Journal of Artificial Intelligence Research 78 (2023) 143-166 Submitted 02/2023; published 09/2023

Clustering what Matters: Optimal Approximation for Clustering with Outliers

Akanksha Agrawal AKANKSHA@CSE.IITM.AC.IN
Indian Institute Of Technology, Chennai 600036, India

Tanmay Inamdar TANINAMDAR@GMAIL.COM
University of Bergen, Bergen 5007, Norway

Saket Saurabh SAKET@IMSC.RES.IN
Institute of Mathematical Sciences, 4th Cross St, Chennai 600113, India
University of Bergen, Bergen 5007, Norway

Jie Xue JIEXUE@NYU.EDU
NYU Shanghai, 1555 Shiji Blvd, Pudong, Shanghai, China, 200122a

Abstract

Clustering with outliers is one of the most fundamental problems in Computer Science. Given a set X of n points and two numbers k, m, the clustering with outliers aims to exclude m points from X and partition the remaining points into k clusters that minimizes a certain cost function. In this paper, we give a general approach for solving clustering with outliers, which results in a fixed-parameter tractable (FPT) algorithm in k and m—i.e., an algorithm with running time of the form $f(k, m) \cdot n^{O(1)}$ for some function f—that almost matches the approximation ratio for its outlier-free counterpart. As a corollary, we obtain FPT approximation algorithms with optimal approximation ratios for k-Median and k-Means with outliers in general and Euclidean metrics. We also exhibit more applications of our approach to other variants of the problem that impose additional constraints on the clustering, such as fairness or matroid constraints.

1. Introduction

Clustering is a family of problems that aims to group a given set of objects in a meaningful way—the exact "meaning" may vary based on the application. These are fundamental problems in Computer Science with applications ranging across multiple fields like pattern recognition, machine learning, computational biology, bioinformatics and social science. Thus, these problems have been a subject of extensive studies in the field of Algorithm Design (and its sub-fields), see for instance, the surveys on this topic (and references therein) (Xu & Tian, 2015; Rokach, 2009; Blömer, Lammersen, Schmidt, & Sohler, 2016).

Two of the central clustering problems are k-Median and k-Means. In the standard k-Median problem, we are given a set X of n points, and an integer k, and the goal is to find a set $C^* \subseteq X$ of at most k *centers*, such that the following cost function is minimized over all subsets C of size at most k.

$$\mathsf{cost}(X, C) \coloneqq \sum_{p \in X} \min_{c \in C} \mathsf{d}(p, c)$$

In k-Means, the objective function instead contains the sum of squares of distances.

Often real world data are contaminated with a small amount of noise and these noises can substantially change the clusters that we obtain using the underlying algorithm. To circumvent the issue created by such noises, there are several studies of clustering problems with outliers, see for instance, (Chen, 2008; Krishnaswamy, Li, & Sandeep, 2018; Goyal, Jaiswal, & Kumar, 2020; Feng, Zhang, Huang, Xu, & Wang, 2019; Friggstad, Khodamoradi, Rezapour, & Salavatipour, 2019a; Almanza, Epasto, Panconesi, & Re, 2022).

In outlier extension of the k-Median problem, which we call k-MedianOut, we are also given an additional integer $m \geq 0$ that denotes the number of *outliers* that we are allowed to drop. We want to find a set C of at most k centers, and a set $Y \subseteq X$ of at most m outliers, such that $\mathsf{cost}(X \setminus Y, C) := \sum_{p \in X \setminus Y} \min_{c \in C} \mathsf{d}(p, c)$ is minimized over all (Y, C) satisfying the requirements. Observe that the cost of clustering for k-MedianOut equals the sum of distances of each point to its nearest center, after excluding a set of m points from consideration [1]. We remark that in a similar spirit we can define the outlier version of the k-Means problem, which we call k-MeansOut.

In this paper, we will focus on approximation algorithms. An algorithm is said to have an approximation ratio of α, if it is guaranteed to return a solution of cost no greater than α times the optimal cost, while satisfying all other conditions. That is, the solution must contain at most k centers, and drop m outliers. If the algorithm is randomized, then it must return such a solution with *high probability*, i.e., probability at least $1 - n^{-c}$ for some $c \geq 1$.

For a fixed set C of centers, the set of m outliers is automatically defined, namely the set of m points that are farthest from C (breaking ties arbitrarily). Thus, an optimal clustering for k-MedianOut, just like k-Median, can be found in $n^{O(k)}$ time by enumerating all center sets. On the other hand, we can enumerate all $n^{O(m)}$ subsets of outliers, and reduce the problem directly to k-Median. Other than these straightforward observations, there are several non-trivial approximations known for k-MedianOut, which we discuss in a subsequent paragraph.

1.1 Our Results

In this work, we describe a general framework that reduces a *clustering with outliers* problem (such as k-MedianOut or k-MeansOut) to its outlier-free counterpart in an approximation-preserving fashion. More specifically, given an instance $\mathcal{I}$ of k-MedianOut, our reduction runs in time $f(k, m, \epsilon) \cdot n^{O(1)}$, and produces multiple instances of k-Median, such that a β-approximation for *at least* one of the produced instances of k-Median implies a $(\beta + \epsilon)$-approximation for the original instance $\mathcal{I}$ of k-MedianOut. This is the main result of our paper.

Our framework does not rely on the specific properties of the underlying metric space. Thus, for special metrics, such as Euclidean spaces, or shortest-path metrics induced by sparse graph classes, for which FPT $(1 + \epsilon)$-approximations are known for k-Median, our framework implies matching approximation for k-MedianOut. Finally, our framework is quite versatile in that one can extend it to obtain approximation-preserving FPT reductions

1. Our results actually hold for a more general formulation of k-Median, where the set of candidate centers may be different from the set X of points to be clustered. We consider this general setting in the technical sections.

for related *clustering with outliers* problems, such as k-MeansOut, and clustering problems with *fair outliers* (such as (Bandyapadhyay, Inamdar, Pai, & Varadarajan, 2019; Jia, Sheth, & Svensson, 2020)), and Matroid Median with Outliers. We conclude by giving a partial list of the corollaries of our reduction framework. The running time of each algorithm is $f(k, m, \epsilon) \cdot n^{O(1)}$ for some function f that depends on the problem and the setting. Next to each result, we also cite the result that we use as a black box to solve the outlier-free clustering problem.

- $(1 + \frac{2}{e} + \epsilon) \approx (1.74 + \epsilon)$-approximation (resp. $1 + \frac{8}{e} + \epsilon$)-approximation) for k-MedianOut (resp. k-MeansOut) in general metrics (Cohen-Addad, Saulpic, & Schwiegelshohn, 2021). These approximations are tight even for $m = 0$, under a reasonable complexity theoretic hypothesis, as shown in the same paper.
- $(1 + \epsilon)$-approximation for k-MedianOut and k-MeansOut in (i) metric spaces of constant doubling dimensions, which includes Euclidean spaces of constant dimension, (ii) metrics induced by graphs of bounded treewidth, and (iii) metrics induced by graphs that exclude a fixed graph as a minor (such as planar graphs). (Cohen-Addad et al., 2021).
- $(2 + \epsilon)$-approximation for Matroid Median with Outliers in general metrics, where k refers to the *rank* of the matroid. (Cohen-Addad, Gupta, Kumar, Lee, & Li, 2019)
- $(1 + \frac{2}{e} + \epsilon)$-approximation for Colorful k-Median in general metrics, where m denotes the total number of outliers across all color classes (Cohen-Addad et al., 2019). The preceding two problems are orthogonal generalizations of k-MedianOut, and are formally defined in Section 4.

1.2 Our Techniques

Our reduction is inspired from the following seemingly simple observation that relates k-MedianOut and k-Median. Let $\mathcal{I}$ be an instance of k-MedianOut, where we want to find a set C of k centers, such that the sum of distances of all except at most m points to the nearest center in C is minimized. By treating the outliers in an optimal solution for $\mathcal{I}$ as *virtual centers*, one obtains a solution for $(k+m)$-Median *without outliers* whose cost is at most the optimal cost of $\mathcal{I}$. In other words, the optimal cost of an appropriately defined instance $\widetilde{\mathcal{I}}$ of $(k+m)$-Median is a *lower bound* on the optimal cost of $\mathcal{I}$. Since k-Median is a well-studied problem, at this point, one would hope that it is sufficient to restrict the attention to $\widetilde{\mathcal{I}}$. That is, if we obtain a solution (i.e., a set of $k+m$ centers) for $\widetilde{\mathcal{I}}$, can then be modified to obtain a solution (i.e., a set of k centers and m outliers) for $\mathcal{I}$. However, it is unclear whether one can do such a modification without blowing up the cost for $\mathcal{I}$. Nevertheless, this connection between $\widetilde{\mathcal{I}}$ and $\mathcal{I}$ turns out to be useful, but we need several new ideas to exploit it.

As in before, we start with a constant approximation for $\widetilde{\mathcal{I}}$, and perform a sampling similar to (Chen, 2009) to obtain a weighted set of points. This set is obtained by dividing the set of points connected to each center in the approximate solution into *concentric rings*, such that the "error" introduced in the cost by treating all points in the ring as identical

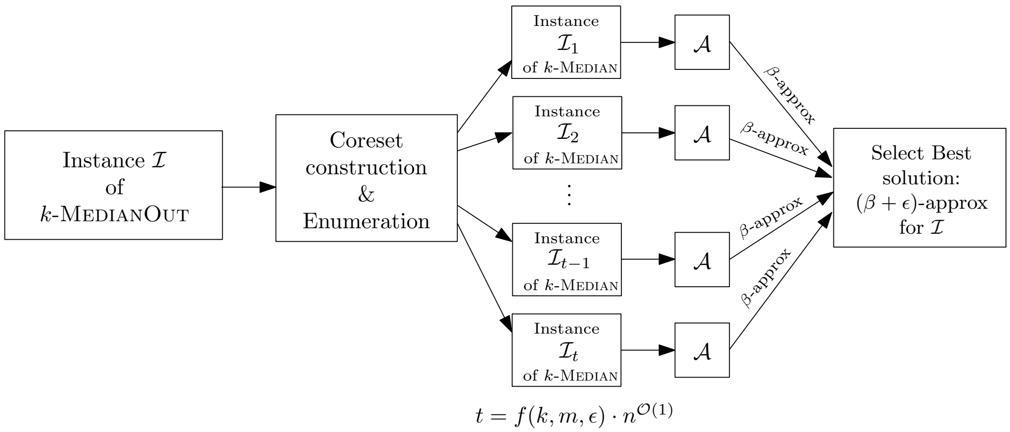

Figure 1: A conceptual overview of our approximation-preserving reduction for k-MEDIANOUT to k-MEDIAN. Here, $\mathcal{A}$ denotes a β-approximation algorithm for k-MEDIAN, which we use as a black box on each of the instances $\mathcal{I}_1, \ldots, \mathcal{I}_t$ of k-MEDIAN.

is negligible. Then, we sample $O((k + m) \log n/\epsilon)$ points from each ring, and give each point an appropriate weight. We then prove a crucial concentration bound (cf. Lemma 1), which informally speaking relates the connection cost of original set of points in a ring, and the corresponding weighted sample. In particular, for *any set of k centers*, with good probability, the difference between the original and the weighted costs is "small", *even after excluding at most m outliers from both sets.* Intuitively speaking, this concentration bound holds because the sample size is large enough compared to both k and m. Then, by taking the union of all such samples, we obtain a weighted set S of $O(((k + m) \log n/\epsilon)^2)$ points that preserves the connection cost to any set of k centers, even after excluding m outliers with at least a constant probability. Then, we enumerate all sets Y of size m from S, and solve the resulting k-MEDIAN instance induced on $S \setminus Y$. Finally, we argue that at least one of the resulting instances $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_t$ (we show that $t = f(k, m, \epsilon) \cdot n^{O(1)}$ for some function f) will have the property that, a β-approximation for $\mathcal{I}'$ implies a $(\beta + \epsilon)$-approximation for $\mathcal{I}$. See Figure 1 for a conceptual flowchart of the algorithm.

1.3 Related Work

The first constant approximation for k-MEDIANOUT was given by (Chen, 2008) for some large constant. More recently, (Krishnaswamy et al., 2018; ?) gave constant approximations based on iterative LP rounding technique, and the 6.387-approximation by latter is currently the best known approximation. These approximation algorithms run in polynomial time in n. (Krishnaswamy et al., 2018) also give the best known polynomial approximations for related problems of k-MEANSOUT and MATROID MEDIAN.

Now we turn to FPT approximations, which is also the setting for our results. To the best of our knowledge, there are three works in this setting, (Feng et al., 2019; Goyal et al., 2020; Statman, Rozenberg, & Feldman, 2020). The idea of relating k-Median with m Outliers to $(k+m)$-Median that we discuss above is also present in these works. Even though it is not stated explicitly, the approach of (Statman et al., 2020) can be used to obtain FPT approximations in general metrics; albeit with a worse approximation ratio. However, by using additional properties of Euclidean k-MedianOut/k-MeansOut (where one is allowed to place centers anywhere in $\mathbb{R}^d$) their approach yields a $(1+\epsilon)$-approximation in FPT time. (Goyal et al., 2020) design approximation algorithms with ratio of $3+\epsilon$ for k-MedianOut (resp. $9+\epsilon$ for k-MedianOut) in time $((k+m)/\epsilon)^{O(k)} \cdot n^{O(1)}$. Thus, our approximation ratios of $1+\frac{2}{e}+\epsilon$ for k-MeansOut, and $1+\frac{8}{e}+\epsilon$ for k-MeansOut improve on these results – albeit with a slightly worse FPT running time. Furthermore, our result is essentially an *approximation-preserving reduction* from k-MedianOut to k-Median in the same metric, which yields $(1+\epsilon)$-approximations in some special settings as discussed earlier. On the other hand, it seems that a loss of $3+\epsilon$ (resp. $9+\epsilon$) in the approximation guarantee is inherent to the algorithm of (Goyal et al., 2020).

On the lower bound side, (Guha & Khuller, 1999) showed it is NP-hard to approximate k-Median (and thus k-MedianOut) within a factor $1+\frac{2}{e}-\epsilon$ for any $\epsilon > 0$. Recently, (Cohen-Addad et al., 2019) strengthened this result assuming Gap-ETH, and showed that an $(1+\frac{2}{e}-\epsilon)$-approximation algorithm must take at least $n^{k^{g(\epsilon)}}$ time for some function $g()$.

Bicriteria approximations relax the strict requirement of using at most k centers, or dropping at most m outliers, in order to give improved approximation ratios, or efficiency (or both). For k-MedianOut, (Charikar, Khuller, Mount, & Narasimhan, 2001) gave a $4(1+1/\epsilon)$-approximation, while dropping $m(1+\epsilon)$ outliers. (Gupta, Kumar, Lu, Moseley, & Vassilvitskii, 2017) gave a constant approximation based on local search for k-MeansOut that drops $O(km\log(n\Delta))$ outliers, where Δ is the diameter of the set of points. (Friggstad, Rezapour, & Salavatipour, 2019b) gave a $(25+\epsilon)$-approximation that uses $k(1+\epsilon)$ centers but only drops m outliers. In Euclidean spaces, they also give a $(1+\epsilon)$-approximation that returns a solution with $k(1+\epsilon)$ centers.

2. Preliminaries

Basic notions. Let (Γ, d) be a metric space, where Γ is a finite set of points, and $\mathsf{d} : \Gamma \times \Gamma \to \mathbb{R}$ is a distance function satisfying symmetry and triangle inequality. For any finite set $S \subseteq \Gamma$ and a point $p \in \Gamma$, we let $\mathsf{d}(p, S) \coloneqq \min_{s \in S} \mathsf{d}(p, C)$, and let $\mathsf{diam}(S) \coloneqq \max_{x,y \in S} \mathsf{d}(x, y)$. For two non-empty sets $S, C \subseteq \Gamma$, let $\mathsf{d}(S, C) = \min_{p \in S} \mathsf{d}(p, S) = \min_{p \in S} \min_{c \in C} \mathsf{d}(p, c)$. For a point $p \in \Gamma, r \geq 0$, and a set $C \subseteq \Gamma$, let $B_C(p, r) = \{q \in C : \mathsf{d}(p, q) \leq r\}$. Let T be a finite (multi)set of n real numbers, for some positive integer n, and let $1 \leq m \leq n$. Then, we use the notation $\mathsf{sum}_{\sim m}(T)$ to denote the sum of $n-m$ smallest values in T (including repetitions in case of a multi-set).

The k-median problem. In the k-Median problem, an instance is a triple $\mathcal{I} = (X, F, k)$, where X and F are finite sets of points in some metric space (Γ, d), and $k \geq 1$ is an integer. The points in X are called *clients*, and the points in F are called *facilities* or *centers*. The task is to find a subset $C \subseteq F$ of size at most k that minimizes the cost function

$$\mathsf{cost}(X, C) \coloneqq \sum_{p \in X} \mathsf{d}(p, C).$$

The *size* of an instance $\mathcal{I} = (X, F, k)$ is defined as $|\mathcal{I}| = |X \cup F|$, which we denote by n.

k-median with outliers. The input to k-MEDIANOUT contains an additional integer $0 \le m \le n$, and thus an instance is given by a 4-tuple $\mathcal{I} = (X, F, k, m)$. Let $C \subseteq F$ be a set of facilities. We define $\mathsf{cost}_m(X, C) := \mathsf{sum}_{\sim m}\{\mathsf{cost}(p, C) : p \in X\}$, i.e., the sum of $n - m$ smallest distances of points in X to the set of centers C. The goal is to find a set of centers C minimizing $\mathsf{cost}_m(X, C)$ over all sets $C \subseteq F$ of size at most k. Given a set $C \subseteq F$ of centers, we denote the corresponding solution by (Y, C), where $Y \subseteq X$ is a set of m outlier points in X with largest distances realizing $\mathsf{cost}_m(X, C)$. Given an instance $\mathcal{I}$ of k-MEDIANOUT, we use $\mathsf{OPT}(\mathcal{I})$ to denote the value of an optimal solution to $\mathcal{I}$.

Weighted sets and random samples. During the course of the algorithm, we will often deal with *weighted sets* of points. Here, $S \subseteq X$ is a weighted set, with each point $p \in S$ having integer weight $w(p) \ge 0$. For any set $C \subseteq F$ and $1 \le m \le |S|$, define $\mathsf{wcost}_m(S, C) := \mathsf{sum}_{\sim m}\{d(p, C) \cdot w(p) : p \in S\}$. A *random sample* of a finite set S refers to a random subset of S. Throughout this paper, random samples are always generated by picking points *uniformly* and *independently*.

3. k-Median with Outliers

In this section, we give our FPT reduction from k-MEDIANOUT to the standard k-MEDIAN problem. Formally, we shall prove the following theorem.

Theorem 1. *Suppose there exists a β-approximation algorithm for k-MEDIAN with running time $T(n, k)$, and a τ-approximation algorithm for $k+m$-MEDIAN with polynomial running time, where β and τ are constants. Then there exists a $(\beta + \epsilon)$-approximation algorithm for k-MEDIANOUT with running time $\left(\frac{k+m}{\epsilon}\right)^{O(m)} \cdot T(n, k) \cdot n^{O(1)}$, where n is the instance size and m is the number of outliers.*

Combining the above theorem with the known $(1 + \frac{2}{e} + \epsilon)$-approximation k-median algorithm (Cohen-Addad et al., 2019) that runs in $(k/\epsilon)^{O(k)} \cdot n^{O(1)}$ time, we directly have the following result.

Corollary 1. *There exists a $(1 + \frac{2}{e} + \epsilon)$-approximation algorithm for k-MEDIANOUT with running time $\left(\frac{k+m}{\epsilon}\right)^{O(m)} \cdot \left(\frac{k}{\epsilon}\right)^{O(k)} \cdot n^{O(1)}$, where n is the instance size and m is the number of outliers.*

The rest of this section is dedicated to proving Theorem 1. Let $\mathcal{I} = (X, F, k, m)$ be an instance of k-MEDIANOUT. We define a $(k + m)$-MEDIAN instance $\mathcal{I}' = (X, F \cup X, k + m)$, where in addition to the original set of facilities, there is a facility co-located with each client. We have the following observation.

Observation 1. $\mathsf{OPT}(\mathcal{I}') \le \mathsf{OPT}(\mathcal{I})$*, i.e., the value of an optimal solution to $\mathcal{I}'$ is a lower bound on the value of an optimal solution to $\mathcal{I}$.*

Proof. Let (Y^*, C^*) be an optimal solution to $\mathcal{I}$ realizing the value $\mathsf{OPT}(\mathcal{I})$. We define a solution (Y', C') for $\mathcal{I}'$ as follows: let $Y' = \emptyset$, and $C' = C^* \cup Y^*$. That is, the set of centers C' is obtained by adding a facility co-located with each outlier point from Y^*, and the set of outliers is empty. Now we argue about the costs. Since $C^* \subseteq C'$, for each point $p \in Y^*$, $\mathsf{d}(p, C') \leq \mathsf{d}(p, C^*)$. On the other hand, for each $q \in X \setminus Y^*$, $\mathsf{d}(q, C') = 0$, since there is a co-located center in C^*. This implies that $\mathsf{cost}_0(X, C') \leq \mathsf{cost}_m(X, C)$. Since the solution (Y', C') is feasible for the instance $\mathcal{I}'$, it follows that $\mathsf{OPT}(\mathcal{I}')$ is no larger than the cost $\mathsf{cost}_0(X, C')$. □

Now, we use τ-approximation algorithm guaranteed by the theorem, for the instance $\mathcal{I}'$, and obtain a set of at most $k' \leq k + m$ centers A such that $\mathsf{cost}_0(X, A) \leq \tau \cdot \mathsf{OPT}(\mathcal{I}') \leq \tau \cdot \mathsf{OPT}(\mathcal{I})$. By assumption, running this algorithm takes polynomial time. Let $R = \frac{\mathsf{cost}_0(X,A)}{\tau n}$ be a lower bound on average radius, and $\phi = \lceil \log(\tau n) \rceil$. For each center $c_i \in A$, let $X_i \subseteq X$ denote the set of points whose closest center in A is c_i. By arbitrarily breaking ties, we can assume that the sets X_i are disjoint, i.e., $\{X_i\}_{1 \leq i \leq k'}$ forms a partition of X. Now we further partition each X_i into smaller groups such that the points in each group have similar distances to c_i. Specifically, we define

$$X_{i,j} \coloneqq \begin{cases} B_{X_i}(c_i, R) & \text{if } j = 0, \\ B_{X_i}(c_i, 2^j R) \setminus B_{X_i}(c_i, 2^{j-1} R) & \text{if } j \geq 1. \end{cases}$$

Let $s = \frac{c\tau^2}{\epsilon^2}(m + k \ln n + \ln(1/\lambda))$, for some large enough constant c. We define a weighted set of points $S_{i,j} \subseteq X_{i,j}$ as follows. If $|X_{i,j}| \leq s$, then we say $X_{i,j}$ is *small*. In this case, define $S_{i,j} \coloneqq X_{i,j}$ and let the weight $w_{i,j}$ of each point $p \in S_{i,j}$ be 1. Otherwise, $|X_{i,j}| > s$ and we say that $X_{i,j}$ is *large*. In this case, we take a random sample $S_{i,j} \subseteq X_{i,j}$ of size s. We set the weight of every point in $S_{i,j}$ to be $w_{i,j} = |X_{i,j}|/|S_{i,j}|$. For convenience, assume the weights $w_{i,j}$ to be integers [2]. Finally, let $S = \bigcup_{i,j} S_{i,j}$. The set S can be thought of as an ϵ-coreset for the k-MedianOut instance $\mathcal{I}$. Even though we do not define this notion formally, the key properties of S will be proven in Lemma 2 and 3. Thus, we will often informally refer to S as a *coreset*.

Proposition 1. *We have $|S| = O(((k+m)\log n/\epsilon)^2)$ if λ is a constant.*

Proof. For any $p \in X$, $\mathsf{d}(p, A) \leq \mathsf{cost}_0(X, A) = \tau n \cdot R \leq 2^\phi R$. Therefore, for any $c_i \in A$ and $j > \phi$, $X_{i,j'} = \emptyset$, and $X_i = \bigcup_{j=0}^{\phi} X_{i,j}$. It follows that the number of non-empty sets $X_{i,j}$ is at most $|A| \cdot (1 + \log(\tau n)) = O((k+m)\log n)$, since $|A| \leq k+m$ and τ is a constant. For each non-empty $X_{i,j}$, $|S_{i,j}| \leq 2s = O((m + k \log n)/\epsilon^2)$, if λ is a constant. Since $S = \bigcup_{i,j} S_{i,j}$, the claimed bound follows. □

Proposition 2. *(Chen, 2009; Haussler, 1992) Let $M \geq 0$ and η be fixed constants, and let $h(\cdot)$ be a function defined on a set V such that $\eta \leq h(p) \leq \eta + M$ for all $p \in V$. Let $U \subseteq V$ be a random sample of size s, and $\delta > 0$ be a parameter. If $s \geq \frac{M^2}{2\delta^2}\ln(2/\lambda)$, then*

$$\Pr\left[\left|\frac{h(V)}{|V|} - \frac{h(U)}{|U|}\right| \geq \delta\right] \leq \lambda,$$

where $h(U) \coloneqq \sum_{u \in U} h(u)$, and $h(V) \coloneqq \sum_{v \in V} h(v)$.

2. We defer the discussion on how to ensure the integrality of the weights to Section 3.1.

Lemma 1. *Let (Γ, d) be a metric space, $V \subseteq \Gamma$ be a finite set of points, $\lambda', 0 < \xi < 1$, $q \geq 0$ be parameters, and define $s' = \frac{4}{\xi^2}\left(q + \ln \frac{2}{\lambda'}\right)$. Suppose $U \subseteq V$ is a random sample of size s'. Then for any fixed finite set $C \subseteq F$ with probability at least $1 - \lambda'$ it holds that for any $0 \leq t \leq q$,*

$$|\mathsf{cost}_t(V, C) - \mathsf{wcost}_{t'}(U, C)| \leq \xi |V| (\mathsf{diam}(V) + \mathsf{d}(V, C)),$$

where $t' = \lfloor t|U|/|V| \rfloor$ and $w(u) = |V|/|U|$ for all $u \in U$.

Proof. Throughout the proof, we fix the set $C \subseteq F$ and $0 \leq t \leq q$ as in the statement of the lemma. Next, we define the following notation. For each $v \in V$, let $h(v) \coloneqq d(v, C)$, and let $h(V) \coloneqq \sum_{v \in V} h(v)$, and $h(U) \coloneqq \sum_{u \in U} h(u)$. Analogously, let $h'(V) \coloneqq \mathsf{cost}_t(V, C)$, and $h'(U) \coloneqq \mathsf{cost}_{t'}(U, C)$. Let $\eta(V) \coloneqq \min_{v \in V} \mathsf{d}(v, C)$, and $\eta(U) \coloneqq \min_{u \in U} \mathsf{d}(u, C)$. We summarize a few properties about these definitions in the following observation.

Observation 2. *The following inequalities hold.*

- $\left(t\frac{|U|}{|V|} - 1\right) \leq t' \leq t\frac{|U|}{|V|}$
- $h'(V) \leq h(V) - t \cdot \eta(V) \leq h(V)$, *and* $h'(V) \geq h(V) - t \cdot (\eta(V) + \mathsf{diam}(V))$
- $h'(U) \leq h(U)$, *and* $h'(U) \geq h(U) - t\frac{|U|}{|V|} \cdot (\eta(U) + \mathsf{diam}(U))$
- $\eta(V) \leq \eta(U) \leq \eta(V) + \mathsf{diam}(V)$

Proof. The first item is immediate from the definition $t' = \lfloor t|U|/|V| \rfloor$. Consider the second item. For each $v \in V$, let $g(v) \coloneqq \mathsf{d}(v, C) - \eta(V)$. Let $V' \subseteq V$ denote a set of points of size t that have the t largest distances to the centers in C. By triangle inequality, we get that $\mathsf{d}(v, C) \leq \mathsf{d}(v, v^*) + \mathsf{d}(v^*, C) \leq \mathsf{diam}(V) + \eta(V)$, where $v^* \in V$ is a point realizing the minimum distance $\eta(V)$ to the set of centers C. This implies that $g(v) \leq \mathsf{diam}(V)$ for all $v \in V$. Now, observe that

$$\begin{aligned} h(V) &= h'(V) + \sum_{v \in V'} (\eta(V) + g(v)) && \text{(Since } h'(V) \text{ excludes the distances of points in } V') \\ &= h'(V) + t \cdot \eta(V) + \sum_{v \in V'} g(v) && (1) \\ &\geq h'(V) + t \cdot \eta(V) && (g(v) \geq 0 \text{ for all } v \in V) \end{aligned}$$

By rearranging the last inequality, we get the first part of the second item. To see the second part, observe that (1) implies that $h(V) \leq h'(V) + t \cdot \eta(V) + t \cdot \mathsf{diam}(V)$, since $g(v) \leq \mathsf{diam}(V)$ for all $v \in V$.

The proof of the third item is analogous to the proof of the first item. In addition, we need to combine the inequalities from the first item of the observation. We omit the details. The fourth item follows from the fact that $U \subseteq V$, and via triangle inequality. ☐

By applying Proposition 2 with $\eta = \eta(V), M = \mathsf{diam}(V)$ and $\delta = \xi M/2$, we know with probability at most λ',

$$\left| \frac{\sum_{v \in V} \mathsf{d}(v, C)}{|V|} - \frac{\sum_{u \in U} \mathsf{d}(u, C)}{|U|} \right| \geq \frac{\xi}{2} \mathsf{diam}(V).$$

Recall that, $h(V) = \sum_{v \in V} \mathsf{d}(v, C)$ and $h(U) = \sum_{u \in U} \mathsf{d}(u, C)$. Thus, with probability at least $1 - \lambda'$, we have that

$$\left| \frac{h(V)}{|V|} - \frac{h(U)}{|U|} \right| \leq \frac{\xi}{2} \mathsf{diam}(V). \tag{2}$$

Now, we prove the following technical claim.

Claim 1. *Suppose (2) holds. Then we have,*

$$\left| \frac{h'(V)}{|V|} - \frac{h'(U)}{|U|} \right| \leq \xi \cdot (\mathsf{diam}(V) + \mathsf{d}(V, C)) \tag{3}$$

Proof. We suppose that (2) holds, and show that (3) holds with probability 1. First, consider,

$$\begin{aligned}
\frac{h'(U)}{|U|} - \frac{h'(V)}{|V|} &\leq \frac{h(U)}{|U|} - \frac{h(V)}{|V|} + \frac{t \cdot (\eta(V) + \mathsf{diam}(V))}{|V|} && \text{(From Observation 2, Part 2)} \\
&\leq \frac{\xi}{2} \mathsf{diam}(V) + \frac{t}{|V|} \cdot (\eta(V) + \mathsf{diam}(V)) && \text{(From (2))} \\
&\leq \frac{\xi}{2} \mathsf{diam}(V) + \frac{\xi}{2} \cdot (\eta(V) + \mathsf{diam}(V)) && (4)
\end{aligned}$$

where the last inequality follows from the assumption that $|V| \geq s' \geq \frac{4q}{\xi} \geq \frac{4t}{\xi}$. Now, consider

$$\begin{aligned}
\frac{h'(V)}{|V|} - \frac{h'(U)}{|U|} &\leq \frac{h(V)}{|V|} - \frac{h(U)}{|U|} - \frac{t\eta(V)}{|V|} + \frac{t \frac{|U|}{|V|} \cdot (\eta(U) + \mathsf{diam}(U))}{|U|} \\
& && \text{(From Observation 2, Part 3)} \\
&\leq \frac{\xi}{2} \mathsf{diam}(V) - \frac{t \cdot \eta(V)}{|V|} + \frac{t \cdot \eta(U)}{|V|} + \frac{t \cdot \mathsf{diam}(U)}{|V|} && \text{(From (2))} \\
&\leq \frac{\xi}{2} \mathsf{diam}(V) - \frac{t \cdot \eta(V)}{|V|} + \frac{t \cdot (\eta(V) + \mathsf{diam}(V)) + t \cdot \mathsf{diam}(U)}{|V|} \\
& && \text{(From Observation 2, Part 4)} \\
&\leq \frac{\xi}{2} \mathsf{diam}(V) + \frac{2t \cdot \mathsf{diam}(V)}{|V|} && (\mathsf{diam}(U) \leq \mathsf{diam}(V)) \\
&\leq \xi \mathsf{diam}(V) && (5)
\end{aligned}$$

where the last inequality follows from the assumption that $|V| \geq s' \geq \frac{4q}{\xi} \geq \frac{4t}{\xi}$. $\square$

Thus, from Claim 1, we know that since (2) holds with probability at least $1 - \lambda'$, the following inequality also holds with probability at least $1 - \lambda'$.

$$\left| h'(V) - h'(U) \cdot \frac{|V|}{|U|} \right| \leq \xi |V| \cdot (\mathsf{diam}(V) + \mathsf{d}(V, C)).$$

The preceding inequality is equivalent to the one in the lemma, because $h'(V) = \mathsf{cost}_t(V, C)$, and $h'(U) \cdot \frac{|V|}{|U|} = \frac{|V|}{|U|} \cdot \mathsf{cost}_{t'}(U, C) = \mathsf{wcost}_{t'}(U, C)$. Finally, notice that Claim 1 holds when the h' function is defined with respect to any choice of $t \in \{0, 1, \dots, q\}$. Therefore, with probability at least $1 - \lambda'$, the inequality in the lemma holds for any $t \in \{0, 1, \dots, q\}$, which completes the proof. $\square$

Next, we show the following observation, whose proof is identical to an analogous proof in (Chen, 2008).

Observation 3. *The following inequalities hold.*

- $\sum_{i,j} |X_{i,j}| 2^j R \leq 3 \cdot \mathsf{cost}_0(X, A) \leq 3\tau \cdot \mathsf{OPT}(\mathcal{I})$.
- $\sum_{i,j} |X_{i,j}| \mathsf{diam}(X_{i,j}) \leq 6 \cdot \mathsf{cost}_0(X, A) \leq 6\tau \cdot \mathsf{OPT}(\mathcal{I})$.

Proof. For any $p \in X_{i,j}$, it holds that $2^j R \leq \max\{2\mathsf{d}(p, A), R\} \leq 2\mathsf{d}(p, A) + R$. Therefore,

$$\begin{aligned}
\sum_{i,j} |X_{i,j}| \cdot 2^j R &\leq \sum_{i,j} \sum_{p \in X_{i,j}} 2^j R \\
&\leq \sum_{i,j} \sum_{p \in X_{i,j}} 2\mathsf{d}(p, A) + R \\
&= 2 \sum_{p \in X} \mathsf{d}(p, A) + |X| \cdot |R| \\
&= 2 \cdot \mathsf{cost}_0(X, A) + n|R| \\
&\leq 3 \cdot \mathsf{cost}_0(X, A) && \text{(By definition of } R\text{)} \\
&\leq 3\tau \mathsf{OPT}(\mathcal{I}') \leq 3\tau \mathsf{OPT}(\mathcal{I}).
\end{aligned}$$

We also obtain the second item by observing that $\mathsf{diam}(X_{i,j}) \leq 2 \cdot 2^j \cdot R$. □

Next, we show that the following lemma, which informally states that the union of the sets of sampled points approximately preserve the cost of clustering w.r.t. *any* set of at most k centers, *even after* excluding at most m outliers overall.

Lemma 2. *The following statement holds with probability at least $1 - \lambda/2$: For all sets $C \subseteq F$ of size at most k, and for all sets of non-negative integers $\{m_{i,j}\}_{i,j}$ such that $\sum_{i,j} m_{i,j} \leq m$,*

$$\left| \sum_{i,j} \mathsf{cost}_{m_{i,j}}(X_{i,j}, C) - \sum_{i,j} \mathsf{wcost}_{t_{i,j}}(S_{i,j}, C) \right| \leq \epsilon \cdot \sum_{i,j} \mathsf{cost}_{m_{i,j}}(X_{i,j}, C) \tag{6}$$

where $t_{i,j} = \lfloor m_{i,j}/w_{i,j} \rfloor$.

Proof. Fix an arbitrary set $C \subseteq F$ of at most k centers, and the integers $\{m_{i,j}\}_{i,j}$ such that $\sum_{i,j} m_{i,j} \leq m$ as in the statement of the lemma. For each $i = 1, \ldots, |A|$, and $0 \leq j \leq \phi$, we invoke Lemma 1 by setting $V = X_{i,j}$, and $U = S_{i,j}$, $\xi = \frac{\epsilon}{8\tau}$, $\lambda' = n^{-k}\lambda/(4(k+m)(1+\phi))$, and $q = m$. This implies that, the following inequality holds with probability at least $1 - \lambda'$ for each set $X_{i,j}$, and the corresponding $m_{i,j} \leq m$,

$$\begin{aligned}
&\left| \mathsf{cost}_{m_{i,j}}(X_{i,j}, C) - \mathsf{wcost}_{t_{i,j}}(S_{i,j}, C) \right| \\
&\leq \frac{\epsilon}{8\tau} |X_{i,j}| (\mathsf{diam}(X_{i,j}) + \mathsf{d}(X_{i,j}, C))
\end{aligned} \tag{7}$$

Note that the sample size required in order for this inequality to hold is

$$\begin{aligned} s' &= \left\lceil \frac{4}{\xi^2}\left(m + \ln\left(\frac{2}{\lambda'}\right)\right)\right\rceil \\ &= \left\lceil 4\left(\frac{8\tau}{\epsilon}\right)^2 \cdot \left(m + \ln\left(\frac{8n^k(k+m)(1+\phi)}{\lambda}\right)\right)\right\rceil \leq s. \end{aligned}$$

For any i, j, if $X_{i,j} < s$ (i.e., $X_{i,j}$ is *small*), then the sample $S_{i,j}$ is equal to $X_{i,j}$, and each point in $S_{i,j}$ has weight equal to 1. This implies that $\mathsf{cost}_{m_{i,j}}(X_{i,j}, C) = \mathsf{wcost}_{t_{i,j}}(S_{i,j}, C)$ for all such $X_{i,j}$, and their contribution to the right hand side of inequality (6) is zero. Thus, it suffices to restrict the sum on the right hand side of (6) over *large* sets $X_{i,j}$'s. Let $\mathcal{L}$ consist of all pairs (i, j) such that $X_{i,j}$ is large. We have the following claim.

Claim 2. $\sum_{(i,j)\in\mathcal{L}} |X_{i,j}|\mathsf{d}(X_{i,j}, C) \leq 2\mathsf{cost}_m(X, C)$.

Proof. Let Y denote the farthest m points in X from the set of centers C. Now, fix $(i, j) \in \mathcal{L}$ and let $q_{i,j} := |X_{i,j} \cap Y| \leq m$ denote the number of outliers in $X_{i,j}$. Since $|X_{i,j}| \geq 2m \geq 2q_{i,j}$, the set $X_{i,j} \setminus Y$ is non-empty, and all points $X_{i,j} \setminus Y$ contribute towards $\mathsf{cost}_m(X, C)$. That is,

$$\sum_{(i,j)\in\mathcal{L}} \sum_{p\in X_{i,j}\setminus Y} \mathsf{d}(p, C) \leq \mathsf{cost}_m(X, C) \tag{8}$$

For any $p \in X_{i,j} \setminus Y$, $\mathsf{d}(X_{i,j}, C) \leq \mathsf{d}(p, C)$ from the definition. Therefore,

$$\begin{aligned} &\sum_{(i,j)\in\mathcal{L}} |X_{i,j}| \cdot \mathsf{d}(X_{i,j}, C) \\ &\leq \sum_{(i,j)\in\mathcal{L}} 2|X_{i,j} \setminus Y| \cdot \mathsf{d}(X_{i,j}, C) \\ &\leq 2 \cdot \sum_{(i,j)\in\mathcal{L}} \sum_{p\in X_{i,j}\setminus Y} \mathsf{d}(p, C) \\ &\leq 2 \cdot \mathsf{cost}_m(X, C) \end{aligned}$$

Here, to see the second inequality, see that $|X_{i,j}| \geq 2q_{i,j}$, which implies that $|X_{i,j}| - q_{i,j} \leq 2(|X_{i,j}| - q_{i,j})$. The last inequality follows from (8). □

Thus, by revisiting (6) and (7), we get:

$$\begin{aligned} &\sum_{(i,j)\in\mathcal{L}} \left|\mathsf{cost}_{m_{i,j}}(X_{i,j}, C) - \mathsf{wcost}_{t_{i,j}}(S_{i,j}, C)\right| \\ &\leq \frac{\epsilon}{8\tau} \sum_{(i,j)\in\mathcal{L}} |X_{i,j}|(\mathsf{diam}(X_{i,j}) + \mathsf{d}(X_{i,j}, C)) && \text{(From (7))} \\ &\leq \frac{\epsilon}{8\tau} \cdot (6\tau \cdot \mathsf{OPT}(\mathcal{I}) + 2\mathsf{cost}_m(X, C)) && \text{(From Obs. 3 and Claim 2)} \\ &= \frac{\epsilon}{8\tau}(8\tau \cdot \mathsf{cost}_m(X, C)) = \epsilon \cdot \mathsf{cost}_m(X, C) \end{aligned}$$

Where, in the last inequality, since C is an arbitrary set of at most k centers, $\mathsf{OPT}(\mathcal{I}) \leq \mathsf{cost}_m(X, C)$. Note that the preceding inequality holds for a fixed set C of centers with probability at least $1 - |A| \cdot (1+\phi)\lambda' = 1 - n^{-k}\lambda/2$, which follows from taking the union bound over all sets $X_{i,j}$, $1 \leq i \leq |A| \leq k+m$, and $0 \leq j \leq \phi$.

Since F has at most n^k subsets of size at most k, the statement of the lemma follows from taking a union bound. □

Now we are ready to prove Theorem 1. We enumerate every subset $T \subseteq S$ of size at most m. For each T, we compute a β-approximation solution for the (weighted) k-median instance $(S \backslash T, F, k)$. Theorem 1 only assumes the existence of a β-approximation algorithm for unweighted k-median, which cannot be applied to weighted point sets. However, we can transform $S \backslash T$ to an equivalent unweighted sets R, which contains, for each $x \in S \backslash T$, $w(x)$ copies of (unweighted) x, where $w(x)$ is the weight of x in $S \backslash T$. It is clear that $\mathsf{wcost}(S \backslash T, C) = \mathsf{cost}(R, C)$ for all $C \subseteq F$. Thus, we can apply the β-approximation k-Median algorithm on (R, F, k) to compute a center set $C \subseteq F$ of size k such that $\mathsf{wcost}(S \setminus T, C) \leq \beta \cdot \mathsf{wcost}(S \setminus T, C')$ for any $C' \subseteq F$ of size k. We do this for all $T \subseteq S$ of size at most m. Let $\mathcal{C}$ denote the set of all center sets C computed. We pick a center set $C^* \subseteq \mathcal{C}$ that minimizes $\mathsf{cost}_m(X, C^*)$, and return (Y^*, C^*) as the solution where $Y^* \subseteq X$ consists of the m points in X farthest to the center set C^*.

Lemma 3. *With probability at least $1 - \frac{\lambda}{2}$, for all $C \subseteq F$ of size k we have*

$$\mathsf{cost}_m(X, C^*) \leq \frac{1+\epsilon}{1-\epsilon} \cdot \beta \mathsf{cost}_m(X, C).$$

Proof. The statement in Lemma 2 holds with probability at least $1 - \lambda/2$. Thus, it suffices to assume the statement in Lemma 2, and show $\mathsf{cost}_m(X, C^*) \leq (1+\epsilon)^2 \beta \cdot \mathsf{cost}_m(X, C)$ for any $C \subseteq F$ of size k. Fix a subset $C \subseteq F$ of size k. Let $Y \subseteq X$ consist of the m points in X farthest to C, and define $m_{i,j} = |Y \cap X_{i,j}|$. Set $t_{i,j} = \lfloor m_{i,j}/w_{i,j} \rfloor$. Note that $\mathsf{cost}_m(X, C) = \sum_{i,j} \mathsf{cost}_{m_{i,j}}(X_{i,j}, C)$. Furthermore, by Lemma 2, we have

$$\begin{aligned} \sum_{i,j} \mathsf{wcost}_{t_{i,j}}(S_{i,j}, C) &\leq (1+\epsilon) \cdot \sum_{i,j} \mathsf{cost}_{m_{i,j}}(X_{i,j}, C) \\ &= (1+\epsilon) \cdot \mathsf{cost}_m(X, C). \end{aligned} \tag{9}$$

Now let $T_{i,j} \subseteq S_{i,j}$ consist of the $t_{i,j}$ points in $S_{i,j}$ farthest to C, and define $T = \bigcup_{i,j} T_{i,j}$. Since $|T| \leq m$, T is considered by our algorithm and thus there exists a center set $C' \in \mathcal{C}$ that is a β-approximation solution for the (weighted) k-median instance $(S \backslash T, F, k)$. We have

$$\mathsf{wcost}(S \backslash T, C') \leq \beta \cdot \mathsf{wcost}(S \backslash T, C) = \beta \sum_{i,j} \mathsf{wcost}_{t_{i,j}}(S_{i,j}, C). \tag{10}$$

Note that $\mathsf{wcost}(S \backslash T, C') \geq \sum_{i,j} \mathsf{wcost}_{t_{i,j}}(S_{i,j}, C')$. Furthermore, by applying Lemma 2 again, we have $\sum_{i,j} \mathsf{wcost}_{t_{i,j}}(S_{i,j}, C') \geq (1-\epsilon) \cdot \sum_{i,j} \mathsf{cost}_{m_{i,j}}(X_{i,j}, C')$. It then follows that

$$(1-\epsilon) \cdot \mathsf{cost}_m(X_{i,j}, C') \leq (1-\epsilon) \cdot \sum_{i,j} \mathsf{cost}_{m_{i,j}}(X_{i,j}, C') \leq \mathsf{wcost}(S \backslash T, C'). \tag{11}$$

Finally, we have $\mathsf{cost}_m(X, C^*) \le \mathsf{cost}_m(X, C')$ by the construction of C^*. Combining this with (9), (10), and (11), we have $\mathsf{cost}_m(X, C^*) \le \frac{1+\epsilon}{1-\epsilon} \cdot \beta \mathsf{cost}_m(X, C)$, which completes the proof. $\square$

By choosing $\lambda > 0$ to be a sufficiently small constant, and by appropriately rescaling ϵ [3], the above lemma shows that our algorithm outputs a $(\beta + \epsilon)$-approximation solution with a constant probability. By repeating the algorithm a logarithmic number of rounds, we can guarantee the algorithm succeeds with high probability. The number of subsets $T \subseteq S$ of size at most m is bounded by $|S|^{O(m)}$, which is $\left(\frac{(k+m)\log n}{\epsilon}\right)^{O(m)}$ by Proposition 1. Note that $(\log n)^{O(m)} \le \max\{m^{O(m)}, n^{O(1)}\}$. Thus, the number of subsets $T \subseteq S$ of size at most m is bounded by $f(k, m, \epsilon) \cdot n^{O(1)}$, where $f(k, m, \epsilon) = \left(\frac{k+m}{\epsilon}\right)^{O(m)}$. Thus, we need to call the β-approximation k-Median algorithm $f(k, m, \epsilon) \cdot n^{O(1)}$ times, which takes $f(k, m, \epsilon)n^{O(1)} \cdot T(n, k)$ time overall. The first call of the algorithm for obtaining a τ-approximation to the $(k+m)$-Median instance takes polynomial time. Besides this, the other parts of our algorithm can all be done in polynomial time. This completes the proof of Theorem 1.

3.1 Ensuring Integral Weights in the Coreset

Recall that in order to obtain the set $S_{i,j}$ from a *large* $X_{i,j}$, we sample s points uniformly and independently at random (with replacement), and give each point in $S_{i,j}$ the weight $w_{i,j} = \frac{|X_{i,j}|}{|S_{i,j}|}$. In the main body of the proof, we assumed that the quantity $w_{i,j}$ is integral for the sake of simplicity. However, in general $\frac{|X_{i,j}|}{|S_{i,j}|}$ may not be an integer. Here, we describe how to modify this construction to ensure integral weights.

To this end, let $X_{i,j}^{(1)} \subseteq X_{i,j}$ be an arbitrary subset of size $|X_{i,j}| \mod s$, and let $X_{i,j}^{(2)} = X_{i,j} \setminus Y_{i,j}$. From this time onward, we *treat* $X_{i,j}^{(1)}$ and $X_{i,j}^{(2)}$ as two separate sets of the form $X_{\cdot,\cdot}$, and proceed with the construction of the coreset.

In particular, observe that $|X_{i,j}^{(1)}| < s$, i.e., it is *small*, and $|X_{i,j}^{(2)}| = t \cdot s$ for some positive integer t, and thus $X_{i,j}^{(2)}$ is *large*. Therefore, we let $S_{i,j}^{(1)} \leftarrow X_{i,j}^{(1)}$, and each point is added with weight 1. On the other hand, to obtain $S_{i,j}^{(2)}$, we sample s points uniformly and independently at random from $X_{i,j}^{(2)}$, and set the weight of each point to be $|X_{i,j}^{(2)}|/s$, which is an integer. From this point onward, we proceed with exactly the same analysis as in the original proof, i.e., we treat $X_{i,j}^{(1)}$ as a *small* set, and $X_{i,j}^{(2)}$ as a large set in the analysis. Since for the small sets, the sampled set is equal to the original set, their contribution to the left hand side of the following inequality in the statement of Lemma 2, is equal to zero.

$$\left| \sum_{i,j} \mathsf{cost}_{m_{i,j}}(X_{i,j}, C) - \sum_{i,j} \mathsf{wcost}_{t_{i,j}}(S_{i,j}, C) \right| \le \epsilon \cdot \sum_{i,j} \mathsf{cost}_{m_{i,j}}(X_{i,j}, C)$$

Therefore, the analysis of Lemma 2 goes through without any modifications. The only other minor change is that the number of points in the coreset S, which is obtained by taking the

3. Since Lemma 3 implies a $\beta(1 + O(\epsilon))$-approximation, and β is a constant, it suffices to redefine $\epsilon = \epsilon/c$ for some large enough constant c to get the desired result.

union of all $S_{\cdot,\cdot}$, is now at most twice the previous bound, which is easily absorbed in the big-oh notation.

4. Extensions

4.1 k-Means with Outliers

This is similar to k-MedianOut, except that the cost function is the sum of *squares* of distances of all except m outlier points to a set of k facilities. This generalizes the well-known k-Means problem. Here, the main obstacle is that, the squares of distances do not satisfy triangle inequality, and thus it does not form a metric. However, they satisfy a *relaxed* version of triangle inequality (i.e., $\mathsf{d}(p,q)^2 \leq 2(\mathsf{d}(p,r)^2 + \mathsf{d}(r,q)^2)$). This technicality makes the arguments tedious, nevertheless, we can follow the same approach as for k-MedianOut, to obtain optimal FPT approximation schemes. Our technique implies an optimal $(1 + \frac{8}{e} + \epsilon)$-approximation for k-MeansOut (using the result of (Cohen-Addad et al., 2019) as a black-box), improving upon polynomial-time 53.002-approximation from (Krishnaswamy et al., 2018), and $(9 + \epsilon)$-approximation from (Goyal et al., 2020) in time FPT in k, m and ϵ.

In fact, using our technique, we can get improved approximation guarantees for (k, z)-Clustering with Outliers, where the cost function involves z-th power of distances, where $z \geq 1$ is fixed for a problem. Note that the cases $z = 1$ and $z = 2$ correspond to k-MedianOut and k-MeansOut respectively. We give the details for (k, z)-Clustering with Outliers in Appendix A.

4.2 Matroid Median with Outliers

A *matroid* is a pair $\mathcal{M} = (F, \mathcal{S})$, where F is a ground set, and $\mathcal{S}$ is a collection of subsets of F with the following properties: (i) $\emptyset \in \mathcal{S}$, (ii) If $A \in \mathcal{S}$, then for every subset $B \subseteq A$, $B \in \mathcal{S}$, and (iii) For any $A, B \in \mathcal{S}$ with $|B| < |A|$, there exists an $b \in B \setminus A$ such that $B \cup \{b\} \in \mathcal{S}$. The rank of a matroid $\mathcal{M}$ is the size of the largest independent set in $\mathcal{S}$. Using the definition of matroid, it can be easily seen that all inclusion-wise maximal independent sets (called *bases*) have the same size.

An instance of Matroid Median with Outliers is given by $(X, F, \mathcal{M}, m)$, where $\mathcal{M} = (F, \mathcal{S})$ is a matroid with rank k defined over a finite ground set F, and X, F are sets of clients and facilities, belonging to a finite metric space (Γ, d). The objective is to find a set $C \subseteq F$ of facilities that minimizes $\mathsf{cost}_m(X, C)$, and $C \in \mathcal{S}$, i.e., C is an independent set in the given matroid. Note that an explicit description of a matroid of rank k may be as large as n^k. Therefore, we assume that we are given an efficient *oracle access* to the matroid $\mathcal{M}$. That is, we are provided with an algorithm $\mathcal{A}$ that, given a candidate set $S \subseteq F$, returns in time $T(\mathcal{A})$ (which is assumed to be polynomial in $|F|$), returns whether $S \in \mathcal{I}$.

We can adapt our approach to Matroid Median with Outliers in a relatively straightforward manner. Recall that our algorithm needs to start with an instance of outlier-free problem (i.e., Matroid Median) that provides a lower bound on the optimal cost of the given instance. To this end, given an instance $\mathcal{I} = (X, F, \mathcal{M} = (F, \mathcal{S}), m)$ of Matroid Median with Outliers, we define an instance $\mathcal{I}' = (X, F, \mathcal{M}', 0)$ of Matroid

Median with 0 Outliers (i.e., Matroid Median), where $\mathcal{M}' = (F \cup X, \mathcal{S}')$ is defined as follows. $\mathcal{S}' = \{Y \cup C : Y \subseteq X \text{ with } |Y| \leq m \text{ and } C \subseteq F \text{ with } C \in \mathcal{S}\}$. That is, each independent set of $\mathcal{M}'$ is obtained by taking the union of an independent set of facilities from $\mathcal{M}$, and a subset of X of size at most m. It is straightforward to show that $\mathcal{M}'$ satisfies all three axioms mentioned above, and thus is a matroid over the ground set $F \cup X$. In particular, it is the direct sum of $\mathcal{M}$ and a uniform matroid $\mathcal{M}_m$ over X of rank m (i.e., where any subset of X of size at most m is independent). Note that using the oracle algorithm $\mathcal{A}$, we can simulate an oracle algorithm to test whether a candidate set $C \subseteq F \cup X$ is independent in $\mathcal{M}'$. Therefore, using a $(2+\epsilon)$-approximation for Matroid Median (Cohen-Addad et al., 2019) in time FPT in k and ϵ, we can find a set $A \subseteq F \cup X$ of size at most $k+m$ that we can use to construct a coreset. The details about enumeration are similar to that for k-MedianOut, and are thus omitted.

4.3 Colorful k-Median

This is an orthogonal generalization of k-MedianOut to ensure a certain notion of *fairness* in the solution (see (Jia et al., 2020)). Suppose the set of points X is partitioned into ℓ different colors $X_1 \uplus X_2 \uplus \ldots \uplus X_\ell$. We are also given the corresponding number of outliers $m_1, m_2, \ldots, m_\ell$. The goal is to find a set of at most facilities C to minimize the connection cost of all except at most m_t outliers from each color class X_t, i.e., we want to minimize the cost function: $\sum_{t=1}^{\ell} \mathsf{cost}_{m_t}(X_t, C)$. This follows a generalizations of the well-known k-Center problem introduced in (Bandyapadhyay et al., 2019) and (Anegg, Angelidakis, Kurpisz, & Zenklusen, 2020; Jia et al., 2020) , called Colorful k-Center. Similar generalization of Facility Location has also been studied in (Chekuri, Inamdar, Quanrud, Varadarajan, & Zhang, 2022).

Using our ideas, we can find an FPT approximation parameterized by k, $m = \sum_{t=1}^{\ell} m_t$, and ϵ. To this end, we sample sufficiently many points from each color class X_t separately, and argue that it preserves the cost appropriately. The technical details follow the same outline as that for k-Median with m Outliers. In particular, during the enumeration phase—just like that for k-MedianOut—we obtain several instances of k-Median. That is, our algorithm is *color-agnostic* after constructing the coreset. Thus, we obtain a tight $(1 + \frac{2}{e} + \epsilon)$-approximation for this problem. This is the first non-trivial true approximation for this problem – previous work (Gupta, Moseley, & Zhou, 2021) only gives a *pseudo-approximation*, i.e., a solution with cost at most a constant times that of an optimal cost, but using slightly more than k facilities.

4.4 A Combination of Above Generalizations

Our technique also works for a combination of the aforementioned generalizations that are orthogonal to each other. To consider an extreme example, consider Colorful Matroid Median with ℓ different color classes (a similar version for k-Center objective has been recently studied by (Anegg, Koch, & Zenklusen, 2022)), where we want to find a set of facilities that is independent in the given matroid, in order to minimize the sum of distances of all except m_t outlier points for each color class X_t. By using a combination of the ideas mentioned above, one can get FPT approximations for such generalizations.

5. Concluding Remarks

In this paper, we give a reduction from k-MedianOut to k-Median that runs in time FPT in k, m, and ϵ, and preserves the approximation ratio up to an additive ϵ factor. As a consequence, we obtain improved FPT approximations for k-MedianOut in general as well as special kinds of metrics, and these approximation guarantees are known to be tight in general. Furthermore, our technique is versatile in that it also gives improved approximations for related clustering problems, such as k-MeansOut, Matroid Median with Outliers, and Colorful k-Median, among others.

The most natural direction is to improve the FPT running time while obtaining the tight approximation ratios. More fundamentally, perhaps, is the question whether we need an FPT dependence on the number of outliers, m; or whether it is possible to obtain approximation guarantees for k-MedianOut matching that for k-Median, with a running time that is FPT in k and ϵ alone.

Acknowledgments

The authors are listed in the alphabetical order of the last names. An extended abstract of the paper appeared in AAAI 2023.

We thank Rajni Dabas for bringing a few minor typos to our attention.

T. Inamdar is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 819416). S. Saurabh is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 819416) and Swarnajayanti Fellowship (No DST/SJF/MSA01/2017-18).

References

Almanza, M., Epasto, A., Panconesi, A., & Re, G. (2022). k-clustering with fair outliers. In Candan, K. S., Liu, H., Akoglu, L., Dong, X. L., & Tang, J. (Eds.), *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, pp. 5–15. ACM.

Anegg, G., Angelidakis, H., Kurpisz, A., & Zenklusen, R. (2020). A technique for obtaining true approximations for k-center with covering constraints. In Bienstock, D., & Zambelli, G. (Eds.), *Integer Programming and Combinatorial Optimization - 21st International Conference, IPCO 2020, London, UK, June 8-10, 2020, Proceedings*, Vol. 12125 of *Lecture Notes in Computer Science*, pp. 52–65. Springer.

Anegg, G., Koch, L. V., & Zenklusen, R. (2022). Techniques for generalized colorful k-center problems. In Chechik, S., Navarro, G., Rotenberg, E., & Herman, G. (Eds.), *30th Annual European Symposium on Algorithms, ESA 2022, September 5-9, 2022, Berlin/Potsdam, Germany*, Vol. 244 of *LIPIcs*, pp. 7:1–7:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Bandyapadhyay, S., Inamdar, T., Pai, S., & Varadarajan, K. R. (2019). A constant approximation for colorful k-center. In Bender, M. A., Svensson, O., & Herman, G. (Eds.), *27th Annual European Symposium on Algorithms, ESA 2019, September 9-11, 2019,* *Munich/Garching, Germany*, Vol. 144 of *LIPIcs*, pp. 12:1–12:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Blömer, J., Lammersen, C., Schmidt, M., & Sohler, C. (2016). Theoretical analysis of the k-means algorithm–a survey. In *Algorithm Engineering*, pp. 81–116. Springer.

Charikar, M., Khuller, S., Mount, D. M., & Narasimhan, G. (2001). Algorithms for facility location problems with outliers. In Kosaraju, S. R. (Ed.), *Proceedings of the Twelfth Annual Symposium on Discrete Algorithms, January 7-9, 2001, Washington, DC, USA*, pp. 642–651. ACM/SIAM.

Chekuri, C., Inamdar, T., Quanrud, K., Varadarajan, K., & Zhang, Z. (2022). Algorithms for covering multiple submodular constraints and applications. *Journal of Combinatorial Optimization*, 1–32.

Chen, K. (2008). A constant factor approximation algorithm for k-median clustering with outliers. In Teng, S. (Ed.), *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2008, San Francisco, California, USA, January 20-22, 2008*, pp. 826–835. SIAM.

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.

Cohen-Addad, V., Gupta, A., Kumar, A., Lee, E., & Li, J. (2019). Tight FPT approximations for k-median and k-means. In Baier, C., Chatzigiannakis, I., Flocchini, P., & Leonardi, S. (Eds.), *46th International Colloquium on Automata, Languages, and Programming, ICALP 2019, July 9-12, 2019, Patras, Greece*, Vol. 132 of *LIPIcs*, pp. 42:1–42:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Cohen-Addad, V., Saulpic, D., & Schwiegelshohn, C. (2021). A new coreset framework for clustering. In Khuller, S., & Williams, V. V. (Eds.), *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pp. 169–182. ACM.

Feng, Q., Zhang, Z., Huang, Z., Xu, J., & Wang, J. (2019). Improved algorithms for clustering with outliers. In Lu, P., & Zhang, G. (Eds.), *30th International Symposium on Algorithms and Computation, ISAAC 2019, December 8-11, 2019, Shanghai University of Finance and Economics, Shanghai, China*, Vol. 149 of *LIPIcs*, pp. 61:1–61:12. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

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

Friggstad, Z., Rezapour, M., & Salavatipour, M. R. (2019b). Local search yields a PTAS for k-means in doubling metrics. *SIAM J. Comput.*, *48*(2), 452–480.

Goyal, D., Jaiswal, R., & Kumar, A. (2020). FPT approximation for constrained metric k-median/means. In Cao, Y., & Pilipczuk, M. (Eds.), *15th International Symposium on Parameterized and Exact Computation, IPEC 2020, December 14-18, 2020, Hong Kong, China (Virtual Conference)*, Vol. 180 of *LIPIcs*, pp. 14:1–14:19. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Guha, S., & Khuller, S. (1999). Greedy strikes back: Improved facility location algorithms. *J. Algorithms*, *31*(1), 228–248.

Gupta, A., Moseley, B., & Zhou, R. (2021). Structural iterative rounding for generalized k-median problems. In Bansal, N., Merelli, E., & Worrell, J. (Eds.), *48th International Colloquium on Automata, Languages, and Programming, ICALP 2021, July 12-16, 2021, Glasgow, Scotland (Virtual Conference)*, Vol. 198 of *LIPIcs*, pp. 77:1–77:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Gupta, S., Kumar, R., Lu, K., Moseley, B., & Vassilvitskii, S. (2017). Local search methods for k-means with outliers. *Proc. VLDB Endow.*, *10*(7), 757–768.

Haussler, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inf. Comput.*, *100*(1), 78–150.

Jia, X., Sheth, K., & Svensson, O. (2020). Fair colorful k-center clustering. In Bienstock, D., & Zambelli, G. (Eds.), *Integer Programming and Combinatorial Optimization - 21st International Conference, IPCO 2020, London, UK, June 8-10, 2020, Proceedings*, Vol. 12125 of *Lecture Notes in Computer Science*, pp. 209–222. Springer.

Krishnaswamy, R., Li, S., & Sandeep, S. (2018). Constant approximation for k-median and k-means with outliers via iterative rounding. In Diakonikolas, I., Kempe, D., & Henzinger, M. (Eds.), *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pp. 646–659. ACM.

Rokach, L. (2009). A survey of clustering algorithms. In *Data mining and knowledge discovery handbook*, pp. 269–298. Springer.

Statman, A., Rozenberg, L., & Feldman, D. (2020). k-means: Outliers-resistant clustering+++. *Algorithms*, *13*(12), 311.

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*(2), 165–193.

Appendix A. (k, z)-clustering with Outliers

Let $z \geq 1$ be a fixed real that is not part of the input of the problem.
The input of the (k, z)-Clustering problem is an instance $\mathcal{I} = ((\Gamma, \mathsf{d}), X, F, k)$, where (Γ, d) is a metric space, $X \subseteq \Gamma$ is a (finite) set of n points, called *points* or *clients*, $F \subseteq \Gamma$ is a set of *facilities*, and k is a positive integer. The task is to find a set $C \subseteq F$ of facilities (called *centers*) in F that minimizes the following cost function:

$$\mathsf{cost}(X, C) \coloneqq \sum_{p \in X} \mathsf{cost}(p, C)$$

where $\mathsf{cost}(p, C) \coloneqq (\mathsf{d}(p, C))^z$.

(k, z)-clustering with m outliers. Here, the input contains an additional integer $1 \leq m \leq n$, and the goal is to find a set $X' \subseteq X$ of $n - m$ points, such that $\mathsf{cost}(X', C)$ is minimized (over choices all of X' and C). Here, the set $X \setminus X'$ of at most m points corresponds to the set of *outliers*. In another notation, we want to find a set $C \subseteq F$ of at most k centers that minimizes $\mathsf{cost}_m(X,C) \coloneqq \mathsf{sum}_{\sim m}\{\mathsf{cost}(p,C) : p \in X\}$, i.e., the sum of $n-m$ smallest distances of points in X to the set of centers C.

First, we state a few properties about the z-th powers of distances, which will be subsequently useful in the analysis.

Proposition 3. *Let $P, C \subseteq \Gamma$ be non-empty finite subsets of points. For any point $p \in P$, the following holds:*

- $\mathsf{d}(P,C)^z \le \mathsf{cost}(p,C) \le (\mathsf{d}(P,C) + \mathsf{diam}(P))^z \le 2^z \cdot (\mathsf{d}(P,C)^z + \mathsf{diam}(P)^z)$
- $(\mathsf{d}(p,C) - \mathsf{d}(P,C))^z \le (\mathsf{diam}(P))^z$

Proof. Let $p^* \in P$ be a point realizing the smallest distance $\mathsf{d}(P,C)$. It follows that for any $p \in P$,

$$\begin{aligned}
\mathsf{d}(P,C) = \mathsf{d}(p^*,C) &\le \mathsf{d}(p,C) \\
&\le \mathsf{d}(p,p^*) + \mathsf{d}(p^*,C) && \text{(by triangle inequality)} \\
&\le \mathsf{diam}(P) + \mathsf{d}(P,C) && (\mathsf{d}(p,p^*) \le \mathsf{diam}(P)) \\
&\le 2\max\{\mathsf{diam}(P), \mathsf{d}(P,C)\}
\end{aligned}$$

Now, by taking the z-th power of each term, we get the first inequality, which follows from $\max\{a,b\} \le a+b$.

Note that the first and third line in the preceding chain of inequalities implies that $(\mathsf{d}(p,C) - \mathsf{d}(P,C)) \le \mathsf{diam}(P)$. Note that both sides of the inequality are non-negative. Thus, by taking the z-th power of both sides, the second item follows. □

Consider an instance $\mathcal{I} = ((\Gamma,\mathsf{d}), X, F, k, m)$ be an instance of (k,z)-Clustering with m Outliers. We define an instance $\mathcal{I}' = ((\Gamma,\mathsf{d}), X, F \cup X, k+m, 0)$ of $(k+m,z)$-Clustering (without outliers), where in addition to the original set of facilities, there is a facility co-located with each client. The following observation and its proof is analogous to Observation 1, and thus we omit the proof.

Observation 4. $\mathsf{OPT}(\mathcal{I}') \le \mathsf{OPT}(\mathcal{I})$*, i.e., the value of an optimal solution to $\mathcal{I}'$ is a lower bound on the value of an optimal solution to $\mathcal{I}$.*

The following definitions and the construction of the coreset is analogous to that for k-median with m outliers, with appropriate modifications needed for z-th power of distances. First, we assume that there exists a τ-approximation algorithm for (k,z)-clustering problem that runs in polynomial time, where $\tau = O(1)$. Then, by using this τ-approximation algorithm for the instance $\mathcal{I}'$, we obtain a set of at most $k' \le k+m$ centers A such that $\mathsf{cost}_0(X,A) \le \tau \cdot \mathsf{OPT}(\mathcal{I}') \le \tau \cdot \mathsf{OPT}(\mathcal{I})$. Let $R = \left(\frac{\mathsf{cost}_0(X,A)}{\tau n}\right)^{1/z}$ be a lower bound on average radius, and let $\phi = \lceil \log(\tau n) \rceil$. For each $c_i \in A$, let $X_i \subseteq X$ denote the set of points whose closest center in A is c_i. By arbitrarily breaking ties, we assume that the sets X_i are disjoint, i.e., the sets $\{X_i\}_{1 \le i \le k'}$ form a partition of X. Now, we define the set of *rings* centered at each center c_i as follows.

$$X_{i,j} \coloneqq \begin{cases} B_{X_i}(c_i, R) & \text{if } j = 0 \\ B_{X_i}(c_i, 2^j R) \setminus B_{X_i}(c_i, 2^{j-1} R) & \text{if } j \ge 1 \end{cases}$$

Let $s = \frac{c\tau^2 2^{(c'z)}}{\epsilon^2}(m + k\ln n + \ln(1/\lambda))$, for some large enough constants c, c'. We define a weighted set of points $S_{i,j} \subseteq X_{i,j}$ as follows. If $|X_{i,j}| \leq s$, then say that $X_{i,j}$ is *small*, and let $S_{i,j} \coloneqq X_{i,j}$, and let the weight of each point $p \in S_{i,j}$ be 1. Otherwise, if $|X_{i,j}| > s$, then say that $X_{i,j}$ is *large*. In this case, let $Y_{i,j} \subseteq X_{i,j}$ be an arbitrary subset of size $|X_{i,j}| \mod q$. We add each point $q \in Y_{i,j}$ to $S_{i,j}$ with weight 1. Furthermore, we sample s points uniformly at random (with replacement) $X_{i,j} \setminus Y_{i,j}$, and add to the set $S_{i,j}$ with weight equal to $\frac{|X_{i,j} \setminus Y_{ij}|}{s}$, which is an integer. Thus, we assume that $|S_{i,j}| \leq 2s$, and the weight of every point in $S_{i,j}$ is an integer. Finally, let $S = \bigcup_{i,j} S_{i,j}$.

Lemma 4. *Let (Γ, d) be a metric space, and let $V \subseteq \Gamma$ be a finite set of points. Let $\lambda', \xi > 0$, $q \geq 0$, be parameters, and define $s' = \frac{4}{\xi^2}\left(q + \ln\frac{2}{\lambda'}\right)$. If $|V| \geq s'$, and U is a sample of s' points picked uniformly and independently at random from V, with each point of U having weight $|V|/|U|$, such that the total weight $w(U)$ is equal to $|V|$, then for any fixed finite set $C \subseteq \Gamma$, and for any $0 \leq t \leq q$, with probability at least $1 - \lambda'$ it holds that*

$$|\mathsf{cost}_t(V, C) - \mathsf{wcost}_{t'}(U, C)| \leq 2^{2z+2}\xi|V| \cdot (\mathsf{diam}(V)^z + \mathsf{d}(V, C)^z), \tag{12}$$

where $t' = \lfloor t|U|/|V| \rfloor$.

Proof. Throughout the proof, we fix the set C and $0 \leq t \leq q$ as in the statement of the lemma. Next, we define the following notation. For all $v \in V$, let $h(v) = \mathsf{cost}(v, C) = \mathsf{d}(v, C)^z$, and let $h(V) \coloneqq \sum_{v \in V} h(v)$, and $h(U) \coloneqq \sum_{u \in U} h(u)$. Analogously, let $h'(V) \coloneqq \mathsf{cost}_t(V, C)$, and $h'(U) \coloneqq \mathsf{cost}_{t'}(U, C)$, i.e., sum of all except t (resp. t') largest h-values. Let $\eta(V) \coloneqq \min_{v \in V} \mathsf{d}(v, C)^z$, and $\eta(U) \coloneqq \min_{u \in U} \mathsf{d}(u, C)$. We summarize a few properties about these definitions in the following observation, which is analogous to Observation 2.

Observation 5.

- $\left(t\frac{|U|}{|V|} - 1\right) \leq t' \leq t\frac{|U|}{|V|}$
- *For any $p \in P$, $(\eta(V)^z \leq \mathsf{d}(p, C)^z = \mathsf{cost}(p, C)^z \leq \eta(V)^z + 2^z(\eta(V)^z + \mathsf{diam}(V))$*
- $h'(V) \leq h(V) - t \cdot \eta(V)^z \leq h(V)$, *and* $h'(V) \geq h(V) - 2^z \cdot t \cdot (\eta(V)^z + \mathsf{diam}(V)^z)$
- $h'(U) \leq h(U)$, *and* $h'(U) \geq h(U) - 2^z \cdot t\frac{|U|}{|V|} \cdot (\eta(U)^z + \mathsf{diam}(U)^z)$
- $\eta(V)^z \leq \eta(U)^z \leq 2^z(\eta(V)^z + \mathsf{diam}(V)^z)$

Proof. The first item is immediate from the definition $t' = \lfloor t|U|/|V| \rfloor$.

Consider the second item. For each $v \in V$, let $g(v) \coloneqq \mathsf{d}(v, C) - \eta(V)$. Let $V' \subseteq V$ denote a set of points of size t that have the t largest distances to the centers in C. From Proposition 3, we get that for any $p \in V$, $\eta(V)^z \leq \mathsf{cost}(p, C)^z \leq 2^z \cdot (\eta(V)^z + \mathsf{diam}(P)^z)$. This implies that $g(v) \leq \mathsf{diam}(V)$ for all $v \in V$. Now, observe that

$$\begin{aligned} h(V) &= h'(V) + \sum_{v \in V'} (\eta(V) + g(v))^z \quad \text{(Since } h'(V) \text{ excludes the distances of points in } V'\text{)} \\ &\geq h'(V) + t \cdot \eta(V)^z \qquad (g(v) \geq 0 \text{ for all } v \in V) \end{aligned}$$

By rearranging the last inequality, we get the first part of the third item. Also note that the first inequality also implies that $h(V) \le h'(V) + 2^z t \cdot \eta(V) + 2^z \sum_{v \in V} g(v)^z$, via Proposition 3. Then, by recalling that $g(v) \le \mathsf{diam}(V)$ for all $v \in V$, the second part of the third item follows.

The proof of the fourth item is analogous to that of the third item. In addition, we need to combine the inequalities from the first item of the observation. We omit the details. The fifth item follows from the fact that $U \subseteq V$, and via triangle inequality. $\square$

Let $\eta = \eta(V)^z$, $M = 2^{2z+2}(\eta(V)^z + \mathsf{diam}(V)^z)$, and $\delta = \xi M/2$. Then, the second item of Observation 5 implies that $\eta \le h(v) \le \eta + M$ for all $v \in V$. Then, Proposition 2 implies that,

$$\begin{aligned}
&\Pr\left[\left|\frac{\sum_{v\in V}\mathsf{cost}(v,C)}{|V|} - \frac{\sum_{u\in U}\mathsf{cost}(u,C)}{|U|}\right| \ge \frac{\xi}{2}2^z(\eta(V)^z) + \mathsf{diam}(V)^z)\right]\\
&= \Pr\left[\left|\frac{h(V)}{|V|} - \frac{h(U)}{|U|}\right| \ge \delta\right]\\
&\le \lambda'.
\end{aligned}$$

Thus, with probability at least $1 - \lambda'$, we have that

$$\left|\frac{h(V)}{|V|} - \frac{h(U)}{|U|}\right| \le \frac{\xi}{2} \cdot M \tag{13}$$

In the rest of the proof, we condition on this event, and assume that (13) holds, and show that the inequality in the lemma holds with probability 1. First, consider,

$$\begin{aligned}
\frac{h'(U)}{|U|} - \frac{h'(V)}{|V|} &\le \frac{h(U)}{|U|} - \frac{h(V)}{|V|} + \frac{2^z \cdot t \cdot (\eta(V)^z + \mathsf{diam}(V)^z)}{|V|} && \text{(From Obs. 5)}\\
&\le \frac{\xi}{2}M + \frac{t \cdot M}{|V|} && \text{(From (13))}\\
&\le \xi M && (14)
\end{aligned}$$

where the last inequality follows from the assumption that $|V| \ge s' \ge \frac{4q}{\xi} \ge \frac{4t}{\xi}$. Now, consider

$$\begin{aligned}
&\frac{h'(V)}{|V|} - \frac{h'(U)}{|U|}\\
&\le \frac{h(V)}{|V|} - \frac{h(U)}{|U|} + \frac{2^z \cdot t\frac{|U|}{|V|} \cdot (\eta(U)^z + \mathsf{diam}(V)^z)}{|U|} && \text{(From Obs. 5, Part 4)}\\
&\le \frac{\xi}{2}M + \frac{2^z \cdot t \cdot \eta(U)^z}{|V|} + \frac{2^z \cdot t \cdot \mathsf{diam}(V)^z}{|V|} && \text{(From (2)}\\
&\le \frac{\xi}{2}M + \frac{2^{2z} \cdot t \cdot (\eta(V)^z + \mathsf{diam}(V)^z) + t \cdot 2^z \cdot \mathsf{diam}(V)^z}{|V|} && \text{(From Obs. 2, Part 5)}\\
&\le \frac{\xi}{2}M + \frac{2^{2z+1} \cdot t \cdot (\eta(V)^z + \mathsf{diam}(V)^z)}{|V|} && (\text{Since } |V| \ge s' \ge \tfrac{4q}{\xi} \ge \tfrac{4t}{\xi})\\
&= \frac{\xi}{2}M + \frac{\xi}{2}M = \xi M && (15)
\end{aligned}$$

Note that (14) and (15) hold with probability 1, conditioned on the inequality (13) holding, which happens with probability at least $1-\lambda'$. Therefore, the following inequality holds with probability at least $1-\lambda'$:

$$\left|h'(V) - h'(U)\cdot\frac{|V|}{|U|}\right| \le 2^{2z+2}\xi\cdot|V|\cdot(\mathsf{diam}(V)^z + \mathsf{d}(V,C)^z) \tag{16}$$

where we recall that $\eta(V) = \mathsf{d}(V,C)$. The preceding inequality is equivalent to the inequality in the lemma, by recalling that $h'(V) = \mathsf{cost}_t(V,C)$, and $h'(U)\cdot\frac{|V|}{|U|} = \frac{|V|}{|U|}\cdot\mathsf{cost}_{t'}(U,C) = \mathsf{wcost}_{t'}(U,C)$, since the weight of every sampled point in U is equal to $|V|/|U|$. This concludes the proof of the lemma. □

Next, we show the following claim.

Claim 3.

- $\sum_{i,j}|X_{i,j}|(2^jR)^z \le (1+2^z)\cdot\mathsf{cost}_0(X,A) \le (1+2^z)\tau\cdot\mathsf{OPT}(\mathcal{I})$.
- $\sum_{i,j}|X_{i,j}|\mathsf{diam}(X_{i,j})^z \le 2^z(1+2^z)\cdot\mathsf{cost}_0(X,A) \le 2^{2z+1}\cdot\tau\cdot\mathsf{OPT}(\mathcal{I})$.

Proof. For any $p\in X_{i,j}$, it holds that $2^jR \le \max\{2\mathsf{d}(p,A), R\} \le 2\mathsf{d}(p,A)+R$.

$$\begin{aligned}
\sum_{i,j}|X_{i,j}|\cdot(2^jR)^z &\le \sum_{i,j}\sum_{p\in X_{i,j}}(2^jR)^z \\
&\le \sum_{i,j}\sum_{p\in X_{i,j}}(2\mathsf{d}(p,A)+R)^z \\
&= 2^z\sum_{p\in X}\mathsf{d}(p,A)^z + |X|\cdot R^z \\
&= 2^z\cdot\mathsf{cost}_0(X,A) + n\cdot R^z \\
&\le (1+2^z)\cdot\mathsf{cost}_0(X,A) && \text{(By definition of } R\text{)} \\
&\le (1+2^z)\cdot\tau\cdot\mathsf{OPT}(\mathcal{I}') \\
&\le (1+2^z)\cdot\tau\cdot\mathsf{OPT}(\mathcal{I}). && \text{(From Obs. 4)}
\end{aligned}$$

We also obtain the second item by observing that $\mathsf{diam}(X_{i,j}) \le 2\cdot 2^j\cdot R$, and using an analogous argument. □

Next, we show that the following lemma, which informally states that the union of the sets of sampled points approximately preserve the cost of clustering w.r.t. *any* set of at most k centers, *even after* excluding at most m outliers overall.

Lemma 5. *The following statement holds with probability at least $1-\lambda/2$:*
For all sets $C\subseteq F$ of size at most k, and for all sets of non-negative integers $\{m_{i,j}\}_{i,j}$ such that $\sum_{i,j}m_{i,j}\le m$,

$$\left|\sum_{i,j}\mathsf{cost}_{m_{i,j}}(X_{i,j},C) - \sum_{i,j}\mathsf{wcost}_{m'_{i,j}}(S_{i,j},C)\right| \le \epsilon\cdot\sum_{i,j}\mathsf{cost}_{m_{i,j}}(X_{i,j},C) \tag{17}$$

where $t_{i,j} = \lfloor m_{i,j}/w_{i,j}\rfloor$.

Proof. Fix an arbitrary set C of at most k centers and the integers $\{m_{i,j}\}_{i,j}$ such that $\sum_{i,j} m_{i,j} \le m$ as in the statement of the lemma. For each $i = 1, \dots, |A|$, and $0 \le j \le \phi$, we invoke Lemma 4 by setting $V \leftarrow X_{i,j}$, and $U \leftarrow S_{i,j}$, $\xi \leftarrow \frac{\epsilon}{2^{9z}\tau}$, $\lambda' \leftarrow n^{-k}\lambda/(4(k+m)(1+\phi))$, and $q \leftarrow m$. This implies that, the following inequality holds with probability at least $1-\lambda'$ for each set $X_{i,j}$, and for the corresponding $m_{i,j} \le m$:

$$\left|\mathsf{cost}_{m_{i,j}}(X_{i,j}, C) - \mathsf{wcost}_{t_{i,j}}(S_{i,j}, C)\right| \le \frac{\epsilon}{2^{9z}\tau} 2^{2z+2} |X_{i,j}|(\mathsf{diam}(X_{i,j}) + \mathsf{d}(X_{i,j}, C)) \tag{18}$$

Note that for any i, j, if $X_{i,j} < s$, i.e., $X_{i,j}$ is *small*, then the sample $S_{i,j}$ is equal to $X_{i,j}$, and each point in $S_{i,j}$ has weight equal to 1. This implies that $\mathsf{cost}_{t_{i,j}}(X_{i,j}, C) = \mathsf{wcost}_{t'_{i,j}}(S_{i,j}, C)$ for all such $X_{i,j}$, the contribution to the right hand side of inequality (18) is zero. Thus, it suffices to restrict the sum on the right hand side of (18) over *large* sets $X_{i,j}$'s. We have the following claim about the large sets $X_{i,j}$, the proof of which is analogous to that of Claim 2, and is therefore omitted.

Claim 4. $\sum_{i,j: X_{i,j} \text{ is large}} \mathsf{d}(X_{i,j}, C)^z \le 2\mathsf{cost}_m(X, C)$.

Thus, by revisiting (18), we get:

$$\begin{aligned}
&|\mathsf{cost}_m(X, C) - \Delta(C)| \\
&\le \sum_{i,j: X_{i,j} \text{ is large}} \left|\mathsf{cost}_{t_{i,j}}(X_{i,j}, C) - \mathsf{wcost}_{t'_{i,j}}(S_{i,j}, C)\right| && (19)\\
&\le \frac{\epsilon}{2^{9z}\tau} \sum_{i,j: X_{i,j} \text{ is large}} 2^{2z+2} \cdot |X_{i,j}|(\mathsf{diam}(X_{i,j})^z + \mathsf{d}(X_{i,j}, C)^z) \\
& && \text{(By setting } q_{i,j} \leftarrow t_{i,j} \text{ and } q'_{i,j} \leftarrow t'_{i,j} \text{ in (18))}\\
&\le \frac{\epsilon}{2^{9z}\tau} \cdot (2^{2z+2} \cdot 2^{2z+1}\tau \mathsf{OPT}(\mathcal{I}) + 2^{2z+3}\mathsf{cost}_m(X, C) && \text{(From Claim 3 and Claim 4)}\\
&\le \frac{\epsilon}{2^{9z}\tau}(2^{9z} \cdot \tau \cdot \mathsf{cost}_m(X, C)) = \epsilon \cdot \mathsf{cost}_m(X, C))
\end{aligned}$$

Where, the last inequality follows from the fact that since C is an arbitrary set of at most k centers, $\mathsf{OPT}(\mathcal{I}) \le \mathsf{cost}_m(X, C)$. Note that the preceding inequality holds for a fixed set C of centers with probability at least $1 - |A| \cdot (1+\phi)\lambda' = 1 - n^{-k}\lambda/2$, which follows from taking the union bound over all sets $X_{i,j}$, $1 \le i \le |A| \le k+m$, and $0 \le j \le \phi$.

Since there are at most n^k subsets C of F size at most k, the statement of the lemma follows from taking a union bound. $\square$

Once we obtain a coreset S satisfying Lemma 5, we can perform a similar enumeration of sets of size at most m, and obtain $\left(\frac{k+m}{\epsilon}\right)^{O(m)} \cdot n^{O(1)}$ instances of (k,z)-Clustering. We call a β-approximation on each of these instances, and each call takes time $T(n,k)$. The subsequent analysis is identical to that for k-MedianOut which can be used to show an analogous version of Lemma 3. We omit the details, and conclude this section with the following theorem, which generalizes Theorem 1.

Theorem 2. *Let $z \ge 1$ be a fixed constant. Suppose there exists a β-approximation algorithm for (k,z)-Clustering with running time $T(n,k)$ for some constant $\beta \ge 1$, and there exists a τ-approximation algorithm for (k,z)-Clustering that runs in polynomial* *time, where* $\tau = O(1)$*. Then there exists a* $(\beta + \epsilon)$*-approximation algorithm for* (k, z)*-* Clustering with Outliers*, with running time* $\left(\frac{k+m}{\epsilon}\right)^{O(m)} n^{O(1)} \cdot T(n, k)$*, where* n *is the instance size and* m *is the number of outliers.*