A More Efficient Reduction from Outlier-Aware to Outlier-Free k-Median

Zhen Zhang[1,2], **Han Peng**[1,2*], **Limei Liu**[1,2], **Junyu Huang**[3], **Xiaolong Li**[1,2*], **Qilong Feng**[3]

[1]School of Advanced Interdisciplinary Studies, Hunan University of Technology and Business, China
[2]Xiangjiang Laboratory, China
[3]School of Computer Science and Engineering, Central South University, China
{zz, Han.Peng, seagullm, lxl}@hutb.edu.cn, junyuhuangcsu@foxmail.com, csufeng@mail.csu.edu.cn

Abstract

Given a non-negative integer ℓ, the k-median with outliers problem extends the standard k-median problem by allowing the removal of up to ℓ points and minimizing the clustering cost over the remaining ones. Algorithmic development in this setting remains an active area of research due to its relevance in processing noisy data. In this paper, we present a sampling-based reduction from the k-median with outliers problem to its outlier-free counterpart. The reduction incurs a multiplicative overhead of $(k\ell^{-1} + \varepsilon^{-1})^{O(\ell)}$ in the running time: it yields $(k\ell^{-1} + \varepsilon^{-1})^{O(\ell)}$ outlier-free instances, a solution to one of which can be directly transformed into a solution to the original instance with an arbitrarily small loss in the approximation ratio. This improves upon the previously known reduction with an overhead of $((k + \ell)\varepsilon^{-1})^{O(\ell)}$. As applications, we obtain faster fixed-parameter tractable (FPT) algorithms with tight approximation guarantees for the k-median with outliers problem under various metric spaces. Furthermore, our approach naturally generalizes to constrained variants of the problem where additional constraints are imposed on the cluster sizes, and yields similar improvements in their FPT approximations.

Introduction

Center-based clustering problems are ubiquitous in various fields involving data analysis and processing. Given a set of points in a metric space and a positive integer k, the goal of these problems is to identify at most k centers and assign each point to its nearest center so as to minimize a specified objective function that quantifies the clustering quality. Among such problems, the *k-median* (k-MED) problem is one of the most widely studied, which minimizes the sum of distances between points and their corresponding centers. Solving the k-MED problem facilitates the identification of underlying data distributions and the extraction of representative prototypes. As a result, algorithms for this problem have been extensively investigated and have become fundamental tools in a wide range of applications; see, e.g., (Cohen-Addad et al. 2019, 2023; Gowda et al. 2023).

Despite its popularity, the k-MED problem is inherently sensitive to noise, as a small number of *outliers* located far from the bulk of the data can substantially affect the value of its objective function. Allowing the removal of such outliers can often enhance the robustness of the clustering process. Motivated by this, there has been considerable interest in the k-MED *with outliers* (k-MEDOUT) problem. Given a non-negative integer ℓ, this problem is designed to discard at most ℓ points (regarded as outliers) and minimize the clustering cost of the remaining points. The problem can be formally defined as follows.

Definition 1 (k-MEDOUT) *An instance $((\mathcal{X}, d), \mathcal{P}, \mathcal{F}, \ell, k)$ of the k-MEDOUT problem is specified by a metric d over a set $\mathcal{X}$ of points, a subset $\mathcal{P} \subseteq \mathcal{X}$ consisting of the points to be clustered, a subset $\mathcal{F} \subseteq \mathcal{X}$ consisting of permissible locations for centers, and integers $\ell \in [0, |\mathcal{P}|]$ and $k \in [1, |\mathcal{F}|]$. A feasible solution $(\mathcal{O}, \mathcal{C})$ to the instance is defined by a subset $\mathcal{O} \subseteq \mathcal{P}$ of no more than ℓ outliers and a subset $\mathcal{C} \subseteq \mathcal{F}$ of no more than k centers. The cost of the solution is $\sum_{p \in \mathcal{P} \setminus \mathcal{O}} \min_{c \in \mathcal{C}} d(p, c)$. The objective is to find a feasible solution that minimizes this cost.*

Solving the k-MEDOUT problem enables a more justifiable identification of outliers, as it is guided by the underlying cluster structure. This, in turn, leads to the formation of more cohesive clusters. Due to its effectiveness in handling noisy data, the problem has attracted considerable attention from both theoretical and practical perspectives. In particular, the development of its approximation algorithms remains a highly active line of research. The current best-known polynomial-time approximation guarantee for the k-MEDOUT problem is the iterative rounding-based ratio of $6.994 + \varepsilon$ (Gupta, Moseley, and Zhou 2021). There has also been work on developing bi-criteria approximation algorithms for the problem, which permit limited violations of the upper bound on the number of centers or outliers in exchange for real-world applicability (Huang, Liu, and Ding 2024) or better approximation guarantees (Charikar et al. 2001; Friggstad et al. 2019; Cohen-Addad, Feldmann, and Saulpic 2021; Wu et al. 2024).

A frequently adopted approach to simplifying the k-MEDOUT problem is to restrict attention to instances where the maximum numbers of centers and outliers (i.e., k and ℓ) are significantly smaller than the input size. In this context, *fixed-parameter tractable* (FPT) approximation algorithms, parameterized by k and ℓ, become viable. Such al-

*Corresponding author

Approx.	Time	Tech.	Ref.
$O(1)$	$(k+\ell)^{O(k+\ell)}n \log n$	Coreset construction	(Feldman and Schulman 2012)
$3+\varepsilon$	$\left((k+\ell)\varepsilon^{-1}\right)^{O(k)} n$	Sampling	(Goyal, Jaiswal, and Kumar 2020; Chen et al. 2023)
$1+2e^{-1}+\varepsilon$	$(k+\ell)^{O(\ell)}k^{O(k)}\varepsilon^{-O(k+\ell)}n^{O(1)}$	Reduction to k-MED	(Agrawal et al. 2023; Jaiswal and Kumar 2023)
$1+2e^{-1}+\varepsilon$	$k^{O(k)}\varepsilon^{-O(k+\ell)}n^{O(1)}$	Reduction to k-MED	This work

Table 1: FPT approximation algorithms for the k-MEDOUT problem.

gorithms run in time $f(k,\ell,\varepsilon)n^{O(1)}$, where $f()$ is an arbitrary computable function, n denotes the number of input points, and ε is an arbitrary constant in $(0,1)$. By leveraging the small values of ℓ and k, FPT approximation algorithms that achieve much better approximation ratios than polynomial-time algorithms have been developed (Feldman and Schulman 2012; Goyal, Jaiswal, and Kumar 2020; Chen et al. 2023; Agrawal et al. 2023; Jaiswal and Kumar 2023), as summarized in Table 1.

The best-known FPT approximation algorithm for the k-MEDOUT problem, given in (Agrawal et al. 2023; Jaiswal and Kumar 2023), is obtained by reducing the problem to its outlier-free counterpart (namely, the standard k-MED problem) with an arbitrarily small loss in approximation guarantees. Specifically, Agrawal et al. (2023) performed such an *approximation-preserving reduction* by constructing $((k+\ell)\varepsilon^{-1})^{O(\ell)}n^{O(1)}$ outlier-free instances, and showed that an α-approximation solution to one of them implies an $\alpha(1+\varepsilon)$-approximation solution to the original instance. Consequently, the reduction in (Agrawal et al. 2023) incurs multiplicative overheads of $((k+\ell)\varepsilon^{-1})^{O(\ell)}n^{O(1)}$ and $1+\varepsilon$ in the running time and approximation ratio of the considered outlier-free algorithm, respectively. The overhead in the running time was later improved to $((k+\ell)\varepsilon^{-1})^{O(\ell)}$ by Jaiswal and Kumar (2023). When combined with existing FPT approximation algorithm for the k-MED problem (Cohen-Addad et al. 2019), these reductions yield a $(1+2e^{-1}+\varepsilon)$-approximation for the k-MEDOUT problem, matching the known lower bound on the approximation ratios achievable by FPT algorithms for the problem (Cohen-Addad et al. 2019).

Our Results

When parameterized by k and ℓ, the reduction from the *max k-coverage* problem presented in (Cohen-Addad et al. 2019) implies that, under the gap-exponential time hypothesis (Chalermsook et al. 2020), no FPT algorithm can achieve an approximation ratio better than $1+2e^{-1}$ for the k-MEDOUT problem, even in the case where $\ell=0$. This hardness result suggests that further improving the approximation ratios of FPT algorithms for the k-MEDOUT problem is unlikely. Nevertheless, reducing the running time required to achieve such tight approximations, for example by optimizing the parameter-dependent term $f(k,\ell,\varepsilon)$, remains an open possibility. This motivates our work, in which we improve the approximation-preserving reductions given in (Agrawal et al. 2023; Jaiswal and Kumar 2023) to develop a faster tight FPT approximation algorithm.

Our main technical contribution is a sampling-based method for identifying a set of points close to the outliers in an optimal solution, which serve as *anchor points* for locating the outliers. Leveraging these anchor points, we identify and remove the outliers, and thereby reduce the k-MEDOUT problem to its outlier-free counterpart with a multiplicative overhead of $(k\ell^{-1}+\varepsilon^{-1})^{O(\ell)}$ in the running time, as given in Table 2 and Theorem 1.

Theorem 1 *Given an instance $\mathcal{I}=((\mathcal{X},d),\mathcal{P},\mathcal{F},\ell,k)$ of the k-MEDOUT problem satisfying $|\mathcal{P}\cup\mathcal{F}|=n$ and $\ell>0$, an α-approximation algorithm for the k-MED problem that runs in time $T(n-\ell,k)$ on instances with $n-\ell$ points and at most k centers, and a constant $\varepsilon\in(0,1)$, there exists an $\alpha(1+\varepsilon)$-approximation algorithm for $\mathcal{I}$ with running time $\tau\cdot T(n-\ell,k)+\tau(k\ell\varepsilon^{-1})^{O(1)}+O(n(k+\ell\varepsilon^{-1}))$, where $\tau=(k\ell^{-1}+\varepsilon^{-1})^{O(\ell)}$.*

Theorem 1 suggests that any FPT or polynomial-time approximation algorithm for the k-MED problem can be transformed into an FPT approximation algorithm with a nearly identical approximation ratio for the k-MEDOUT problem. In particular, by combining Theorem 1 with the $(1+2e^{-1}+\varepsilon)$-approximation algorithm for the k-MED problem running in $(k\varepsilon^{-1})^{O(k)}n^{O(1)}$ time given in (Cohen-Addad et al. 2019), and incorporating a simple symbolic analysis of the overhead introduced by our reduction, we obtain a tight FPT approximation algorithm for the k-MEDOUT problem, which runs in $k^{O(k)}\varepsilon^{-O(k+\ell)}n^{O(1)}$ time. As shown in Table 1 and Corollary 1, this algorithm improves upon the previously known results with the same approximation guatantees given in (Agrawal et al. 2023; Jaiswal and Kumar 2023), saving a factor of $(k+\ell)^{O(\ell)}$ in the running time.

Corollary 1 *Given an instance $\mathcal{I}=((\mathcal{X},d),\mathcal{P},\mathcal{F},\ell,k)$ of the k-MEDOUT problem satisfying $|\mathcal{P}\cup\mathcal{F}|=n$ and a constant $\varepsilon\in(0,1)$, there exists a $(1+2e^{-1}+\varepsilon)$-approximation algorithm for $\mathcal{I}$ with running time $k^{O(k)}\varepsilon^{-O(k+\ell)}n^{O(1)}$.*

In addition to its applicability to general metric spaces, our reduction, in line with previously known reductions (Agrawal et al. 2023; Jaiswal and Kumar 2023), can also be combined with approximation algorithms for the k-MED problem under specialized metrics, including the $(1+\varepsilon)$-approximation algorithms under metrics with constant doubling dimensions (Cohen-Addad, Feldmann, and Saulpic 2021), Euclidean metrics (Huang, Li, and Wu 2024), bounded-treewidth metrics (Cohen-Addad, Saulpic, and Schwiegelshohn 2021), and minor-free metrics (Cohen-Addad, Saulpic, and Schwiegelshohn 2021), as well as

Overhead	Ref.
$O(n^\ell)$	Folklore (brute force)
$\big((k+\ell)\varepsilon^{-1}\big)^{O(\ell)} n^{O(1)}$	(Agrawal et al. 2023)
$\big((k+\ell)\varepsilon^{-1}\big)^{O(\ell)}$	(Jaiswal and Kumar 2023)
$(k\ell^{-1}+\varepsilon^{-1})^{O(\ell)}$	This work

Table 2: Multiplicative overheads in the running time incurred by reductions to the k-MED problem.

the 1.999-approximation algorithm under the Ulam metric (Chakraborty, Das, and Krauthgamer 2023). As in the general setting, the resulting algorithms preserve the approximation guarantees of their outlier-free counterparts and, due to the efficiency of our reduction, are notably faster than those derived from previously known reductions (Agrawal et al. 2023; Jaiswal and Kumar 2023).

Our approach accommodates arbitrary selections of outliers and therefore applies to instances where additional constraints preclude straightforwardly designating the farthest points from the centers as outliers. We show that this approach can be easily extended to yield approximation-preserving reductions for constrained generalizations of the k-MEDOUT problem, in which additional constraints are imposed on the size of each cluster. Notably, similar reductions for these constrained variants have also been proposed in (Jaiswal and Kumar 2023; Dabas, Gupta, and Inamdar 2025). However, as in the unconstrained setting, our reductions are more efficient, yielding algorithms that save a factor of $(k+\ell)^{O(\ell)}$ in the running time.

Comparison with Earlier Work

The added complexity of identifying outliers makes the k-MEDOUT problem more challenging than its outlier-free counterpart, which motivates research on reducing the former to the latter. A natural approach is to enumerate all subsets of size ℓ from the input, treat each as a candidate set of outliers, and solve the outlier-free clustering problem on the remaining points. This entails solving $O(n^\ell)$ outlier-free instances and thus incurs a multiplicative overhead of $O(n^\ell)$ in the running time. Ideas for improving upon this brute-force approach have been proposed, as summarized in Table 2. Agrawal et al. (2023) constructed the reduction for k-MEDOUT based on a solution to an outlier-free instance with $k+\ell$ centers. They defined a set of concentric rings around each center, and sampled $((k+\ell)\varepsilon^{-1}\log n)^2$ points within the rings. These points form a coreset that closely approximates the set of points to be clustered, ensuring that the multiplicative overhead in the running time, incurred by enumerating all subsets of size ℓ for outlier identification, is bounded by $((k+\ell)\varepsilon^{-1})^{O(\ell)}n^{O(1)}$. In a similar vein, Jaiswal and Kumar (2023) sampled $\varepsilon^{-O(1)}\ell^{O(1)}$ points with probabilities weighted according to their distance to the nearest of the $k+\ell$ centers derived from the outlier-free instance, and showed that a near-optimal set of outliers can be found by enumerating subsets of a set comprising the sampled points and the neighboring points of the centers. This implies an approximation-preserving reduction for the k-MEDOUT problem with a multiplicative overhead of $((k+\ell)\varepsilon^{-1})^{O(\ell)}$ in the running time.

In this paper, we aim to further improve the efficiency of approximation-preserving reduction for the k-MEDOUT problem. We obtain $O(k+\ell)$ centers by solving an outlier-free instance (as in (Agrawal et al. 2023; Jaiswal and Kumar 2023)) and use them to initialize D-sampling. Specifically, the algorithm iteratively selects points with probability proportional to their distance to the nearest point among the centers and previously sampled points. While sharing algorithmic similarities with the reductions in (Agrawal et al. 2023; Jaiswal and Kumar 2023) (which also sample points based on a $(k+\ell)$-center solution), our approach no longer relies on sampling to identify candidate outliers. Instead, it samples to find anchor points, namely, points sufficiently close to the outliers in an optimal solution that can serve as anchors to locate them. This idea enables us to further reduce the required sampling size to linear in ℓ. Guided by these anchor points, we give an approximation-preserving reduction for the k-MEDOUT problem that incurs a multiplicative overhead of $(k\ell^{-1}+\varepsilon^{-1})^{O(\ell)}$ in the running time, which is significantly smaller than the overheads of the previous reductions (Agrawal et al. 2023; Jaiswal and Kumar 2023). As a result, our reduction yields a faster FPT approximation algorithm with a tight approximation guarantee, as previously pointed out.

Preliminaries

We consider a metric space $(\mathcal{X}, d)$ with distance function d. Given a point $a \in \mathcal{X}$ and a subset $\mathcal{B} \subseteq \mathcal{X}$, let $d(a,\mathcal{B}) = \min_{b\in\mathcal{B}} d(a,b)$ denote the distance from a to its nearest point in $\mathcal{B}$. Given a positive integer k and two subsets $\mathcal{B}_1, \mathcal{B}_2 \subseteq \mathcal{X}$ with $\min\{|\mathcal{B}_1|, |\mathcal{B}_2|\} \geq k$, let $\mathrm{opt}_k(\mathcal{B}_1,\mathcal{B}_2) = \min_{\mathcal{C}\subseteq\mathcal{B}_2 \wedge |\mathcal{C}|=k} \sum_{b\in\mathcal{B}_1} d(b,\mathcal{C})$ denote the minimum k-clustering cost of $\mathcal{B}_1$ using centers selected from $\mathcal{B}_2$. Given a positive integer i, define $[i] = \{1,\cdots,i\}$.

The following result, known as Chernoff bound, provides an upper bound on the probability that the sum of independent binary random variables deviates significantly below its expectation.

Lemma 1 (Chernoff (1952)) *Given two real numbers $p, \lambda \in (0,1)$ and t independent binary random variables $v_1,\ldots,v_t$ with $\Pr[v_i = 1] = p$ for each $i \in [t]$, we have $\Pr[\sum_{i=1}^t v_i < (1-\lambda)pt] < e^{-\lambda^2 pt/2}$.*

The k-MEDOUT problem reduces to the k-MED problem when $\ell = 0$. In this outlier-free setting, we can construct constant-factor bi-criteria approximation solutions in linear time, as described in the following lemma.

Lemma 2 (Chen (2009); Wei (2016)) *Given an instance $((\mathcal{X},d),\mathcal{P},\mathcal{F},0,k)$ of the k-MEDOUT problem, a bi-criteria approximation solution $(\emptyset,\mathcal{C})$ with $|\mathcal{C}| = O(k)$ and $\sum_{p\in\mathcal{P}} d(p,\mathcal{C}) = O(\mathrm{opt}_k(\mathcal{P},\mathcal{F}))$ can be found in $O(|\mathcal{P}\cup\mathcal{F}|\cdot k)$ time.*

Given a set of anchor points for locating the outliers, the goal is to select, for each anchor point, a prescribed number of nearby points as outliers in the approximation solution,

Algorithm 1: Identifying anchor points via sampling

Input: A real number $\epsilon \in (0,1)$ and an instance $\mathcal{I} = ((\mathcal{X}, d), \mathcal{P}, \mathcal{F}, \ell, k)$ of the k-MEDOUT problem with $\ell > 0$.
Output: A subset $\mathcal{A} \subseteq \mathcal{P} \cup \mathcal{F}$

1: Construct a bi-criteria approximation solution $(\emptyset, \mathcal{A})$ to instance $((\mathcal{X}, d), \mathcal{P}, \mathcal{F} \cup \mathcal{P}, 0, k+\ell)$ using Lemma 2, and let β be the ratio of its cost to the optimum.
2: **for** $i \leftarrow 1$ to $\lceil 2\ell\beta\epsilon^{-1} \rceil$ **do**
3: Sample a point $p \in \mathcal{P}$ with probability proportional to $d(p, \mathcal{A})$.
4: $\mathcal{A} \leftarrow \mathcal{A} \cup \{p\}$
5: **return** $\mathcal{A}$.

while ensuring that no point is selected more than once. This selection task is modeled via the *unit-supply transportation* problem, which is defined as follows.

Definition 2 (Unit-Supply Transportation) *An instance $(\mathcal{S}, \mathcal{D}, \mu, d)$ of the unit-supply transportation problem is specified by a set $\mathcal{S}$ of supply points, a set $\mathcal{D}$ of demand points, a demand function $\mu : \mathcal{D} \to \mathbb{Z}_{\geq 0}$, and a distance function $d : \mathcal{S} \times \mathcal{D} \to \mathbb{R}_{\geq 0}$. A feasible solution to the instance is a transportation mapping $\phi : \mathcal{S} \to \mathcal{D}$ satisfying $|\phi^{-1}(a)| = \mu(a)$ for each $a \in \mathcal{D}$. The cost of the solution is $\sum_{b \in \mathcal{S}} d(b, \phi(b))$. The objective is to find a feasible solution of minimum cost.*

It is well known that the unit-supply transportation problem can be solved in polynomial time, based on the total unimodularity of the constraint matrix in its linear programming formulation.

Lemma 3 (Schrijver (1998)) *An instance $(\mathcal{S}, \mathcal{D}, \mu, d)$ of the unit-supply transportation problem can be exactly solved in $(|\mathcal{S}| \cdot |\mathcal{D}|)^{O(1)}$ time.*

The Algorithm for Identifying Anchor Points

In this section we describe how to identify the anchor points used to locate the outliers, as outlined in Algorithm 1. The algorithm takes as input a real number $\epsilon \in (0,1)$ and an instance $\mathcal{I} = ((\mathcal{X}, d), \mathcal{P}, \mathcal{F}, \ell, k)$ of the k-MEDOUT problem, where $\ell > 0$. It constructs a set $\mathcal{A}$ by combining the centers obtained from a β-approximation solution to the outlier-free instance $((\mathcal{X}, d), \mathcal{P}, \mathcal{F} \cup \mathcal{P}, 0, k+\ell)$ with $\lceil 2\ell\beta\epsilon^{-1} \rceil$ additional points sampled from $\mathcal{P}$. This section establishes that $\mathcal{A}$ contains the desired anchor points lying sufficiently close to the outliers.

Intuitively, Lemma 2 provides an upper bound on the total distance from all points in $\mathcal{P}$, including the outliers, to their nearest centers in the solution to the outlier-free instance $((\mathcal{X}, d), \mathcal{P}, \mathcal{F} \cup \mathcal{P}, 0, k+\ell)$. Accordingly, Algorithm 1 includes these centers as anchor points. However, this bound alone is insufficient to support an approximation-preserving reduction with an arbitrarily small loss in the approximation ratio. To remedy this, the algorithm additionally samples $\lceil 2\ell\beta\epsilon^{-1} \rceil$ anchor points from $\mathcal{P}$, where each is selected with probability proportional to its distance from the current set of anchor points. This ensures that the outliers not well covered by the initial $O(k+\ell)$ centers are likely to be close to at least one of the anchor points. These insights lead to the proof of the following lemma, where we consider $\mathcal{O}$ as the set of outliers.

Lemma 4 *Given a subset $\mathcal{O} \subseteq \mathcal{P}$ with $|\mathcal{O}| \leq \ell$, inequality $\sum_{o \in \mathcal{O}} d(o, \mathcal{A}) \leq \epsilon \cdot \mathrm{opt}_k(\mathcal{P} \backslash \mathcal{O}, \mathcal{F})$ holds with probability at least $1 - e^{-1/4}$.*

Proof We prove the lemma by showing that points near the outliers not well covered by the set of centers obtained in step 1 of Algorithm 1 can be sampled in step 3 with high probability. Let $\mathcal{A}_0$ denote this center set. For each $i \in [\lceil 2\ell\beta\epsilon^{-1} \rceil]$, let $\mathcal{A}_i$ be the union of $\mathcal{A}_0$ and the points sampled in the first i iterations of the algorithm. The approximation ratio of the solution $(\emptyset, \mathcal{A}_0)$ to the outlier-free instance considered in step 1 suggests that its cost is close to $\mathrm{opt}_k(\mathcal{P} \backslash \mathcal{O}, \mathcal{F})$, as stated in the following claim.

Claim 1 $\sum_{p \in \mathcal{P}} d(p, \mathcal{A}_0) \leq \beta \cdot \mathrm{opt}_k(\mathcal{P} \backslash \mathcal{O}, \mathcal{F})$.

We break the analysis into the following two cases:

(1) There exists an integer $i \in [0, \lceil 2\ell\beta\epsilon^{-1} \rceil - 1]$ satisfying

$$\sum_{o \in \mathcal{O}} d(o, \mathcal{A}_i) \leq \frac{\epsilon}{\beta} \sum_{p \in \mathcal{P}} d(p, \mathcal{A}_i);$$

(2) For each integer $i \in [0, \lceil 2\ell\beta\epsilon^{-1} \rceil - 1]$, it holds that

$$\sum_{o \in \mathcal{O}} d(o, \mathcal{A}_i) > \frac{\epsilon}{\beta} \sum_{p \in \mathcal{P}} d(p, \mathcal{A}_i).$$

For case (1), let i be the smallest integer that satisfies the condition. We have

$$\begin{aligned} \sum_{o \in \mathcal{O}} d(o, \mathcal{A}) &\leq \sum_{o \in \mathcal{O}} d(o, \mathcal{A}_i) \\ &\leq \frac{\epsilon}{\beta} \sum_{p \in \mathcal{P}} d(p, \mathcal{A}_i) \\ &\leq \frac{\epsilon}{\beta} \sum_{p \in \mathcal{P}} d(p, \mathcal{A}_0) \\ &\leq \epsilon \cdot \mathrm{opt}_k(\mathcal{P} \backslash \mathcal{O}, \mathcal{F}), \end{aligned}$$

where the first and third steps follow from the fact that $\mathcal{A}_0 \subseteq \mathcal{A}_i \subseteq \mathcal{A}$, and the last step is due to Claim 1. This completes the proof of Lemma 4 for case (1).

We now focus on case (2), where the outliers in $\mathcal{O} \backslash \mathcal{A}_i$ remain far from the points in $\mathcal{A}_i$ for each $i \in [0, \lceil 2\ell\beta\epsilon^{-1} \rceil - 1]$. In each such iteration, Algorithm 1 samples a point $p \in \mathcal{P}$ with probability proportional to $d(p, \mathcal{A}_i)$, and adds it to $\mathcal{A}_i$ to form $\mathcal{A}_{i+1}$. This suggests that inequality $|\mathcal{A}_i \cap \mathcal{O}| < |\mathcal{A}_{i+1} \cap \mathcal{O}|$ holds with probability

$$\frac{\sum_{o \in \mathcal{O} \backslash \mathcal{A}_i} d(o, \mathcal{A}_i)}{\sum_{p \in \mathcal{P}} d(p, \mathcal{A}_i)} = \frac{\sum_{o \in \mathcal{O}} d(o, \mathcal{A}_i)}{\sum_{p \in \mathcal{P}} d(p, \mathcal{A}_i)} > \frac{\epsilon}{\beta}. \quad (1)$$

We consider $\lceil 2\ell\beta\epsilon^{-1} \rceil$ independent binary random variables $v_0, \ldots, v_{\lceil 2\ell\beta\epsilon^{-1} \rceil - 1}$ with $\Pr[v_i = 1] = \epsilon\beta^{-1}$ for each $i \in [0, \lceil 2\ell\beta\epsilon^{-1} \rceil - 1]$. It can be shown that

$$\Pr[|\mathcal{A} \cap \mathcal{O}| < |\mathcal{O}|] \leq \Pr[|\mathcal{A} \cap \mathcal{O}| - |\mathcal{A}_0 \cap \mathcal{O}| < \ell]$$

Algorithm 2: The algorithm for the k-MEDOUT problem

Input: A real number $\epsilon \in (0,1)$, an instance $\mathcal{I} = ((\mathcal{X}, d), \mathcal{P}, \mathcal{F}, \ell, k)$ of the k-MEDOUT problem with $\ell > 0$, and an algorithm `Clustering` applicable only when $\ell = 0$
Output: A solution $(\mathcal{O}^\dagger, \mathcal{C}^\dagger)$ to $\mathcal{I}$

1: Let $\mathcal{A}$ be the set returned by Algorithm 1 with $(\epsilon, \mathcal{I})$ as input.
2: Let $\mathcal{S} \subseteq \mathcal{P}$ be the union of the ℓ nearest neighbors (self included) in $\mathcal{P}$ of each point in $\mathcal{A}$.
3: Let v be a virtual point with $d(v,p) = 0$ for each $p \in \mathcal{S}$.
4: $\mathcal{H} \leftarrow \emptyset$.
5: **for** each function $\mu : \mathcal{A} \cup \{v\} \to \mathbb{Z}_{\geq 0}$ that satisfies $\sum_{a \in \mathcal{A}} \mu(a) = \ell$ and $\mu(v) = |\mathcal{S}| - \ell$ **do**
6: Solve the instance $(\mathcal{S}, \mathcal{A} \cup \{v\}, \mu, d)$ of the unit-supply transportation problem using Lemma 3, and let $\phi : \mathcal{S} \to \mathcal{A} \cup \{v\}$ be the solution.
7: $\mathcal{O} \leftarrow \bigcup_{a \in \mathcal{A}} \phi^{-1}(a)$.
8: Use `Clustering` to construct a solution $(\emptyset, \mathcal{C})$ to instance $((\mathcal{X}, d), \mathcal{P} \backslash \mathcal{O}, \mathcal{F}, 0, k)$.
9: $\mathcal{H} \leftarrow \mathcal{H} \cup \{(\mathcal{O}, \mathcal{C})\}$.
10: **return** $(\mathcal{O}^\dagger, \mathcal{C}^\dagger) \leftarrow \arg\min_{(\mathcal{O},\mathcal{C}) \in \mathcal{H}} \sum_{p \in \mathcal{P} \backslash \mathcal{O}} d(p, \mathcal{C})$.

$$< \Pr\left[\sum_{i=0}^{\lceil 2\ell\beta\epsilon^{-1}\rceil - 1} v_i < \ell\right]$$
$$< e^{-1/4},$$

where the first step is due to the fact that $|\mathcal{O}| \leq \ell$, the second step follows from the fact that

$$\Pr[|\mathcal{A}_i \cap \mathcal{O}| < |\mathcal{A}_{i+1} \cap \mathcal{O}|] > \epsilon\beta^{-1} = \Pr[v_i = 1]$$

for each $i \in [0, \lceil 2\ell\beta\epsilon^{-1}\rceil - 1]$ (due to inequality (1)), and the third step follows from Lemma 1 (with $\lambda = 1/2$). Therefore, with probability at least $1 - e^{-1/4}$, we have $|\mathcal{A} \cap \mathcal{O}| = |\mathcal{O}|$, which implies that $\mathcal{O} \subseteq \mathcal{A}$ and hence $\sum_{o \in \mathcal{O}} d(o, \mathcal{A}) = 0$. This establishes the validity of Lemma 4 for case (2). □

The Anchor Points-Based Algorithm for the k-MEDOUT Problem

In this section we solve the k-MEDOUT problem based on the anchor points. Our approach is presented in Algorithm 2 and illustrated in Figure 1, which considers a real number $\epsilon \in (0,1)$, an instance $\mathcal{I} = ((\mathcal{X}, d), \mathcal{P}, \mathcal{F}, \ell, k)$ with $|\mathcal{P} \cup \mathcal{F}| = n$ and $\ell > 0$, and an algorithm `Clustering` applicable only to the case where $\ell = 0$. The algorithm constructs a set $\mathcal{A}$ of anchor points using Algorithm 1, and forms the candidate set $\mathcal{S} \subseteq \mathcal{P}$ of outliers as the union of the ℓ nearest neighbors of each anchor point. To identify a subset of this candidate set that is close to an optimal set of outliers, the algorithm iterates over all $|\mathcal{A}|$-tuples $(\mu_1, \cdots, \mu_{|\mathcal{A}|})$ of non-negative integers satisfying $\sum_{i=1}^{|\mathcal{A}|} \mu_i = \ell$, which correspond to guesses of the number of outliers captured by each anchor point. For each such tuple, it maps the corresponding number of candidate outliers to each anchor point by solving an instance of the unit-supply transportation problem. Specifically, it introduces a demand point v with $\mu(v) = |\mathcal{S}| - \ell$ and $d(v,p) = 0$ for each $p \in \mathcal{S}$ in the instance, and considers the candidates in $\mathcal{S}$ not mapped to v in the corresponding solution as outliers. Given the identified outliers, the algorithm removes them from $\mathcal{P}$ and executes algorithm `Clustering` on the remaining points to obtain the k centers. Based on these iterations, Algorithm 2 constructs a candidate set $\mathcal{H}$ of solutions and finally returns the solution $(\mathcal{O}^\dagger, \mathcal{C}^\dagger) \in \mathcal{H}$ that minimizes the cost for $\mathcal{I}$.

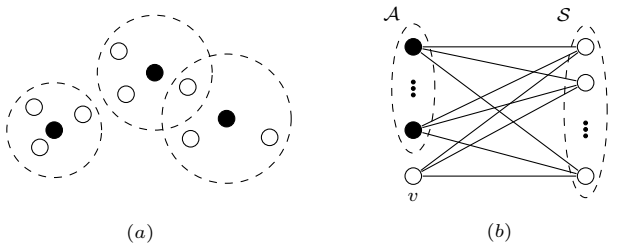

Figure 1: (a): The ℓ nearest neighbors of each anchor point in $\mathcal{A}$ (anchor points are shown in black) are collected into the set $\mathcal{S}$ of candidates for outliers; (b): An instance of the unit-supply transportation problem is constructed on a bipartite graph formed by the points in $\mathcal{A} \cup \mathcal{S} \cup \{v\}$, and is solved to select outliers from $\mathcal{S}$.

The following lemma ensures that Algorithm 2 can capture a subset of ℓ points close to the outliers, regardless of how the outliers are distributed.

Lemma 5 *Given a subset $\mathcal{O} \subseteq \mathcal{P}$ with $|\mathcal{O}| = \ell$, the following event occurs with probability at least $1 - e^{-1/4}$: There exists a solution $(\mathcal{O}', \mathcal{C}') \in \mathcal{H}$ such that a bijection $\gamma : \mathcal{O} \backslash \mathcal{O}' \to \mathcal{O}' \backslash \mathcal{O}$ satisfying $\sum_{o \in \mathcal{O} \backslash \mathcal{O}'} d(o, \gamma(o)) \leq 2\epsilon \cdot \mathrm{opt}_k(\mathcal{P} \backslash \mathcal{O}, \mathcal{F})$ can be constructed.*

Proof For each anchor point $a \in \mathcal{A}$, let $\mathcal{O}(a) = \{o \in \mathcal{O} : \arg\min_{a' \in \mathcal{A}} d(a', o) = a\}$ denote the subset of outliers in $\mathcal{O}$ captured by a. Let $(\mathcal{O}', \mathcal{C}')$ be the solution added to $\mathcal{H}$ in the iteration where each value $|\mathcal{O}(a)|$ is correctly guessed (namely $\mu(a) = |\mathcal{O}(a)|$ for each $a \in \mathcal{A}$) and the instance $\mathcal{I}' = (\mathcal{S}, \mathcal{A} \cup \{v\}, \mu, d)$ of the unit-supply transportation problem is constructed with the desired setting of the demand function. Let $\phi : \mathcal{S} \to \mathcal{A} \cup \{v\}$ denote the corresponding solution to $\mathcal{I}'$. By constructing a bijection between $\mathcal{O}(a)$ and $\phi^{-1}(a)$ for each $a \in \mathcal{A}$, we can obtain a tight bijection between $\mathcal{O}$ and $\mathcal{O}'$ with high probability, as stated in the following claim.

Claim 2 *With probability at least $1 - e^{-1/4}$, there exists a bijection $\tilde{\gamma} : \mathcal{O} \to \mathcal{O}'$ with $\sum_{o \in \mathcal{O}} d(o, \tilde{\gamma}(o)) \leq 2\epsilon \cdot \mathrm{opt}_k(\mathcal{P} \backslash \mathcal{O}, \mathcal{F})$.*

We map each $o \in \mathcal{O} \setminus \mathcal{O}'$ to $\mathcal{O}' \setminus \mathcal{O}$ by iteratively applying the bijection $\tilde{\gamma}$ stated in Claim 2 until the image lies in $\mathcal{O}' \setminus \mathcal{O}$. We define a sequence $(\tilde{\gamma}_0(o), \tilde{\gamma}_1(o), \tilde{\gamma}_2(o), \cdots)$ where $\tilde{\gamma}_0(o) = o$ and $\tilde{\gamma}_i(o) = \tilde{\gamma}(\tilde{\gamma}_{i-1}(o))$ for each $i \geq 1$. Given that $|\mathcal{O} \setminus \mathcal{O}'| = |\mathcal{O}' \setminus \mathcal{O}|$ and $\tilde{\gamma} : \mathcal{O} \to \mathcal{O}'$ is a bijection, this process is guaranteed to terminate at some index t

satisfying $\tilde{\gamma}_t(o) \in \mathcal{O}' \setminus \mathcal{O}$. We then define $\gamma(o) = \tilde{\gamma}_t(o)$, and denote this index by $t(o) = t$. It can be shown that

$$\begin{aligned}\sum_{o\in\mathcal{O}\setminus\mathcal{O}'} d(o,\gamma(o)) \le& \sum_{o\in\mathcal{O}\setminus\mathcal{O}'}\sum_{i=1}^{t(o)} d(\tilde{\gamma}_{i-1}(o), d(\tilde{\gamma}_i(o))\\ \le& \sum_{o\in\mathcal{O}} d(o,\tilde{\gamma}(o))\\ \le& 2\epsilon\cdot \mathrm{opt}_k(\mathcal{P}\setminus\mathcal{O},\mathcal{F}),\end{aligned}$$

where the first step follows from triangle inequality, the second step follows from the bijectivity of $\tilde{\gamma}$, and the third step is due to Claim (2). Consequently, γ is a bijection from $\mathcal{O}\setminus\mathcal{O}'$ to $\mathcal{O}'\setminus\mathcal{O}$ satisfying the statement of Lemma 5. □

We now analyze Algorithm 2 to show the correctness of Theorem 1.

Proof (of Theorem 1) We prove the theorem by showing that Algorithm 2 returns the desired approximation solution within the claimed time bound. We first analyze the approximation guarantee of the algorithm. Let $(\mathcal{O}^*, \mathcal{C}^*)$ be an optimal solution to $\mathcal{I}$, let $\mathrm{opt} = \sum_{p\in\mathcal{P}\setminus\mathcal{O}^*} d(p,\mathcal{C}^*)$ denote the cost of $(\mathcal{O}^*, \mathcal{C}^*)$, let α denote the approximation ratio of the outlier-free algorithm `Clustering`, and let $(\mathcal{O}', \mathcal{C}') \in \mathcal{H}$ be a solution to $\mathcal{I}$ satisfying the statement of Lemma 5 with respect to $\mathcal{O}^*$, where $\gamma : \mathcal{O}^*\setminus\mathcal{O}' \to \mathcal{O}'\setminus\mathcal{O}^*$ is the corresponding bijection. For each $p \in \mathcal{P}$, let $c(p)$ be the center in $\mathcal{C}^*$ nearest to p. The approximation guarantee of `Clustering` suggests that

$$\begin{aligned}&\sum_{p\in\mathcal{P}\setminus\mathcal{O}'} d(p,\mathcal{C}')\\ \le&\alpha\cdot \mathrm{opt}_k(\mathcal{P}\setminus\mathcal{O}',\mathcal{F})\\ \le&\alpha\cdot \sum_{p\in\mathcal{P}\setminus\mathcal{O}'} d(p,\mathcal{C}^*)\\ =&\alpha\left(\sum_{p\in(\mathcal{P}\setminus\mathcal{O}')\setminus\mathcal{O}^*} d(p,\mathcal{C}^*) + \sum_{o\in\mathcal{O}^*\setminus\mathcal{O}'} d(o,\mathcal{C}^*)\right)\\ \le&\alpha\left(\sum_{p\in(\mathcal{P}\setminus\mathcal{O}')\setminus\mathcal{O}^*} d(p,\mathcal{C}^*) + \sum_{o\in\mathcal{O}^*\setminus\mathcal{O}'} d(o,c(\gamma(o)))\right).\end{aligned} \tag{2}$$

Observe that

$$\begin{aligned}&\sum_{o\in\mathcal{O}^*\setminus\mathcal{O}'} d(o,c(\gamma(o)))\\ \le& \sum_{o\in\mathcal{O}^*\setminus\mathcal{O}'} d(o,\gamma(o)) + \sum_{o\in\mathcal{O}^*\setminus\mathcal{O}'} d(\gamma(o),\mathcal{C}^*)\\ \le& 2\epsilon\cdot\mathrm{opt}_k(\mathcal{P}\setminus\mathcal{O}^*,\mathcal{F}) + \sum_{o\in\mathcal{O}'\setminus\mathcal{O}^*} d(o,\mathcal{C}^*)\\ =& 2\epsilon\cdot\mathrm{opt} + \sum_{o\in\mathcal{O}'\setminus\mathcal{O}^*} d(o,\mathcal{C}^*),\end{aligned} \tag{3}$$

where the first step follows from triangle inequality, and the second step follows from Lemma 5 and the fact that γ is a bijection from $\mathcal{O}^*\setminus\mathcal{O}'$ to $\mathcal{O}'\setminus\mathcal{O}^*$. Combining inequality (2) with inequality (3), we get

$$\begin{aligned}&\sum_{p\in\mathcal{P}\setminus\mathcal{O}'} d(p,\mathcal{C}')\\ \le&\alpha\left(\sum_{p\in(\mathcal{P}\setminus\mathcal{O}')\setminus\mathcal{O}^*} d(p,\mathcal{C}^*) + \sum_{o\in\mathcal{O}'\setminus\mathcal{O}^*} d(o,\mathcal{C}^*) + 2\epsilon\cdot\mathrm{opt}\right)\\ =&\alpha\left(\sum_{p\in\mathcal{P}\setminus\mathcal{O}^*} d(p,\mathcal{C}^*) + 2\epsilon\cdot\mathrm{opt}\right)\\ =&\alpha(1+2\epsilon)\mathrm{opt}.\end{aligned}$$

This inequality implies that the approximation ratio of Algorithm 2 is $\alpha(1+2\epsilon)$.

We now analyze the running time of Algorithm 2. Let $T(n-\ell, k)$ denote the running time of algorithm `Clustering` on instances with $n-\ell$ points and at most k centers. The set $\mathcal{A}$ of anchor points consists of $O(k+\ell)$ centers and $O(\ell\epsilon^{-1})$ points sampled from $\mathcal{P}$. The set $\mathcal{S}$ of candidates for outliers is formed by collecting the ℓ nearest neighbors of each anchor point. Consequently, we have $|\mathcal{A}| = O(k+\ell\epsilon^{-1})$ and $|\mathcal{S}| \le O(k\ell+\ell^2\epsilon^{-1})$. Identifying the anchor points and candidates for outliers involves solving an instance of the $(k+\ell)$-median problem using the linear-time algorithm in Lemma 2, and computing the distances from the anchor points to all points in $\mathcal{P}$ (used for determining the sampling probabilities in Algorithm 1 and finding the neighbors of the anchor points), which takes $O(n(k+\ell\epsilon^{-1}))$ time. Let τ denote the number of iterations of steps 6–9 in Algorithm 2. In each iteration, the algorithm solves an instance of the unit-supply transportation problem based on Lemma 3 and an instance of the k-MED problem using `Clustering`, which takes $T(n-\ell,k) + (|\mathcal{A}|\cdot|\mathcal{S}|)^{O(1)} \le T(n-\ell,k) + (k\ell\epsilon^{-1})^{O(1)}$ time. Putting everything together, the running time of Algorithm 2 is $\tau\cdot T(n-\ell,k) + \tau(k\ell\epsilon^{-1})^{O(1)} + O(n(k+\ell\epsilon^{-1}))$.

It remains to analyze the value of τ, which is the number of $|\mathcal{A}|$-tuples $(\mu_1,\cdots,\mu_{|\mathcal{A}|})$ of non-negative integers satisfying $\sum_{i=1}^{|\mathcal{A}|}\mu_1 = \ell$. We have

$$\begin{aligned}\tau =& \binom{|\mathcal{A}|+\ell-1}{\ell}\\ <& \frac{(|\mathcal{A}|+\ell-1)^\ell}{\ell!}\\ <& \left(\frac{e(|\mathcal{A}|+\ell-1)}{\ell}\right)^\ell\\ =& (k\ell^{-1}+\epsilon^{-1})^{O(\ell)},\end{aligned}$$

where we use the counting technique of *stars and bars* in the first step, the third step follows from the lower bound of Stirling's approximation, and the last step follows from the fact that $|\mathcal{A}| = O(k+\ell\epsilon^{-1})$.

Letting $\varepsilon = \epsilon/2$, the discussion above implies the correctness of Theorem 1. □

Extensions Under Cluster Size Constraints

In real-world applications involving clustering, it is often necessary to impose constraints on cluster sizes. For example, lower bounds can be used to ensure that the clustering results satisfy the k-*anonymity* principle (Arutyunova and Schmidt 2021), and upper bounds can help prevent load imbalance among cluster centers. In this section, we consider the generalization of the k-MEDOUT problem incorporating such cluster size constraints, which is referred to as the *size-constrained* k-MEDOUT (SC-k-MEDOUT) problem and defined as follows.

Definition 3 (SC-k-MEDOUT) *An instance* $((\mathcal{X}, d), \mathcal{P}, \mathcal{F}, \ell, k)$ *of the* k-MEDOUT *problem can be extended to its size-constrained variant* $((\mathcal{X}, d), \mathcal{P}, \mathcal{F}, \ell, k, \mu_1, \mu_2)$ *by incorporating two mappings* $\mu_1, \mu_2 : \mathcal{F} \to \mathbb{Z}_{\geq 0}$*, where* $\mu_1(c) \leq \mu_2(c)$ *for each* $c \in \mathcal{F}$*. A feasible solution* $(\mathcal{O}, \mathcal{C}, \varphi)$ *to the variant is specified by a subset* $\mathcal{O} \subseteq \mathcal{P}$ *of no more than* ℓ *outliers, a subset* $\mathcal{C} \subseteq \mathcal{F}$ *of no more than* k *centers, and a mapping* $\varphi : \mathcal{P} \setminus \mathcal{O} \to \mathcal{C}$ *with* $|\varphi^{-1}(c)| \in [\mu_1(c), \mu_2(c)]$ *for all* $c \in \mathcal{C}$*. The cost of the solution is* $\sum_{p \in \mathcal{P} \setminus \mathcal{O}} d(p, \varphi(p))$*, and the objective is to find a feasible solution minimizing this cost.*

A key distinction between the size-constrained and unconstrained k-MEDOUT problems lies in the structural properties of their respective optimal solutions. In the unconstrained setting, each point is assigned to its nearest center in an optimal solution, and the set of outliers consists of the ℓ points farthest from their respective nearest centers. In contrast, for the SC-k-MEDOUT problem, constructing solutions solely based on point-to-center distances may violate feasibility, and the optimal solutions no longer exhibit the aforementioned structural regularities. Encouragingly, our approach to identifying a near-optimal set of outliers makes no assumptions on the spatial distribution of outliers in optimal solutions, which allows the outliers to be arbitrary points, as established in Lemma 5. This enables us to effortlessly extend our approach from the unconstrained case to the size-constrained setting. Indeed, our algorithm for the SC-k-MEDOUT problem retains the overall structure of Algorithm 2, except that the set $\mathcal{H}$ of candidate solutions is constructed using a subroutine applicable to the size-constrained setting, and the final solution is selected based on its cost under the cluster size constraints. Analyzing the performance of this variant of Algorithm 2 yields the following theorem.

Theorem 2 *Given an instance* $\mathcal{I} = ((\mathcal{X}, d), \mathcal{P}, \mathcal{F}, \ell, k, \mu_1, \mu_2)$ *of the* SC-k-MEDOUT *problem satisfying* $|\mathcal{P} \cup \mathcal{F}| = n$ *and* $\ell > 0$*, an* α*-approximation algorithm for the outlier-free counterpart of the problem that runs in time* $T(n - \ell, k)$ *on instances with* $n - \ell$ *points and at most* k *centers, and a constant* $\varepsilon \in (0, 1)$*, there exists an* $\alpha(1 + \varepsilon)$*-approximation algorithm for* $\mathcal{I}$ *with running time* $\tau \cdot T(n - \ell, k) + \tau(k\ell\varepsilon^{-1})^{O(1)} + O(n(k + \ell\varepsilon^{-1}))$*, where* $\tau = (k\ell^{-1} + \varepsilon^{-1})^{O(\ell)}$.

Notably, the applicability of Theorem 2 hinges on the availability of a suitable approximation algorithm for solving the outlier-free counterpart of the considered instance. In the general case where each center is associated with both non-uniform lower and upper bounds, no efficient approximation algorithms are currently known. However, several special cases of the problem have been studied, for which fixed-parameter approximation algorithms (with parameter k) are known. These special cases include

(i) the *capacitated* k-MED problem, where each center is associated only with an upper bound on the cluster size (Adamczyk et al. 2019; Cohen-Addad and Li 2019; Goyal, Jaiswal, and Kumar 2020; Bandyapadhyay, Fomin, and Simonov 2024),

(ii) the *lower-bounded* k-MED problem, where each center is associated only with a lower bound (Goyal, Jaiswal, and Kumar 2020; Bandyapadhyay, Fomin, and Simonov 2024),

(iii) and the *balanced* k-MED problem, where all centers are associated with uniform upper and lower bounds (Ding 2020; Kong, Zhang, and Feng 2023).

In each of these cases, Theorem 2 can be combined with an existing outlier-free algorithm to yield an approximation algorithm for the respective special case of the SC-k-MEDOUT problem.

Conclusions

In this paper, we study the k-MEDOUT problem under the setting where the maximum numbers of outliers and centers are treated as fixed parameters. By locating the outliers in an optimal solution based on a carefully chosen set of anchor points, we reduce the k-MEDOUT problem to its outlier-free counterpart. This reduction leads to a faster FPT algorithm with a tight approximation guarantee. Moreover, our approach is applicable to variants of the problem that impose additional constraints on the cluster sizes, and yields similar improvements in the FPT approximation results.

Given the known lower bound on the approximation ratios achievable by FPT algorithms for the k-MEDOUT problem (Cohen-Addad et al. 2019), further improvement from the perspective of approximation guarantees for the problem is unlikely. Nevertheless, it remains an open question whether comparable approximation guarantees can be attained through more efficient algorithms. Notably, the existing inapproximability result (Cohen-Addad et al. 2019) is established in the outlier-free case, and thus does not preclude the possibility of achieving comparable approximation guarantees for the k-MEDOUT problem using FPT algorithms parameterized solely by k. This suggests a promising direction for future research in the design of more efficient parameterized algorithms. Moreover, the reduction presented in this paper, along with the reductions in (Agrawal et al. 2023; Jaiswal and Kumar 2023), relies on randomized procedures and therefore introduces randomness into the resulting algorithms. Understanding whether deterministic constructions can achieve similar guarantees forms another natural direction to explore. Finally, it is worth investigating whether the anchor point-based techniques developed in this work can be extended to obtain analogous improvements under other clustering objectives, such as k-*means* and k-*center*.

Acknowledgments

This work was supported by National Natural Science Foundation of China (62202161), Open Project of Xiangjiang Laboratory (23XJ01001), National Natural Science Foundation of China (62432016, 62502545), Science and Technology Innovation Program of Hunan Province (2025RC3207), Scientific Research Fund of Hunan Provincial Education Department (23B0592), Innovation Fund of QiYuan Lab (2022-JCJQ-LA-001-088), and Key Research and Development Program of Hunan Province (2024JK2007).

References

Adamczyk, M.; Byrka, J.; Marcinkowski, J.; Meesum, S. M.; and Wlodarczyk, M. 2019. Constant-factor FPT approximation for capacitated k-median. In *Proceedings of the 27th Annual European Symposium on Algorithms (ESA)*, volume 144, 1:1–1:14.

Agrawal, A.; Inamdar, T.; Saurabh, S.; and Xue, J. 2023. Clustering what matters: Optimal approximation for clustering with outliers. In *Proceeding of the 37th AAAI Conference on Artificial Intelligence (AAAI)*, 6666–6674.

Arutyunova, A.; and Schmidt, M. 2021. Achieving anonymity via weak lower bound constraints for k-median and k-means. In *Proceedings of the 38th International Symposium on Theoretical Aspects of Computer Science (STACS)*, volume 187, 7:1–7:17.

Bandyapadhyay, S.; Fomin, F. V.; and Simonov, K. 2024. On coresets for fair clustering in metric and Euclidean spaces and their applications. *Journal of Computer and System Sciences*, 142: 103506.

Chakraborty, D.; Das, D.; and Krauthgamer, R. 2023. Clustering permutations: New techniques with streaming applications. In *Proceedings of the 14th Innovations in Theoretical Computer Science Conference (ITCS)*, volume 251, 31:1–31:24.

Chalermsook, P.; Cygan, M.; Kortsarz, G.; Laekhanukit, B.; Manurangsi, P.; Nanongkai, D.; and Trevisan, L. 2020. From gap-exponential time hypothesis to fixed parameter tractable inapproximability: Clique, dominating set, and more. *SIAM Journal on Computing*, 49(4): 772–810.

Charikar, M.; Khuller, S.; Mount, D. M.; and Narasimhan, G. 2001. Algorithms for facility location problems with outliers. In *Proceedings of the 12th Annual Symposium on Discrete Algorithms (SODA)*, 642–651.

Chen, K. 2009. On coresets for k-median and k-means clustering in metric and Euclidean spaces and their applications. *SIAM Journal on Computing*, 39(3): 923–947.

Chen, X.; Han, L.; Xu, D.; Xu, Y.; and Zhang, Y. 2023. k-median/means with outliers revisited: A simple FPT Approximation. In *Proceedings of the 29th International Conference on Computing and Combinatorics (COCOON)*, volume 14423, 295–302.

Chernoff, H. 1952. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23(4): 493–507.

Cohen-Addad, V.; Feldmann, A. E.; and Saulpic, D. 2021. Near-linear time approximation schemes for clustering in doubling metrics. *Journal of the ACM*, 68(6): 44:1–44:34.

Cohen-Addad, V.; Grandoni, F.; Lee, E.; and Schwiegelshohn, C. 2023. Breaching the 2 LMP approximation barrier for facility location with applications to k-median. In *Proceedings of the 34th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 940–986.

Cohen-Addad, V.; Gupta, A.; Kumar, A.; Lee, E.; and Li, J. 2019. Tight FPT approximations for k-median and k-means. In *Proceedings of the 46th International Colloquium on Automata, Languages, and Programming (ICALP)*, volume 132, 42:1–42:14.

Cohen-Addad, V.; and Li, J. 2019. On the fixed-parameter tractability of capacitated clustering. In *Proceedings of the 46th International Colloquium on Automata, Languages, and Programming (ICALP)*, volume 132, 41:1–41:14.

Cohen-Addad, V.; Saulpic, D.; and Schwiegelshohn, C. 2021. A new coreset framework for clustering. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, 169–182.

Dabas, R.; Gupta, N.; and Inamdar, T. 2025. FPT approximation for capacitated clustering with outliers. *Theoretical Computer Science*, 1027: 115026.

Ding, H. 2020. Faster balanced clusterings in high dimension. *Theoretical Computer Science*, 842: 28–40.

Feldman, D.; and Schulman, L. J. 2012. Data reduction for weighted and outlier-resistant clustering. In *Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 1343–1354.

Friggstad, Z.; Khodamoradi, K.; Rezapour, M.; and Salavatipour, M. R. 2019. Approximation schemes for clustering with outliers. *ACM Transactions on Algorithms*, 15(2): 26:1–26:26.

Gowda, K. N.; Pensyl, T. W.; Srinivasan, A.; and Trinh, K. 2023. Improved bi-point rounding algorithms and a golden barrier for k-median. In *Proceedings of the 34th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 987–1011.

Goyal, D.; Jaiswal, R.; and Kumar, A. 2020. FPT approximation for constrained metric k-median/means. In *Proceedings of the 15th International Symposium on Parameterized and Exact Computation (IPEC)*, volume 180, 14:1–14:19.

Gupta, A.; Moseley, B.; and Zhou, R. 2021. Structural iterative rounding for generalized k-median problems. In *Proceedings of the 48th International Colloquium on Automata, Languages, and Programming (ICALP)*, volume 198, 77:1–77:18.

Huang, J.; Liu, W.; and Ding, H. 2024. Bi-criteria sublinear time algorithms for clustering with outliers in high dimensions. In *Proceedings of the 30th International Conference on Computing and Combinatorics (COCOON)*, volume 15161, 91–103.

Huang, L.; Li, J.; and Wu, X. 2024. On optimal coreset construction for Euclidean (k, z)-clustering. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC)*, 1594–1604.

Jaiswal, R.; and Kumar, A. 2023. Clustering what matters in constrained settings: Improved outlier to outlier-free reductions. In *Proceeding of the 34th International Symposium on Algorithms and Computation (ISAAC)*, volume 283, 41:1–41:16.

Kong, X.; Zhang, Z.; and Feng, Q. 2023. On parameterized approximation algorithms for balanced clustering. *Journal of Combinatorial Optimization*, 45(1): 49.

Schrijver, A. 1998. *Theory of linear and integer programming*. Chichester, England: John Wiley & Sons.

Wei, D. 2016. A constant-factor bi-criteria approximation guarantee for k-means++. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 604–612.

Wu, C.; Möhring, R. H.; Wang, Y.; Xu, D.; and Zhang, D. 2024. Approximation algorithms for robust clustering problems using local search techniques. In *Proceedings of the 18th Annual Conference on Theory and Applications of Models of Computation (TAMC)*, volume 14637, 197–208.