m}}\|^2$.

Let $\mathcal{T}(B) = \{(q_1, w_1), (q_2, w_2)\}$ denote this coreset.

Proof: We will construct these weighted points through a sequence of steps. Let the leftmost point in B be p_l and the rightmost point be p_r.

- For every point $p \in B$ to the right of $\overline{\mathrm{m}}$, we add a point at the rightmost extreme of B with weight $\frac{\|p\overline{\mathrm{m}}\|}{\|p_r\overline{\mathrm{m}}\|}$. Clearly, $\frac{\|p\overline{\mathrm{m}}\|}{\|p_r\overline{\mathrm{m}}\|}\|p_r\overline{\mathrm{m}}\|^2 \ge \|p\overline{\mathrm{m}}\|^2$. Similarly for every point $p \in B$ to the left of $\overline{\mathrm{m}}$ we add a point at the leftmost extreme of B with weight $\frac{\|p\overline{\mathrm{m}}\|}{\|p_l\overline{\mathrm{m}}\|}$. This results into weighted points p_l, p_r. Furthermore, we have $\overline{\mathrm{m}}(p_l, p_r) = \overline{\mathrm{m}}$, $\widehat{\mathcal{E}}(\{p_l, p_r\}) \ge \widehat{\mathcal{E}}(B)$, and $w_{p_l} + w_{p_r} \le |B|$.
- Now, we scale up the weights so that $w_{p_l} + w_{p_r} = n$. Note that this does not change the mean, and only increases $\widehat{\mathcal{E}}(\{p_l, p_r\})$.
- Finally, consider the scaled set $C(t) = \{(p_l \cdot t + (1-t)\overline{\mathrm{m}}, w_{p_l}), (p_r \cdot t + (1-t)\overline{\mathrm{m}}, w_{p_r})\}$. Clearly, $C(t)$ for $t \in [0,1]$ has $\overline{\mathrm{m}}(C(t)) = \overline{\mathrm{m}}$. Furthermore, $C(1)$ is just the current two weighed points, and $C(0)$ is just one point at $\overline{\mathrm{m}}$. Thus, pick $t^* \in [0,1]$, such that $\widehat{\mathcal{E}}(C(t^*)) = \widehat{\mathcal{E}}(B)$. This is possible, since $\widehat{\mathcal{E}}(C(1)) \ge \widehat{\mathcal{E}}(B)$.

  Clearly, $C(t^*)$ is the required coreset.

■

Let $\mathcal{S}(P) = \cup_{B\in\mathcal{B}(P)}\mathcal{T}(B)$ be the constructed coreset for P.

3.2.2 Correctness

The following claim is well known [KMN+02, HS04].

Lemma 3.3 *Let B be a set of points in $\mathbb{R}^d$, then for any $q \in \mathbb{R}^d$, we have $\mu_q(B) = |B|\|q\overline{\mathrm{m}}\|^2 + \widehat{\mathcal{E}}(B)$,*

Lemma 3.4 *Let B be a set of points in $\mathbb{R}^d$, and let $\mathcal{T} = \mathcal{T}(B)$, and q any point in $\mathbb{R}^d$. Then $\mu_q(B) = \mu_q(\mathcal{T})$.*

Proof: We have $\mu_q(B) = |B|\|q\overline{\mathrm{m}}\|^2 + \widehat{\mathcal{E}}(B)$, and $\mu_q(\mathfrak{T}) = w(\mathfrak{T})\|q\overline{\mathrm{m}}(\mathfrak{T})\| + \widehat{\mathcal{E}}(\mathfrak{T}) = |B|\|q\overline{\mathrm{m}}\|^2 + \widehat{\mathcal{E}}(B)$, by Lemma 3.2. Thus, $\mu_q(B) = \mu_q(\mathfrak{T})$. ■

Theorem 3.5 *Let P be a set of n points in $\mathbb{R}^d$, such that the points of P all lie on a line ℓ, and let $\mathcal{S}$ be the coreset constructed for it in Section 3.2.1. Then $\mathcal{S}$ is a $(k, \varepsilon/3)$-coreset for k-means clustering of P, for any set of k centers in $\mathbb{R}^d$.*

Proof: The proof is similar to the k-median case. We first rotate space, such that ℓ is on the x-axis. Let $\mathsf{C} = \{\mathsf{c}_1, \ldots, \mathsf{c}_k\}$ be a set of k centers, $\mu_{\mathsf{C}} = \mu_{\mathsf{C}}(P)$ and $\mu'_{\mathsf{C}} = \mu_{\mathsf{C}}(\mathcal{S})$. Let $\mathcal{I}_1, \ldots, \mathcal{I}_k$ be a partition of the line into intervals, such that $\mathcal{I}_i$ is the loci of points closest to c_i out of all the centers in C, for $i = 1, \ldots, k$. The batches of $\mathcal{B}(P)$, and their corresponding coreset points, that lie completely within $\mathcal{I}_i$, do not contribute to the overall error $|\mu_{\mathsf{C}} - \mu'_{\mathsf{C}}|$ by Lemma 3.4.

Thus, the only problematic batches, are the one that contain an endpoint of $\mathcal{I}_1, \ldots, \mathcal{I}_k$. There are at most $k-1$ such batches. Let B be one such batch. Assume that the interval $\mathcal{I}(B)$ intersects $\mathcal{I}_1, \ldots, \mathcal{I}_t$, and let $V_i = \mathcal{I}_i \cap B$, for $i = 1, \ldots, t$. Let $\overline{\mathrm{m}} = \overline{\mathrm{m}}(B)$ and let $\mathcal{S}_B = \mathfrak{T}(B)$. We partition $\mathcal{S}_B$ into portions corresponding the sets $V_1, \ldots, V_t$. Formally, $\mathcal{S}_i$ is a set of the two points of $\mathcal{S}_B$, re-weighted such that $w(\mathcal{S}_i) = |V_i|$, for $i = 1, \ldots, t$. We have, by Lemma 3.3, that

$$\begin{aligned} \mu_{\mathsf{C}}(\mathcal{S}_B) &= \sum_i \mu_{\mathsf{C}}(\mathcal{S}_i) \le \sum_i \mu_{\mathsf{c}_i}(\mathcal{S}_i) = \sum_i \left(\widehat{\mathcal{E}}(\mathcal{S}_i) + |V_i| \|\mathsf{c}_i\overline{\mathrm{m}}\|^2 \right) \\ &= \sum_i \left(\widehat{\mathcal{E}}(\mathcal{S}_i) + \sum_{p \in V_i} \|\mathsf{c}_i\overline{\mathrm{m}}\|^2 \right) = \widehat{\mathcal{E}}(\mathcal{S}_B) + \left(\sum_i \sum_{p \in V_i} \|\mathsf{c}_i\overline{\mathrm{m}}\|^2 \right). \end{aligned}$$

Let $P_i = \left\{ p \in V_i \;\middle|\; \|\mathsf{c}_i p\|^2 - \|p\overline{\mathrm{m}}\|^2 \ge 0 \right\}$ and $N_i = V_i \setminus P_i$. Since $\|\mathsf{c}_i p\| - \|p\overline{\mathrm{m}}\| \le \|\mathsf{c}_i\overline{\mathrm{m}}\|$, we have

$$\begin{aligned} \sum_i \sum_{p \in V_i} \|\mathsf{c}_i\overline{\mathrm{m}}\|^2 &\ge \sum_i \sum_{p \in P_i} (\|\mathsf{c}_i p\| - \|p\overline{\mathrm{m}}\|)^2 \\ &\ge \sum_i \sum_{p \in P_i} (\|\mathsf{c}_i p\| - \|p\overline{\mathrm{m}}\|)^2 + \sum_i \sum_{p \in N_i} \left(\|\mathsf{c}_i p\|^2 - \|p\overline{\mathrm{m}}\|^2 \right) \\ &\ge \sum_i \sum_{p \in V_i} \|\mathsf{c}_i p\|^2 - 2 \sum_i \sum_{p \in V_i} \|\mathsf{c}_i p\| \cdot \|p\overline{\mathrm{m}}\| - \sum_i \sum_{p \in V_i} \|p\overline{\mathrm{m}}\|^2 \\ &\ge \mu_{\mathsf{C}}(B) - 2 \sum_i \sum_{p \in V_i} \|\mathsf{c}_i p\| \cdot \|p\overline{\mathrm{m}}\| - \widehat{\mathcal{E}}(B). \end{aligned}$$

We also have $\|\mathsf{c}_i p\| + \|p\overline{\mathrm{m}}\| \ge \|\mathsf{c}_i\overline{\mathrm{m}}\|$ and so,

$$\sum_i \sum_{p \in V_i} \|\mathsf{c}_i\overline{\mathrm{m}}\|^2 \le \sum_i \sum_{p \in V_i} (\|\mathsf{c}_i p\| + \|p\overline{\mathrm{m}}\|)^2 \le \mu_{\mathsf{C}}(B) + 2 \sum_i \sum_{p \in V_i} \|\mathsf{c}_i p\| \cdot \|p\overline{\mathrm{m}}\| + \widehat{\mathcal{E}}(B).$$

We conclude that $\left| \sum_i \sum_{p \in V_i} \|\mathsf{c}_i\overline{\mathrm{m}}\|^2 - \mu_{\mathsf{C}}(B) \right| \le 2 \sum_i \sum_{p \in V_i} \|\mathsf{c}_i p\| \cdot \|p\overline{\mathrm{m}}\| + \widehat{\mathcal{E}}(B).$

This gives us,

$$
\begin{aligned}
|\mu_{\mathsf{C}}(\mathcal{S}_B) - \mu_{\mathsf{C}}(B)| &\leq 2\sum_i \sum_{p \in V_i} \|\mathsf{c}_i p\| \cdot \|p\overline{\mathrm{m}}\| + 2\widehat{\mathcal{E}}(B) \\
&\leq 2\widehat{\mathcal{E}}(B) + 2\sqrt{\sum_i \sum_{p \in V_i} \|\mathsf{c}_i p\|^2} \sqrt{\sum_i \sum_{p \in V_i} \|p\overline{\mathrm{m}}\|^2} \\
&\leq 2\widehat{\mathcal{E}}(B) + 2\sqrt{\mu_{\mathsf{C}}(B)}\sqrt{\widehat{\mathcal{E}}(B)},
\end{aligned}
$$

by the Cauchy-Swartz inequality. By construction $\widehat{\mathcal{E}}(B) \leq (\varepsilon^2/100k^2)\mu_{\mathrm{opt}}(P,k)$. Thus,

$$
\begin{aligned}
|\mu_{\mathsf{C}}(\mathcal{S}_B) - \mu_{\mathsf{C}}(B)| &\leq 2\frac{\varepsilon^2}{100k^2}\mu_{\mathrm{opt}}(P,k) + 2\frac{\varepsilon}{10k}\sqrt{\mu_{\mathsf{C}}(B)\mu_{\mathrm{opt}}(P,k)} \\
&\leq 2\frac{\varepsilon^2}{100k^2}\mu_{\mathrm{opt}}(P,k) + 2\frac{\varepsilon}{10k} \cdot \frac{\mu_{\mathsf{C}}(B) + \mu_{\mathrm{opt}}(P,k)}{2} \\
&\leq \frac{\varepsilon}{5k}\mu_{\mathrm{opt}}(P,k) + \frac{\varepsilon}{10k}\mu_{\mathsf{C}}(B).
\end{aligned}
$$

Since there are $k-1$ border batches, we conclude that

$$
|\mu_{\mathsf{C}}(\mathcal{S}) - \mu_{\mathsf{C}}(P)| \leq \frac{\varepsilon}{5}\mu_{\mathrm{opt}}(P,k) + \frac{\varepsilon}{10}\mu_{\mathsf{C}}(P) \leq \frac{\varepsilon}{3}\mu_{\mathsf{C}}(P),
$$

as required. ■

3.3 Extending to higher dimension

Again we use a similar approach to the one we used for k-means. We calculate an approximation $\mu_{\mathrm{opt}}(P,k) \leq A \leq c\mu_{\mathrm{opt}}(P,k)$, where $c > 1$ is a constant. Then we partition the point set P into sets $P_1, P_2, \ldots P_k$ with P_i consisting of points in the area of control of $\mathsf{c}_i \in A$. Then we draw $O(\frac{1}{\varepsilon^{d-1}})$ lines through each of the centers of A as before and snap the points of P_i onto the closest line around c_i. We compute a coreset for every line using the algorithm of Section 3.2.1. This gives us $O(\frac{k^2}{\varepsilon^2})$ points selected for the coreset on every line, thus the total size of the resulting coreset $\mathcal{S}$ is $O(\frac{k^3}{\varepsilon^{d+1}})$.

The resulting set is the required coreset. The proof is an easy extension of the one dimensional case. Indeed, the snapping into the lines introduces a multiplicative error smaller than $\varepsilon/3$. The coreset construction introduces an error of similar magnitude, by Theorem 3.5. Since this is a straightforward extension of our previous discussion, we omit any further details.

Theorem 3.6 *Given a set P of n points in $\mathbb{R}^d$, one can compute a (k,ε)-coreset for P for k-means clustering of size $O(k^3/\varepsilon^{d+1})$.*

Matoušek showed that there exists an ε-approximate centroid set of size $O(n\varepsilon^{-d}\log(1/\varepsilon))$. Interestingly enough, his construction is weight insensitive. In particular, using a $(k,\varepsilon/2)$-coreset $\mathcal{S}$ in his construction, results in a ε-approximate centroid set of size $O\big(|\mathcal{S}|\,\varepsilon^{-d}\log(1/\varepsilon)\big)$.

Theorem 3.7 *Given a set P of n points in $\mathbb{R}^d$, one can compute a (k, ε)-centroid set for P for k-means clustering of size $O(k^3/\varepsilon^{2d+1} \log(1/\varepsilon))$.*

Theorem 3.7 slightly improves (as far as the dependency of k is concerned) over the result of Effros and Schulman [ES03] that showed that there exists a centroid set of size $O\big(\varepsilon^{-d-1}(k^4 + k^2\varepsilon^{-2})\big)$. We conjecture that the dependency on ε in the bound on the coreset size in Theorem 3.7 can be further improved.

4 Conclusions

In this paper, we showed the existence of small coresets for the k-means and k-median clustering in $\mathbb{R}^d$, with size independent of n. We believe that this result is quite surprising.

Note, that we had ignored computational issues in this paper. Our techniques do not yield any significant improvement in performance over the approximation algorithms of Har-Peled and Mazumdar [HM04]. As mentioned in the introduction, the results in this papers imply algorithms with running time $O(n + \mathrm{poly}(k, \log n, 1/\varepsilon) + \mathrm{func}(k, \varepsilon))$, where poly denotes a polynomial, and $\mathrm{func}(k, \varepsilon)$ denotes a function that depends only on k and ε (and the dimension d). This however, improves over the results of Har-Peled and Mazumdar [HM04] only for very narrow interval of values of k in the k-means case.

At this point, there are numerous problems for further research. In particular:

1. Can the running time of approximate k-means clustering be improved to be similar to the k-median bounds? Can one do FPTAS for k-median and k-means (in both k and $1/\varepsilon$)? Currently, we can only compute the (k, ε)-coreset in fully polynomial time, but cannot extract the approximation itself from it.

2. Does a coreset exist for the problem of k-median and k-means in high dimensions? There are some partial relevant results [BHI02].

3. Can one improve the bounds on the size of the coresets for k-median and k-mean clustering?

References

[Aro98] S. Arora. Polynomial time approximation schemes for euclidean tsp and other geometric problems. *J. Assoc. Comput. Mach.*, 45(5):753–782, Sep 1998.

[ARR98] S. Arora, P. Raghavan, and S. Rao. Approximation schemes for Euclidean k-median and related problems. In *Proc. 30th Annu. ACM Sympos. Theory Comput.*, pages 106–113, 1998.

[BHI02] M. Bădoiu, S. Har-Peled, and P. Indyk. Approximate clustering via coresets. In *Proc. 34th Annu. ACM Sympos. Theory Comput.*, pages 250–257, 2002.

[CGTS99] M. Charikar, S. Guha, E. Tardos, and D. B. Shmoys. A constant-factor approximation algorithm for the k-median problem. In *Proc. 31st Annu. ACM Sympos. Theory Comput.*, pages 1–10, 1999.

[DF95] R. G. Downey and M. R. Fellows. Fixed-parameter tractability and completeness i: Basic results. *SIAM J. Comput.*, 24(4):873–921, 1995.

[dlVKKR03] W. F. de la Vega, M. Karpinski, C. Kenyon, and Y. Rabani. Approximation schemes for clustering problems. In *Proc. 35th Annu. ACM Sympos. Theory Comput.*, pages 50–58, 2003.

[ES03] M. Effros and L. J. Schulman. Deterministic clustering with data nets. Technical Report TR04-085, Elec. Colloq. Comp. Complexity, 2003. `http://www.eccc.uni-trier.de/eccc-reports/2004/TR04-085/`.

[GNMO00] S. Guha, R. Motwani N. Mishra, and L. O'Callaghan. Clustering data streams. In *Proc. 41th Annu. IEEE Sympos. Found. Comput. Sci.*, pages 359–366, 2000.

[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.

[HS04] S. Har-Peled and B. Sadri. How fast is the k-means method? *Algorithmica*, 2004. To appear.

[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.

[KMN+02] 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. In *Proc. 18th Annu. ACM Sympos. Comput. Geom.*, pages 10–18, 2002.

[KR99] S. G. Kolliopoulos and S. Rao. A nearly linear-time approximation scheme for the euclidean κ-median problem. In *Proc. 7th Annu. European Sympos. Algorithms*, pages 378–389, 1999.

[KSS04] A. Kumar, Y. Sabharwal, and S. Sen. Linear time algorithms for clustering problems in any dimension. manuscript, 2004.

[Mat00] J. Matoušek. On approximate geometric k-clustering. *Discrete Comput. Geom.*, 24:61–84, 2000.

[Mat02] J. Matoušek. *Lectures on Discrete Geometry.* Springer, 2002.

[MP02] R. Mettu and C. G. Plaxton. Optimal time bounds for approximate clustering. In *In Proceedings of the 18th Conference on Uncertainty in Artifical Intelligence*, pages 344–351, 2002.

A Lower Bound on Coreset if Bounding Snapping Error

Let P be a set of n points in $\mathbb{R}^d$. The previous construction of coresets for k-means and k-median clustering by Har-Peled and Mazumdar [HM04], worked by finding a set $\mathcal{S}$, such that $\nu_{\mathcal{S}}(P) \leq \varepsilon \nu_{\mathrm{opt}}(P, k)$. This property by itself is sufficient to guarantee that $\mathcal{S}$ is (k, ε)-coreset for P. Surprisingly, the following theorem shows that, in the worst case, any set with this property must be large (i.e., size dependent on n).

Theorem A.1 *There exists a set P of n points in $\mathbb{R}$, such that for any set $\mathcal{S}$, if $\nu_{\mathcal{S}}(P) \leq \varepsilon \nu_{\mathrm{opt}}(P, 1)$ then, $|\mathcal{S}|$ is $\Omega(\frac{\log n}{\varepsilon})$.*

Proof: Consider the points n in P placed on the real line in the following way. There are $n/2^i$ points placed uniformly in the intervals $\mathcal{I}_i = (-2^{i+1}, -2^i) \cup (2^i, 2^{i+1})$ for $i = 1, 2, \ldots \log n$. Now, let S be any weighted coreset for the points of P. Let s_i denote the number of points in $S \cap \mathcal{I}_i$. It is easy to see that the contribution of the points in $\mathcal{I}_i$ towards $\mathcal{E}$ is minimized when the points of $S \cap \mathcal{I}_i$ are uniformly distributed in $\mathcal{I}_i$ and in this case the contribution is $\geq n/(4s_i)$. Also the origin is a median in this case and $\nu_{\mathrm{opt}}(P, 1) \leq 2n \log n$. Hence,

$$\frac{1}{4} \sum_{i=1}^{\log n} \frac{n}{s_i} \leq \mathcal{E} \leq \varepsilon \nu_{\mathrm{opt}}(P, 1) \leq \varepsilon 2n \log n.$$

This gives us,

$$\sum_{i=1}^{\log n} \frac{1}{s_i} \leq 8\varepsilon \log n,$$

implying that

$$|S| = \sum_{i=1}^{\log n} s_i \geq \frac{\log n}{8\varepsilon}.$$

■

This testifies that our more involved analysis (i.e., Theorem 2.9) to get a better coreset of size independent of n is indeed necessary. In particular, our improved coreset construction works since it guarantees that the errors introduced by snapping the points to the coreset cancel themselves out when considering any set of k medians.