The Power of Uniform Sampling for Coresets

(Full version available at arXiv:2209.01901)

Vladimir Braverman*
Johns Hopkins University
vladimir.braverman@gmail.com

Vincent Cohen-Addad
Google Research, Switzerland
vcohenad@gmail.com

Shaofeng H.-C. Jiang‡
Peking University
shaofeng.jiang@pku.edu.cn

Robert Krauthgamer§
Weizmann Institute of Science
robert.krauthgamer@weizmann.ac.il

Chris Schwiegelshohn¶
Aarhus University
cschwiegelshohn@gmail.com

Mads Bech Toftrup
Aarhus University
toftrup@cs.au.dk

Xuan Wu
Johns Hopkins University
wu3412790@gmail.com

***Abstract*—Motivated by practical generalizations of the classic k-median and k-means objectives, such as clustering with size constraints, fair clustering, and Wasserstein barycenter, we introduce a meta-theorem for designing coresets for constrained-clustering problems. The meta-theorem reduces the task of coreset construction to one on a bounded number of ring instances with a much-relaxed additive error. This reduction enables us to construct coresets using uniform sampling, in contrast to the widely-used importance sampling, and consequently we can easily handle constrained objectives. Notably and perhaps surprisingly, this simpler sampling scheme can yield coresets whose size is independent of n, the number of input points.**

Our technique yields smaller coresets, and sometimes the first coresets, for a large number of constrained clustering problems, including capacitated clustering, fair clustering, Euclidean Wasserstein barycenter, clustering in minor-excluded graph, and polygon clustering under Fréchet and Hausdorff distance. Finally, our technique yields also smaller coresets for 1-median in low-dimensional Euclidean spaces, specifically of size $\tilde{O}(\varepsilon^{-1.5})$ in $\mathbb{R}^2$ and $\tilde{O}(\varepsilon^{-1.6})$ in $\mathbb{R}^3$.

***Index Terms*—coresets, clustering, Wasserstein barycenter, fair clustering, capacitated clustering**

I. Introduction

Over the last 20 years, coresets and in particular coresets for clustering problems have received substantial attention. At a high a level, a coreset for a data set P with respect to a set of queries $\mathcal{Q}$ with query-evaluation function $f_P : \mathcal{Q} \to \mathbb{R}_+$, is a data set S with a corresponding $f_S : \mathcal{Q} \to \mathbb{R}_+$ that approximates the evaluation function for every query. Typically, S is a (small) reweighted subset of P, and the function f_S is defined similarly to f_P. For many clustering problems, P is the input set, each query $q \in \mathcal{Q}$ is a candidate center set and its corresponding $f_P(q)$ is the cost induced by this center set, hence a coreset is just a smaller instance (S instead of P) of the same clustering problem.

The quality of a coreset S is usually measured by its size (the number of distinct points) and by the type of query evaluations that it approximates. For example, a natural query for a clustering problem is the cost induced by greedily assigning every point in P to its closest neighbor in the center set q, aggregated over the points in P. A prime example is the Euclidean (k, z)-clustering problem in dimension d, in which $P \subset \mathbb{R}^d$ is the input, $z > 0$ is a parameter (typically fixed), each query is a center set $C \subset \mathbb{R}^d$ of size k, and the query evaluation is the cost function

$$\mathrm{cost}_z(P, C) := \sum_{x \in P} (\mathrm{dist}(x, C))^z, \tag{1}$$

where $\mathrm{dist}(x, y) := \|x - y\|_2$ and $\mathrm{dist}(x, C) := \min_{c \in C} \mathrm{dist}(x, c)$. The special case $z = 2$ is the widely studied Euclidean k-MEANS problem. Following a long line of research [1]–[16], it is now known that Euclidean k-MEANS admits an ε-coreset S of size $\tilde{O}(k\varepsilon^{-2} \min\{k, \varepsilon^{-2}\})$ [16], [17], where an *ε-coreset* means that for every center set C, the cost of P and that of S are within a $(1 \pm \varepsilon)$-factor.

The most immediate approach to construct a coreset is to sample a subset of the input (and reweight its points appropriately), and the main challenge is to find a sampling distribution that works well. A natural starting point is *uniform sampling*, however without further assumptions or preprocessing steps, it is easy to construct instances where uniform sampling requires so many samples that it achieves no significant space savings.[1]

Instead, state-of-the-art algorithms for computing coresets are typically based on non-uniform sampling. These algorithms, initiated by Chen's seminal paper [4] and codified in

Shaofeng H.-C. Jiang is partially supported by a national key R&D program of China No. 2021YFA1000900.

* Work partially supported by ONR Award N00014-18-1-2364.

‡ Research partially supported by a startup fund from Peking University, and the Advanced Institute of Information Technology, Peking University.

§ Work partially supported by ONR Award N00014-18-1-2364, by the Israel Science Foundation grant #1086/18, by a Weizmann-UK Making Connections Grant, by a Minerva Foundation grant, and the Weizmann Data Science Research Center.

¶ Work partially supported by an Independent Research Fund Denmark (DFF) Sapere Aude Research Leader grant No 1051-00106B.

[1] Consider a one-dimensional input $P \subset \mathbb{R}$ with n points at 0 and a single point at 1. For $k = 1$, by placing a center at 0, the only point incurring a positive cost is the point at 1. However, a uniform sample is unlikely to pick the point at 1 unless the sample size is $\Omega(n)$.

2575-8454/22/$31.00 ©2022 IEEE
DOI 10.1109/FOCS54457.2022.00051

their modern form under the name *sensitivity sampling* by Feldman and Langberg [6], draw $|S|$ points from the same probability distribution $(p_x : x \in P)$, and reweight every sample x inverse proportionally to its sampling probability, namely, $w_S(x) = \frac{1}{|S| \cdot p_x}$. The sampling probability p_x is set proportionally to the *sensitivity* of x, which is the maximum possible relative contribution of p to any query evaluation. For example, for Euclidean k-MEANS, this is $s(x) := \sup_{|C|=k} \frac{(\mathrm{dist}(x,C))^2}{\mathrm{cost}(P,C)}$. The sensitivity sampling framework has become an enormously successful and popular method for many additional problems, including kernel methods [18], [19], low-rank approximation [20], linear regression [21], [22], and logistic regression [23], [24].

Unfortunately, not all problems are easily expressed in the sensitivity framework. Consider, for example, clustering with size constraints, which loosely means constraining the number of points served by every center. For example, limiting the centers to each serve at most T points is known as capacitated clustering with uniform capacity T. Constrained clustering introduces a number of technical issues that make it difficult to generalize the analysis for coresets in the unconstrained setting. Perhaps the most glaring obstacle is that sensitivity sampling distorts the total weight of the points (it is preserved only in expectation). While it is easy to preserve the total weight by rescaling the weights in S so that $\sum_{x \in S} w_S(x) = |P|$, size constraints usually require the total weight to be preserved for many subsets, which cannot be achieved under the same scaling. Indeed, directly applying the sensitivity sampling framework to capacitated clustering can result in additive error proportional to the diameter, which is generally unaffordable.

In contrast, uniform sampling can avoid the aforementioned issue, by running it on top of some preprocessing, like Chen's [4] metric decomposition, and indeed it has been applied to obtain coresets for size-constrained clustering problems, including capacitated and fair clustering [25], [26]. While uniform sampling only yields a coreset with additive error for each part in the decomposition, this additive error can essentially compensate for the weight-distortion issue. However, the framework of [4] also has a number of drawbacks compared to the subsequent sensitivity-sampling methods. For example, its coreset size grows (at least linearly) with $\log |P|$, regardless of the metric space and objective function. In contrast, sensitivity sampling, when applicable, can yield coreset size that depends only on k and ε. Thus, improving our ability to apply uniform sampling in coreset constructions has been an important open problem for more than a decade.

A. Our Results

We propose an improved coreset framework that preprocesses the data so that uniform sampling is applicable. Similarly to Chen's method [4], the key step is a reduction to ring instances, defined as follows. A point set R is called a *ring* if all its points are at distance in the range $[r, 2r]$ from some center point c (for some $r > 0$). Our main result is the following meta-theorem (see Theorem III.2 for a formal statement):

Theorem I.1. *Assume that for rings, uniform sampling produces a coreset of size T with* additive *error at most $\varepsilon r|R|$; then for every input P one can construct a coreset of size $O(T \cdot k^2/\varepsilon)$.*

This result generalizes to (k, z)-CLUSTERING, as defined in (1) and more formally in Definition II.2.

This result has a number of applications. To begin with, it allows us to obtain the first coresets whose size is *independent of* $|P|$ for the aforementioned problems of clustering with size constraints. This includes:

- A coreset of size $\mathrm{poly}(k/\varepsilon)$ for a *capacitated* version of Euclidean k-MEDIAN and k-MEANS (see Section II for the definition), which improves over the size bound $\mathrm{poly}(k/\varepsilon \cdot \log |P|)$ of Cohen-Addad and Li [25]. See Theorem IV.5 for details.
- A coreset of size $\mathrm{poly}(k/\varepsilon)$ for *fair* Euclidean k-MEDIAN and k-MEANS (see Section II for the definition), which improves over the size bounds $\mathrm{poly}(k/\varepsilon \cdot \log |P|)$ of Bandyapadhyay, Fomin and Simonov [26], and $\mathrm{poly}(k/\varepsilon^d)$ of Huang, Jiang and Vishnoi [27]. See Theorem IV.7 for details.

Moreover, our framework has applications to other clustering problems. On a technical level, a natural approach to proving that a randomly chosen subset S is a coreset is to first make sure that, with high probability, S preserves the cost for a *single* arbitrary center set, and then apply a union bound over all possible center sets. Since the number of possible center sets can be huge, and even infinite, the space is often discretized by a certain "net" before applying a union bound. For instance, a recent approach established an ε-approximate centroid set, as defined by [17]; however, only a few techniques are known to find such an ε-approximate centroid of small size. A popular alternative to this net approach is to bound VC-dimension[2] of the function space $\{f_x(C) := w(x) \cdot \mathrm{dist}(x, C)\}_x$, where $w(x)$ is related to the probability of sampling x. In general, this $w(x)$ is non-uniform over all $x \in P$, and this is particularly true for coresets constructed via sensitivity sampling. Let us refer to the special case where $w(x)$ is uniform (i.e., takes a single value over all $x \in P$) as the *uniform* function space. Our framework for coreset construction is based on uniform function spaces, which turns out to be a crucial difference with major advantages compared to the sensitivity-sampling framework. Indeed, the advantage of relying on uniform function spaces is two-fold.

1) For several problems, we know how to bound the VC-dimension of the uniform function space, but not that of the non-uniform function space. Examples include the shortest-path metric in planar graphs [28] and the

[2]Strictly speaking, the correct term here is pseudo-dimension, because VC-dimension is defined for a range space (i.e., set system), and we use here VC-dimension as a generic term for sake of exposition. The technical sections use the correct mathematical terminology, which is often the shattering dimension. It is well-known that these three terms are closely related.

Fréchet distance [29]. Our new framework leads to new/improved coreset results for such clustering problems.

2) The uniform function space has a simpler structure and may have a smaller VC-dimension bound. Consider for instance the widely studied range space induced by halfspaces in the Euclidean plane ($\mathbb{R}^d$ for $d = 2$); the VC-dimension of its uniform range space is known to be exactly $d+1 = 3$, whereas for the non-uniform range space the known upper bound is only $3d + 1 = 7$ [30, Lemma 3.3]. This leads directly to better bounds on the coreset size. In particular, when the VC-dimension is low, one can plug in at a key step of the analysis, a bound from discrepancy theory [31] about ε-approximation, which beats the usual ε^{-2} factor.

These advantages lead to new coreset results in several different metric spaces:

- A coreset of size $\tilde{O}(\varepsilon^{-1.5})$ for geometric median in dimension 2, and of size $\tilde{O}(\varepsilon^{-1.6})$ for dimension 3 (see Corollary 5.8 of the full version [32]). The previously known coreset size for these problems was $\tilde{O}(\varepsilon^{-2})$ due to [6].
- A coreset of size $\tilde{O}(\varepsilon^{-2}d\ell)$ for the p-Wasserstein barycenter (see Theorem 5.10 of the full version [32]). This is the k-MEDIAN problem for $k = 1$, in a metric space over all probability distributions that are supported on at most ℓ points in $\mathbb{R}^d$. The p-Wasserstein distance between two distributions D_1 and D_2 is the p-th moment of the minimum cost matching between the distributions (i.e., edge weights represent Euclidean distance raised to power p, and the total cost is raised to power $1/p$). This improves over the previous bound $O(\varepsilon^{-2}d^4\ell^8)$, due to [33].
- A coreset of size $\tilde{O}(|H| \cdot \mathrm{poly}(k/\varepsilon))$ for k-MEDIAN in shortest-path graph metrics that are induced by graphs excluding a fixed minor H (see Section 5.3 of the full version [32]). This improves over a previous bound $\tilde{O}(f(|H|) \cdot \mathrm{poly}(k/\varepsilon))$, due to [15], where f is not specified but is at least doubly exponential.
- A coreset of size $\mathrm{poly}(kd\ell/\varepsilon \cdot \log m)$ for k-MEDIAN under Fréchet and Hausdorff distances (see Section 5.3 of the full version [32]). In this problem, also known as (k, ℓ)-MEDIAN, the data set comprises of polygonal curves in $\mathbb{R}^d$, each with at most m line segments, and the center curves are restricted to at most ℓ line segments. This is the first coreset whose size is independent of the number of input curves, improving over [34].

These new results highlight the flexibility of our framework and we expect that it will have additional applications.

B. Our Techniques

We outline our main technical novelty in obtaining the meta-theorem (formalized in Theorem III.2 that reduces the coreset-construction problem into only $\tilde{O}(k^2/\epsilon)$ ring instances, in which uniform sampling is applicable.

For sake of presentation, let us focus on $z = 1$ (i.e., k-MEDIAN). The proof of the meta-theorem combines several known geometric techniques for constructing coresets, that originally cannot give a coreset with size bound $\mathrm{poly}(k/\epsilon)$. Our algorithm first finds an $(O(1), O(1))$-bicriteria approximation C^* with $|C^*| = O(k)$ centers[3], then partitions the data accordingly into $O(k)$ clusters, and then further partitions each cluster into rings with exponentially-increasing radii, similarly to the steps in [4]. The issue with this partition, as noted also in [4], is that it creates $O(\log n)$ rings, which eventually introduces an $O(\log n)$ factor in the coreset size. To bypass this, we identify in each cluster a set of $\tilde{O}(k/\epsilon)$ high-cost rings (and thus $\tilde{O}(k^2/\epsilon)$ rings in total), for which the points inside contribute significantly to the objective. Call these high cost rings *marked*, and the remaining rings *unmarked*. Consecutive unmarked rings (i.e., between two marked rings) are merged into in at most $\tilde{O}(k/\epsilon)$ *unmarked groups*. The $\tilde{O}(k^2/\epsilon)$ marked rings are handled as in [4] using uniform sampling. The remaining issue is how to construct coresets for the unmarked groups. An unmarked group can be a union of multiple consecutive rings, and since points do not have a similar distance to the cluster center, uniform sampling is no longer applicable. However, by our construction, each unmarked group has a small contribution to the cost and we show that a simple two-point geometric construction can already serve as a coreset for the entire group. Such a two-point coreset is much more powerful than it appears to be. In particular, it even satisfies a property that we call assignment-preserving (see Definition III.1 and a similar formulation in prior work [26], [27], [35]), and hence can serve as a coreset for clustering with capacity and fairness constraints.

Let $c_i \in C^*$ and let P_i denote the cluster with center c_i. Technically, the construction of the unmarked groups and their two-point coresets is done by interpreting the entire cluster as a one-dimensional instance (by taking $\mathrm{dist}(x, c_i)$ for each point $x \in P_i$), and then applying on the unmarked rings a known greedy-bucketing construction for dimension one [3]. To construct the two-points coreset for a group G, let $p^{\mathrm{close}}, p^{\mathrm{far}} \in G$ be a closest point and a furthest point, respectively, from the center c_i. Then for every point $x \in G$, represent the distance $\mathrm{dist}(x, c_i)$ as a convex combination of $\mathrm{dist}(p^{\mathrm{close}}, c_i)$ and $\mathrm{dist}(p^{\mathrm{far}}, c_i)$, namely, find $\lambda_x \in [0, 1]$ such that $\mathrm{dist}(x, c_i) = \lambda_x \,\mathrm{dist}(p^{\mathrm{close}}, c_i) + (1 - \lambda_x)\,\mathrm{dist}(p^{\mathrm{far}}, c_i)$. Now let the coreset for G be $S := \{p^{\mathrm{close}}, p^{\mathrm{far}}\}_w$ with weights $w(p^{\mathrm{close}}) = \sum_{x\in G} \lambda_x$ and $w(p^{\mathrm{far}}) = \sum_{x\in G}(1-\lambda_x)$. Obviously, S has only two distinct points and it preserves the total weight and the cost with respect to c_i as the entire G.

It remains to analyze the error between our two-point coreset S and the group G with respect to an arbitrary center set, even with capacity constraints. Fix a center set C with $|C| = k$ and capacity constraint $\Gamma : C \to \mathbb{R}_+$ that prescribes the number of points connected to each center $c \in C$ (see Definition II.3 for formal definition). We first observe that due

[3] An (α, β)-bicriteria approximation for a clustering problem is a set of at most $\beta \cdot k$ centers that has cost $\alpha \cdot \mathrm{OPT}_k$, where OPT_k is the optimal cost of clustering using k centers.

to the triangle inequality and our grouping method, the cost of clustering S approximates that of G within an additive error, namely, $|\operatorname{cost}(G,C,\Gamma)-\operatorname{cost}(S,C,\Gamma)|\le \tilde{O}(\frac{\epsilon}{k})\cdot\operatorname{cost}(P_i,c_i)$ (see Definition II.4 and Lemma III.5). However, as the cluster P_i has $\tilde{O}(\frac{k}{\epsilon})$ unmarked groups, its cumulative error is bounded by $\tilde{O}(\frac{k}{\epsilon})\cdot\tilde{O}(\frac{\epsilon}{k})\cdot\operatorname{cost}(P_i,c_i)=\tilde{O}(\operatorname{cost}(P_i,c_i))$, which exceeds our intended error bound $\tilde{O}(\epsilon)\cdot\operatorname{cost}(P_i,c_i)$. To reduce the number of groups that can suffer an additive error, we further divide the unmarked groups into colored groups and uncolored groups with respect to C. In particular, we call a ring "important" if it contains any center from C. We "color" $O(\log\frac{1}{\epsilon})$ neighboring rings of each important ring and "color" all the groups that contain at least one colored ring. This way, we obtain at most $O(k\log\frac{1}{\epsilon})$ colored groups. We let these $O(k\log\frac{1}{\epsilon})$ colored groups suffer the additive error, and this time the total error from them is bounded by $\tilde{O}(\epsilon)\cdot\operatorname{cost}(P_i,c_i)$.

It remains to bound the error for the uncolored groups, and crucially, in Lemma III.6 we show these groups do not suffer an additive error but only a multiplicative error. A key observation is that if a group G is not colored (with respect to C), then every $c\in C$ is either too far from all the points in G or too close to the cluster center c_i. Based on this observation, we surprisingly find that when the group is not colored, our simple two-points coresets S can already serve as an assignment-preserving coreset *without* additive error. This Lemma III.6 is one of the main technical lemmas that deal with the assignment constraint, and its proof requires very careful explicit constructions for the assignments of the two-point coreset S and the group G.

C. Additional Related Work

Although the coreset paradigm is most often applied to clustering problems, there are actually several other applications, see the surveys [36], [37] for further pointers to the literature. Restricting attention to coresets for clustering, the most common setting is that of a Euclidean space, but there are many results also for other metric spaces. To streamline the presentation, we focus here on the results for k-MEDIAN. For general n-point metrics, [6] gave coresets of size $O(\frac{k\log n}{\epsilon^2})$, and for general metrics with bounded doubling dimension d, [38] designed a coreset of size $O(\frac{k^3d}{\epsilon^2})$, which was later improved by [17] to $\tilde{O}(\frac{kd}{\epsilon^2})$. Another line of research addresses the shortest-path metrics of graphs, and notably, $\operatorname{poly}(k/\epsilon)$-size coresets for k-MEDIAN were obtained for graphs of bounded treewidth, planar graphs, and more generally excluded-minor graphs [15], [17], [39]. For an empirical evaluation of these algorithms, we refer to [40].

Coresets for even more general clustering problems, i.e., beyond (k,z)-CLUSTERING, received significant attention as well. Apart from the capacity and fairness constrained clustering that are studied in this paper, coresets were designed also for ordered weighted clustering [41], for clustering with outliers [38], [42], for training Gaussian mixture models [43], [44], for time-series clustering [45], and many other related problems. Another interesting generalization is clustering of sets of points in $\mathbb{R}^d$ (instead of points), including arbitrary finite sets [46], lines [47], and axis-align affine subspaces [30].

II. PRELIMINARIES

We use $\mathbb{R}_+$ to denote set $\{x\ge 0\mid x\in\mathbb{R}\}$. A weighted set S is associated with a weight function $w_S: S\to\mathbb{R}_+$. We interpret an unweighted set S as a weighted set with unit weight, i.e., $w_S(\cdot)=1$. For some weight function $w_S: S\to\mathbb{R}_+$ and $T\subseteq S$, define $w_S(T):=\sum_{x\in T}w_S(x)$. We assume there is an underlying metric $M(X,\operatorname{dist})$ throughout the paper. This metric space may not be finite; for instance, it can be Euclidean space $(\mathbb{R}^d,\ell_2)$. For a point $x\in X$ and a point set $C\subseteq X$, let $\operatorname{dist}(x,C):=\min_{c\in C}\operatorname{dist}(x,c)$. For $u\in X, 0\le a<b$, let $\operatorname{ring}(u,a,b):=\{x\in X: a<\operatorname{dist}(x,u)\le b\}$ be the set of points within distance between a and b from u.

We need the following generalized triangle inequalities which are well-known tools for studying (k,z)-CLUSTERING. Variants of these inequalities can be found in multiple related papers [8], [11], [17], [48].

Lemma II.1 (Generalized triangle inequality)**.** *Let $a,b,c\in X$ and $z\ge 1$. For every $0<t\le 1$, the following inequalities hold.*

1) *(Corollary A.2 of [48])*
$$\operatorname{dist}(a,b)^z\le(1+t)^{z-1}\operatorname{dist}(a,c)^z+\big(1+\frac{1}{t}\big)^{z-1}\operatorname{dist}(b,c)^z$$
2) *(Claim 5 of [11])*
$$|\operatorname{dist}(a,c)^z-\operatorname{dist}(b,c)^z|\le t\cdot\operatorname{dist}(a,c)^z+(\frac{3z}{t})^{z-1}\operatorname{dist}(a,b)^z.$$

Definition II.2 (Coresets for (k,z)-CLUSTERING)**.** Given a weighted data set $P\subseteq X$, for $C\subseteq X$ with $|C|\le k$, define the cost for (k,z)-CLUSTERING as
$$\operatorname{cost}_z(P,C):=\sum_{x\in P}w_P(x)\cdot(\operatorname{dist}(x,C))^z.$$
For $0<\epsilon<1$, a weighted set S such that $S\subseteq P$ is an ϵ-coreset for (k,z)-CLUSTERING if
$$\forall C\subseteq X,|C|\le k,\qquad \operatorname{cost}_z(S,C)\in(1\pm\epsilon)\cdot\operatorname{cost}_z(P,C),$$

The following definition of assignment constraints generally captures the constraints in fair clustering and capacitated clustering, and our key notion of assignment-preserving coresets is defined with respect to it. Similar notions of assignment constraints and assignment-preserving coresets have also been considered in previous works which study fair clustering [26], [27], [35].

Definition II.3 (Assignment constraints and assignment functions)**.** Given a weighted set $P\subseteq X$ and $C\subseteq X$, an assignment constraint is a function $\Gamma: C\to\mathbb{R}_+$ such that $\sum_{c\in C}\Gamma(c)=w_P(P)$, and we call an assignment function $\sigma: P\times C\to\mathbb{R}_+$ consistent with Γ, denoted as $\sigma\sim\Gamma$, if $\forall c\in C$, $\sigma(P,c):=\sum_{p\in P}\sigma(p,c)=\Gamma(c)$. For $P_1\subseteq P$ and $C_1\subseteq C$, we define
$$\operatorname{cost}_z^\sigma(P_1,C_1):=\sum_{x\in P_1}\sum_{c\in C_1}\sigma(x,c)\cdot(\operatorname{dist}(x,c))^z$$

as the connection cost between P_1 and C_1 under σ.

Definition II.4 ((k,z)-CLUSTERING with assignment constraints)**.** Given a weighted data set $P \subseteq X$, a center set $C \subseteq X$ with $|C| \leq k$, and an assignment constraint $\Gamma : C \to \mathbb{R}_+$ the objective for (k,z)-CLUSTERING with assignment constraint Γ is defined as

$$\mathrm{cost}_z(P, C, \Gamma) := \min_{\sigma: P\times C\to\mathbb{R}_+, \sigma\sim\Gamma} \mathrm{cost}_z^{\sigma}(P, C).$$

Definition II.5 (Assignment-preserving coresets for (k,z)-CLUSTERING)**.** Let P be a weighted dataset. A weighted subset $S \subseteq P$ is an assignment-preserving ϵ-coreset for (k,z)-CLUSTERING, if $w_P(P) = w_S(S)$, and for every $C \subseteq X$ with $|C| \leq k$ and assignment constraint $\Gamma : C \to \mathbb{R}_+$,

$$\mathrm{cost}_z(P, C, \Gamma) \in (1 \pm \epsilon) \cdot \mathrm{cost}_z(S, C, \Gamma).$$

We make an observation in Fact II.6 that an assignment-preserving coreset is as well a coreset for (unconstrained) clustering.

Fact II.6. *For $P \subseteq X$, if $S \subseteq P$ is an assignment-preserving ϵ-coreset for (k,z)-CLUSTERING on P, then S is an ϵ-coreset for (k,z)-CLUSTERING on P.*

Moreover, this definition of assignment-preserving coresets generally captures many capacity-constrained clustering problems. For instance, in capacitated clustering, the goal is to minimize the (k,z)-CLUSTERING objective subject to the constraint that each center is assigned by at most a certain number of data points. Coresets for capacitated clustering have been considered in [25] and our notion of assignment-preserving coresets captures their definition.

Fair clustering: In (α,β)-fair (k,z)-CLUSTERING ([49], [50]), a data set P, a collections of groups (not necessary disjoint) $P_1, P_2, ..., P_m \subseteq P$ and two constraints vectors $\alpha, \beta \in [0,1]^m$ are given. The objective is to find an assignment σ from P to C such that for every group P_i and every center $c \in C$,

$$\frac{\sigma(P_i, c)}{\sigma(P, c)} \in [\alpha_i, \beta_i].$$

It has been well known that the requirement of (α,β)-fair (k,z)-CLUSTERING can be expressed as a collection of assignment constraints [26], [27], [35], [51]. Following the reduction in [27], an algorithm that constructs assignment-preserving coresets for (k,z)-CLUSTERING implies coresets algorithm for (α,β)-fair (k,z)-CLUSTERING (See Section IV for more details).

III. NEW FRAMEWORK

Definition III.1 (Assignment-preserving coresets with additive error)**.** Given a data set $P \subseteq X$, a subset $S \subseteq P$ is called an assignment-preserving (ϵ, A)-coreset for (k,z)-CLUSTERING on P, if for every $C \subseteq X$ with $|C| \leq k$ and every assignment constraint $\Gamma : C \to \mathbb{R}_+$,

$$|\,\mathrm{cost}_z(P, C, \Gamma) - \mathrm{cost}_z(S, C, \Gamma)| \leq \epsilon \cdot (\mathrm{cost}_z(P, C, \Gamma) + A).$$

The main idea of our new framework (Theorem III.2) is to reduce constructing coresets on general datasets, to the special case of constructing coresets on datasets that belong to *rings*. Note that for the rings, we only consider coresets with an additional *additive* error (Definition III.1), which seems to be weaker than the relative-error coresets that we aim for. However, by a standard argument, this actually suffices to imply a coreset for the entire dataset without the additive error (see Appendix A of the full version [32]).

Theorem III.2. *There is an algorithm that given dataset $P \subseteq X$, center $c \in X$, $0 < \epsilon < 1$, computes a 2-partition $\{W, Z\}$ of P and a weighted point set $S \subseteq P$ of size $2^{O(z\log z)} \cdot \tilde{O}(k\epsilon^{-z})$, such that*

1) *W consists of $2^{O(z\log z)} \cdot \tilde{O}(k\epsilon^{-z})$ rings $\{R_i\}_i$ where each $R_i \subseteq \mathrm{ring}(c, r_i, 2r_i)$ for some $r_i > 0$, and*
2) *S is an assignment-preserving $(\epsilon, \mathrm{cost}_z(P,c))$-coreset for (k,z)-CLUSTERING on Z,*

running in time $\tilde{O}(|P|k)$.

Note that the assignment-preserving coreset S for the Z part can be constructed even in general metrics. Moreover, this assignment-preserving coreset is very general (see Fact II.6), and it can be used as a coreset for all clustering problems that we consider in this paper. Hence, in order to obtain a full coreset, it only remains to construct coresets for W, which merely consists of $2^{O(z\log z)} \cdot \tilde{O}(k\epsilon^{-z})$ *rings*. Therefore, this theorem essentially reduces the coreset construction for a general data set to ring datasets. In particular, if one can obtain a coreset (with additive error) of size $T(\epsilon, k, z)$ for each ring, then one can construct a coreset of size $2^{O(z\log z)} \cdot \tilde{O}(k^2\epsilon^{-z}) \cdot T(\epsilon, k, z)$ for the entire dataset.

Improved bound for $k = 1$: For the special case of $(1,z)$-CLUSTERING (noting that when $k = 1$ the assignment constraints become trivial and it is equivalent to the unconstrained version), we have a better argument that yields an improved dependence in ϵ.

Theorem III.3. *There is an algorithm that given dataset $P \subseteq X$, center $c \in X$, $0 < \epsilon < 1$, computes a 2-partition $\{W, Z\}$ of P and a weighted point set $S \subseteq P$ of size 3, such that*

1) *W consists of $O(\log \frac{z}{\epsilon})$ rings $\{R_i\}_i$ where each $R_i \subseteq \mathrm{ring}(c, r_i, 2r_i)$ for some $r_i > 0$, and*
2) *S is an $(\epsilon, \mathrm{cost}_z(P,c))$-coreset for $(1,z)$-CLUSTERING on Z,*

running in time $\tilde{O}(|P|k)$.

The power of uniform sampling: Due to the uniform nature of the ring datasets, we can show that the very simple uniform sampling already suffices for constructing coresets for (k,z)-CLUSTERING on ring datasets (See Section 5 of the full version [32]). This new construction based on uniform sampling further reduces the construction of coresets into bounding the *uniform* shattering dimension of the ball range space induced by the metric space. The uniform shattering dimension is both easier to analyze, and wider considered in the literature than the much more involved weighted shattering

dimension used in previous works, which in turn results in several new and/or improved coreset size bounds.

A. Proof of Theorem III.2

We provide a sketch of the main algorithm in Algorithm 1 and present details of each step in the corresponding paragraph.

Algorithm 1 Algorithm Outline for Theorem III.2

1: set $t \leftarrow \lceil 2 + \log \frac{24zk}{\epsilon} \rceil$, $\mathrm{err} \leftarrow \left(\frac{\epsilon}{6z}\right)^z \cdot \frac{\mathrm{cost}_z(P,c)}{kt}$, and $\mathbb{Z}^* \leftarrow \mathbb{Z} \cup \{-\infty\}$
2: decompose P into rings $P_i \leftarrow P \cap \mathrm{ring}(c, 2^{i-1}, 2^i)$ ($i \in \mathbb{Z}^*$) ▷ call P_i heavy if $\mathrm{cost}_z(P_i, c) \geq \mathrm{err}$
3: mark all heavy rings
4: merge consecutive unmarked rings to obtain $2^{O(z \log z)} \cdot \tilde{O}(k\epsilon^{-z})$ many groups such that each of the groups has cost at most err, as in Lemma III.4
5: construct a two-points coreset for each group produced in the last step
6: let W be the union of marked rings, let $Z \leftarrow P \setminus W$, and let S include the union of coresets obtained in the last step

Ring decomposition: Set $t := \lceil 2 + \log \frac{24zk}{\epsilon} \rceil$, $\mathrm{err} := \left(\frac{\epsilon}{6z}\right)^z \cdot \frac{\mathrm{cost}_z(P,c)}{kt}$, and $\mathbb{Z}^* \leftarrow \mathbb{Z} \cup \{-\infty\}$. If $c \in P$, add c into both Z and S in advance, and let $P \leftarrow P \setminus \{c\}$. Decompose P into rings $\{P_i \mid i \in \mathbb{Z}^*\}$, where for $i \in \mathbb{Z}$,

$$P_i := P \cap \mathrm{ring}(c, 2^{i-1}, 2^i)$$

and if $i = -\infty$, $P_i := P \cap \{c\}$.

Since at most $|P|$ rings are non-empty, we can easily compute the above decomposition in near-linear time.

Call a ring j *heavy* if $\mathrm{cost}_z(P_j, c) \geq \mathrm{err}$. So the number of heavy rings is at most $\frac{\mathrm{cost}_z(P,c)}{\mathrm{err}} = 2^{O(z \log z)} \cdot \tilde{O}(k\epsilon^{-z})$. We mark all heavy rings. Call a ring *unmarked* if it is not a marked ring.

Defining the partition: Now, we define Z as the set of points belong to the unmarked rings, and define W as the marked rings. Clearly, W consists of $2^{O(z \log z)} \cdot \tilde{O}(k\epsilon^{-z})$ (marked) rings. Hence, it remains to construct an $(\epsilon, \mathrm{cost}_z(P,c))$-coreset for Z, the unmarked rings.

Re-grouping unmarked rings: Observe that unmarked rings can be grouped into $2^{O(z \log z)} \cdot \tilde{O}(k\epsilon^{-z})$ buckets of consecutive rings, due to the fact that there are at most $2^{O(z \log z)} \cdot \tilde{O}(k\epsilon^{-z})$ heavy rings. Denote these buckets as $B_1, B_2, \ldots$, where each B_i consists of a collection of consecutive unmarked rings. For every bucket B_i, we apply the following Lemma III.4 to further group B_i into $\mathcal{G}_i$. Let $\mathcal{G} := \bigcup_i \mathcal{G}_i$ be the set of all groups of unmarked rings.

Lemma III.4. *For every bucket B_i, rings in B_i can be grouped into $\mathcal{G}_i$ where each $G \in \mathcal{G}_i$ consists of consecutive unmarked rings, such that $\forall G \in \mathcal{G}_i$, $\mathrm{cost}_z(G,c) \leq \mathrm{err}$. Furthermore, $|\bigcup_i \mathcal{G}_i| \leq 2^{O(z \log z)} \cdot \tilde{O}(k\epsilon^{-z})$.*

Proof: Fix some B_i. We start with constructing a grouping $\mathcal{G}'_i$ of B_i. Initialize $\mathcal{G}'_i := \emptyset$ as the tentative result. List points p in B_i in the increasing order of $\mathrm{dist}(p,c)$. Starting from the first element in B_i, greedily pick a maximal subset G' (in order) such that $\mathrm{cost}_z(G', c) \leq \mathrm{err}$. Here, we allow G' to include points fractionally. Keep on picking such G' and add it into $\mathcal{G}'_i$, until all elements in B_i are picked.

We construct $\mathcal{G}_i$ from $\mathcal{G}'_i$ as follows. Observe that each $G' \in \mathcal{G}'_i$ can *partially* intersect at most two rings from B_i. Now, examine each $G' \in \mathcal{G}'_i$, for every partially intersected ring R, designate this entire R as a new group and include it in $\mathcal{G}_i$, remove the partially intersected rings from G', and include all rings in the remaining G' as a new group to $\mathcal{G}_i$. Eventually, remove the empty or duplicated groups from $\mathcal{G}_i$, if any.

By construction, $\forall G \in \mathcal{G}_i$, either G consists of a single unmarked ring which implies $\mathrm{cost}_z(G,c) \leq \mathrm{err}$, or G consists of several consecutive unmarked rings such that $\mathrm{cost}_z(G,c) \leq \mathrm{err}$.

Hence, it remains to bound $|\bigcup_i \mathcal{G}_i|$. Since every $G' \in \mathcal{G}'_i$ creates at most 3 groups in $\mathcal{G}_i$, we have $|\mathcal{G}_i| \leq 3 \cdot |\mathcal{G}'_i|$. Observe that in every $\mathcal{G}'_i$, there is at most one G' such that $\mathrm{cost}_z(G,c) < \mathrm{err}$ and all other G' satisfy $\mathrm{cost}_z(G,c) = \mathrm{err}$. Therefore, since there are at most $2^{O(z \log z)} \cdot \tilde{O}(k\epsilon^{-z})$ buckets, over all i, we have

$$\left|\bigcup_i \mathcal{G}_i\right| \leq 2^{O(z \log z)} \cdot \tilde{O}(k\epsilon^{-z}) + \frac{\mathrm{cost}(P,c)}{\mathrm{err}} = 2^{O(z \log z)} \cdot \tilde{O}(k\epsilon^{-z}).$$

This finishes the proof of Lemma III.4. ∎

Two-points coresets construction for unmarked groups: To construct the coreset S for the unmarked rings, we first construct a two-points coreset for each group of rings $G_i \in \mathcal{G}$, and then take the union of them.

For every group $G_i \in \mathcal{G}$, we construct a coreset of only two points using the following steps. Let $p_i^{\mathrm{close}}, p_i^{\mathrm{far}} \in G_i$ be the closest and furthest points to c (breaking ties consistently), respectively. Then for every $p \in G_i$, $\mathrm{dist}^z(p,c)$ can be represented by $\mathrm{dist}^z(p,c) = \lambda_p \cdot \mathrm{dist}^z(p_i^{\mathrm{close}}, c) + (1 - \lambda_p) \cdot \mathrm{dist}^z(p_i^{\mathrm{far}}, c)$, such that $\lambda_p \in [0,1]$. Then define the weight $w(p_i^{\mathrm{close}}) := \sum_{p \in G_i} \lambda_p$, and similarly the weight of $w(p_i^{\mathrm{far}}) := \sum_{p \in G_i} (1 - \lambda_p)$. Note that $w(p_i^{\mathrm{close}}) + w(p_i^{\mathrm{far}}) = w_P(G_i)$ and $\mathrm{cost}_z(\{p_i^{\mathrm{close}}, p_i^{\mathrm{far}}\}_w, c) = \mathrm{cost}_z(G_i, c)$, where $\{p_i^{\mathrm{close}}, p_i^{\mathrm{far}}\}_w$ denotes the two-point coreset.

Size analysis: Since the unmarked rings are partitioned into $|\mathcal{G}| \leq 2^{O(z \log z)} \cdot \tilde{O}(k\epsilon^{-z})$ groups and each of the group creates two coreset points in S, the total number of coreset points is bounded by $|S| \leq 2^{O(z \log z)} \cdot \tilde{O}(k\epsilon^{-z})$.

Error analysis: Observe that the requirement of $|Z| = w_S(S)$ follows immediately from the construction, hence we focus on analyzing the coreset cost error.

In the following, we fix a center set C with $|C| = k$, and an assignment constraint $\Gamma : C \to \mathbb{R}_+$. We call a ring P_i *important* if there exists $u \in C$, $2^{i-1} < \mathrm{dist}(u,c) \leq 2^i$. Namely, P_i is important if $C \cap \mathrm{ring}(c, 2^{i-1}, 2^i) \neq \emptyset$. We color a ring P_j if there exists an important ring P_i such that $|i - j| \leq$

t. Namely, for every important ring, we color its $2t$ neighbors and itself.

We call a group $G_i \in \mathcal{G}$ *colored* if it contains any colored ring. We call a group *uncolored* if it is not colored. Colored and uncolored groups have totally different behaviors in producing coreset error. We bound the coresets errors producing by colored and uncolored groups, by using Lemma III.5 and Lemma III.6 respectively.

Lemma III.5. *Let G_i be an unmarked group. Let $\sigma : G_i \times C \to \mathbb{R}_+$ and $\pi : D_i \times C \to \mathbb{R}_+$ both consistent with Γ. Then we have*

$$\begin{aligned} & |\mathrm{cost}_z^\sigma(G_i, C) - \mathrm{cost}_z^\pi(D_i, C)| \\ \leq \ & \epsilon \cdot \mathrm{cost}_z^\sigma(G_i, C) + \frac{\epsilon}{3kt} \cdot \mathrm{cost}_z(P, c). \end{aligned}$$

Lemma III.5 can be proved by triangle inequality and due to page limitation, we do not include the proof here but refer to the full version [32].

Note that Lemma III.5 works for both colored and uncolored groups. However, the number of uncolored groups can be much larger than the number of colored groups. Thus we must use a refined analysis Lemma III.6 to bound the error of these uncolored groups.

Lemma III.6. *Let $G_i \in \mathcal{G}$ denote an unmarked and uncolored group. Let $D_i := \{p_i^{\mathrm{close}}, p_i^{\mathrm{far}}\}_w$ be the two-point coreset constructed for G_i. The following holds.*

1) *For every $\sigma : G_i \times C \to \mathbb{R}_+$ consistent with Γ (see Definition II.3 for the relevant definitions), there exists $\pi : D_i \times C \to \mathbb{R}_+$ consistent with Γ, such that $\mathrm{cost}_z^\pi(D_i, C) \leq (1+\epsilon) \cdot \mathrm{cost}_z^\sigma(G_i, C)$.*
2) *For $\pi : D_i \times C \to \mathbb{R}_+$ consistent with Γ, there exists $\sigma : G_i \times C \to \mathbb{R}_+$ consistent with Γ, such that $\mathrm{cost}_z^\sigma(G_i, C) \leq (1+\epsilon)\,\mathrm{cost}_z^\pi(D_i, C)$.*

In particular, let σ and π denote the optimal assignment consistent with Γ from G_i and D_i to C respectively, then

$$|\,\mathrm{cost}_z^\sigma(G_i, C) - \mathrm{cost}_z^\pi(D_i, C)| \leq O(\epsilon) \cdot \mathrm{cost}_z^\sigma(G_i, C).$$

Proof: Divide C into $C_{\mathrm{close}} := \{u \in C \mid \mathrm{dist}(u, c) < \mathrm{dist}(p_i^{\mathrm{close}}, c)\}$ and $C_{\mathrm{far}} := C \setminus C_{\mathrm{close}}$. Recall that G_i is uncolored. We need the following inequalities to characterize the distances between G_i and C_{far} or C_{close}.

Lemma III.7. *The following inequalities hold.*

1) *For every $u \in C_{\mathrm{close}}$, $\mathrm{dist}(p_i^{\mathrm{close}}, c) > \frac{9z}{\epsilon}\,\mathrm{dist}(u, c)$.*
2) *For every $u \in C_{\mathrm{far}}$, $\mathrm{dist}(p_i^{\mathrm{far}}, c) < \frac{\epsilon}{24z} \cdot \mathrm{dist}(u, c)$.*
3) *For every $x \in G_i$ and $u \in C_{\mathrm{far}}$, $\mathrm{dist}(p_i^{\mathrm{far}}, c) < \frac{\epsilon}{12z} \cdot \mathrm{dist}(x, u)$.*
4) *For every $x \in G_i$ and $u \in C_{\mathrm{far}}$, $\mathrm{dist}^z(x, u) \in (1 \pm \frac{\epsilon}{6}) \cdot \mathrm{dist}^z(u, c)$.*
5) *For every $x \in G_i$ and $u \in C_{\mathrm{close}}$, $\mathrm{dist}^z(x, u) \in (1 \pm \frac{\epsilon}{6}) \cdot \mathrm{dist}^z(x, c)$.*

Proof: For item (1) and (2), assume u is in P_j, so P_j is important. As G_i is not colored, by definition, we know that for every $P_{i'} \subseteq G_i$, $|i' - j| > t$. So if $u \in C_{\mathrm{close}}$, we have

$$\mathrm{dist}(p_i^{\mathrm{close}}, c) \geq 2^{t-1} \cdot \mathrm{dist}(p_j^{\mathrm{far}}, c) > \frac{9z}{\epsilon} \cdot \mathrm{dist}(u, c).$$

On the other hand, if $u \in C_{\mathrm{far}}$, we have

$$\mathrm{dist}(p_i^{\mathrm{far}}, c) \leq 2^{-t+1} \cdot \mathrm{dist}(p_j^{\mathrm{close}}, c) \leq \frac{\epsilon}{24z} \cdot \mathrm{dist}(u, c).$$

Now we prove item (3). By item (2), $\max_{x \in G_i} \mathrm{dist}(x, c) = \mathrm{dist}(p_i^{\mathrm{far}}, c) < \frac{\epsilon}{24z} \cdot \mathrm{dist}(u, c)$, by triangle inequality we know that for every $x \in G_i$,

$$\begin{aligned} \mathrm{dist}(x, u) \ &\geq \ \mathrm{dist}(u, c) - \mathrm{dist}(x, c) > \big(\frac{24z}{\epsilon} - 1\big)\,\mathrm{dist}(p_i^{\mathrm{far}}, c) \\ &> \ \frac{12z}{\epsilon} \cdot \mathrm{dist}(p_i^{\mathrm{far}}, c). \end{aligned}$$

So we know that for every $u \in C_{\mathrm{far}}$ and $x \in G_i$, $\mathrm{dist}(p_i^{\mathrm{far}}, c) < \frac{\epsilon}{12z} \cdot \mathrm{dist}(x, u)$.

Now we prove item (4). By the triangle inequality and item (2) we know that,

$$\begin{aligned} |\,\mathrm{dist}(x, u) - \mathrm{dist}(u, c)| \ &\leq \ \mathrm{dist}(x, c) \leq \mathrm{dist}(p_i^{\mathrm{far}}, c) \\ &< \ \frac{\epsilon}{24z} \cdot \mathrm{dist}(u, c). \end{aligned}$$

Thus $\mathrm{dist}(x, u) \in \big(1 \pm \frac{\epsilon}{24z}\big) \cdot \mathrm{dist}(u, c)$. Thus $\mathrm{dist}^z(x, u) \in \big(1 \pm \frac{\epsilon}{6}\big) \cdot \mathrm{dist}^z(u, c)$.

Now we prove item (5). By the triangle inequality and item (1), we know that

$$\begin{aligned} |\,\mathrm{dist}(x, u) - \mathrm{dist}(x, c)| \ &\leq \ \mathrm{dist}(u, c) < \frac{\epsilon}{9z} \cdot \mathrm{dist}(p_i^{\mathrm{close}}, c) \\ &\leq \ \frac{\epsilon}{9z} \cdot \mathrm{dist}(x, c). \end{aligned}$$

Thus $\mathrm{dist}(x, u) \in \big(1 \pm \frac{\epsilon}{9z}\big) \cdot \mathrm{dist}(x, c)$. Thus $\mathrm{dist}^z(x, u) \in (1 \pm \frac{\epsilon}{6}) \cdot \mathrm{dist}^z(x, c)$. This finishes the proof of Lemma III.7. ∎

We return to the proof of Lemma III.6, we prove item (1) and item (2) separately.

Proof of item (1): Recall that for every $x \in G_i$, there exists a unique λ_x such that $\mathrm{dist}^z(x, c) = \lambda_x\,\mathrm{dist}^z(p_i^{\mathrm{close}}, c) + (1-\lambda_x)\,\mathrm{dist}^z(p_i^{\mathrm{far}}, c)$, and we have set $w(p_i^{\mathrm{close}}) = \sum_{x \in G_i} \lambda_x$ and $w(p_i^{\mathrm{far}}) = \sum_{x \in G_i}(1 - \lambda_x)$.

Consider some $\sigma : G_i \times C \to \mathbb{R}_+$ consistent with Γ. We define $\pi : D_i \times C \to \mathbb{R}_+$, such that for every $u \in C$, $\pi(p_i^{\mathrm{close}}, u) = \sum_{x \in G_i} \lambda_x \sigma(x, u)$ and $\pi(p_i^{\mathrm{far}}, u) = \sum_{x \in G_i}(1 - \lambda_x)\sigma(x, u)$. We note that for every $u \in C$,

$$\pi(D_i, u) = \sigma(G_i, u).$$

Hence π is consistent with Γ. It remains to prove that for every $u \in C$,

$$\begin{aligned} & \pi(p_i^{\mathrm{close}}, u)\,\mathrm{dist}^z(p_i^{\mathrm{close}}, u) + \pi(p_i^{\mathrm{far}}, u)\,\mathrm{dist}^z(p_i^{\mathrm{far}}, u) \\ \leq \ & (1+\epsilon) \sum_{x \in G_i} \sigma(x, u) \cdot \mathrm{dist}^z(x, u). \end{aligned}$$

- If $u \in C_{\mathrm{far}}$, by item (3) of Lemma III.7, we know that for every $u \in C_{far}$ and $x \in G_i$, $\mathrm{dist}(p_i^{\mathrm{far}}, c) <$

$\frac{\epsilon}{12z} \cdot \mathrm{dist}(x,u)$. So by the generalized triangle inequality Lemma II.1 we have

$$\begin{aligned}
&\pi(p_i^{\mathrm{close}}, u)\,\mathrm{dist}^z(p_i^{\mathrm{close}}, u) + \pi(p_i^{\mathrm{far}}, u)\,\mathrm{dist}^z(p_i^{\mathrm{far}}, u)\\
&= \sum_{x\in G_i} \big(\lambda_x \sigma(x,u)\,\mathrm{dist}^z(p_i^{\mathrm{close}}, u)\\
&\quad + (1-\lambda_x)\,\sigma(x,u)\,\mathrm{dist}^z(p_i^{\mathrm{far}}, u)\big)\\
&= \sum_{x\in G_i} \sigma(x,u)\big(\lambda_x\,\mathrm{dist}^z(p_i^{\mathrm{close}}, u)\\
&\quad + (1-\lambda_x)\,\mathrm{dist}^z(p_i^{\mathrm{far}}, u)\big)\\
&\le \sum_{x\in G_i} \sigma(x,u)\big(\lambda_x\big((1+\frac{\epsilon}{2})\,\mathrm{dist}^z(x,u)\\
&\quad + (6z/\epsilon)^{z-1}\,\mathrm{dist}^z(p_i^{\mathrm{close}}, x)\big)\\
&\quad + (1-\lambda_x)\,\big((1+\frac{\epsilon}{2})\,\mathrm{dist}^z(x,u)\\
&\quad + (6z/\epsilon)^{z-1}\,\mathrm{dist}^z(p_i^{\mathrm{far}}, x)\big)\big)\\
&\le \big(1+\frac{\epsilon}{2}\big)\left(\sum_{x\in G_i}\sigma(x,u)\cdot\mathrm{dist}^z(x,u)\right)\\
&\quad + \big(\frac{6z}{\epsilon}\big)^{z-1}\cdot\sum_{x\in G_i}\sigma(x,u)\cdot\big(2\,\mathrm{dist}(p_i^{\mathrm{far}}, c)\big)^z\\
&\le \big(1+\frac{\epsilon}{2}\big)\left(\sum_{x\in G_i}\sigma(x,u)\cdot\mathrm{dist}^z(x,u)\right)\\
&\quad + \big(\frac{6z}{\epsilon}\big)^{z-1}\cdot\sum_{x\in G_i}\sigma(x,u)\cdot\big(\frac{\epsilon}{6z}\cdot\mathrm{dist}(x,u)\big)^z\\
&\le \big(1+\frac{\epsilon}{2}\big)\left(\sum_{x\in G_i}\sigma(x,u)\cdot\mathrm{dist}^z(x,u)\right)\\
&\quad + \frac{\epsilon}{2}\cdot\sum_{x\in G_i}\sigma(x,u)\cdot\mathrm{dist}^z(x,u)\\
&\le (1+\epsilon)\sum_{x\in G_i}\sigma(x,u)\cdot\mathrm{dist}^z(x,u).
\end{aligned}$$

- If $u \in C_{\mathrm{close}}$, we observe that by construction,

$$\begin{aligned}
&\pi(p_i^{\mathrm{close}}, u)\,\mathrm{dist}^z(p_i^{\mathrm{close}}, c) + \pi(p_i^{\mathrm{far}}, u)\,\mathrm{dist}^z(p_i^{\mathrm{far}}, c)\\
&= \sum_{x\in G_i}\big(\lambda_x\sigma(x,u)\,\mathrm{dist}^z(p^{\mathrm{close}}, c) +\\
&\quad (1-\lambda_x)\sigma(x,u)\,\mathrm{dist}^z(p^{\mathrm{far}}, c)\big)\\
&= \sum_{x\in G_i}\sigma(x,u)\,\mathrm{dist}^z(x,c).
\end{aligned}$$

It suffices to show that replacing c with u produces affordable error. We refer to the full version [32] for the details.

Proof of item (2): Consider $\pi : D_i \times C \to \mathbb{R}_+$ consistent with Γ. We need to construct $\sigma : G_i \times C \to \mathbb{R}_+$ so that $\mathrm{cost}_z^\sigma(G_i, C) \le (1+\epsilon)\,\mathrm{cost}_z^\pi(D_i, C)$. We find such σ by considering the following linear program,

$$\begin{aligned}
\text{minimize}\quad & \sum_{x\in G_i}\sum_{u\in C}\sigma(x,u)\cdot\mathrm{dist}^z(x,u) && \\
\text{subject to}\quad & \sigma(x,u) \ge 0 && \forall x\in G_i, u\in C,\\
& \sum_{u\in C}\sigma(x,u) = 1 && \forall x\in G_i,\\
& \sum_{x\in G_i}\sigma(x,u) = \Gamma(u) && \forall u\in C
\end{aligned}$$

The above linear programming is clearly a feasible min-cost flow problem as there must exist $\sigma : G_i \times C \to \mathbb{R}_+$ consistent with Γ. Let σ denote the optimal solution of the LP. It suffices to show $\mathrm{cost}_z^\sigma(G_i, C) \le (1+\epsilon)\,\mathrm{cost}_z^\pi(D_i, C)$. We need the following Lemma III.8.

Lemma III.8. *The following inequalities for σ hold.*

1) *For every $u \in C_{\mathrm{far}}$,*

$$\begin{aligned}
&\sum_{x\in G_i}\sigma(x,u)\,\mathrm{dist}^z(x,u)\\
&\le \Big(1+\frac{\epsilon}{2}\Big)\big(\pi(p_i^{\mathrm{close}}, u)\,\mathrm{dist}^z(p_i^{\mathrm{close}}, u)\\
&\quad + \pi(p_i^{\mathrm{far}}, u)\,\mathrm{dist}^z(p_i^{\mathrm{far}}, u)\big).
\end{aligned}$$

2) *We have the following for C_{close},*

$$\begin{aligned}
&\sum_{x\in G_i}\sigma(x, C_{\mathrm{close}})\,\mathrm{dist}^z(x,c)\\
&\le \pi(p_i^{\mathrm{close}}, C_{\mathrm{close}})\,\mathrm{dist}^z(p_i^{\mathrm{close}}, c)\\
&\quad + \pi(p_i^{\mathrm{far}}, C_{\mathrm{close}})\,\mathrm{dist}^z(p_i^{\mathrm{far}}, c) + \frac{\epsilon}{3}\cdot\mathrm{cost}_z^\pi(D_i, C).
\end{aligned}$$

3) $\mathrm{cost}_z^\sigma(G_i, C_{\mathrm{close}}) \le \big(1+\frac{\epsilon}{2}\big)\,\mathrm{cost}_z^\pi(D_i, C_{\mathrm{close}}) + \frac{\epsilon}{2}\cdot\mathrm{cost}_z^\pi(D_i, C)$.

Due to page limitation, we skip the proof of Lemma III.8 but refer to the full version [32].

Now we are ready to prove Lemma III.6. Item (1) and Item (3) of Lemma III.8 imply $\mathrm{cost}_z^\sigma(G_i, C_{\mathrm{far}}) \le \big(1+\frac{\epsilon}{2}\big)\,\mathrm{cost}_z^\pi(D_i, C_{\mathrm{far}})$ and $\mathrm{cost}_z^\sigma(G_i, C_{\mathrm{close}}) \le \big(1+\frac{\epsilon}{2}\big)\cdot\mathrm{cost}_z^\pi(D_i, C_{\mathrm{close}}) + \frac{\epsilon}{2}\cdot\mathrm{cost}_z^\pi(D_i, C)$. Combining with them, we have

$$\begin{aligned}
\mathrm{cost}_z^\sigma(G_i, C) &= \mathrm{cost}_z^\sigma(G_i, C_{\mathrm{close}}) + \mathrm{cost}_z^\sigma(G_i, C_{\mathrm{far}})\\
&\le \big(1+\frac{\epsilon}{2}\big)\cdot\mathrm{cost}_z^\pi(D_i, C_{\mathrm{close}}) + \big(1+\frac{\epsilon}{2}\big)\cdot\\
&\quad \mathrm{cost}_z^\pi(D_i, C_{\mathrm{far}}) + \frac{\epsilon}{2}\cdot\mathrm{cost}_z^\pi(D_i, C)\\
&= (1+\epsilon)\cdot\mathrm{cost}_z^\pi(D_i, C).
\end{aligned}$$

Thus we have proved Lemma III.6. ∎

Concluding the error analysis: Now we are ready to finish the error analysis for the coreset S on the unmarked groups Z. It can be simply done by combing Lemma III.5 and Lemma III.6. Recall that by construction there are at most $k(2t+1)$ colored rings and every colored group contains at least one colored rings. Since groups contain disjoint rings, we know that there are at most $k(2t+1)$ many colored groups. Let

σ denote the optimal assignment consistent with Γ from Z to C. Recall that $Z = \bigcup_{i:G_i\in\mathcal{G}} G_i$. Let Γ_i denote the assignment constraints such that $\forall u \in C, \Gamma_i(u) = \sum_{x\in G_i} \sigma(x,u)$. Let π_i denote the optimal assignment from D_i to C consistent with Γ_i. By Lemma III.5 and Lemma III.6 we have,

$$\begin{aligned}
&\mathrm{cost}_z(S,C,\Gamma)\\
\le &\sum_{i:G_i\in\mathcal{G}} \mathrm{cost}_z(D_i,C,\Gamma_i)\\
= &\sum_{i:G_i\in\mathcal{G}} \mathrm{cost}_z^{\pi_i}(D_i,C)\\
\le &(1+\epsilon)\cdot\sum_{i:G_i\in\mathcal{G}} \mathrm{cost}_z^{\sigma}(G_i,C)\\
&+k(2t+1)\cdot\frac{\epsilon}{3kt}\cdot\mathrm{cost}_z(P,c)\\
\le &(1+\epsilon)\cdot\mathrm{cost}_z(Z,C,\Gamma)+\epsilon\cdot\mathrm{cost}_z(P,c).
\end{aligned}$$

Similarly, we have that

$$\mathrm{cost}_z(Z,C,\Gamma)\le(1+\epsilon)\cdot\mathrm{cost}_z(S,C,\Gamma)+\epsilon\cdot\mathrm{cost}_z(P,c)$$

and conclude that

$$|\mathrm{cost}_z(Z,C,\Gamma)-\mathrm{cost}_z(S,C,\Gamma)|\le O(\epsilon)\cdot\mathrm{cost}_z(Z,C,\Gamma).$$

It remains to scale ϵ. □

B. Proof of Theorem III.3

Theorem III.9 (Restatement of Theorem III.3)**.** *There is an algorithm that given dataset $P\subseteq X$, center $c\in X$, $0<\epsilon<1$, computes a 2-partition $\{W,Z\}$ of P and a weighted point set $S\subseteq P$ of size 3, such that*

1) *W consists of $O(\log\frac{z}{\epsilon})$ rings $\{R_i\}_i$ where each $R_i\subseteq\mathrm{ring}(c,r_i,2r_i)$ for some $r_i>0$, and*
2) *S is an $(\epsilon,\mathrm{cost}_z(P,c))$-coreset for $(1,z)$-*CLUSTERING *on Z,*

running in time $\tilde{O}(|P|k)$.

Ring decomposition: Let $r:=\left(\frac{\mathrm{cost}_z(P,c)}{|P|}\right)^{1/z}$ denote the average cost of P. We decompose P into 3 groups.

- $P_{\mathrm{close}}=\{p\in P\mid d(p,c)<\frac{\epsilon}{6z}\cdot r\}$.
- $P_{\mathrm{far}}=\{p\in P\mid d(p,c)>\frac{120z}{\epsilon^2}r\}$.
- $P_{\mathrm{main}}=P\setminus(P_{\mathrm{close}}\bigcup P_{\mathrm{far}})$.

Define $W:=P_{\mathrm{main}}$, and it is clear that W can be covered by a union of $O(\log\frac{z}{\epsilon})$ rings of the form $\mathrm{ring}(c,a,2a)$ for some $a\ge 0$. Define $Z:=P\setminus W$, then $Z=P_{\mathrm{close}}\cup P_{\mathrm{far}}$. It remains to define an $(\epsilon,\mathrm{cost}_z(P,c))$-coreset S for Z.

Constructing coreset S: Recall that $Z=P_{\mathrm{close}}\cup P_{\mathrm{far}}$, so we construct coresets for P_{close} and P_{far} separately, and take the union of them.

- For P_{close}, we add to S a single coreset point c with weight $w(c):=|P_{\mathrm{close}}|$. Note that if one insist looking for a subset of P as coreset, one can replace c with the closet point $c_{\min}\in P_{\mathrm{close}}$ to c. It only remains to scale ϵ.
- For P_{far}, let $p^{\mathrm{far}},p^{\mathrm{close}}\in P_{\mathrm{far}}$ denote the further and closest point to c. For every $x\in P_{\mathrm{far}}$ there is a unique λ_x such that $\mathrm{dist}^z(x,c)=\lambda_x\cdot\mathrm{dist}^z(p^{\mathrm{close}},c)+(1-\lambda_x)\cdot\mathrm{dist}^z(p^{\mathrm{far}},c)$. We add p^{far} and p^{close} to S and set the weight as $w(p^{\mathrm{close}})=\sum_{x\in P_{\mathrm{far}}}\lambda_x$ and $w(p^{\mathrm{far}})=\sum_{x\in P_{\mathrm{far}}}(1-\lambda_x)$. Note that $w(p^{\mathrm{close}})+w(p^{\mathrm{far}})=|P_{\mathrm{far}}|$ and $\mathrm{cost}_z(\{p^{\mathrm{close}},p^{\mathrm{far}}\}_w,c)=\mathrm{cost}_z(P_{\mathrm{far}},c)$.

Clearly $|S|=3$ and we argue the error bound in the following.

Error analysis: Due to page limitation, we refer to Section 3.2 of the full version [32] for the error analysis.

IV. Assignment-preserving Coresets for $\mathbb{R}^d$ Rings

In this section, we show how to construct assignment-preserving coresets for k-MEDIAN. For simplicity, throughout this section, we use $\mathrm{cost}(\cdot)$ to represent $\mathrm{cost}_1(\cdot)$.

Theorem IV.1. *Let $c\in\mathbb{R}^d$, $r>0$, and $P\subseteq\mathrm{ring}(c,r,2r)$ be a dataset with $|P|=n$. Let $D\subseteq P$ be a uniform sample of size $m=\tilde{O}(\frac{k}{\epsilon^5}\log\delta^{-1})$ and re-weight D such that $\forall x\in D, w_D(x):=\frac{n}{m}$. Then with probability at least $1-\delta$, D is an assignment-preserving (ϵ,nr)-coreset for k-*MEDIAN.

We can assume the input dimension $d=\tilde{O}(\epsilon^{-2}\log k)$, by applying the iterative size reduction technique introduced in recent paper [15] which is based on a terminal version of Johnson-Lindenstrauss Lemma [52].[4] Thus it suffices to prove Theorem IV.1 with target coreset size $m=\tilde{O}(\frac{kd}{\epsilon^3})$.

The following lemma shows that it suffices to bound $|\mathrm{cost}(P,C,\Gamma)-\mathrm{cost}(D,C,\Gamma)|$, for a k-point center set C with assignment constraint Γ such that the total mass of assignment for the "far" portion of C is small.

Lemma IV.2. *Let P and D be the dataset and coreset in Theorem IV.1. Let $C\subseteq\mathbb{R}^d, |C|=k$ and $\Gamma:C\to\mathbb{R}_+$ be an assignment constraint such that $\sum_{u\in C}\Gamma(u)=n$. Let $C_{\mathrm{far}}=\{u\in C\mid\mathrm{dist}(u,c)>5kr/\epsilon^2\}$. If $\Gamma(C_{\mathrm{far}})=\sum_{u\in C_{\mathrm{far}}}\Gamma(u)>\epsilon n/k$,*

$$\left|\mathrm{cost}(P,C,\Gamma)-\mathrm{cost}(D,C,\Gamma)\right|<\epsilon\,\mathrm{cost}(P,C,\Gamma).$$

Proof: Recall that P and D are both subsets of $\mathrm{ring}(c,r,2r)$. Thus we have

$$\max\{\mathrm{cost}(P,c),\mathrm{cost}(D,c)\}\le 2nr$$

As $\Gamma(C_{\mathrm{far}})\ge\epsilon n/k$, at least $\epsilon n/k$ points in P must have connection cost at least $5kr/\epsilon^2-2r>4kr/\epsilon^2$. So we know that $\mathrm{cost}(P,C,\Gamma)\ge\epsilon n/k\cdot 4kr/\epsilon^2=4nr/\epsilon$. So by triangle inequality we know that

$$\begin{aligned}
\left|\mathrm{cost}(P,C,\Gamma)-\mathrm{cost}(D,C,\Gamma)\right| &\le \mathrm{cost}(P,c)+\mathrm{cost}(D,c)\\
&\le 4nr\\
&< \epsilon\,\mathrm{cost}(P,C,\Gamma).
\end{aligned}$$

[4] Strictly speaking, the iterative size reduction technique in [15] is designed for classical ϵ-coresets instead of our assignment-preserving (ϵ,A)-coresets for k-MEDIAN. The algorithm in [15] iteratively construct ϵ_{i+1}-coreset on ϵ_i-coreset with carefully chosen ϵ_i's. Here, since we only require the argument work for a *fixed* $A=nr$, we can apply the reduction in an identical way by iteratively constructing (ϵ_{i+1},A)-coreset on (ϵ_i,A)-coreset with the same set of ϵ_i's.

■

Lemma IV.3 ([25, Lemma 13]). *Let $C \subseteq \mathbb{R}^d, |C| = k$ and $\Gamma : C \to \mathbb{R}_+$ be an assignment constraint such that $\sum_{u\in C}\Gamma(u) = n$. Let Q be a uniform sample of $P \subseteq \mathrm{ring}(c, r, 2r)$ with size $m = \tilde{O}(\epsilon^{-3}\log\delta^{-1})$ and re-weighted by $\forall x \in Q, w_Q(x) = n/m$. Then with probability $1-\delta$,*

$$|\operatorname{cost}(Q, C, \Gamma) - \operatorname{cost}(P, C, \Gamma)| \le \varepsilon nr.$$

Lemma IV.3 is a concentration inequality for a fixed center set with capacity constraints, given by [25]. To show the coreset property holds for all possible center sets, we carefully construct a discretization $\mathcal{F}$ of centers and the assignment constraints.

Definition of $\mathcal{F}$: Let N denote an ϵr-net of the ball $B(c, \frac{5kr}{\epsilon^2})$. So $|N| \le (\frac{k}{\epsilon})^{O(d)}$. Let $t := \lceil \frac{5k^2}{\epsilon^3} \rceil$ and $H := \{i \cdot \frac{n}{t} \mid i = 0, 1, \ldots, t\}$ denote the set of multiples of $\frac{n}{t}$ that do not exceed n. Let $N \times H$ denote the set of weighted points x such that $x \in N$ and $w(x) \in H$. We define

$$\mathcal{F} := \{(C, \Gamma) \mid C \subseteq N, |C| \le k, \Gamma(C) \subseteq H, \Gamma(C) = n\}.$$

Note that $|\mathcal{F}| \le \left(\frac{k}{\epsilon}\right)^{O(kd)} \cdot (\frac{k}{\epsilon})^{O(k)}$ and thus $\log|\mathcal{F}| = \tilde{O}(kd\log\frac{k}{\epsilon})$.

In the following lemma, we show that the coreset property on $\mathcal{F}$ implies coreset property on every k-point center set $C \subseteq \mathbb{R}^d$ and assignment Λ with $\Lambda(C_{\mathrm{far}}) \le \epsilon n/k$.

Lemma IV.4. *Let (C, Λ) be a k-point center set in $\mathbb{R}^d$ with assignment constraint Λ such that $\Lambda(C) = n$ and $\Lambda(C_{\mathrm{far}}) \le \epsilon n/k$ where $C_{\mathrm{far}} = \{u \in C \mid \mathrm{dist}(u, c) > 5kr/\epsilon^2\}$, then there exists a k-point center set S with assignment constraint $(S, \Gamma) \in \mathcal{F}$ such that for every weighted set $Q \subseteq \mathrm{ring}(c, r, 2r)$ with $w_Q(Q) = n$,*

$$\operatorname{cost}(Q, C, \Lambda) \in \big(1 \pm O(\epsilon)\big)\big(\operatorname{cost}(Q, S, \Gamma) + \Delta(C)\big) \pm O(\epsilon nr)$$

where $\Delta(C) := \sum_{u\in C_{\mathrm{far}}} \mathrm{dist}(u, c) \cdot \Lambda(u)$.

Proof:

Let $C_{\mathrm{close}} := C \setminus C_{\mathrm{far}}$. For every $u \in C_{\mathrm{close}}$, let $S(u) \in N$ be a net point such that $\mathrm{dist}\big(u, S(u)\big) \le \epsilon r$. Let $u^* \in C_{\mathrm{close}}$ denote the center with largest capacity, namely, $u^* \in \mathrm{argmax}_{u\in C_{\mathrm{close}}}\Lambda(u)$. Clearly $\Lambda(u^*) \ge \frac{\Lambda(C_{\mathrm{close}})}{k} \ge \frac{n}{2k}$.

Recall that $t = \lceil \frac{5k^2}{\epsilon^3} \rceil$. For every $x \in C_{\mathrm{close}} \setminus \{x^*\}$, we let $\Gamma\big(S(x)\big) := \lfloor \frac{t\Lambda(x)}{n} \rfloor \cdot \frac{n}{t}$. We define

$$\Gamma\big(S(x^*)\big) := n - \sum_{x\in C_{\mathrm{close}}\setminus\{x^*\}} \Gamma\big(S(x)\big).$$

As all $\Gamma\big(S(x)\big)$'s are multiples of $\frac{n}{t}$ and sum up to n, we know that $(S, \Gamma) \in \mathcal{F}$. We are ready to prove the lemma.

To simplify the presentation, we observe that it suffices to assume $S = C_{\mathrm{close}}$. To see this, recall that $\forall u \in C_{\mathrm{close}}, \mathrm{dist}(u, S(u)) \le \epsilon r$, thus replacing every u with $S(u)$ produces at most ϵnr error, which is affordable.

To prove the upper bound, $\operatorname{cost}(Q, C, \Lambda) \le (1 + O(\epsilon))\big(\operatorname{cost}(Q, S, \Gamma) + \Delta(C)\big) + \epsilon nr$, it suffices to construct an assignment $\sigma' : Q \times C \to \mathbb{R}_+$ that is consistent with Λ so as

$$\operatorname{cost}^{\sigma'}(Q, C) \le (1 + \epsilon)\big(\operatorname{cost}(Q, S, \Gamma) + \Delta(C)\big) + \epsilon nr.$$

Recall that we have assumed w.l.o.g, $S = C_{\mathrm{close}}$. By construction we know that $\Lambda(u) - \Gamma(u) \in [0, \frac{n}{t}]$ for $u \in C_{\mathrm{close}} \setminus \{u^*\}$ and $\Gamma(u^*) - \Lambda(u^*) \in [0, 2\epsilon n/k]$.

To construct σ', we modify the optimal assignment corresponding to $\operatorname{cost}(Q, S, \Gamma)$. Specially, we arbitrarily disconnect $\Gamma(u^*) - \Lambda(u^*)$ mass of points from Q to u^* in $\Gamma(u^*)$ and distribute the mass to $C_{\mathrm{close}} \setminus \{u^*\}$ and C_{far} to satisfy the requirements Λ on them. We claim that by doing this, the connection cost increases by at most

$$(1 + \epsilon)\Delta(C) + \epsilon nr.$$

To see this, we first observe that sending the matching mass from $\mathrm{ring}(c, r, 2r)$ to C_{far} always costs at most $(1 + \epsilon)\Delta(C)$. On the other hand, as $C_{\mathrm{close}} \subseteq B(c, \frac{5kr}{\epsilon^2})$, and we send at most

$$\sum_{u\in C_{\mathrm{close}}\setminus\{u^*\}} \big(\Lambda(u) - \Gamma(u)\big) \le \frac{kn}{t}$$

additional mass to $C_{\mathrm{close}} \setminus \{u^*\}$, the cost in this part increases by at most $\frac{kn}{t} \cdot \frac{5kr}{\epsilon^2} \le \epsilon nr$. Thus we have proved the upper bound.

It remains to prove the lower bound, $\operatorname{cost}(Q, C, \Lambda) \ge \big(1 - O(\epsilon)\big)\big(\operatorname{cost}(Q, S, \Gamma) + \Delta(C)\big) - O(\epsilon nr)$. Let σ denote the optimal assignment for $\operatorname{cost}(Q, C, \Lambda)$, namely, $\operatorname{cost}(Q, C, \Lambda) = \operatorname{cost}^{\sigma}(Q, C)$.

Let $T_{\mathrm{far}} = \sum_{q\in Q}\sum_{u\in C_{\mathrm{far}}} \sigma(q, u)\,\mathrm{dist}(q, u)$ and $T_{\mathrm{close}} = \sum_{q\in Q}\sum_{u\in C_{\mathrm{close}}} \sigma(q, u)\,\mathrm{dist}(q, u)$. So $\operatorname{cost}(Q, C, \Lambda) = T_{\mathrm{far}} + T_{\mathrm{close}}$. We observe that $T_{\mathrm{far}} \ge (1-\epsilon)\Delta(C)$. So we just need to prove $T_{\mathrm{close}} \ge \big(1 - O(\epsilon)\big)\operatorname{cost}(Q, S, \Gamma) - O(\epsilon nr)$. It suffices to construct an assignment $\pi : Q \times S \to \mathbb{R}_+$ that is consistent with Γ and

$$\operatorname{cost}^{\pi}(Q, S) \le \big(1 + O(\epsilon)\big)T_{\mathrm{close}} + O(\epsilon nr).$$

To construct π, we modify σ. Specifically, we arbitrarily disconnect $\Lambda(u) - \Gamma(u)$ mass for every $u \in C_{\mathrm{close}} \setminus \{u^*\}$ and disconnect all mass connecting to C_{far}, and send all those mass to u^*. We note that we have re-allocated at most $2\epsilon n/k$ mass.

Let $y^* \in Q$ be a point such that $\mathrm{dist}(y^*, u^*) \le \frac{\sum_{q\in Q}\sigma(q,u^*)\,\mathrm{dist}(q,u^*)}{\Gamma(u^*)}$. Note that such y^* exists as there is always some point that contributes at most the average. Thus by triangle inequality, for every $x \in Q$,

$$\begin{aligned}
\mathrm{dist}(x, u^*) &\le \mathrm{dist}(x, c) + \mathrm{dist}(c, y^*) + \mathrm{dist}(y^*, u^*) \\
&\le 2r + 2r + \frac{\sum_{q\in Q}\sigma(q, u^*)\,\mathrm{dist}(q, u^*)}{\Gamma(u^*)} \\
&= 4r + \frac{\sum_{q\in Q}\sigma(q, u^*)\,\mathrm{dist}(q, u^*)}{\Gamma(u^*)}
\end{aligned}$$

Thus we know that the re-allocation of mass increases the cost by at most

$$\left(4r+\frac{\sum_{q\in Q}\sigma(q,u^*)\operatorname{dist}(q,u^*)}{\Gamma(u^*)}\right)\cdot\frac{2\epsilon n}{k}\leq O(\epsilon nr)+O(\epsilon\cdot T_{\mathrm{close}})$$

where we have used the fact that $\Gamma(u^*)\geq\frac{n}{2k}$ and $\sum_{q\in Q}\sigma(q,u^*)\operatorname{dist}(q,u^*)\leq T_{\mathrm{close}}$.

So we have constructed such π and thus proved the lower bound. ∎

Proof of Theorem IV.1: Replacing δ with $\frac{\delta}{|\mathcal{F}|}$ in Lemma IV.3. By union bound, Lemma IV.3 and the fact that the uniform sample has size $\tilde{O}(\frac{kd}{\epsilon^3}\cdot\log\delta^{-1})$, we know that w.p. at least $1-\delta$, the coreset property holds for all $(C,\Gamma)\in\mathcal{F}$. By Lemma IV.4, we further know that the coreset property holds for all (C,Γ) such that $\Gamma(C_{\mathrm{far}})\leq\epsilon n/k$. By Lemma IV.2, we know that the coreset property also holds for those (C,Γ) such that $\Gamma(C_{\mathrm{far}})>\epsilon n/k$. ∎

A. ϵ-Coresets for Capacitated and Fair k-MEDIAN

Combing Theorem IV.1 and Theorem III.2 with a standard argument (See Appendix A of the full version [32]), we obtain the algorithm for constructing assignment-preserving ϵ-coresets.

Theorem IV.5. *There is a near-linear time algorithm that takes a data set $P\subseteq\mathbb{R}^d$ and outputs an assignment-preserving ϵ-coreset $D\subseteq P$ with size $|D|=\tilde{O}(\frac{k^3}{\epsilon^6})$ for k-MEDIAN. In particular, this implies an ϵ-coreset for Capacitated k-MEDIAN.*

Fair clustering: Suppose $P\subseteq\mathbb{R}^d$ is a fair k-MEDIAN instance with groups $P_1,\ldots,P_l\subseteq P$. Let Δ denote the number of combinations of groups that one data point can belong to. We note that our assignment-preserving coresets (Definition II.5) matches the case $\Delta=1$. Thanks to a reduction of [27], we can use our assignment-preserving ϵ-coreset to construct ϵ-coreset for fair k-MEDIAN.

Theorem IV.6 ([27, Theorem 4.3])**.** *Suppose there is an algorithm that for any instance $P\subseteq\mathbb{R}^d$ with groups $P_1,\ldots,P_l$, constructs an assignment-preserving ϵ-coreset for k-MEDIAN with probability $1-\delta$ in time $T(|P|,\epsilon,\delta)$. Then there is an algorithm $\mathcal{A}$ that for any fair k-MEDIAN instance P such that P can be partitioned into Δ disjoint groups $P^{(1)},\ldots,P^{(\Delta)}$ where each $P^{(i)}$ consists of points that belong to the same combination of groups, $\mathcal{A}$ constructs an ϵ-coreset on P for fair k-MEDIAN with probability $1-\delta$ in time*

$$\tilde{O}\left(\sum_{i=1}^{\Delta}T\big(|P^{(i)}|,O(\epsilon),O(\delta)\big)\right).$$

We thus can prove the following theorem.

Theorem IV.7. *There is a linear algorithm that constructs an ϵ-coreset for Euclidean fair k-MEDIAN with size $\tilde{O}(\Delta\cdot\frac{k^3}{\epsilon^6})$.*

Extension to Euclidean k-MEANS: Our approach can be extended to obtain an assignment-preserving ϵ-coreset for Euclidean k-MEANS. We refer to our full version [32] for discussion of the details.

REFERENCES

[1] S. Har-Peled and S. Mazumdar, "On coresets for k-means and k-median clustering," in *STOC*. ACM, 2004, pp. 291–300.

[2] G. Frahling and C. Sohler, "Coresets in dynamic geometric data streams," in *STOC*. ACM, 2005, pp. 209–217.

[3] S. Har-Peled and A. Kushal, "Smaller coresets for k-Median and k-Means clustering," *Discret. Comput. Geom.*, vol. 37, no. 1, pp. 3–19, 2007.

[4] K. Chen, "On coresets for k-Median and k-Means clustering in metric and Euclidean spaces and their applications," *SIAM Journal on Computing*, vol. 39, no. 3, pp. 923–947, 2009.

[5] M. Langberg and L. J. Schulman, "Universal epsilon-approximators for integrals," in *SODA*. SIAM, 2010, pp. 598–607.

[6] D. Feldman and M. Langberg, "A unified framework for approximating and clustering data," in *STOC*. ACM, 2011, pp. 569–578, https://arxiv.org/abs/1106.1379.

[7] H. Fichtenberger, M. Gillé, M. Schmidt, C. Schwiegelshohn, and C. Sohler, "BICO: BIRCH meets coresets for k-means clustering," in *ESA*. Springer, 2013, pp. 481–492.

[8] D. Feldman, M. Schmidt, and C. Sohler, "Turning big data into tiny data: Constant-size coresets for k-means, pca, and projective clustering," *SIAM J. Comput.*, vol. 49, no. 3, pp. 601–657, 2020.

[9] V. Braverman, H. Lang, K. Levin, and M. Monemizadeh, "Clustering problems on sliding windows," in *SODA*. SIAM, 2016, pp. 1374–1390.

[10] V. Braverman, G. Frahling, H. Lang, C. Sohler, and L. F. Yang, "Clustering high dimensional dynamic data streams," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 70. PMLR, 2017, pp. 576–585.

[11] C. Sohler and D. P. Woodruff, "Strong coresets for k-median and subspace approximation: Goodbye dimension," in *FOCS*. IEEE Computer Society, 2018, pp. 802–813.

[12] O. Bachem, M. Lucic, and S. Lattanzi, "One-shot coresets: The case of k-clustering," in *AISTATS*, ser. Proceedings of Machine Learning Research, vol. 84. PMLR, 2018, pp. 784–792.

[13] L. Becchetti, M. Bury, V. Cohen-Addad, F. Grandoni, and C. Schwiegelshohn, "Oblivious dimension reduction for k-means: beyond subspaces and the Johnson-Lindenstrauss lemma," in *STOC*. ACM, 2019, pp. 1039–1050.

[14] L. Huang and N. K. Vishnoi, "Coresets for clustering in Euclidean spaces: importance sampling is nearly optimal," in *STOC*. ACM, 2020, pp. 1416–1429.

[15] V. Braverman, S. H. Jiang, R. Krauthgamer, and X. Wu, "Coresets for clustering in excluded-minor graphs and beyond," in *SODA*. SIAM, 2021, pp. 2679–2696.

[16] V. Cohen-Addad, K. G. Larsen, D. Saulpic, and C. Schwiegelshohn, "Towards optimal lower bounds for k-median and k-means coresets," in *STOC*. ACM, 2022, pp. 1038–1051.

[17] V. Cohen-Addad, D. Saulpic, and C. Schwiegelshohn, "A new coreset framework for clustering," in *STOC*. ACM, 2021, pp. 169–182.

[18] J. M. Phillips and W. M. Tai, "Near-optimal coresets of kernel density estimates," *Discret. Comput. Geom.*, vol. 63, no. 4, pp. 867–887, 2020.

[19] S. H. Jiang, R. Krauthgamer, J. Lou, and Y. Zhang, "Coresets for kernel clustering," *CoRR*, vol. abs/2110.02898, 2021.

[20] A. Maalouf, I. Jubran, and D. Feldman, "Fast and accurate least-mean-squares solvers," in *NeurIPS*, 2019, pp. 8305–8316.

[21] L. Huang, K. Sudhir, and N. K. Vishnoi, "Coresets for regressions with panel data," in *NeurIPS*, 2020.

[22] M. Tukan, A. Maalouf, and D. Feldman, "Coresets for near-convex functions," in *NeurIPS*, 2020.

[23] A. Munteanu, C. Schwiegelshohn, C. Sohler, and D. P. Woodruff, "On coresets for logistic regression," in *NeurIPS*, 2018, pp. 6562–6571.

[24] T. Mai, C. Musco, and A. Rao, "Coresets for classification - simplified and strengthened," in *NeurIPS*, 2021, pp. 11 643–11 654.

[25] V. Cohen-Addad and J. Li, "On the fixed-parameter tractability of capacitated clustering," in *ICALP*, ser. LIPIcs, vol. 132. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019, pp. 41:1–41:14, full version: https://arxiv.org/abs/2208.14129.

[26] S. Bandyapadhyay, F. V. Fomin, and K. Simonov, "On coresets for fair clustering in metric and Euclidean spaces and their applications," in *ICALP*, ser. LIPIcs, vol. 198. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021, pp. 23:1–23:15.

[27] L. Huang, S. H. Jiang, and N. K. Vishnoi, "Coresets for clustering with fairness constraints," in *NeurIPS*, 2019, pp. 7587–7598.

[28] N. Bousquet and S. Thomassé, "VC-dimension and Erdős-Pósa property," *Discret. Math.*, vol. 338, no. 12, pp. 2302–2317, 2015.

[29] A. Driemel, A. Nusser, J. M. Phillips, and I. Psarros, "The VC dimension of metric balls under Fréchet and Hausdorff distances," *Discret. Comput. Geom.*, vol. 66, no. 4, pp. 1351–1381, 2021.

[30] V. Braverman, S. H. Jiang, R. Krauthgamer, and X. Wu, "Coresets for clustering with missing values," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 17 360–17 372. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/90fd4f88f588ae64038134f1eeaa023f-Abstract.html

[31] B. Chazelle, *The Discrepancy Method: Randomness and Complexity*, ser. Randomness and Complexity. Cambridge University Press, 2001.

[32] V. Braverman, V. Cohen-Addad, S. H. Jiang, R. Krauthgamer, C. Schwiegelshohn, M. B. Toftrup, and X. Wu, "The power of uniform sampling for coresets," *arXiv*, vol. abs/2209.01901, 2022.

[33] Z. Izzo, S. Silwal, and S. Zhou, "Dimensionality reduction for Wasserstein barycenter," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[34] M. Buchin and D. Rohde, "Coresets for (k, ℓ)-median clustering under the Fréchet distance," in *CALDAM*, ser. Lecture Notes in Computer Science, vol. 13179. Springer, 2022, pp. 167–180.

[35] M. Schmidt, C. Schwiegelshohn, and C. Sohler, "Fair coresets and streaming algorithms for fair k-means," in *WAOA*, ser. Lecture Notes in Computer Science, vol. 11926. Springer, 2019, pp. 232–251.

[36] A. Munteanu and C. Schwiegelshohn, "Coresets-methods and history: A theoreticians design pattern for approximation and streaming algorithms," *Künstliche Intell.*, vol. 32, no. 1, pp. 37–53, 2018.

[37] D. Feldman, *Core-Sets: Updated Survey*. Springer, 2020, pp. 23–44.

[38] L. Huang, S. H. Jiang, J. Li, and X. Wu, "Epsilon-coresets for clustering (with outliers) in doubling metrics," in *FOCS*. IEEE Computer Society, 2018, pp. 814–825.

[39] D. N. Baker, V. Braverman, L. Huang, S. H. Jiang, R. Krauthgamer, and X. Wu, "Coresets for clustering in graphs of bounded treewidth," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 569–579. [Online]. Available: http://proceedings.mlr.press/v119/baker20a.html

[40] C. Schwiegelshohn and O. A. Sheikh-Omar, "An empirical evaluation of k-means coresets," *CoRR*, vol. abs/2207.00966 (to appear at ESA 2022), 2022.

[41] V. Braverman, S. H. Jiang, R. Krauthgamer, and X. Wu, "Coresets for ordered weighted clustering," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 97. PMLR, 2019, pp. 744–753. [Online]. Available: http://proceedings.mlr.press/v97/braverman19a.html

[42] H. Ding, H. Yu, and Z. Wang, "Greedy strategy works for k-center clustering with outliers and coreset construction," in *ESA*, 2019.

[43] O. Bachem, M. Lucic, and A. Krause, "Scalable k-means clustering via lightweight coresets," in *KDD*. ACM, 2018, pp. 1119–1127.

[44] D. Feldman, Z. Kfir, and X. Wu, "Coresets for gaussian mixture models of any shape," *CoRR*, vol. abs/1906.04895, 2019.

[45] L. Huang, K. Sudhir, and N. Vishnoi, "Coresets for time series clustering," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 22 849–22 862. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/c115ba9e04ab27fbbb664f932112246d-Abstract.html

[46] I. Jubran, M. Tukan, A. Maalouf, and D. Feldman, "Sets clustering," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 4994–5005.

[47] Y. Marom and D. Feldman, "k-means clustering of lines for big data," in *NeurIPS*, 2019, pp. 12 797–12 806.

[48] K. Makarychev, Y. Makarychev, and I. P. Razenshteyn, "Performance of johnson-lindenstrauss transform for k-means and k-medians clustering," in *STOC*. ACM, 2019, pp. 1027–1038.

[49] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii, "Fair clustering through fairlets," in *NIPS*, 2017, pp. 5029–5037.

[50] S. K. Bera, D. Chakrabarty, N. Flores, and M. Negahbani, "Fair algorithms for clustering," in *NeurIPS*, 2019, pp. 4955–4966.

[51] M. Böhm, A. Fazzone, S. Leonardi, C. Menghini, and C. Schwiegelshohn, "Algorithms for fair k-clustering with multiple protected attributes," *Oper. Res. Lett.*, vol. 49, no. 5, pp. 787–789, 2021.

[52] S. Narayanan and J. Nelson, "Optimal terminal dimensionality reduction in Euclidean space," in *STOC*. ACM, 2019, pp. 1064–1069.