k-Median/Means with Outliers Revisited: A Simple Fpt Approximation

Xianrun Chen[1], Lu Han[2], Dachuan Xu[3], Yicheng Xu[1](✉), and Yong Zhang[1]

[1] Chinese Academy of Sciences, Shenzhen Institute of Advanced Technology,
Shenzhen, China
yc.xu@siat.ac.cn
[2] Beijing University of Posts and Telecommunications, Beijing, China
[3] Beijing University of Technology, Beijing, China

Abstract. We revisit the classical metric k-median/means with outliers in this paper, whose proposal dates back to (Charikar, Khuller, Mount, and Narasimhan SODA'01). Though good approximation algorithms have been proposed, referring to the state-of-the-art $(6.994+\varepsilon)$-approximation (Gupta, Moseley and Zhou ICALP'21) for k-median with outliers and $(53.002+\varepsilon)$-approximation (Krishnaswamy, Li, and Sandeep SODA'18) for k-means with outliers respectively, we are interested in finding efficient fpt (fixed-parameter tractable) approximations, following a recent research mainstream for constrained clusterings. As our main contribution, we propose a simple but efficient technical framework that yields a $(3+\varepsilon)/(9+\varepsilon)$-approximation for k-median/means with outliers, albeit in $((m+k)/\varepsilon)^{O(k)} \cdot n^{O(1)}$ time. It is notable that our results match with previous result (Goyal, Jaiswal, and Kumar IPEC'20) in terms of ratio and asymptotic running time. But as aforementioned, our technique is much more simplified and straightforward, where instead of considering the whole client set, we restrict ourselves to finding a good approximate facility set for coreset, which can be done easily in fpt time even with provably small loss. Similar idea can be applied to more constrained clustering problems whose coresets have been well-studied.

Keywords: Approximation algorithm · Fixed-parameter tractability · Clustering with outliers · Coreset

1 Introduction

The k-median/means problem is a classical optimization problem in which one must choose k facilities from a given set of candidate locations to serve a set of clients, so as to minimize the total distance cost between the clients and their closest facilities. This problem attracts research interests from various domains, such as computer science, data science, and operations research.

In general metric, both k-median and k-means are NP-hard, more precisely, APX-hard. The k-median is hard to approximate within a factor of $(1+2/e)$,

W. Wu and G. Tong (Eds.): COCOON 2023, LNCS 14423, pp. 295–302, 2024.
https://doi.org/10.1007/978-3-031-49193-1_22

and the k-means is hard to approximate within a factor of $(1 + 8/e)$ under P$\neq$ NP [15,23]. Both problems have been extensively studied in the literature from the perspective of approximation algorithms. The state-of-art approximation algorithm for k-median, to the best of our knowledge, is 2.67059 by Cohen-Addad et al. [8]. Kanungo et al. [24] give a $(9+\varepsilon)$-approximation algorithm for k-means, which is improved to 6.357 by Ahmadian et al. [3].

However, the presence of outliers can significantly affect the solution of the k-median/means, where some clients may not be served by any facility (i.e. outliers) due to various reasons such as geographic constraints or capacity limitations. Excluding outliers may greatly reduce the clustering cost and improve the quality of the clustering. Despite practical considerations, identifying outliers in clustering is a crucial and intriguing part of algorithm design, for example, what we study in this paper, k-median with outliers (k-MedO) and k-means with outliers (k-MeaO), which we will formally define later in definition 1.

Both k-MedO and k-MeaO are more difficult than the vanilla version. In a seminal work, Charikar et al. [5] first introduce the problem of k-MedO in the literature (also called robust k-median), and design a bi-criteria $(1 + \lambda$, $4{+}4/\lambda)$-approximation algorithm that always returns a clustering at cost at most $4 + 4/\lambda$ times the optimum with violation of the outlier constraint by a factor of $1 + \lambda$. No constant approximation algorithm has been proposed for k-MedO until Chen [6] presents the first true constant approximation algorithm, whose approximation ratio is significantly improved to $7.081{+}\varepsilon$ by Krishnaswamy et al. [26] and $6.994{+}\varepsilon$ by Gupta et al. [17] via iterative LP rounding. For k-MeaO, Gupta et al. [18] give a bi-criteria approximation algorithm that outputs a solution with a ratio of 274 using at most $O(m\mathrm{log}n)$ outliers, where m stands for the desired number of outliers and n corresponds to the total count of clients. Krishnaswamy et al. [26] first propose a true constant approximation algorithm of 53.002, which is quite far from the lower bound.

While there is a lot of work focusing on the approximation algorithms for k-MedO and k-MeaO, there is another research mainstream aiming at developing *fixed-parameter tractable* (fpt) approximations, which enables an additional factor of $f(k)$ in the running time. Fpt algorithms have demonstrated their ability to overcome longstanding barriers in the field of approximation algorithms in recent years [11,25], improving the best-known approximation factors in polynomial time for many classic NP-hard problems, e.g., k-vertex separator [27], k-cut [17] and k-treewidth deletion [16].

Coresets turn out to be useful in fpt algorithm design for clustering recently. Coresets are small representative subsets of the original dataset that capture specific geometric structures of the data, which can help to develop existing algorithms. Agarwal et al. [1] first introduce the framework of coreset in computing diameter, width, and smallest bounding box, ball, and cylinder, initializing a research path of the coreset for many other combinatorial problems. In the classic k-median and k-means, Har-Peled and Mazumdar [19] first prove the existence of small coreset for k-median and k-means with size $O(k \log n\varepsilon^{-d})$ in Euclidean metrics and near-optimal size bounds have been obtained in more recent works by Cohen-Addad et al. [10,22]. In general metric, a seminar paper by Chen [7] obtains coresets for k-median and k-means with size $O(k \log n/\varepsilon)$ based on hierarchical sampling. However, coresets for k-MedO and k-MeaO seem to be less understood — previous results either suffer from exponential dependence on $(m + k)$ [12], or violate the constraint of k or m [20]. Recently, Huang et al. [21] present a near-optimal coreset for k-MedO and k-MeaO in Euclidean spaces with size $O((m + k/\varepsilon))$ based on uniform sampling.

Building on the previous work of coresets, there are several fpt results for the clustering problem. For the classic k-median and k-means, $(1+2/e+\varepsilon)$ and $(1+8/e+\varepsilon)$ fpt approximations are obtained by Cohen-Addad et al. [9], which are proven to be tight even in $f(k, \varepsilon) \cdot n^{O(1)}$ time, assuming Gap-ETH. For the k-MedO and k-MeaO, existing work [2,28] mostly overcomes the difficulty of identifying outliers by reducing the k-MedO/k-MeaO into a related $(k + m)$-median/means problem, which leads to an exponential time dependency on the outlier number m. Agrawal et al. [2] present $(1 + 2/e + \varepsilon)$ and $(1 + 8/e + \varepsilon)$ fpt approximations for k-MedO and k-MeaO respectively, in $((k + m)/\varepsilon)^{O(m)} \cdot (k/\varepsilon)^{O(k)} \cdot n^{O(1)}$ time. In addition to coresets, k-means++ [4] is also considered as a dataset reduction for k-MeaO in Euclidean spaces in the literature [13, 28]. Though it is not stated explicitly, Statman et al. [28] yields a $(1 + \varepsilon)$-approximation for Euclidean k-MeaO in fpt time.

Our Contribution. In this paper, we propose a coreset-based technical framework for k-MedO and k-MeaO in general metric. We employ the coreset for k-MedO and k-MeaO as a reduction of search space, which would help us to avoid the exponential time dependency on m. We restrict ourselves to finding a good facility set based on the constructed client coreset with size $O((k+m) \log n/\varepsilon)$. We propose a provably good approximate facility set by finding substitute facilities for *leaders* of clients in coreset, where *leaders* represent the clients with minimum cost in each optimal cluster. Moreover, *leaders* can be found easily in fpt(k) time by enumeration of k-sized subset from the coreset. Based on this idea, we derive a $(3+\varepsilon)$-approximation for k-MedO and a $(9+\varepsilon)$-approximation for the k-MeaO in $((k+m)/\varepsilon)^{O(k)} \cdot n^{O(1)}$ time. It is worth noting that our result improves upon Akanksha et al. [2] in terms of running time, as the outlier count m is consistently much larger than the facility count k in practical but with a constant loss of approximation ratio. Also note that this matches with Goyal et al. [14] in terms of approximation ratio and asymptotic running time, but with a much simplified and straightforward technical framework, which is promising to apply to more (constrained) clustering problems.

The rest of the paper is organized as follows. Section 2 describes the fpt algorithm as well as its analysis in detail. We conclude this paper and provide some interesting directions in Sect. 3.

2 A Coreset-Based Fpt Approximation

For ease of discussion, we provide formal definitions for k-MedO and k-MeaO.

Definition 1. *(k-MedO/k-MeaO) An instance I of the k-median/means problem with outliers contains a tuple $((X \cup F, d), k, m)$, where $(X \cup F, d)$ is a metric space over a set of n points with a function $d(i, j)$ indicating the distance between two points i, j in $X \cup F$. Furthermore, X and F are two disjoint sets referred to as "clients" and "facilities locations", respectively, while k and m are two positive parameters. The objective is to identify a subset S of k facilities in F and simultaneously exclude a subset O of m clients from X to minimize*

$$cost_m(X, S) = \min_{O \subseteq X : |O| = m} \sum_{x \in X \setminus O} d^z(x, S).$$

Here, $z = 1$ corresponds to k-MedO, and $z = 2$ corresponds to k-MeaO.

This definition implies that, for a fixed set S of k open facilities, the set of m outliers can be naturally determined, namely the set of m points that are farthest from S. This reminds us that we can overcome the difficulty of identifying outliers by focusing on selecting a good approximation for the facilities.

As a pre-processing step, we construct a coreset for k-MedO/k-MeaO to reduce the search space, in order to find a good facility set in fpt(k) time.

Definition 2. *(Coreset) Let $\mathcal{S} = \{S_1, S_2, \dots\}$ be a set of candidate solutions. Recall that*

$$cost_m(X, S) = \min_{O \subseteq X : |O| = m} \sum_{x \in X \setminus O} d^z(x, S).$$

Then a weighted subset $C \subseteq X$ is an ε-coreset, if for all candidate solution $S \in \mathcal{S}$ we have

$$|cost_m(X, S) - cost_m(C, S)| \leq \varepsilon \cdot cost_m(X, S).$$

In other words, coreset is a weighted subset for the client set which preserves a good approximation for the whole client set. In this paper, we make use of the coreset construction technique [2], based on a hierarchical sampling idea [7], which is suitable for clustering problems with outliers.

Theorem 1. *(Agrawal et al. [2]) For $\varepsilon > 0$, there exists an algorithm that for each instance $I = ((X \cup F, d), k, m)$ of k-MedO or k-MeaO, outputs an ε-coreset $C \subseteq X$ with size $|C| = O(((k + m) \log n/\varepsilon)^2)$ with constant probability, running in $O(nk) + poly(n)$ time.*

For any instance $I = ((X \cup F, d), k, m)$ and $\varepsilon > 0$, we run the hierarchical sampling in [2] on I to obtain a coreset $C \subseteq X$ with size $O(((k + m) \log n/\varepsilon)^2)$. By the definition of coreset, an α-approximate solution on $I' = ((C \cup F, d), k, m)$ implies a $(1+\varepsilon)\alpha$-approximate solution on I. Therefore, the only thing we need to do is to find a facility subset $S \subseteq F$ of size k to minimize the objective function $\min_{L \subseteq C : |L| = m} \sum_{x \in C \setminus L} d^z(x, S)$ on instance I'. Towards this end, we present our algorithm in Algorithm 1.

Let $F^* = \{f_1^*, f_2^*, \dots f_k^*\}$ be facility set of optimal solution on instance $I' = ((C \cup F, d), k, m)$. For any $f_i^* \in F^*$, let C_i^* be the clients served by facility f_i^*,

Algorithm 1. Coreset-based fpt approximation

```
Require: Instance $I = ((X \cup F, d), k, m)$
  Construct the coreset $C$ based on hierarchical sampling
  Find leaders of $C$ by enumeration
  $S \leftarrow$ the nearest facilities in $F$ for every leader client
  return $S$
```

and denote c_i^* as the *leader* of C_i^* representing the client in C_i^* who is closest to f_i^*. We define *leaders* as all such *leader* clients, thus is a k-sized set.

We will find a good substitute facility for each C_i^* via *leader*, inspired by [9] except that their *leader* is defined on the origin client set in order to deal with capacity constraints. By the definition, we can find *leaders* eventually by enumeration of a k-sized subset, which is allowed in fpt(k) time. We will prove that a good approximate facility set can be found around the *leaders* with a constant approximation guarantee.

Lemma 1. *Algorithm 1 yields a (3+ε)-approximation for k-MedO and a (9+ε)-approximation for k-MeaO.*

Proof. We define f_i for each *leader* client c_i^* as the closest facility in F. As c_i^* is the closest client of f_i^* in C_i^*, it must satisfy that $d(c_i^*, f_i) \leq d(c_i^*, f_i^*) \leq d(c, f_i^*)$ for any client c in C_i^*, which is shown in Fig. 1.

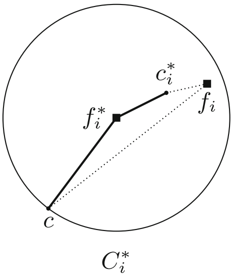

Fig. 1. Substitute facility in optimal cluster C_i^*

Thus, for each c in the coreset,

$$d(c, f_i) \leq d(c, f_i^*) + d(f_i^*, c_i^*) + d(c_i^*, f_i) \leq 3d(c, f_i^*),$$

where the first inequality follows from the triangle inequality. Combined with the definition of coreset (Theorem 1), it holds that

$$cost_m(X, S) \leq \frac{1}{1-\varepsilon} cost_m(C, S) \leq \frac{3^z}{1-\varepsilon} cost_m(C, F^*) \leq \frac{3^z(1+\varepsilon)}{1-\varepsilon} cost_m(X, F^*),$$

which implies a $(3+\varepsilon)$-approximation/$(9+\varepsilon)$-approximation for k-MedO/k-MeaO respectively. Though it is possible that some *leader* clients may share the same closest facility f_i, it will not affect the performance guarantee as it does not hurt to let f_i serve all C_i^* corresponding to these *leader* clients. This concludes the proof of Lemma 1. □

By combining Theorem 1 and Lemma 1 together, we can establish that our algorithm yields a (3+ε)\(9+ε)-approximation with a constant probability. To ensure a high probability of success, we can repeat the algorithm for a logarithmic number of rounds, which leads to the following theorem.

Theorem 2. *For any $\varepsilon > 0$, there exists a (3+ε)-approximation for k-MedO and a (9+ε)-approximation for k-MeaO with high probability, running in $((m+k)/\varepsilon)^{O(k)}\cdot\ n^{O(1)}$ time.*

3 Conclusion

To summarize, we propose a simple unified approach to obtain constant factor approximations for metric k-MedO/k-MeaO in fpt(k) time, more specifically, in $((m+k)/\varepsilon)^{O(k)}\cdot\ n^{O(1)}$ time. It is highlighted that the running time avoids exponential dependency on m, which partially answers (Agrawal et al. AAAI'23) who ask for faster fpt approximations for k-MedO/k-MeaO while obtaining the tight approximation ratios. The proposed approach leverages recent results on coresets for robust clustering, and presents a simple but novel idea to find a good substitute facility set for those *leaders* of coreset. We prove that the substitute facility set can be found easily in fpt(k) time and have provably small loss compared with the optimal facility set in terms of the k-MedO/k-MeaO objective for coreset. We believe similar idea has the potential to apply to a wide range of constrained clustering problems, for example, fair clustering, a recent mainstream in clustering field.

Acknowledgement. Xianrun Chen and Yicheng Xu are supported by Natural Science Foundation of China (No. 12371321), Fundamental Research Project of Shenzhen City (No. JCYJ20210324102012033) and Shenzhen Science and Technology Program (No. CJGJZD20210408092806017). Dachuan Xu is supported by Natural Science Foundation of China (No. 12371320). Yong Zhang is supported by National Key R&D Program of China (No. 2022YFE0196100) and Natural Science Foundation of China (No. 12071460).

References

1. Agarwal, P.K., Har-Peled, S., Varadarajan, K.R.: Approximating extent measures of points. J. ACM **51**(4), 606–635 (2004)
2. Agrawal, A., Inamdar, T., Saurabh, S., Xue, J.: Clustering what matters: optimal approximation for clustering with outliers. CoRR abs/ arXiv: 2212.00696 (2022)
3. Ahmadian, S., Norouzi-Fard, A., Svensson, O., Ward, J.: Better guarantees for k-means and euclidean k-median by primal-dual algorithms. SIAM J. Comput. **49**(4) (2020)

4. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: SODA, pp. 1027–1035. SIAM (2007)
5. Charikar, M., Khuller, S., Mount, D.M., Narasimhan, G.: Algorithms for facility location problems with outliers. In: SODA, pp. 642–651. ACM/SIAM (2001)
6. Chen, K.: A constant factor approximation algorithm for k-median clustering with outliers. In: SODA, pp. 826–835. SIAM (2008)
7. Chen, K.: On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. SIAM J. Comput. **39**(3), 923–947 (2009)
8. Cohen-Addad, V., Grandoni, F., Lee, E., Schwiegelshohn, C.: Breaching the 2 LMP approximation barrier for facility location with applications to k-median. In: SODA, pp. 940–986. SIAM (2023)
9. Cohen-Addad, V., Gupta, A., Kumar, A., Lee, E., Li, J.: Tight FPT approximations for k-median and k-means. In: ICALP. LIPIcs, vol. 132, pp. 42:1–42:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2019)
10. Cohen-Addad, V., Larsen, K.G., Saulpic, D., Schwiegelshohn, C.: Towards optimal lower bounds for k-median and k-means coresets. In: STOC, pp. 1038–1051. ACM (2022)
11. Cohen-Addad, V., Li, J.: On the fixed-parameter tractability of capacitated clustering. In: ICALP. LIPIcs, vol. 132, pp. 41:1–41:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2019)
12. Feldman, D., Schulman, L.J.: Data reduction for weighted and outlier-resistant clustering. In: SODA, pp. 1343–1354. SIAM (2012)
13. Feng, Q., Zhang, Z., Huang, Z., Xu, J., Wang, J.: Improved algorithms for clustering with outliers. In: ISAAC. LIPIcs, vol. 149, pp. 61:1–61:12. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2019)
14. Goyal, D., Jaiswal, R., Kumar, A.: FPT approximation for constrained metric k-median/means. In: IPEC. LIPIcs, vol. 180, pp. 14:1–14:19. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2020)
15. Guha, S., Khuller, S.: Greedy strikes back: improved facility location algorithms. J. Algorithms **31**(1), 228–248 (1999)
16. Gupta, A., Lee, E., Li, J., Manurangsi, P., Wlodarczyk, M.: Losing treewidth by separating subsets. In: SODA, pp. 1731–1749. SIAM (2019)
17. Gupta, A., Moseley, B., Zhou, R.: Structural iterative rounding for generalized k-median problems. In: ICALP. LIPIcs, vol. 198, pp. 77:1–77:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2021)
18. Gupta, S., Kumar, R., Lu, K., Moseley, B., Vassilvitskii, S.: Local search methods for k-means with outliers. Proc. VLDB Endow. **10**(7), 757–768 (2017)
19. Har-Peled, S., Mazumdar, S.: On coresets for k-means and k-median clustering. In: STOC, pp. 291–300. ACM (2004)
20. Huang, L., Jiang, S.H., Li, J., Wu, X.: Epsilon-coresets for clustering (with outliers) in doubling metrics. In: FOCS, pp. 814–825. IEEE Computer Society (2018)
21. Huang, L., Jiang, S.H., Lou, J., Wu, X.: Near-optimal coresets for robust clustering. coRR abs/ arXiv: 2210.10394 (2022)
22. Huang, L., Vishnoi, N.K.: Coresets for clustering in euclidean spaces: importance sampling is nearly optimal. In: STOC, pp. 1416–1429. ACM (2020)
23. Jain, K., Mahdian, M., Saberi, A.: A new greedy approach for facility location problems. In: STOC, pp. 731–740. ACM (2002)
24. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: A local search approximation algorithm for k-means clustering. Comput. Geom. **28**(2–3), 89–112 (2004)

25. Karthik C.S., Laekhanukit, B., Manurangsi, P.: On the parameterized complexity of approximating dominating set. J. ACM **66**(5), 33:1–33:38 (2019)
26. Krishnaswamy, R., Li, S., Sandeep, S.: Constant approximation for k-median and k-means with outliers via iterative rounding. In: STOC, pp. 646–659. ACM (2018)
27. Lee, E.: Partitioning a graph into small pieces with applications to path transversal. Math. Program. **177**(1–2), 1–19 (2019)
28. Statman, A., Rozenberg, L., Feldman, D.: k-means: Outliers-resistant clustering+++. Algorithms **13**(12), 311 (2020)