Poly(k/ε)-Sized Coresets for Fair k-Clustering and Other Cool Cardinality Constraints

Anonymous Authors[1]

Abstract

TODO

1. Preliminaries and Problem Definitions

Throughout this paper, we consider d-dimensional Euclidean spaces, i.e. the distance between two points a and b is defined as $\|a-b\| = \sqrt{\sum_{i=1}^{d}(a_i-b_i)^2}$. We use $[n]$ to the denote the positive integers from 1 to n. The most general problem we consider is the (k,z)-CLUSTERING clustering problem, where, given a point set X, a set of centers S, a positive integer z, and an assignment $s: X \to S$, we have

$$\text{cost}(X,S) := \sum_{p\in X} \|p - s(p)\|^z.$$

Special cases include $z=1$, which corresponds to Euclidean k-median, and $z=2$, which corresponds to Euclidean k-means. For ease of presentation, we present most of our results for Euclidean k-MEDIAN with several constraints, including the ones described above. Our algorithms and analysis can be extended to k-MEANS and the more general (k,z)-CLUSTERING problems. Details can be found in the supplementary material. We focus on assignments resulting from cardinality constraints, i.e. we are also given a mapping $w: C \to [|X|]$ with $\sum_{s_i\in S} w_i = |X|$ and the mapping s is the mapping minimizing the cost such that the number of points served by S_i is w_i. To emphasize that the mapping depends on w, we write $\text{cost}(X,w,S)$. The set containing all feasible solutions (S,w) is denoted by $\mathcal{S}$.

Our aim is to compute coresets for the following evaluation queries.

Coresets for Clustering with Cardinality Constraints Let $X \subseteq \mathbb{R}^d$ be a point set. Let $S \subseteq \mathbb{R}^d$ be a set of at most k centers and let $w: C \to [|X|]$ be a set of cardinalities such that $\sum_{s_i\in S} w_i = |X|$. We say that a weighted subset $D \subseteq X$ is a coreset for clustering with cardinality constraints if for every S and every w, we have

$$|\text{cost}(X,w,S) - \text{cost}(D,w,S)| \le \varepsilon \cdot \text{cost}(X,w,S).$$

We also need something like this:

Definition 1.1. Let $X \subset \mathbb{R}^d$ be a set of points in d-dimensional Euclidean space and let k and z be two positive integers. Let $\varepsilon > 0$ be a precision parameter. Given a set of centers $\mathcal{C}$, a set $\mathbb{S}$ is an *approximate centroid set* for (k,z)-clustering on P if it satisfies the following property.

For every set of k centers $\mathcal{S} \subset \mathbb{R}^d$, there exists $\tilde{\mathcal{S}} \in \mathbb{S}^k$ and a bijection $b: \mathcal{S} \to \tilde{\mathcal{S}}$ such that for all points $p \in P$ and all $s \in \mathcal{S}$ with $\text{cost}(p,s) \le \left(\frac{4z}{\varepsilon}\right)^z \text{cost}(p,\mathcal{C})$, it holds

$$|\text{cost}(p,s) - \text{cost}(p,b(s))| \le \varepsilon \left(\text{cost}(p,s) + \text{cost}(p,\mathcal{C})\right).$$

This is slightly stronger than the definition from (Cohen-Addad et al., 2021). I believe it is necessary to account for cardinalities.

2. Algorithm

To compute a coreset with cardinality constraints, we start with perform the following steps.

1. Find a constant factor bicriteria approximation A[1].
2. For every cluster C_i with center c_i of A, let $\Delta_i := \frac{\text{cost}(C_i,A)}{|C_i|}$ denote the average cost. Define the ring $R_{i,j}$ to be all the points at distance $[2^j \cdot \Delta_i, 2^{j+1} \cdot \Delta_i)$.
3. **Near Points** We denote the set of points in $R_{i,j}$ with $j < 2\log\varepsilon$ by N_i. We add $|C_i \cap N_i|$ copies of c_i to the coreset.
4. **Ring Points** For each $R_{i,j}$ with $\log\varepsilon \le j < 2\log\varepsilon^{-1}$, we sample a set $\Omega(R_{i,j})$ of $\delta = \tilde{O}(\varepsilon^{-3}k\cdot\max(d,\varepsilon^{-2}))$ points uniformly at random, weighted by $\frac{|R_{i,j}\cap C_i|}{\delta}$.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]An (α,β) bicriteria approximation is a solution A such that the cost of A is at most a factor α times the cost of an optimal solution and the number of clusters in A is at most $\beta \cdot k$

5. **Far Points** We define the set of far points in rings $R_{i,j}$ with $2\log\varepsilon^{-1} \le j$ by F_i. We sample a set $\Omega(F_i)$ of $\delta = \tilde{O}(\varepsilon^{-3}k \cdot \max(d, \varepsilon^{-2}))$ points proportionate to their distance to c_i. Weigh each point inversely proportionate to the sampling probability. Finally, uniformly scale the weights up or down such that the sum of weights equals $|F_i|$.

Our aim is to prove the following result.

Theorem 2.1. *For any set of of points $X \subset \mathbb{R}^d$, we can compute in time ... a coreset with cardinality constraints D of size at most ...*

3. Analysis of Near Points

Try to formulate the appropriate lemma here. We would like to argue that for any solution, the following inequality holds.

$$\left|\sum_{p\in N_i}(\text{cost}(p,s(p)) - \text{cost}(c_i,s(p)))\right| \le \varepsilon\cdot(\text{cost}(N_i,S)+\text{cost}(C_i,S)).$$

For the proof, you only have to use the triangle inequality (appropriately generalized if we want it to hold for k-means).

4. Analysis of Ring Points

Outline:

- We would like to argue that for any fixed solution S (i.e. fixed capacities and fixed set of centers) the cost is preserved up to a $(1\pm\varepsilon)$ factor with probability $1-\delta$ if we sample $\varepsilon^{-2-z}\log\delta^{-1}$ many points.
- Enumerate all possible choices of centers.
- Enumerate all possible choices of clusters sizes.
- Explain how to eliminate dependencies on the dimension.

We want to prove

Lemma 4.1. *Let D be subset of δ points, each chosen uniformly at random with replacement from $R_{i,j}$ and weighted by $\frac{|R_{i,j}|}{\delta}$. Then if $\delta > \alpha\cdot\varepsilon^{-3}\cdot(kd+\log 1/\pi)$, we have for all candidate solutions $(S,w)\in\mathcal{S}$ with probability at least $1-\pi$*

$$|\mathit{cost}(R_{i,j},S,w) - \mathit{cost}(D,S,w)| \le \varepsilon\cdot(\mathit{cost}(R_{i,j},S,w)+\mathit{cost}(R_{i,j},\{c_i\}))$$

The following notion will probably be useful.

Let $\mathcal{S}$ be the set of all candidate solutions. Let $\mathcal{S}_{i,j}$ be the set of solutions defined as $\mathcal{S}_{i,j} = \{S' \subset S\in\mathcal{S} \mid \text{cost}(s,c_i) \le \varepsilon^{-O(1)}\cdot 2^j\cdot\Delta_i \wedge s\in S'\}$.

Dealing with a single fixed solution You want to prove an analogue (but perhaps slightly more general? need to check) of Lemma 13 from (Cohen-Addad & Li, 2019). Essentially, we want to show that any fixed solution from $S_{i,j}$, the cost is preserved. To do this, you will need the run of the mill Bernstein's inequality type arguments that you can find either in (Cohen-Addad & Li, 2019), or equivalently from (Cohen-Addad et al., 2021).

Lemma 4.2. *Let D be subset of δ points, each chosen uniformly at random with replacement from $R_{i,j}$ and weighted by $\frac{|R_{i,j}|}{\delta}$. Then if $\delta > \alpha\cdot\varepsilon^{-3}\cdot\log 1/\pi$, we have for any candidate solutions $(S,w)\in\mathcal{S}$ with probability at least $1-\pi$*

$$|\mathit{cost}(R_{i,j},S,w) - \mathit{cost}(D,S,w)| \le \varepsilon\cdot(\mathit{cost}(R_{i,j},S,w)+\mathit{cost}(R_{i,j},\{c_i\}))$$

Discretization of centers The main goal is to discretize the set of potential centers. Let $\mathcal{S}$ be the collection of solutions in our discretization. Here, we want to show that for *any* possible solutions S and *all* points $p\in R_{i,j}$ and $s\in S$, there exists a solution $S'\in\mathcal{S}$ and a bijection $b: S\to S'$ such that if $\|q-s\| \le \varepsilon^{-1}\cdot\|q-c_i\|$ for some $q\in R_{i,j}$ then

$$|\|p-s\| - \|p-b(s)\|| \le \varepsilon\cdot\|p-c_i\|.$$

You do this by casting an $\varepsilon^{O(1)}\cdot 2^j\cdot\Delta_i$-net of the ball centered around c_i with radius roughly (up to constant factors) $\varepsilon^{-1}2^j\cdot\Delta_i$.

Key Lemmas:

Lemma 4.3 (Find this in literature, don't have to prove it yourself)**.** *Let B be the unit ball in d-dimensional Euclidean space. There exists an ε-net of size $\left(1+\frac{2}{\varepsilon}\right)^{-d}$*

Lemma 4.4. *Let $R_{i,j}$ be the set of points at distance $[2^j\cdot\Delta_i, 2^{j+1}\cdot\Delta_i)$ from c_i. Let S be an arbitrary solution. Then there exists an approximate centroid set for $R_{i,j}$ of size $\varepsilon^{-\alpha\cdot kd}$, where α is an absolute constant.*

Discretization of cardinalities Here, we need two arguments. First, we argue that the centers that are at distance at least $\varepsilon^{O(z)}\cdot 2^j\cdot\Delta_i$, we preserve the cost anyway. For the centers that are close, we argue as follows.

1. Centers with small cardinalities (less than $\varepsilon^{O(1)}\cdot\frac{|R_{i,j}|}{k}$ contribute only a negligible amount to the cost).
2. Define an exponential sequence to the base of $(1+\beta)$, where $\beta = \varepsilon^{O(1)}$ (just ε might be possible, the calculation will show).

3. Let $Card := \{\varepsilon^{O(1)} \cdot \frac{|R_{i,j}|}{k} \cdot (1+\beta)^t \| \}$ be the set of candidate cardinalities.

4. Bound the number of distinct cardinalities (should be of the order $\beta^{-1} \log k/\varepsilon$.

5. For a solution S with cardinalities $w : S \to [|R_{i,j}|]$, we replacing the cardinalities with $w(s_i)$ with the largest cardinality in $Card$ that is smaller than $w(s_i)$. Let $\widehat{|R_{i,j}|}$ be the sum of the replacement cardinalities.

6. Argue that a minimum cost assignment of any subset of $\widehat{|R_{i,j}|}$ points of $R_{i,j}$ to the new cardinalities preserves the cost (up to an appropriate term).

7. Argue that the surplus $|R_{i,j}| - \widehat{|R_{i,j}|}$ can be assigned arbitrarily without it affecting the cost.

Key Lemma:

Define the set of integers. $Card := \{\varepsilon^{O(1)} \cdot \frac{|R_{i,j}|}{k} \cdot (1+\beta)^t \| \}$.

First, prove that approximating all solutions with weights from $Card$ (when necessary) is sufficient.

Lemma 4.5. *Let $\mathcal{S}_{i,j,H} = \{S_{i,j} \times H \mid s \in Scost(s, c_i) \leq \varepsilon^{O(1)} \cdot 2^j \cdot \Delta_i\}$ and let $\mathcal{S}_{i,j,F} = \{S_{i,j} \times \mathbb{N} \mid s \in Scost(s, c_i) \geq \varepsilon^{O(1)} \cdot 2^j \cdot \Delta_i\}$. Define $\mathcal{S}' := \mathcal{S}_{i,j,H} \times \mathcal{S}_{i,j,F}$ If for every solution $(S', w') \in \mathcal{S}'$*

$$|cost(R_{i,j}, w', S') - cost(D, w', S')| \leq \varepsilon \cdot cost(R_{i,j}, w, \mathcal{S}),$$

then for every solution $(S, w) \in \mathcal{S}$

$$|cost(R_{i,j}, w', S') - cost(D, w', S')| \leq \varepsilon \cdot cost(R_{i,j}, w, \mathcal{S}).$$

Next, prove that approximating all solution solutions with weights can be done with few samples.

Lemma 4.6. *Let D be subset of δ points, each chosen uniformly at random with replacement from $R_{i,j}$ and weighted by $\frac{|R_{i,j}|}{\delta}$. Then if $\delta > \alpha \cdot \varepsilon^{-3} \cdot (kd + \log 1/\pi)$, we have for all candidate solutions $(S, w) \in \mathcal{S}_{i,j,H}$ with probability at least $1 - \pi$*

$$\begin{aligned} &|cost(R_{i,j}, S, w) - cost(D, S, w)| \\ &\leq \varepsilon \cdot (cost(R_{i,j}, S, w) + cost(R_{i,j}, \{c_i\})) \end{aligned}$$

Finally conclude with a proof of the main lemma from this section.

Removal of d For k-means one can first use PCA. For the others, we have to apply terminal embeddings recursively. Details on how to do this are in (Cohen-Addad et al., 2021) and (Braverman et al., 2021).

References

Braverman, V., Jiang, S. H.-C., Krauthgamer, R., and Wu, X. Coresets for clustering in excluded-minor graphs and beyond. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 2679–2696. SIAM, 2021.

Cohen-Addad, V. and Li, J. On the fixed-parameter tractability of capacitated clustering. In *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*, volume 132, pp. 41–1. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2019.

Cohen-Addad, V., Saulpic, D., and Schwiegelshohn, C. A new coreset framework for clustering. In Khuller, S. and Williams, V. V. (eds.), *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*. ACM, 2021.