Coresets for Clustering in Euclidean Spaces: Importance Sampling Is Nearly Optimal*

Lingxiao Huang
Yale University
USA

Nisheeth K. Vishnoi
Yale University
USA

ABSTRACT

Given a collection of n points in $\mathbb{R}^d$, the goal of the (k, z)-Clustering problem is to find a subset of k "centers" that minimizes the sum of the z-th powers of the Euclidean distance of each point to the closest center. Special cases of the (k, z)-Clustering problem include the k-Median and k-Means problems. Our main result is a unified two-stage importance sampling framework that constructs an ε-coreset for the (k, z)-Clustering problem. Compared to the results for (k, z)-Clustering in [Feldman and Langberg, STOC 2011], our framework saves a $\varepsilon^2 d$ factor in the coreset size. Compared to the results for (k, z)-Clustering in [Sohler and Woodruff, FOCS 2018], our framework saves a $\mathrm{poly}(k)$ factor in the coreset size and avoids the $\exp(k/\varepsilon)$ term in the construction time. Specifically, our coreset for k-Median ($z = 1$) has size $\tilde{O}(\varepsilon^{-4}k)$ which, when compared to the result in [Sohler and Woodruff, FOCS 2018], saves a k factor in the coreset size. Our algorithmic results rely on a new dimension reduction technique that connects two well-known shape fitting problems: subspace approximation and clustering, and may be of independent interest. We also provide a size lower bound of $\Omega\left(k \cdot \min\left\{2^{z/20}, d\right\}\right)$ for a 0.01-coreset for (k, z)-Clustering, which has a linear dependence of size on k and an exponential dependence on z that matches our algorithmic results.

CCS CONCEPTS

• **Theory of computation → Sketching and sampling**; **Facility location and clustering**.

KEYWORDS

Coresets, Clustering, Importance sampling, Dimension reduction, k-means, k-median

ACM Reference Format:
Lingxiao Huang and Nisheeth K. Vishnoi. 2020. Coresets for Clustering in Euclidean Spaces: Importance Sampling Is Nearly Optimal. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing (STOC '20), June 22–26, 2020, Chicago, IL, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3357713.3384296

*Due to space restrictions, some proofs and details are omitted from this version of the paper and available in the full version of this paper on the arXiv.

STOC '20, June 22–26, 2020, Chicago, IL, USA

ACM ISBN 978-1-4503-6979-4/20/06…$15.00
https://doi.org/10.1145/3357713.3384296

1 INTRODUCTION

We study the problem of constructing coresets for (k, z)-Clustering in Euclidean space $\mathbb{R}^d$ where $z \geq 1$ is a constant.

(k, z)-Clustering in $\mathbb{R}^d$. The input is a collection of n points $X \subseteq \mathbb{R}^d$, and the goal is to find a set $C \subseteq \mathbb{R}^d$ of k points, called a *center set*, that minimizes the objective function

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

where, throughout, d^z denotes the Euclidean distance raised to power $z \geq 1$, and $d(x, C) := \min\left\{d(x, c) = \sqrt{\sum_{i \in [d]} (x_i - c_i)^2} : c \in C\right\}$.

This formulation captures classical problems, including the k-Median problem (where $z = 1$) and the k-Means problem (where $z = 2$). Moreover, this formulation can be generalized to weighted point sets where each point $x \in X$ has a weight $u(x)$ and the goal is to compute a k-center set $C \subset \mathbb{R}^d$ that minimizes

$$\mathrm{cost}_z(X, C) := \sum_{x \in X} u(x) \cdot d^z(x, C).$$

The (k, z)-Clustering problem is an essential tool in data analysis and arises in areas such as approximation algorithms, unsupervised learning, and computational geometry [2, 12, 26, 34]. Due to its importance, several approximation algorithms for this clustering problem have been proposed [3, 5, 15, 29].

Coresets. In recent years, a powerful data-reduction technique – coresets – has been used to find approximately optimal clustering in large datasets [18, 19, 21]. Roughly speaking, a coreset is a "compact" summary of the data set, represented by a collection of weighted points, that approximates the clustering objective for every possible choice of center set. Let $\mathcal{C}$ denote the collection of all ordered subsets (repetitions allowed) of $\mathbb{R}^d$ of size k (k-center sets).

Definition 1.1 (**Coreset [18, 25]**). Given a collection $X \subseteq \mathbb{R}^d$ of n weighted points and $\varepsilon \in (0, 1)$, an ε-coreset for (k, z)-Clustering is a subset $S \subseteq \mathbb{R}^d$ with weights $w : S \to \mathbb{R}_{\geq 0}$ such that for any k-center set $C \in \mathcal{C}$, the (k, z)-Clustering objective with respect to C is ε-approximately preserved, i.e.,

$$\sum_{x \in S} w(x) \cdot d^z(x, C) \in (1 \pm \varepsilon) \cdot \mathrm{cost}_z(X, C).$$

Coresets have been extensively studied in Euclidean spaces. For (k, z)-Clustering in $\mathbb{R}^d$, Feldman and Langberg [18] construct an ε-coreset $O(\varepsilon^{-2z} kd \log(k/\varepsilon))$ based on an importance sampling framework. Specifically, the dependence on dimension d can be removed for k-Median [33] and k-Means [6, 7, 19, 33].

However, it was unknown whether we can similarly remove the dependence on d for general (k, z)-Clustering for arbitrary constant $z \geq 1$. Also, the coreset for k-Median [33] has a quadratic dependence of size on k, that does not match the size lower bound. Moreover, the constructions of [6, 33] need a generalized notion of coreset instead of Definition 1.1, which may increase the difficulty of applying existing clustering algorithms on coresets. Thus, it is an important problem to understand whether there exists a unified framework that constructs coresets satisfying Definition 1.1 for general (k, z)-Clustering, with a linear dependence on k, and no dependence on d.

1.1 Our Contributions

The main contribution of this paper is a unified framework that constructs ε-coresets for (k, z)-Clustering of size $\tilde{O}(\min\{\varepsilon^{-2z-2}k, 2^{2z}\varepsilon^{-4}k^2\})$ and a nearly matching lower bound. We first propose a two-stage importance sampling framework that constructs coresets for (k, z)-Clustering (constant $z \geq 1$); summarized in the following theorem.

Theorem 1.2 (**Informal, see Theorem 5.1**). *There exists a randomized algorithm that, given a dataset X of n points in $\mathbb{R}^d$, $\varepsilon \in (0, 1/2)$, constant $z \geq 1$ and integer $k \geq 1$, constructs an ε-coreset of size $\tilde{O}(\min\{\varepsilon^{-2z-2}k, 2^{2z}\varepsilon^{-4}k^2\})$ for (k, z)-Clustering, and runs in time $\tilde{O}(ndk)$.*

We compare our results with existing coreset results for (k, z)-Clustering in Table 1. This is the first result that constructs an ε-coreset for (k, z)-Clustering whose size is independent of d and near-linearly dependent of k. Note that if ε, z are constants, this result saves a d factor compared to prior results [18, 35] and matches the size lower bound. Compared to the result in [33], our coreset saves a $\mathrm{poly}(k)$ factor in the size and can be constructed in polynomial time – avoids the expontential term ($\exp(k/\varepsilon)$) and the dependence of $1/\varepsilon$ ($n\,\mathrm{poly}(k/\varepsilon)$) in the construction time of [33]. Specifically, for k-Median, our result saves a k factor compared to [33]. Our construction applies a unified two-stage importance sampling framework (Algorithm 1). Compared to existing approaches [6, 33] that require to apply projection methods, our construction is simple to implement. Also note that our coreset satisfies (the standard) Definition 1.1, instead of the one that requires offsets as in recent results [6, 33]. Consequently, we can directly combine existing clustering algorithms with our coresets to estimate (k, z)-Clustering objectives. It is an interesting open problem to investigate whether one-stage importance sampling could produce coresets with a comparable size.

We also extend Theorem 1.2 to ℓ_p-metrics whose distance function is $d_p(x, y) = \left(\sum_{i\in[d]} |x_i - y_i|^p\right)^{1/p}$ $(x, y \in \mathbb{R}^d)$ instead of the Euclidean distance in Equation (1); see the following corollary.

Corollary 1.3 (**Informal, see Corollary 5.18**). *There exists a randomized algorithm that, given a dataset X of n points in $\mathbb{R}^d$, $1 \leq p < 2$, $\varepsilon \in (0, 1/2)$, constant $z \geq 1$ and integer $k \geq 1$, constructs an ε-coreset of size $\tilde{O}(\min\{\varepsilon^{-4z-2}k, 2^{4z}\varepsilon^{-4}k^2\})$ for (k, z)-Clustering with ℓ_p-metric, and runs in time $\tilde{O}(ndk)$.*

The main idea is that for $1 \leq p < 2$, there exists an isometric embedding from ℓ_p to ℓ_2^2 [24]. Using this idea, we can reduce the problem of constructing an ε-coreset for (k, z)-Clustering with ℓ_p-metric (Definition 5.17) to constructing an $O(\varepsilon)$-coreset for $(k, 2z)$-Clustering with ℓ_2-metric (Definition 1.1). It is interesting to investigate whether the above corollary can be extended to all constant $p \geq 1$.

We also provide a matching size lower bound (Theorem 1.4).

Theorem 1.4 (**Size lower bound**). *For every $z > 0$ and integers $d, k \geq 1$, there exists a point set X in the Euclidean space $\mathbb{R}^d$ such that any 0.01-coreset for (k, z)-Clustering over X has size $\Omega\left(k \cdot \min\left\{2^{z/20}, d\right\}\right)$.*

To the best of our knowledge, this is the first result that shows that the coreset size for (k, z)-Clustering should exponentially depend on z. However, tight dependence of size on the parameter ε is still unknown. The proof can be found in the full version of this paper.

For the algorithmic result (Theorem 1.2), our main technical contribution is a new dimension reduction technique which reduces the dimension of the space of k-center sets to $\mathrm{poly}(k/\varepsilon)$. Towards this, we develop two new notions: representativeness and weak-coreset (Section 2). Given a collection X in $\mathbb{R}^d$, we first divide all k-center sets into sub-groups by defining "equivalence classes" of k-center sets. Our equivalence classes are induced by some subspace $\Gamma \subseteq \mathbb{R}^d$, in which each class is induced by a k-center set in Γ (Definition 2.1). Based on these equivalence classes, we define a representativeness property (Definition 2.3), i.e., (k, z)-Clustering objectives over X with respect to all k-center sets in an equivalence class are roughly the same. We show that X satisfies the representativeness property with respect to certain $\mathrm{poly}(k/\varepsilon)$-dimensional subspace Γ constructed by [33, Algorithm 1] (Observation 5.13). Moreover, we present a sufficient condition for any weighted subset $S \subseteq X$ such that the representativeness property also holds for S, i.e., (k, z)-Clustering objectives over S with respect to all k-center sets in an equivalence class are roughly the same (Observation 5.14). The sufficient condition roughly requires that S is a "weak-coreset" for (k, z)-subspace approximation over X (Definition 2.5). To satisfy this requirement, we only need the size of S to be $\mathrm{poly}(k/\varepsilon)$ (Theorem 5.10). This enables us to only approximately preserve (k, z)-Clustering objectives in the low-dimensional subspace Γ instead of $\mathbb{R}^d$, which leads to an ε-coreset D for (k, z)-Clustering of size $\mathrm{poly}(k/\varepsilon)$.

Compared to the Feldman-Langberg framework [18], we successfully remove the dependence in coreset size of d (Theorem 5.2). Moreover, by a well-known dimension reduction approach, called "terminal embedding" (Definition 3.2), the dimension of D can be further reduced to $O(\varepsilon^{-2}\log(k/\varepsilon))$. Thus, we can further reduce the coreset size to $\tilde{O}(\varepsilon^{-2z-2}k)$ by applying an importance sampling framework over D (Theorem 5.3). Overall, our dimension reduction technique connects two well-known shape fitting problems: subspace approximation and clustering, and leads to a unified two-stage importance sampling framework for (k, z)-Clustering that removes the dependence of coreset size on d. The geometric observations and notions, including equivalence classes and representativeness property, may be of independent interest.

To establish a nearly tight size lower bound (Theorem 1.4), we construct an instance X of size $\Omega\left(k \cdot \min\left\{2^{z/20}, d\right\}\right)$ such that for any point $x \in X$, there exists a k-center set C_x which satisfies the

[1]The paper did not present this result directly. But their approach can be easily generalized to (k, z)-Clustering.

Table 1: Summary of coreset size for (k,z)-Clustering.

	Reference	Objective	Coreset Size
Upper bounds	[18]	(k,z)-Clustering	$O\left(\varepsilon^{-2z} k\boldsymbol{d}\log(k/\varepsilon)\right)$
	[35]	(k,z)-Clustering	$O\left(2^{2z}\varepsilon^{-2} k\boldsymbol{d}\log(k/\varepsilon)\right)$
	[33]	(k,z)-Clustering	$\mathrm{poly}(k/\varepsilon^z)$ [1]
	This paper	(k,z)-Clustering	$O\left(\min\left\{\varepsilon^{-2z-2}, 2^{2z}\varepsilon^{-4}k\right\}\cdot k\log k\log(k/\varepsilon)\right)$
	[18]	k-Median	$O\left(\varepsilon^{-2}k\boldsymbol{d}\log k\right)$
	[33]	k-Median	$O\left(\varepsilon^{-4}k^2\log k\right)$
	This paper	k-Median	$O\left(\varepsilon^{-4}\cdot k\log k\log(k/\varepsilon)\right)$
	[7]	k-Means	$O\left(\varepsilon^{-3}k^2\log(k/\varepsilon)\right)$
	[6]	k-Means	$O\left(\varepsilon^{-6}k\log^2(k/\varepsilon)\log(1/\varepsilon)\right)$
	This paper	k-Means	$O\left(\varepsilon^{-6}\cdot k\log k\log(k/\varepsilon)\right)$
Lower bounds	[9]	k-Median ($d=1$)	$\Omega\left(\varepsilon^{-1/2}k\right)$
	This paper	(k,z)-Clustering	$\Omega(2^{z/20}k)$

condition that the clustering objective $d^z(x, C_x) \approx \mathrm{cost}_z(X, C_x)$. To gain some intuition about the construction, if $k = 1$ and $d \approx 2^{z/20}$ we let $X = \{e_1, -e_1, \ldots, e_d, -e_d\}$ and observe that $d^z(e_i, -e_i) = 2^z \approx \mathrm{cost}_z(X, -e_i)$ for any $i \in [d]$. Suppose S is a 0.01-coreset and $S \subseteq X$. Then S satisfies that $\mathrm{cost}_z(S, -e_i) \approx \mathrm{cost}_z(X, -e_i)$ for each $i \in [d]$. Intuitively, each e_i should be included in S and, hence, $|S| = O(2^{z/20}k)$. The technical difficulty is that points in a coreset can come from $\mathbb{R}^d \setminus X$. We show that even though e_i may not be included in S, a point close to e_i must be included. This results in a matching lower bound $|X|$.

1.2 Other Related Works

Har-Peled and Mazumdar [21] constructed the first coreset for both k-Median and k-Means, however the size of their coresets depended exponentially on d. Subsequently, Chen [10] improved dependence on the dimension to be polynomial for both k-Median and k-Means. For k-Means, Feldman et al. [19] designed coresets of size independent of d, which was improved by [7] to be $\tilde{O}(\varepsilon^{-3}k^2)$. Recently, the dependence on k has been improved to near-linear by [6]. For k-Median, Sohler and Woodruff [33] show how to remove the dependence on d. Recently, coreset for generalized clustering objective receives attention from the research community, for example, Braverman et al. [9] obtain simultaneous coreset for Ordered k-Median. Coresets for the fair version of k-Median and k-Means have also been investigated [23, 31]. For another special case $z = \infty$, which is the k-Center clustering, an ε-coreset of size $O(\varepsilon^{-d+1}k)$ can be constructed in near-linear time [1, 20]. This size has been proved to be tight for k-Center [D. Feldman, private communication and [9]]. For general (k,z)-Clustering (constant $z \geq 1$), Feldman and Langberg [18] construct an ε-coreset of size $\tilde{O}(\varepsilon^{-2z}kd)$, and recently this result has been generalized to doubling metrics [22].

For general metrics, an ε-coreset for the (k,z)-Clustering problem of size $O(\varepsilon^{-2z}k\log n)$ can be constructed in time $\tilde{O}(nk)$ [18], and the $\log n$ factor is unavoidable [8]. Specifically, for k-means clustering, Braverman et al. [7] show a construction of size $O(\varepsilon^{-2}k\log k\log n)$.

2 OUR NOTIONS OF REPRESENTATIVENESS AND WEAK-CORESET

In this section, we propose new definitions that are important for our dimension reduction technique. Recall that $\mathcal{C}$ denotes the collection of all ordered subsets (repetition allowed) of $\mathbb{R}^d$ of size k (k-center sets). We first define equivalence classes of k-center sets that partition $\mathcal{C}$.

Equivalence classes of k-center sets and representativeness property. Given a subspace $\Gamma \subsetneq \mathbb{R}^d$, we denote by $\pi_\Gamma : \mathbb{R}^d \to \Gamma$ the projection function from $\mathbb{R}^d$ to Γ, i.e., for any $x \in X$,

$$\pi_\Gamma(x) := \arg\min_{y\in\Gamma} d(x,y).$$

If Γ is clear from the context, we may denote π_Γ by π.

Definition 2.1 (**Equivalence relations and equivalence classes induced by a subspace**). Given a subspace $\Gamma \subsetneq \mathbb{R}^d$, we define an equivalence relation $\sim_\Gamma$ between k-center sets as follows: for two k-center sets $C = (c_1, \ldots, c_k)$ and $C' = \left(c'_1, \ldots, c'_k\right)$, we say $C \sim_\Gamma C'$ if and only if for all $i \in [k]$,

$$\pi_\Gamma(c_i) = \pi_\Gamma(c'_i) \quad \text{and} \quad d(c_i, \pi_\Gamma(c_i)) = d(c'_i, \pi_\Gamma(c'_i)).$$

Let Γ' be obtained from Γ by appending an arbitrary dimension $u \in \mathbb{R}^d$ that is orthogonal to Γ.[1] Let $\mathcal{C}_\Gamma$ denote the collection of k-center sets C all of whose points lie in Γ', i.e.,

$$\mathcal{C}_\Gamma := \left\{C = (c_1, \ldots, c_k) \in \mathcal{C} : c_i \in \Gamma' \ \forall i \in [k]\right\}.$$

The relation $\sim_\Gamma$ also induces equivalence classes of k-center sets $\left\{\Delta_C^{(\Gamma)} : C \in \mathcal{C}_\Gamma\right\}$ where each $\Delta_C^{(\Gamma)} := \{C' \in \mathcal{C} : C \sim_\Gamma C'\}$.

In the following, we explain why $\left\{\Delta_C^{(\Gamma)} : C \in \mathcal{C}_\Gamma\right\}$ are equivalence classes induced by Γ. Given a subspace $\Gamma \subsetneq \mathbb{R}^d$, we define a mapping $\phi : \mathbb{R}^d \to \mathbb{R}^{d+1}$ where for any point $x \in \mathbb{R}^d$,

$$\phi(x) := (\pi_\Gamma(x), d(x, \pi_\Gamma(x))).$$

Let $A := \{\phi(C) : C \in \mathcal{C}\}$ denote the collection of images of ϕ with respect to $\mathcal{C}$. By the discussion above, each image $a \in A$ naturally

[1] Here, $\Gamma' = \{ax + bu \mid x \in \Gamma, a \in \mathbb{R}, b \in \mathbb{R}\}$.

corresponds to an equivalence class $\Delta_a^{(\Gamma)} := \{C \in \mathcal{C} : \phi(C) = a\}$. Note that for any $x \in \mathbb{R}^d$, there must exist a point $x' \in \Gamma'$ such that

$$\pi_\Gamma(x) = \pi_\Gamma(x') \quad \text{and} \quad d(x, \pi_\Gamma(x)) = d(x', \pi_\Gamma(x')),$$

since Γ' includes an additional dimension that is orthogonal to Γ. By the above equations, we have that $\phi(x) = \phi(x')$. Thus, for any $a \in A$ and $C' \in \mathcal{C}$ satisfying that $\phi(C') = a$, there must exist a k-center set $C \in \mathcal{C}_\Gamma$ such that $\phi(C) = \phi(C') = a$. This implies that $\Delta_a^{(\Gamma)} = \Delta_C^{(\Gamma)}$. Since $\left\{\Delta_a^{(\Gamma)} : a \in A\right\}$ are equivalence classes induced by Γ, we conclude that $\left\{\Delta_C^{(\Gamma)} : C \in \mathcal{C}_\Gamma\right\}$ are also equivalence classes induced by Γ.

Example 2.2. Let $d = 3$, $k = 1$ and Γ denote the the first axis. Let $x = (x_1, x_2, x_3)$ and $x' = (x_1', x_2', x_3')$ be two 1-center sets in $\mathbb{R}^3$. By Definition 2.1, we have that $c \sim_\Gamma c'$ if and only if

$$x_1 = x_1', \quad \text{and} \quad \sqrt{x_2^2 + x_3^2} = \sqrt{(x_2')^2 + (x_3')^2}.$$

i.e., their first coordinates are the same and their distances to the first axis are the same. Hence, all points with the same x-coordinate form an equivalence class.

Without loss of generality, let Γ' be the plane spanned by the first and the second axes. Then each center $x = (x_1, x_2, 0) \in \Gamma'$ corresponds to an equivalence class

$$\Delta_x^{(\Gamma)} = \left\{x = (x_1, x_2, x_3) \in \mathbb{R}^d : x_1 = x_1, \sqrt{x_2^2 + x_3^2} = |x_2|\right\}.$$

We also propose the following definition that indicates that (k, z)-Clustering objectives within an equivalence class are almost the same.

Definition 2.3 (**Representativeness property**). Given a weighted point set $A \subseteq \mathbb{R}^d$ together with a weight function $w : A \to \mathbb{R}_{\geq 0}$, $\varepsilon \in (0, 1)$ and a subspace Γ, we say that A satisfies the ε-representativeness property with respect to Γ if for any equivalence class Δ_C^Γ and any two k-center sets $C_1, C_2 \in \Delta_C^\Gamma$, the following property holds:

$$\mathrm{cost}_z(A, C_1) \in (1 \pm \varepsilon) \cdot \mathrm{cost}_z(A, C_2).$$

It follows from the definition above that, if both the given dataset X and a weighted point set S satisfy the representativeness property with respect to a certain low-dimensional subspace Γ, then S is a coreset if S approximately preserves all (k, z)-Clustering objectives with respect to k-center sets $C \in \mathcal{C}_\Gamma$. This observation enables us to only consider k-center sets in Γ instead of $\mathbb{R}^d$, which is the key for our dimension reduction technique.

(k, z)-subspace approximation. Let $\mathcal{P}$ denote the collection of all j-flats in $\mathbb{R}^d$ with $j \leq k$, i.e., all subspaces in $\mathbb{R}^d$ of dimension at most k.

Definition 2.4 (**(k, z)-subspace approximation problem**). Given a dataset X in $\mathbb{R}^d$, $z > 0$ and an integer $k \geq 1$, the goal of the (k, z)-subspace approximation problem is to find a subspace $P^\star \in \mathcal{P}$ that minimizes $\sum_{x \in X} d^z(x, P)$ over all $P \in \mathcal{P}$.

Subspace approximation is a well-studied shape fitting problem. Several prior works have focussed on designing approximation algorithms [4, 11, 16] and constructing coresets [14, 18, 33] for subspace approximation. In this paper, we propose the following version of "weak-coreset" for subspace approximation.

Definition 2.5 (**Weak-coreset for (k, z)-subspace approximation**). Given a collection $X \subseteq \mathbb{R}^d$ of n weighted points and an $\varepsilon \in (0, 1)$, an ε-weak-coreset for (k, z)-subspace approximation is a subset $S \subseteq \mathbb{R}^d$ with weights $w : S \to \mathbb{R}_{\geq 0}$ such that

$$\min_{P \in \mathcal{P}} \sum_{x \in S} w(x) \cdot d^z(x, P) \in (1 \pm \varepsilon) \cdot \min_{P \in \mathcal{P}} \sum_{x \in X} d^z(x, P). \qquad (2)$$

Note that a weak-coreset may not approximately preserve all subspace approximation objectives like Definition 1.1. However, we can approximately compute the minimum (k, z)-subspace approximation objective via a weak-coreset. We remark that Definition 2.5 is a different version of a notion in [18], in which a weak-coreset S satisfies that any $(1 + \varepsilon)$-approximate solution for (k, z)-subspace approximation over S is an $(1 + O(\varepsilon))$-approximate solution over X.

3 PRIOR RESULTS ON IMPORTANCE SAMPLING AND TERMINAL EMBEDDINGS

We first introduce two general frameworks for coresets for (k, z)-Clustering that will be used in our algorithm. Both were proposed by Feldman and Langberg [18], and the second one is an improved version by [7]; summarized as follows.

Theorem 3.1 (**Feldman-Langberg Framework [7, 18]**). *Let $\varepsilon, \delta \in (0, 1/2)$, $k \geq 1$ and constant $z \geq 1$. Let $X \subseteq \mathbb{R}^d$ denote a weighted point set of n points together with a weight function $u : X \to \mathbb{R}_{\geq 0}$. Let $C^\star \in \mathcal{C}$ denote a k-center set that is an $O(1)$-approximate solution for (k, z)-Clustering over X. We have two importance sampling frameworks as follows.*

(1) ([18, Theorem 15.5]) Suppose $\sigma : X \to \mathbb{R}_{\geq 0}$ is a sensitivity function satisfying that for any $x \in X$,

$$\sigma(x) \geq \frac{u(x) \cdot d^z(x, C^\star)}{\sum_{y \in X} u(y) \cdot d^z(y, C^\star)},$$

and $\mathcal{G} := \sum_{x \in X} \sigma(x)$. Let $D \subseteq X$ be constructed by taking

$$O\left(\varepsilon^{-2z}(dk \log k + \log(1/\delta))\right)$$

samples, where each sample $x \in X$ is selected with probability $\frac{\sigma(x)}{\mathcal{G}}$ and has weight $w(x) := \frac{\mathcal{G}}{|D| \cdot \sigma(x)}$. For each $c \in C^\star$, let $w(c) := (1 + 10\varepsilon) \cdot \sum_{x \in X_c} u(x) - \sum_{x \in D \cap X_c} w(x)$ where X_c is the collection of points in X whose closest point in $C^\star$ is c. Then, with probability at least $1 - \delta$, $S := D \cup C^\star$ is an ε-coreset for (k, z)-Clustering over X.[2]

(2) ([7, Theorem 5.2]) Suppose $\sigma : X \to \mathbb{R}_{\geq 0}$ is a sensitivity function satisfying that for any $x \in X$,

$$\sigma(x) \geq \sup_{C \in \mathcal{C}} \frac{u(x) \cdot d^z(x, C)}{\sum_{y \in X} u(y) \cdot d^z(y, C)},$$

and $\mathcal{G} := \sum_{x \in X} \sigma(x)$. Let $S \subseteq X$ be constructed by taking

$$O\left(\varepsilon^{-2}\mathcal{G}(dk \log \mathcal{G} + \log(1/\delta))\right)$$

[2]This conclusion is based on [18, Theorem 15.5]. We discuss the theorem in Section 6.

*samples, where each sample $x \in X$ is selected with probability $\frac{\sigma(x)}{\mathcal{G}}$ and has weight $w(x) := \frac{\mathcal{G}}{|S|\cdot\sigma(x)}$. Then, with probability at least $1-\delta$, S is an ε-coreset for (k,z)-*Clustering *over X.*

The Feldman-Langberg framework applies one-staged importance sampling and is easy to implement. We only need to compute an approximate solution $C^\star$, sensitivities $\sigma(x)$, and take samples with probability proportional to $\sigma(x)$. For the first framework, note that we can further guarantee that the output S is a subset of X by adding an additional constraint that $C^\star \subset X$.[3] For the second framework, the total sensitivity $\mathcal{G}$ is shown to be $O(2^{2z}k)$ by [35] and, hence, the resulting coreset size is quadratic in k by plugging the value of $\mathcal{G} = O(2^{2z}k)$ in the size bound $O\left(\varepsilon^{-2}\mathcal{G}(dk\log\mathcal{G} + \log(1/\delta))\right)$ in Theorem 3.1.

Terminal embedding. We introduce the definition of terminal embedding that is useful for our dimension reduction result. Roughly speaking, a terminal embedding projects a point set $X \subseteq \mathbb{R}^d$ to a low-dimensional space such that all pairwise distances between X and $\mathbb{R}^d$ are approximately preserved.

*Definition 3.2 (***Terminal embedding***).* Let $\varepsilon \in (0,1)$ and $X \subseteq \mathbb{R}^d$ be a collection of n points. A mapping $f : \mathbb{R}^d \to \mathbb{R}^m$ is called a terminal embedding of X if for any $x \in X$ and $y \in \mathbb{R}^d$,

$$d(x,y) \le d(f(x), f(y)) \le (1+\varepsilon)d(x,y).$$

Note that a terminal embedding must be a one-to-one mapping over X by definition. The following is a recent result on terminal embedding.

Theorem 3.3 (**Small terminal embeddings [30]**). *Let $\varepsilon \in (0,1)$ and $X \subseteq \mathbb{R}^d$ be a collection of n points. There exists a terminal embedding with a target dimension $m = O(\varepsilon^{-2}\log n)$.*

It follows from this theorem that, if $|X| = \mathrm{poly}(k/\varepsilon)$, then there is a terminal embedding of target dimension $O\left(\varepsilon^{-2}\log(k/\varepsilon)\right)$, which is independent of d. This is useful for analyzing the correctness of the second stage of our framework.

4 TECHNICAL OVERVIEW

In this section, we will show how to prove our algorithmic result (Theorem 1.2) and how to achieve a nearly matching size lower bound (Theorem 1.4). We first introduce a new dimension reduction technique that is useful for our algorithmic result.

Dimension reduction. For simplicity, we take the k-Median problem ($z=1$) as an example and let $X \subseteq \mathbb{R}^d$ be the input point set. By Theorem 3.1, there exists an ε-coreset for k-Median of size $O(\varepsilon^{-2}dk\log k)$. However, the coreset size contains a factor d and our goal is to construct a coreset that does not depend on d.

To this end, a commonly used approach is called *dimension reduction* [7, 13, 19, 33]. Roughly speaking, we would like to show that it suffices to only consider all k-center sets in some low-dimensional space instead of $\mathbb{R}^d$, and that enables us to remove the dependence on d in the coreset size. One can try the dimension reduction approach proposed by [27], however, this can be shown not to work. We explain the details in the full version.

Representativeness property can allow to remove the dependence on d. Our second attempt is motivated by the work of [33] and projects X to a low-dimensional subspace such that all k-Median objectives can be estimated by the projections of X. For a subspace $\Gamma \subsetneq \mathbb{R}^d$, recall that Γ' is obtained from Γ by appending an arbitrary dimension in $\mathbb{R}^d$ that is orthogonal to Γ. Also recall that $\mathcal{C}_\Gamma$ denotes the collection of k-center sets $C \subset \Gamma'$, i.e.,

$$\mathcal{C}_\Gamma := \left\{C = (c_1,\ldots,c_k) : c_i \in \Gamma'\ \forall i, C \in \mathcal{C}\right\}$$

Now suppose we have a subspace $\Gamma \subsetneq \mathbb{R}^d$ of dimension $\mathrm{poly}(k/\varepsilon)$ and a weighted point set D, together with a weight function $u : D \to \mathbb{R}_{\ge 0}$, that approximately preserves all k-Median objectives with respect to k-center sets in $\mathcal{C}_\Gamma$, i.e., for any k-center set $C \in \mathcal{C}_\Gamma$,

$$\sum_{x\in D} u(x)\cdot d(x,C) \in (1\pm\varepsilon)\cdot \mathrm{cost}_1(X,C). \tag{3}$$

By Theorem 3.1, the size of D can be upper bounded by $\mathrm{poly}(k/\varepsilon)$. If we can show that Inequality (3) holds for all k-center sets in $\mathcal{C}$, then we obtain an ε-coreset D of size independent of d. Moreover, by the result on terminal embeddings (Theorem 5.3), we can further reduce the dimension to $O\left(\varepsilon^{-2}\log(k/\varepsilon)\right)$. This observation leads to an ε-coreset S for k-Median of size $\tilde{O}(\varepsilon^{-4}k)$ by applying the first framework in Theorem 3.1 to D. In the following, we show how to construct sufficient conditions for Γ and D such that Inequality (3) holds for all k-center sets in $\mathcal{C}$, which leads to Theorem 5.2.

Our key idea is to construct a Γ such that both X and D satisfy the representativeness property with respect to Γ. Recall that an equivalence class $\Delta^\Gamma_{C'}$ for a center set $C' \in \mathcal{C}_\Gamma$ is the collection of center sets $C \in \mathcal{C}$ such that $C \sim_\Gamma C'$. Now suppose we have that Inequality (3) holds for all k-center sets in $\mathcal{C}_\Gamma$. Consequently, for any $C \in \Delta^\Gamma_{C'}$,

$$\begin{aligned}\sum_{x\in D} u(x)\cdot d(x,C) \approx \sum_{x\in D} u(x)\cdot d(x,C') \\ \approx \mathrm{cost}_1(X,C') \approx \mathrm{cost}_1(X,C).\end{aligned}$$

Then D is an ε-coreset for k-Median in $\mathbb{R}^d$. Hence, we focus on proving the representativeness property.

Given a subset $A \subseteq \mathbb{R}^d$, let $\mathrm{Conv}(A)$ denote the convex hull of A. Recall that OPT_1 denotes the optimal k-Median objective over X. Sohler and Woodruff [33, Algorithm 1] show how to construct a subspace Γ of dimension $O(\varepsilon^{-2}k)$ such that for any k-center set $C \in \mathcal{C}$, letting π_C denote the projection from X to $\mathrm{Conv}(\Gamma \cup C)$,

$$\sum_{x\in X} d(x,\pi_\Gamma(x)) - d(x,\pi_C(x)) = O(\varepsilon^2)\cdot \mathsf{OPT}_1. \tag{4}$$

By [33] (summarized in Lemma 5.4), this implies that for any k-center set $C \in \mathcal{C}$,

$$\sum_{x\in X}\left(d^2(\pi_\Gamma(x),C) + d^2(x,\pi_\Gamma(x))\right)^{1/2} \in (1\pm\varepsilon)\cdot \mathrm{cost}_1(x,C), \tag{5}$$

i.e., we can use Γ to approximately preserve all k-Median objectives. Note that for any two center sets C_1, C_2 in an equivalence class, we

[3] Suppose $C^\star$ is a k-center set that is an $O(1)$-approximate solution for (k,z)-Clustering over X. Then $\left\{\arg\min_{x\in X} c_i^\star : i \in [n]\right\}$ is a k-center set that is an $O(2^z)$-approximate solution for (k,z)-Clustering over X. Hence, we can add this additional constraint.

have

$$\sum_{x\in X}\left(d^2(\pi_\Gamma(x),C_1)+d^2(x,\pi_\Gamma(x))\right)^{1/2}$$
$$=\sum_{x\in X}\left(d^2(\pi_\Gamma(x),C_2)+d^2(x,\pi_\Gamma(x))\right)^{1/2}.$$

Combining the above equation with Inequality (5), we conclude that X satisfies the representativeness property with respect to Γ (Observation 5.13). The technical difficulty is to show that D also satisfies the representativeness property with respect to Γ.

A failed attempt. By a similar construction as [33, Algorithm 1], we can construct a subspace Γ for any given D that satisfies for any k-center set $C\in\mathcal{C}$,

$$\sum_{x\in D}u(x)\cdot\left(d^2(\pi_\Gamma(x),C)+d^2(x,\pi_\Gamma(x))\right)^{1/2} \in(1\pm\varepsilon)\cdot\mathrm{cost}_1(D,C), \tag{6}$$

i.e., we can also use Γ to approximately preserve k-Median objectives over the weighted point set D. Similarly, we conclude that D also satisfies the representativeness property with respect to Γ. Then assuming Inequality (3) holds for all k-center sets $C\subset\Gamma'$, D is an $O(\varepsilon)$-coreset for k-Median and we are done. However, this assumption only has a guarantee when Γ is independent of the choice of D. Thus, our task is to construct a subspace Γ satisfying Inequalities (4) and (6), and meanwhile independent on the choice of D.

Weak coreset for subspace approximation implies the representativeness property for D. By [33, Algorithm 1], we construct a subspace Γ of dimension $\mathrm{poly}(k/\varepsilon)$ that satisfies Inequality (4) and only depends on X, which implies that Γ is independent on the choice of D. The remaining task is to find a sufficient condition for D such that Inequality (6) holds. Again by [33] (summarized in Lemma 5.4), Inequality (6) holds if for any k-center set $C\in\mathcal{C}$,

$$\sum_{x\in D}u(x)\cdot(d(x,\pi_\Gamma(x))-d(x,\pi_C(x)))=O(\varepsilon^2)\cdot\mathsf{OPT}_1. \tag{7}$$

Since Inequality (4) holds, for the above inequality it suffices to guarantee that

$$\sum_{x\in D}u(x)\cdot(d(x,\pi_\Gamma(x))-d(x,\pi_C(x))) \approx\sum_{x\in X}d(x,\pi_\Gamma(x))-d(x,\pi_C(x)). \tag{8}$$

Our first assumption is that the total projection distance to Γ is approximately preserved by D:

$$\sum_{x\in D}u(x)\cdot d(x,\pi_\Gamma(x))\approx\sum_{x\in X}d(x,\pi_\Gamma(x)).$$

Then for Inequality (8), it suffices to ensure that

$$\min_{C\in\mathcal{C}}\sum_{x\in D}u(x)\cdot d(x,\pi_C(x))\approx\min_{C\in\mathcal{C}}\sum_{x\in X}d(x,\pi_C(x)). \tag{9}$$

Interestingly, this relates to the subspace approximation problem by regarding X as a point set in space $\Gamma^\perp$. From this viewpoint, Inequality (9) can be reduced to ensuring that D is a weak-coreset for $(k,1)$-subspace approximation over X (Definition 2.5). Overall, roughly, a weak-coreset D for $(k,1)$-subspace approximation satisfies the representativeness property with respect to Γ.

We still need to verify that the size of a weak-coreset D can be independent of dimension d. By applying [32], we know that D is a weak-coreset if D approximately preserves all $(k,1)$-subspace approximation objectives in the collection $\mathcal{P}'$ of all k-flats spanned by at most $\tilde{O}(\varepsilon^{-1}k^2)$ points from X (Lemma 5.11), i.e., D is an ε-coreset for $(k,1)$-subspace approximation over X with respect to $\mathcal{P}'$. By [35, Theorem 4], it follows that the coreset size $|D|$ only depends on k,ε and the "function dimension" of $\mathcal{P}'$[4] Since the function dimension of $\mathcal{P}'$ is $\mathrm{poly}(k/\varepsilon)$ by [35, Theorem 4], the size of a weak-coreset $|D|$ can be also upper bounded by $\mathrm{poly}(k/\varepsilon)$, which is independent of d (Theorem 5.10).

Overall, we develop a two-stage importance sampling framework that constructs an ε-coreset for k-Median of size $\tilde{O}(\varepsilon^{-4}k)$ (Algorithm 1). In the first stage, we construct a weighted point set D of size $\mathrm{poly}(k/\varepsilon)$ that is a coreset for k-Median in Γ' and an ε-weak-coreset for $(k,1)$-subspace approximation in $\Gamma^\perp$. By the discussion above, D is an ε-coreset for k-Median in $\mathbb{R}^d$ of size $\mathrm{poly}(k/\varepsilon)$ (Theorem 5.2). In the second stage, we further construct an ε-coreset S over D of size $\tilde{O}(\varepsilon^{-4}k)$ using the result on terminal embeddings (Theorem 5.3), where S is also an $O(\varepsilon)$-coreset for k-Median over X (Theorem 1.2).

For general $z>1$, the proof is similar to $z=1$. The difference is that in the first stage of Algorithm 1, we need the weighted point set D to be a coreset for (k,z)-Clustering in Γ' and an ε-weak-coreset for (k,z)-subspace approximation in $\Gamma^\perp$. To achieve this, the number of samples $|D|$ should be $z^{O(z)}\cdot\mathrm{poly}(k/\varepsilon)$, which is still independent of d. Again, using the result on terminal embeddings, we can construct a coreset for (k,z)-Clustering of size $\tilde{O}(\varepsilon^{-2z-2}k)$ (Theorem 1.2).

Size lower bounds. We also provide a nearly matching size lower bound of coresets for (k,z)-Clustering (Theorem 1.4), by constructing a point set $X\subset\mathbb{R}^d$ in $\mathbb{R}^d$ such that any 0.01-coreset for (k,z)-Clustering over X has size $\Omega\left(k\cdot\min\left\{2^{z/20},d\right\}\right)$. The main idea is to ensure that for any point $x\in X$, there exists a k-center set C_x satisfying that the clustering objective $d^z(x,C_x)\approx\mathrm{cost}_z(X,C_x)$. Intuitively, we need to include a close point for each $x\in X$ in any coreset such that $\mathrm{cost}_z(X,C_x)$ can be approximately preserved.

We discuss a simple case that $k=1$ and $d\approx 2^{z/20}$, and show how to construct such a bad instance X. The general case that $k\geq 1$ can be proved by making k copies of X in which all copies are far away from each other. We let $X=\{e_1,-e_1,\ldots,e_d,-e_d\}$ and observe that for each $e_i\in X$,

$$d^z(e_i,-e_i)\approx\mathrm{cost}_z(X,-e_i).$$

Suppose S is a 0.01-coreset for (k,z)-Clustering together with weights $w(x)$. If we restrict that $S\subseteq X$, we can conclude that $S=X$ which implies that $|S|=2^{d+1}=\Omega(2^{z/20})$. Intuitively, if there exists $e_i\in X\setminus S$, then it is unlikely that the (k,z)-Clustering objective with respect to center $-e_i$ can be approximately preserved by S. The obstacle is that points in a coreset can come from outside of X.

[4]Since this paper only uses function dimension as a black box, we do not present the definition. We refer interested readers to [18, Definition 6.4] or [7, Definition 4.5] for concrete definitions.

We divide $\mathbb{R}^d$ into $|X|$ Voronoi cells $(P_x)_x$ induced by X, where each P_x is the collection of points in $\mathbb{R}^d$ whose closest point in X is x. If $|S| < |X|$, there must exist a Voronoi cell that does not contain points in S, say P_{e_1} without loss of generality. The key idea is to show that S can not approximately preserve the (k,z)-Clustering objectives with respect to center 0 and center $-e_1$ simultaneously, which implies that the size of S should be at least $|X| = 2d$.

Let H denote the unit l_2-ball centered at the origin. On the one hand, we can show that the contribution $\sum_{x\in S\cap H} w(x)\cdot d^z(x,-e_1)$ is tiny compared to $\sum_{x\in X} d^z(x,-e_1) \approx 2^z$, since each point $x \in H \setminus P_{e_1}$ satisfies that

$$d^z(x,-e_1) \le 2^{0.9z}$$

and the total weights $\sum_{x\in S\cap H} w(x)$ can be shown at most $2^{0.05z}$. On the other hand, for point $x \in S \setminus (H \cup P_{e_1})$, we can verify that

$$\frac{d^z(x,-e_1)}{d^z(x,0)} \le 2^{0.9z}.$$

This gap is much smaller compared to $\frac{\sum_{x\in X} d^z(x,-e_1)}{\sum_{x\in X} d^z(x,0)} \approx 2^{0.95z}$, which implies that S can not approximately preserve the (k,z)-Clustering objectives with respect to center 0 and center $-e_1$ simultaneously. This indicates that any 0.01-coreset for k-Median over X has size at least $|X| = 2d$, which proves Theorem 1.4.

5 OUR ALGORITHM AND MAIN THEOREM

Let $X \subseteq \mathbb{R}^d$ denote a collection of n points in $\mathbb{R}^d$. Let OPT_z denote the optimal (k,z)-Clustering objective over X. The main theorem of this paper is the following.

Theorem 5.1 (**Coreset for (k,z)-Clustering; near-linear size in k**). *There exists a randomized algorithm that, for a given dataset X of n points in $\mathbb{R}^d$, $\varepsilon,\delta \in (0,0.5)$, constant $z \ge 1$ and integer $k \ge 1$, with probability at least $1-\delta$, constructs an ε-coreset for (k,z)-Clustering of size*

$$O\left(\min\left\{\varepsilon^{-2z-2}, 2^{2z}\varepsilon^{-4}k\right\} k\log k \log\frac{k}{\varepsilon\delta}\right)$$

and runs in time

$$O\left(ndk + nd\log(n/\delta) + k^2\log^2 n + \log^2(1/\delta)\log^2 n\right).$$

We propose a unified two-stage importance sampling framework for (k,z)-Clustering; see Algorithm 1. Note that Algorithm 1 provides an ε-coreset of size $\tilde{O}(\varepsilon^{-2z-2}k)$. By applying another importance sampling approach in the second stage (see discussion in Remark 5.9), we can achieve an ε-coreset of size $\tilde{O}(2^{2z}\varepsilon^{-4}k^2)$. This gives the size bound stated in Theorem 5.1.

5.1 Proof of the Main Theorem 5.1

The proof of Theorem 5.1 relies on the following two theorems. The first theorem shows that the first stage of Algorithm 1 constructs a coreset for (k,z)-Clustering of size $N_1 = \mathrm{poly}(k, 1/\varepsilon)$ with high probability. The second one is a size reduction theorem that constructs a coreset of size $N_2 = \tilde{O}(\varepsilon^{-2z-2}k)$ based on the output of the first stage.

Theorem 5.2 (**First stage of Algorithm 1**). *For every dataset X of n points in $\mathbb{R}^d$, $\varepsilon,\delta \in (0,0.5)$, constant $z \ge 1$ and integer $k \ge 1$, with probability at least $1-\delta/2$, the first stage of Algorithm 1 constructs an ε-coreset D_1 for (k,z)-Clustering, and runs in time*

$$O\left(ndk + nd\log(n/\delta) + k^2\log^2 n + \log^2(1/\delta)\log^2 n\right).$$

Theorem 5.3 (**Second stage of Algorithm 1**). *Let $X \subseteq \mathbb{R}^d$ be a collection of n points, $\varepsilon,\delta \in (0,1/2)$, constant $z \ge 1$ and integer $k \ge 1$. Suppose the first stage of Algorithm 1 constructs an ε-coreset D_1 for (k,z)-Clustering. Then with probability at least $1-\delta/2$, the second stage of Algorithm 1 outputs an $O(\varepsilon)$-coreset S for (k,z)-Clustering, and runs in time $O(nd)$.*

The proof of Theorem 5.2 can be found in Section 5.2 and the proof of Theorem 5.3 can be found in Section 5.3. Observe that Theorem 5.1 is a direct corollary of Theorems 5.2 and 5.3.

Proof of Theorem 5.1. The overall running time is a direct corollary of Theorems 5.2 and 5.3.

By Theorem 5.2, with probability at least $1-\delta/2$, (D_1,u) is an ε-coreset for (k,z)-Clustering of size N_1. Then by Theorem 5.3,

Algorithm 1: Coreset construction for (k,z)-Clustering

Input: A dataset X of n points in $\mathbb{R}^d$, $\varepsilon,\delta \in (0,1/2)$, constant $z \ge 1$ and an integer $k \ge 1$.
Output: A point set $S \subseteq \mathbb{R}^d$ together with a weight function $w : S \to \mathbb{R}_{\ge 0}$.

/* The first importance sampling stage */

$N_1 \leftarrow O\left((168z)^{10z}\varepsilon^{-5z-15}k^5\log\frac{k}{\delta}\right)$;

Compute a k-center set $C^\star \subseteq \mathbb{R}^d$ as an α-approximation of the (k,z)-Clustering problem over X ($\alpha = O(1)$);

For each $x \in X$, compute $c^\star(x)$ to be the closest point to x in $C^\star$ (ties are broken arbitrarily). For each $c \in C^\star$, denote X_c to be the set of points $x \in X$ with $c^\star(x) = c$;

For each $x \in X$, let $\sigma_1(x) \leftarrow 2^{2z+2}\alpha^2\left(\frac{d^z(x,c^\star(x))}{\mathrm{cost}_z(X,C^\star)} + \frac{1}{|X_{c^\star(x)}|}\right)$;

Pick a non-uniform random sample D_1 of N_1 points from X, where each $x \in X$ is selected with probability $\frac{\sigma_1(x)}{\sum_{y\in X}\sigma_1(y)}$. For each $x \in D_1$, let $u(x) \leftarrow \frac{\sum_{y\in X}\sigma_1(y)}{|D_1|\cdot\sigma_1(x)}$;

/* The second importance sampling stage */

$N_2 \leftarrow O\left(\varepsilon^{-2z-2}k\log k\log\frac{k}{\varepsilon\delta}\right)$;

For each $c \in C^\star$, compute D_c to be the set of points in D_1 whose closest point in $C^\star$ is c (ties are broken arbitrarily);

For each $x \in D_1$, let $\sigma_2(x) \leftarrow \frac{u(x)\cdot d^z(x,C^\star)}{\sum_{y\in D_1} u(y)\cdot d^z(y,C^\star)}$;

Pick a non-uniform random sample D_2 of N_2 points from X, where each $x \in X$ is selected with probability $\frac{\sigma_2(x)}{\sum_{y\in D_1}\sigma_2(y)}$. For each $x \in D_2$, let $w(x) \leftarrow \frac{\sum_{y\in D_1}\sigma_2(y)}{|D_2|\cdot\sigma_2(x)}$. For each $c \in C^\star$, let $w(c) \leftarrow (1+10\varepsilon)\sum_{x\in D_c} u(x) - \sum_{x\in D_2\cap D_c} w(x)$;

$S \leftarrow D_2 \cup C^\star$;

Output (S,w);

with probability at least $1-\delta$, the output S is an ε-coreset which completes the proof. □

5.2 Analyzing the First Stage of Algorithm 1

In this section, we prove Theorem 5.2 that provides a theoretical guarantee for the first stage. Given a subset $A \subseteq \mathbb{R}^d$, let $\mathrm{Conv}(A)$ denote the convex hull of A. For preparation, we have the following lemmas. The first lemma shows that there exists a subspace Γ such that the projections of X to Γ can be used to estimate all (k,z)-Clustering objectives. Note that [33] only considers unweighted point sets, but it can be easily generalized to weighted point sets.

Lemma 5.4 (**Restatement of [33, Lemma 6 and Theorem 10]**). *Let $A \subseteq \mathbb{R}^d$ be a weighted point set together with a weight function $w: A \to \mathbb{R}_{\geq 0}$. Let OPT_z be the optimal weighted (k,z)-clustering objective over A, $\varepsilon \in (0,0.5)$, constant $z \geq 1$, and Γ be a subspace of $\mathbb{R}^d$. Suppose for any k-center set $C \in \mathcal{C}$, we have*

$$\sum_{x\in A} w(x)\cdot\left(d^z(x,\pi(x)) - d^z(x,\pi_C(x))\right) \leq \frac{\varepsilon^{z+3}}{3\cdot(84z)^{2z}}\cdot \mathsf{OPT}_z, \quad (10)$$

where π and π_C denote the projection from A to Γ and $\mathrm{Conv}(\Gamma\cup C)$ respectively. Then for any k-center set $C \in \mathcal{C}$, the following inequality holds

$$\sum_{x\in A} w(x)\cdot\left(d^2(\pi(x),C) + d^2(x,\pi(x))\right)^{z/2} \in (1\pm\varepsilon)\cdot\sum_{x\in A} w(x)\cdot d^z(x,C). \quad (11)$$

we have the following lemma showing that D_1 is a coreset in an arbitrary low dimensional subspace with high probability.

Lemma 5.5 (**D_1 is a coreset in a low-dimensional subspace**). *Suppose Γ is an arbitrary subspace of dimension $O\left((84z)^{2z}\varepsilon^{-z-3}k\right)$ in $\mathbb{R}^d$. Let Γ' be obtained from Γ by appending an arbitrary dimension in $\mathbb{R}^d$ that is orthogonal to Γ. With probability at least $1-\delta/4$, for any k-center set $C \subset \Gamma'$ we have*

$$\sum_{x\in D_1} u(x)\cdot d^z(x,C) \in (1\pm\varepsilon)\cdot \mathrm{cost}_z(X,C).$$

Proof. Let $\varepsilon' = \frac{\varepsilon^{z+3}}{6\cdot(84z)^{2z}}$ and $m = O(k/\varepsilon')$ be the dimension of Γ. We first have the following claim.

Claim 5.6. *For any $x \in X$,*

$$\sup_{C\in\mathcal{C}} \frac{d^z(x,C)}{\mathrm{cost}_z(X,C)} \leq 2^z\cdot\frac{d^z(x,c^\star(x))}{\mathsf{OPT}_z} + 2^{2z+1}\alpha\cdot\frac{1}{|X_{c^\star(x)}|}.$$

Proof. The proof idea is similar to [35, Theorem 7]. We first note that d^z satisfies the relaxed triangle inequality, i.e., for any $x, x', x'' \in \mathbb{R}^d$, we have

$$d^z(x,x'') \leq 2^z\cdot\left(d^z(x,x') + d^z(x',x'')\right). \quad (12)$$

Then for any $x \in X$ and any k-center set $C \in \mathcal{C}$, we have

$$\begin{aligned} & d^z(x,C) \\ \leq\ & 2^z\left(d^z(x,c^\star(x)) + d^z(c^\star(x),C)\right) \\ & \text{(Ineq. (12))} \\ \leq\ & 2^z\cdot d^z(x,c^\star(x)) + \frac{2^z}{|X_{c^\star(x)}|}\cdot\sum_{y\in X} d^z(c^\star(y),C) \\ & \text{(Defn. of } X_{c^\star(x)}) \\ \leq\ & 2^z\cdot d^z(x,c^\star(x)) \\ & + \frac{2^z}{|X_{c^\star(x)}|}\cdot\sum_{y\in X} 2^z\cdot\left(d^z(c^\star(x),x) + d^z(x,C)\right) \\ & \text{(Ineq. (12))} \\ =\ & 2^z\cdot d^z(x,c^\star(x)) \\ & + \frac{2^{2z}}{|X_{c^\star(x)}|}\cdot\left(\mathrm{cost}_z(X,C^\star) + \mathrm{cost}_z(X,C)\right). \\ & \text{(Defn. of } \mathrm{cost}_z) \end{aligned} \quad (13)$$

Thus, we have that

$$\begin{aligned} & \frac{d^z(x,C)}{\mathrm{cost}_z(X,C)} \\ \leq\ & 2^z\cdot\frac{d^z(x,c^\star(x))}{\mathrm{cost}_z(X,C)} + \frac{2^{2z}}{|X_{c^\star(x)}|}\cdot\left(1+\frac{\mathrm{cost}_z(X,C^\star)}{\mathrm{cost}_z(X,C)}\right) \\ & \text{(Ineq. (13))} \\ \leq\ & 2^z\cdot\frac{d^z(x,c^\star(x))}{\mathsf{OPT}_z} + \frac{2^{2z}}{|X_{c^\star(x)}|}\cdot(1+\alpha) \\ & \text{(Defn. of } C^\star) \\ \leq\ & 2^z\cdot\frac{d^z(x,c^\star(x))}{\mathsf{OPT}_z} + 2^{2z+1}\alpha\cdot\frac{1}{|X_{c^\star(x)}|}, \end{aligned}$$

which implies the claim since C is arbitrary. □

Then for any $x \in X$,

$$\begin{aligned} & \sup_{C\in\mathcal{C}}\frac{d^z(x,C)}{\mathrm{cost}_z(X,C)} \\ \leq\ & 2^z\cdot\frac{d^z(x,c^\star(x))}{\mathsf{OPT}_z} + 2^{2z+1}\alpha\cdot\frac{1}{|X_{c^\star(x)}|} \\ & \text{(Claim 5.6)} \\ \leq\ & 2^z\alpha\cdot\frac{d^z(x,c^\star(x))}{\mathrm{cost}_z(X,C^\star)} + 2^{2z+1}\alpha\cdot\frac{1}{|X_{c^\star(x)}|} \\ & \text{(Defn. of } C^\star) \\ \leq\ & \sigma_1(x). \end{aligned}$$

Also note that

$$\begin{aligned} & \sum_{x\in X}\sigma_1(x) \\ =\ & 2^{2z+2}\alpha^2\cdot\sum_{x\in X}\left(\frac{d^z(x,c^\star(x))}{\mathrm{cost}_z(X,C^\star)} + \frac{1}{|X_{c^\star(x)}|}\right) \\ \leq\ & 2^{2z+2}\alpha^2\cdot(1+k) \qquad (|C^\star| = k) \\ \leq\ & 2^{2z+3}\alpha^2 k. \end{aligned}$$

Thus, we have

$$N_1 = \Omega\left(\frac{\sum_{x\in X}\sigma_1(x)}{(\varepsilon')^2}\cdot(km\log(\sum_{x\in X}\sigma_1(x)) + \log\frac{1}{\delta})\right),$$

Then by Theorem 3.1, we complete the proof. □

Next, we give the main technical lemma. It indicates that if a subspace Γ satisfies Inequality (10), then clustering objectives over D_1 can be estimated by the projections of D_1 to Γ, similar to Inequality (11). The proof can be found in Section 5.4.

Lemma 5.7 (**Γ preserves (k,z)-Clustering objectives over D_1**). *Suppose Γ is a subspace of $\mathbb{R}^d$ satisfying that $C^\star \subset \Gamma$ and for any k-center set $C \in \mathcal{C}$,*

$$\sum_{x\in X}\left(d^z(x,\pi(x)) - d^z(x,\pi_C(x))\right) = \frac{\varepsilon^{z+3}}{3\cdot(84z)^{2z}}\cdot \mathsf{OPT}_z,$$

where π and π_C denote the projection from X to Γ and $\mathrm{Conv}(\Gamma\cup C)$ respectively. Let D_1 together with u be the weighted point set obtained by the first stage of Algorithm 1. With probability at least $1-\delta/4$, for any k-center set $C\in\mathcal{C}$,

$$\sum_{x\in D_1} u(x)\cdot d^z(x,C) \in (1\pm 2\varepsilon)\cdot\sum_{x\in D_1} u(x)\cdot\left(d^2(\pi(x),C) + d^2(x,\pi(x))\right)^{z/2}.$$

By the above lemmas, we are ready to prove Theorem 5.2. The proof idea is to first show the existence of a subspace Γ that satisfies Inequality (10) for $A = X$ in Lemma 5.4. By Lemma 5.4, we can prove that X satisfies the representativeness property with respect to Γ. Similarly, we can also show that D_1 satisfies the representativeness property with respect to Γ by Lemma 5.7. Recall that Γ' is obtained from Γ by appending an arbitrary dimension in $\mathbb{R}^d$ that is orthogonal to Γ. Finally, by Lemma 5.5, D_1 is an ε-coreset for (k,z)-Clustering in Γ'. Combining with these properties, we can conclude that D_1 is an ε-coreset for (k,z)-Clustering in $\mathbb{R}^d$, which proves the theorem.

Proof of Theorem 5.2. Let $\varepsilon' = \frac{\varepsilon^{z+3}}{6\cdot(84z)^{2z}}$. By [33, Algorithm 1], there exists a subspace Γ satisfying the following properties:

P1. Γ satisfies that for any $C\in\mathcal{C}$,

$$\sum_{x\in X}\left(d^z(x,\pi(x)) - d^z(x,\pi_C(x))\right) \le \varepsilon'\cdot\mathsf{OPT}_z/2,$$

where π and π_C denote the projection from X to Γ and $\mathrm{Conv}(\Gamma\cup C)$ respectively.

P2. Γ is of dimension $O(k/\varepsilon')$.

Since the dimension of Γ' is $O(k/\varepsilon')$, we have that with probability at least $1-\delta/4$, D_1 is an ε-coreset for (k,z)-Clustering in Γ' by Lemma 5.5. It means that for any k-center set $C\subset\Gamma'$,

$$\sum_{x\in D_1} u(x)\cdot d^z(x,C) \in (1\pm\varepsilon)\cdot\mathrm{cost}_z(X,C). \tag{14}$$

Moreover, with probability at least $1-\delta/4$, for any k-center set $C\in\mathcal{C}$,

$$\sum_{x\in X}\left(d^2(\pi(x),C) + d^2(x,\pi(x))\right)^{z/2} \in (1\pm\varepsilon)\cdot\mathrm{cost}_z(X,C) \tag{15}$$

by Lemma 5.4, and

$$\sum_{x\in D_1} u(x)\cdot d^z(x,C) \in (1\pm\varepsilon)\sum_{x\in D_1} u(x)\cdot\left(d^2(\pi(x),C) + d^2(x,\pi(x))\right)^{z/2}. \tag{16}$$

by Lemma 5.7. Then we have the following claim.

Claim 5.8. *Both X and D_1 satisfy the 2ε-representativeness property with respect to Γ.*

Proof. For any equivalence class Δ_C^Γ and any two k-center sets $C_1, C_2\in\Delta_C^\Gamma$, we have

$$\begin{aligned}
&\mathrm{cost}_z(X,C_1)\\
\in\ & (1\pm\varepsilon)\cdot\sum_{x\in X}\left(d^2(\pi(x),C_1) + d^2(x,\pi(x))\right)^{z/2}\\
&(\text{Ineq. }(15))\\
\in\ & (1\pm\varepsilon)\cdot\sum_{x\in X}\left(d^2(\pi(x),C_2) + d^2(x,\pi(x))\right)^{z/2}\\
&(\text{Defition } 2.1)\\
\in\ & (1\pm 2\varepsilon)\cdot\mathrm{cost}_z(X,C_2).\\
&(\text{Ineq. }(15))
\end{aligned}$$

By the same argument, Inequality (16) implies that D_1 also satisfies the ε-representativeness property. This completes the proof. □

Now we are ready to prove the theorem. Given a k-center set $C\in\mathcal{C}$, suppose C belongs to the equivalence class $\Delta_{C'}^\Gamma$ for some $C'\in\Gamma'$.

$$\begin{aligned}
&\sum_{x\in D_1} u(x)\cdot d^z(x,C)\\
\in\ & (1\pm 2\varepsilon)\cdot\sum_{x\in D_1} u(x)\cdot d^z(x,C') && (\text{Claim } 5.8)\\
\in\ & (1\pm 2\varepsilon)\cdot\mathrm{cost}_z(X,C') && (\text{Ineq. }(14))\\
\in\ & (1\pm 4\varepsilon)\cdot\mathrm{cost}_z(X,C), && (\text{Claim } 5.8)
\end{aligned}$$

which completes the proof of correctness by letting $\varepsilon' = O(\varepsilon)$.

For the running time, it costs $O\left(ndk + nd\log(n/\delta) + k^2\log^2 n + \log^2(1/\delta)\log^2 n\right)$ time to construct a $2^{O(z)}$-approximate solution $C^\star$ [28],[5] $O(ndk)$ time to compute all X_c and $\sigma_1(x)$, and $O(N_1) = O(n)$ time to construct D_1. Hence, we prove for the overall running time. □

5.3 Analyzing the Second Stage of Algorithm 1

Next, we prove the reduction theorem (Theorem 5.3) that provides a theoretical guarantee for the second stage. The main idea is to apply the result on terminal embeddings such that the dimension is further reduced to $O\left(\varepsilon^{-2}\log(k/\varepsilon)\right)$.

Proof. It costs $O(N_1 dk) = O(ndk)$ time to compute all D_c and $\sigma_2(x)$, and $O(N_2) = O(n)$ time to construct S. Hence, we only need

[5] [28] only discuss k-median, but their construction can be easily generalized to (k,z)-Clustering by the relaxed triangle inequality of d^z.

to focus on the correctness. Since we suppose that the ouput D_1 of the first stage is an ε-coreset over X, we have that

$$\sum_{x\in D_1} u(x)\cdot d^z(x,C^\star) \le (1+\varepsilon)\cdot \mathrm{cost}_z(X,C^\star) \le 2\alpha\cdot \mathrm{OPT}_z.$$

Hence, $C^\star$ is also an $O(1)$-approximation of the (k,z)-Clustering problem over D_1. Let $f:\mathbb{R}^d\to\mathbb{R}^m$ be a terminal embedding of D_1 where $m = O(z^2\varepsilon^{-2}\log N_1)$. By Theorem 3.3, we have that for any $x\in D_1$ and $y\in\mathbb{R}^d$,

$$\begin{aligned} d^z(x,y) \le\ & d^z\left(f(x),f(y)\right) \\ \le\ & \left(1+\frac{\varepsilon}{10z}\right)^z\cdot d^z(x,y) \le (1+\varepsilon)\cdot d^z(x,y). \end{aligned} \tag{17}$$

Hence, we have for any set $A\subseteq\mathbb{R}^d$,

$$d^z(x,A) \le d^z\left(f(x),f(A)\right) \le (1+\varepsilon)\cdot d^z(x,A). \tag{18}$$

Then $f(C^\star)$ is an $O(1)$-approximation of the (k,z)-Clustering problem over the weighted point set $f(D_1)$ with weights $u(x)$. By Theorem 3.1, with probability at least $1-\delta/2$, $f(S)$ together with weights $w(x)$ is an ε-coreset for (k,z)-Clustering over $(f(D_1),u)$ since $N_2 = \Omega\left(\varepsilon^{-2z}(km\log k+\log(1/\delta))\right)$. Then it suffices to prove that S together with weights $w(x)$ is an $O(\varepsilon)$-coreset for (k,z)-Clustering over X. For any k-center set $C\in\mathcal{C}$, we have the following

P1. $\sum_{x\in D_1} u(x)\cdot d^z(x,C)\in(1\pm\varepsilon)\cdot\mathrm{cost}_z(X,C)$ by the assumption of the theorem.
P2. $\sum_{x\in S} w(x)\cdot d^z\left(f(x),f(C)\right)\in(1\pm\varepsilon)\cdot\sum_{x\in D_1}u(x)\cdot d^z\left(f(x),f(C)\right)$ by the definition of S and Theorem 3.1.
P3. $\sum_{x\in D_1} u(x)\cdot d^z\left(f(x),f(C)\right)\in(1\pm\varepsilon)\cdot\sum_{x\in D_1}u(x)\cdot d^z(x,C)$ by Inequality (18).
P4. $\sum_{x\in S} w(x)\cdot d^z(x,C)\in(1\pm\varepsilon)\cdot\sum_{x\in S}w(x)\cdot d^z\left(f(x),f(C)\right)$ by Inequality (18).

Combining the above properties, we have that

$$\begin{aligned} &\sum_{x\in S} w(x)\cdot d^z(x,C) \\ \in\ & (1\pm\varepsilon)\cdot\sum_{x\in S} w(x)\cdot d^z(f(x),f(C)) && \text{(P4)} \\ \in\ & (1\pm 2\varepsilon)\cdot\sum_{x\in D_1} u(x)\cdot d^z(f(x),f(C)) && \text{(P2)} \\ \in\ & (1\pm 3\varepsilon)\cdot\sum_{x\in D_1} u(x)\cdot d^z(x,C) && \text{(P3)} \\ \in\ & (1\pm 4\varepsilon)\cdot\mathrm{cost}_z(X,C). && \text{(P1)} \end{aligned}$$

which completes the proof. □

Remark 5.9. *In the second stage of Algorithm 1, we apply the first framework stated in Theorem 3.1. This is because we want to reduce the dependence of size on k to be linear. In the case that ε is small, we can apply the second framework stated in Theorem 3.1 instead. By Theorem 3.1, the coreset size should be $O\left((2^{2z}\varepsilon^{-2}k\cdot(km\log k+\log(1/\delta))\right)$ where $m=O(\varepsilon^{-2}\log(N_1/\varepsilon))$ by Theorem 5.3. This provides us an ε-coreset of size*

$$O(2^{2z}\varepsilon^{-4}k^2\log(k/\varepsilon)\log(k/\varepsilon\delta)).$$

5.4 Proof of the Main Technical Lemma 5.7

For preparation, we introduce the following theorem showing the existence of a weak-coreset S for (k,z)-subspace approximation over X of size independent of n,d. Recall that $\mathcal{P}$ is the collection of all j-flats in $\mathbb{R}^d$ with $j\le k$, i.e., all subspaces in $\mathbb{R}^d$ of dimension at most k.

Theorem 5.10 (**Weak-coreset for subspace approximation**). *Given a dataset X of n points in $\mathbb{R}^d$, $\varepsilon,\delta\in(0,0.5)$, constant $z\ge 1$ and integer $k\ge 1$, suppose $\sigma: X\to\mathbb{R}_{\ge 0}$ is a sensitivity function satisfying that*

$$\sigma(x)\ge\sup_{P\subseteq\mathcal{P}}\frac{d^z(x,P)}{\sum_{y\in X}d^z(y,P)}$$

for each $x\in X$. Let $\mathcal{G}=\sum_{x\in X}\sigma(x)$ denote the total sensitivity. Suppose $S\subseteq X$ is constructed by taking

$$O\left(\frac{\mathcal{G}^2}{\varepsilon^2}\cdot(\varepsilon^{-1}k^3\log(k/\varepsilon)+\log(1/\delta))\right)$$

samples, where each sample $x\in X$ is selected with probability $\frac{\sigma(x)}{\mathcal{G}}$ and has weight $w(x):=\frac{\mathcal{G}}{|S|\cdot\sigma(x)}$. Then with probability at least $1-\delta$, S is an ε-weak-coreset for the (k,z)-subspace approximation problem over X.

Actually, the above construction implies an algorithm to compute a nearly optimal solution for the (k,z)-subspace approximation problem over X; see discussion in Remark 5.12. To prove the theorem, we need the following lemma based on [17, Theorem 9]. It indicates that a nearly optimal solution for (k,z)-subspace approximation exists in some low dimensional space.

Lemma 5.11 (**Existence of approximate k-flats in low dimensional subspaces**). *Given a weighted dataset X of n points together with weights $u(x)$ in $\mathbb{R}^d$, $\varepsilon\in(0,0.5)$, constant $z\ge 1$ and integer $k\ge 1$, there exists a k-flat P that is spanned by at most $O(\varepsilon^{-1}k^2\log(k/\varepsilon))$ points from X, such that*

$$\sum_{x\in X}u(x)\cdot d^z(x,P)\le(1+\varepsilon)\cdot\min_{P'\in\mathcal{P}}\sum_{x\in X}u(x)\cdot d^z(x,P').$$

Proof. By [32, Theorem 1.3], there exists a collection $D\subseteq X$ of $O(\varepsilon^{-1}k\log(1/\varepsilon)$ points such that the spanned subspace of D contains a k-flat P satisfying that

$$\sum_{x\in X}u(x)\cdot d^z(x,P)\le(1+\varepsilon)^{(k+1)z}\cdot\min_{P'\in\mathcal{P}}\sum_{x\in X}u(x)\cdot d^z(x,P').$$

Replacing $\varepsilon'=O(\varepsilon/zk)$, we complete the proof. □

We are ready to prove the theorem.

Proof of Theorem 5.10. Denote $\mathcal{P}'$ to be the collection of all k-flats that are spanned by at most $O\left(\varepsilon^{-1}k^2\log(k/\varepsilon)\right)$ points from X. By [18, Lemma 8.2], the function dimension[6] of $(X,\mathcal{P}')$ is $O\left(\varepsilon^{-1}k^3\log(k/\varepsilon)\right)$. Then by [35, Theorem 4], with probability at least $1-\delta$, for any k-flat $P\in\mathcal{P}'$,

$$\sum_{x\in S}w(x)\cdot d^z(x,P)\in(1\pm\varepsilon)\cdot\sum_{x\in X}d^z(x,P). \tag{19}$$

[6]Since this paper only uses function dimension as a black box, we do not present the definition. We refer interested readers to [18, Definition 6.4] or [7, Definition 4.5] for concrete definitions.

Then we have

$$
\begin{aligned}
& \min_{P\in\mathcal{P}} \sum_{x\in S} w(x)\cdot d^z(x,P) \\
\geq\ & (1-\varepsilon)\cdot \min_{P\in\mathcal{P}'} \sum_{x\in S} w(x)\cdot d^z(x,P) \\
& \text{(Lemma 5.11)} \\
\geq\ & (1-\varepsilon)^2\cdot \min_{P\in\mathcal{P}'} \sum_{x\in X} d^z(x,P) \\
& \text{(Ineq. (19))} \\
\geq\ & (1-\varepsilon)^2\cdot \min_{P\in\mathcal{P}} \sum_{x\in X} d^z(x,P).
\end{aligned}
$$

We also have

$$
\begin{aligned}
& \min_{P\in\mathcal{P}} \sum_{x\in S} w(x)\cdot d^z(x,P) \\
\leq\ & \min_{P\in\mathcal{P}'} \sum_{x\in S} w(x)\cdot d^z(x,P) \\
\leq\ & (1+\varepsilon)\cdot \min_{P\in\mathcal{P}'} \sum_{x\in X} d^z(x,P) \\
& \text{(Ineq. (19))} \\
\leq\ & (1+\varepsilon)^2\cdot \min_{P\in\mathcal{P}} \sum_{x\in X} w(x)\cdot d^z(x,P). \\
& \text{(Lemma 5.11)}
\end{aligned}
$$

Letting $\varepsilon' = O(\varepsilon)$, we complete the proof. □

Remark 5.12. *Theorem 5.10 actually provides an approach to compute a $(1+\varepsilon)$-approximate solution for the (k,z)-subspace approximation problem. Suppose $P^\star \in \mathcal{P}'$ is a k-flat satisfying that*

$$\sum_{x\in S} w(x)\cdot d^z(x,P^\star) \leq (1+\varepsilon)\cdot \min_{P\in\mathcal{P}'} \sum_{x\in S} w(x)\cdot d^z(x,P).$$

Then by the above proof, we directly have

$$
\begin{aligned}
(1-\varepsilon)\cdot \min_{P\in\mathcal{P}} \sum_{x\in X} d^z(x,P) \leq\ & \sum_{x\in S} w(x)\cdot d^z(x,P^\star) \\
\leq\ & (1+\varepsilon)^3\cdot \min_{P\in\mathcal{P}} \sum_{x\in X} d^z(x,P),
\end{aligned}
$$

which indicates that $\sum_{x\in S} w(x)\cdot d^z(x,P^\star)$ is a $(1\pm O(\varepsilon))$-approximation of the (k,z)-subspace approximation objective $\min_{P\in\mathcal{P}} \sum_{x\in X} d^z(x,P)$. Moreover, since $P^\star\in\mathcal{P}'$, we also have that

$$\sum_{x\in S} w(x)\cdot d^z(x,P^\star) \in (1\pm\varepsilon) \sum_{x\in X} d^z(x,P^\star)$$

by Inequality (19). Thus, $P^\star$ is a $(1+O(\varepsilon))$-approximate solution for the (k,z)-subspace approximation problem.

Now we can prove the main lemma.

Proof of Lemma 5.7. Let $\varepsilon' = \frac{\varepsilon^{z+3}}{6\cdot(84z)^{2z}}$. Recall that we have Γ is a subspace of $\mathbb{R}^d$ satisfying that $C^\star \subset \Gamma$ and for any k-center set $C\in\mathcal{C}$,

$$\sum_{x\in X}\left(d^z(x,\pi(x)) - d^z(x,\pi_C(x))\right) = \frac{\varepsilon'}{2}\cdot \mathsf{OPT}_z,$$

We first have the following observations

$$
\begin{aligned}
\sum_{x\in X} d^z(x,\pi(x)) \leq\ & \sum_{x\in X} d^z(x,C^\star) && (C^\star\in\Gamma) \\
\leq\ & \alpha\cdot \mathsf{OPT}_z, && (\text{Defn. of } C^\star)
\end{aligned}
\tag{20}
$$

and

$$
\begin{aligned}
\sigma_1(x) >\ & 2^{2z+2}\alpha^2\cdot \frac{d^z(x,c^\star(x))}{\mathrm{cost}_z(X,C^\star)} && (\text{Defn. of } \sigma_1(x)) \\
\geq\ & \frac{2^{2z+2}\alpha\cdot d^z(x,\pi(x))}{\mathsf{OPT}_z}, && (C^\star\in\Gamma)
\end{aligned}
\tag{21}
$$

and

$$
\begin{aligned}
\sum_{x\in X}\sigma_1(x) =\ & 2^{2z+2}\alpha^2 \sum_{x\in X}\left(\frac{d^z(x,c^\star(x))}{\mathrm{cost}_z(X,C^\star)} + \frac{1}{|X_{c^\star(x)}|}\right) \\
\leq\ & 2^{2z+2}\alpha^2\cdot(1+k) \qquad (|C^\star| = k) \\
\leq\ & 2^{2z+3}\alpha^2 k.
\end{aligned}
\tag{22}
$$

For a k-center set $C\in\mathcal{C}$, recall that π_C is the projection from X to $\mathrm{Conv}(\Gamma\cup C)$. We claim that

$$
\begin{aligned}
& \min_{C\in\mathcal{C}} \sum_{x\in D_1} u(x)\cdot d^z(x,\pi_C(x)) \\
\geq\ & \min_{C\in\mathcal{C}} \sum_{x\in X} d^z(x,\pi_C(x)) - \frac{\varepsilon'\cdot\mathsf{OPT}_z}{2}.
\end{aligned}
\tag{23}
$$

Let $\widehat{C}\in\mathcal{C}$ denote the k-center set such that $\sum_{x\in X} d^z(x,\pi_C(x))$ is minimized. To prove this inequality, we consider two cases. If $\sum_{x\in X} d^z(x,\pi_{\widehat{C}}(x)) \leq \frac{\varepsilon'\cdot\mathsf{OPT}_z}{2}$, we directly have

$$
\begin{aligned}
& \min_{C\in\mathcal{C}} \sum_{x\in D_1} u(x)\cdot d^z(x,\pi_C(x)) \\
\geq 0 \geq\ & \min_{C\in\mathcal{C}} \sum_{x\in X} d^z(x,\pi_C(x)) - \frac{\varepsilon'\cdot\mathsf{OPT}_z}{2}.
\end{aligned}
$$

Otherwise, suppose $\sum_{x\in X} d^z(x,\pi_{\widehat{C}}(x)) > \varepsilon'\cdot\mathsf{OPT}_z/2$. Since $C^\star\subseteq\Gamma$, we have that for any k-center set $C\in\mathcal{C}$, $\sum_{x\in X} d^z(x,\pi_C(x)) \leq \mathrm{cost}_z(X,C^\star)$. We regard X as a point set in $\Gamma^\perp$ (i.e., the orthogonal complement of Γ). Then each k-center set $C\in\mathcal{C}$ corresponds to a subspace $H\subseteq\Gamma^\perp$ of dimension at most k, satisfying that $\mathrm{Conv}(\Gamma\cup C) = \mathrm{Conv}(\Gamma\cup H)$. This enables us to apply Theorem 5.10 to $\Gamma^\perp$. We set $\sigma(x)$ in Theorem 5.10 as follows:

$$
\begin{aligned}
& \sigma(x) \\
:=\ & \frac{\sigma_1(x)}{2^{2z+2}\alpha^2}\cdot\frac{\mathrm{cost}_z(X,C^\star)}{\sum_{x\in X} d^z(x,\pi_{\widehat{C}}(x))} \\
=\ & \frac{d^z(x,c^\star(x))}{\sum_{x\in X} d^z(x,\pi_{\widehat{C}}(x))} + \frac{\mathrm{cost}_z(X,C^\star)}{|X_{c^\star(x)}|\cdot\left(\sum_{x\in X} d^z(x,\pi_{\widehat{C}}(x))\right)} \\
& (\text{Defn. of } \sigma_1(x)) \\
\geq\ & \sup_{C\in\mathcal{C}} \frac{d^z(x,\pi_C(x))}{\sum_{x\in X} d^z(x,\pi_C(x))}. \\
& (C^\star\in\Gamma \text{ and Defn. of } \widehat{C})
\end{aligned}
$$

Note that the sampling distribution with respect to σ is exactly the same as to σ_1. Moreover, we have

$$\begin{aligned}
\mathcal{G} := & \sum_{x\in X}\sigma(x) \\
= & \sum_{x\in X}\frac{d^z(x,c^\star(x))}{\sum_{x\in X} d^z(x,\pi_{\widehat{C}}(x))} \\
& + \frac{\mathrm{cost}_z(X,C^\star)}{|X_{c^\star(x)}|\cdot\left(\sum_{x\in X} d^z(x,\pi_{\widehat{C}}(x))\right)} \\
= & \frac{(k+1)\cdot \mathrm{cost}_z(X,C^\star)}{\sum_{x\in X} d^z(x,\pi_{\widehat{C}}(x))} \qquad (|C^\star|=k) \\
\leq & \frac{\alpha(k+1)\cdot \mathsf{OPT}_z}{\varepsilon'\cdot\mathsf{OPT}_z/2} \qquad \left(\sum_{x\in X} d^z(x,\pi_{\widehat{C}}(x)) > \frac{\varepsilon'\cdot\mathsf{OPT}_z}{2}\right) \\
= & \frac{2\alpha(k+1)}{\varepsilon'}.
\end{aligned}$$

Hence, $N_1 = \Omega\left(\frac{\mathcal{G}^2}{(\varepsilon')^2}\cdot((\varepsilon')^{-1}k^3\log\frac{k}{\varepsilon'}+\log\frac{1}{\delta})\right)$ as stated in Theorem 5.10. By Theorem 5.10, we have that with probability at least $1-\delta/8$,

$$\begin{aligned}
& \min_{C\in\mathcal{C}}\sum_{x\in D_1} u(x)\cdot d^z(x,\pi_C(x)) \\
\geq\ & (1-\frac{\varepsilon'}{2\alpha})\cdot\min_{C\in\mathcal{C}}\sum_{x\in X} d^z(x,\pi_C(x)) \\
\geq\ & \min_{C\in\mathcal{C}}\sum_{x\in X} d^z(x,\pi_C(x)) - \frac{\varepsilon'}{2\alpha}\cdot\mathrm{cost}_z(X,C^\star). \\
\geq\ & \min_{C\in\mathcal{C}}\sum_{x\in X} d^z(x,\pi_C(x)) - \frac{\varepsilon'\cdot\mathsf{OPT}_z}{2},
\end{aligned}$$

which completes the proof of Inequality (23).

Next, we prove that with probability at least $1-\delta/8$, the following property holds:

$$\sum_{x\in D_1} u(x)\cdot d^z(x,\pi(x)) \leq \sum_{x\in X} d^z(x,\pi(x)) + \frac{\varepsilon'\cdot\mathsf{OPT}_z}{2}. \tag{24}$$

For each sample $x\in D_1$, we note that

$$\begin{aligned}
& |D_1|\cdot u(x)\cdot d^z(x,\pi(x)) \\
= & \frac{\sum_{y\in X}\sigma_1(y)}{\sigma_1(x)}\cdot d^z(x,\pi(x)) \\
\geq & \frac{2^{2z+3}\alpha^2 k}{\frac{2^{2z+2}\alpha\cdot d^z(x,\pi(x))}{\mathsf{OPT}_z}}\cdot d^z(x,\pi(x)) \qquad (25)\\
& \text{(Ineqs. (21) and (22))} \\
= & 2\alpha k\cdot\mathsf{OPT}_z.
\end{aligned}$$

Then by Hoeffding's inequality, we have that

$$\begin{aligned}
& \Pr\left[\left|\sum_{x\in X} d^z(x,\pi(x)) - \sum_{x\in D_1} u(x)\cdot d^z(x,\pi(x))\right| \geq \frac{\varepsilon'\cdot\mathsf{OPT}_z}{2}\right] \\
\leq\ & 2\cdot\exp\left(-\frac{2(\frac{\varepsilon'\cdot\mathsf{OPT}_z}{2})^2}{N_1\cdot(2\alpha k\cdot\mathsf{OPT}_z)^2}\right) \qquad \text{(Ineq. (25))} \\
\leq\ & \frac{\delta}{8}, \qquad \text{(value of } N_1)
\end{aligned}$$

which completes the proof of Inequality (24).

Now we are ready to prove the lemma. With probability at least $1-\delta/4$, Inequalities (23) and (24) hold (union bound). Then for any k-center set $C\in\mathcal{C}$,

$$\begin{aligned}
& \sum_{x\in D_1} u(x)\cdot d^z(x,\pi(x)) - \sum_{x\in D_1} u(x)\cdot d^z(x,\pi_C(x)) \\
\leq\ & \sum_{x\in D_1} u(x)\cdot d^z(x,\pi(x)) - \min_{C'\in\mathcal{C}}\sum_{x\in D_1} u(x)\cdot d^z(x,\pi_{C'}(x)) \\
\leq\ & \sum_{x\in X} d^z(x,\pi(x)) + \frac{\varepsilon'\cdot\mathsf{OPT}_z}{2} \\
& - \min_{C'\in\mathcal{C}}\sum_{x\in X} d^z(x,\pi_{C'}(x)) + \frac{\varepsilon'\cdot\mathsf{OPT}_z}{2} \\
& \text{(Ineqs. (23) and (24))} \\
\leq\ & \sum_{x\in X} d^z(x,\pi(x)) - \sum_{x\in X} d^z(x,\pi_C(x)) + \varepsilon'\cdot\mathsf{OPT}_z \\
\leq\ & 2\varepsilon'\cdot\mathsf{OPT}_z. \\
& \text{(by assumption)}
\end{aligned}$$

By Lemma 5.4, we complete the proof of Lemma 5.7. ◻

5.5 Geometric Observations

Note that the first stage of Algorithm 1 is almost the same to the second framework stated in Theorem 3.1 except that the coreset size N_1 is independent of d. In this section, we discuss the geometric observations that makes N_1 samples enough for an ε-coreset.

Construct a subspace $\Gamma\subsetneq\mathbb{R}^d$ of dimension poly(k/ε) by [33, Algorithm 1], which leads to Inequality (15) by Lemma 5.4. Recall that Γ' is obtained from Γ by appending an arbitrary dimension in $\mathbb{R}^d$ that is orthogonal to Γ. Also recall that $\mathcal{C}_\Gamma$ denotes the collection of k-center sets $C\subset\Gamma'$. We have the following geometric observations implying that we only need to approximately preserve all (k,z)-Clustering objectives with respect to k-center sets in Γ' instead of the whole $\mathcal{C}$. This reduces the function dimension of k-center sets from $O(dk)$ to poly(k/ε). The first observation follows from Claim 5.8.

Observation 5.13 (**Representativeness property for** X). *X satisfies the ε-representativeness property with respect to Γ.*

Moreover, the representativeness property can be generalized to subsets of X that are weak-coresets for the (k,z)-subspace approximation problem.

Observation 5.14 (**Representativeness property for weighted subsets of** X). *Let S be a weighted subset of X together with a weight function $w: S\to\mathbb{R}_{\geq 0}$ and $\varepsilon' = \frac{\varepsilon^{z+3}}{6\cdot(84z)^{2z}}$. S satisfies the ε-representativeness property with respect to Γ if the following holds:*

(1) S is an ε'-weak-coreset for the (k,z)-subspace approximation problem in $\Gamma^\perp$.
(2) S approximately preserves the l_z-subspace approximation objective with respect to Γ, i.e.,

$$\sum_{x\in S} w(x)\cdot d^z(x,\Gamma) \in \sum_{x\in X} d^z(x,\Gamma) \pm \varepsilon'\cdot\mathsf{OPT}_z.$$

Proof. By the proof of Lemma 5.7, these two conditions imply that for any k-center set $C \in \mathcal{C}$,

$$\sum_{x\in S} w(x)\cdot d^z(x,\pi(x)) - \sum_{x\in S} w(x)\cdot d^z(x,\pi_C(x)) \le O(\varepsilon')\cdot \mathsf{OPT}_z,$$

where π and π_C denote the projection from X to Γ and $\mathrm{Conv}(\Gamma\cup C)$ respectively. Then by Lemma 5.7, Inequality (16) holds. By Claim 5.8, we complete the proof. □

Now suppose we have a weighted subset $S \subseteq X$ that satisfies the ε-representativeness property. By Definition 2.3, if S approximately preserves the (k,z)-Clustering objective for some k-center set $C \in \mathcal{C}_\Gamma$ over X, then we directly have that S approximately preserves all (k,z)-Clustering objectives with respect to k-center sets within the whole equivalence class Δ_C^Γ. Hence, we only need to consider those k-center sets in Γ instead of $\mathbb{R}^d$ and conclude the following corollary. The corollary indicates that coreset for clustering in low dimensional subspace plus weak-coreset for subspace approximation implies coreset for clustering in $\mathbb{R}^d$.

Corollary 5.15 (**Dimension reduction for (k,z)-Clustering**). *For every dataset X of n points in $\mathbb{R}^d$, $\varepsilon,\delta \in (0,0.5)$, constant $z \ge 1$ and integer $k \ge 1$, let $\varepsilon' = \frac{\varepsilon^{z+3}}{6\cdot(84z)^{2z}}$. There exists a subspace $\Gamma \subsetneq \mathbb{R}^d$ of dimension $O(k/\varepsilon')$ such that for any weighted point set $S \subseteq X$ together with a weight function $w : S \to \mathbb{R}_{\ge 0}$, S is an $O(\varepsilon)$-coreset for (k,z)-Clustering if*

(1) S is an ε-coreset for (k,z)-Clustering in subspace Γ';
(2) S is an ε'-weak-coreset for the (k,z)-subspace approximation problem in $\Gamma^\perp$.
(3) S approximately preserves the l_z-subspace approximation objective with respect to Γ, i.e.,

$$\sum_{x\in S} w(x)\cdot d^z(x,\Gamma) \in \sum_{x\in X} d^z(x,\Gamma) \pm \varepsilon'\cdot \mathsf{OPT}_z.$$

In fact, the above corollary can be generalized to other shape fitting problems. The main reason is that Lemma 5.4 not only holds for k-center sets but also holds for an arbitrary non-empty set that is contained in a k-dimensional subspace by [33, Theorem 10]. For instance, if we consider $\mathcal{P}$ that is the collection of all j-flats ($j \le k$), then Corollary 5.15 can be translated to a dimension reduction result for subspace approximation as follows.

Corollary 5.16 (**Dimension reduction for subspace approximation**). *For every dataset X of n points in $\mathbb{R}^d$, $\varepsilon,\delta \in (0,0.5)$, constant $z \ge 1$ and integer $k \ge 1$, let $\varepsilon' = \frac{\varepsilon^{z+3}}{6\cdot(84z)^{2z}}$. There exists a subspace $\Gamma \subsetneq \mathbb{R}^d$ of dimension $O(k/\varepsilon')$ such that for any weighted point set $S \subseteq X$ together with a weight function $w : S \to \mathbb{R}_{\ge 0}$, S is an $O(\varepsilon)$-coreset for (k,z)-subspace approximation if*

(1) S is an ε-coreset for (k,z)-subspace approximation in subspace Γ';
(2) S is an ε'-weak-coreset for the (k,z)-subspace approximation problem in $\Gamma^\perp$.
(3) S approximately preserves the l_z-subspace approximation objective with respect to Γ, i.e.,

$$\sum_{x\in S} w(x)\cdot d^z(x,\Gamma) \in \sum_{x\in X} d^z(x,\Gamma) \pm \min_{P\in\mathcal{P}} \sum_{x\in X} d^z(x,P).$$

Similarly, by the Feldman-Langberg framework, this corollary provides an ε-coreset for (k,z)-subspace approximation of size $\mathrm{poly}(k/\varepsilon)$, which matches the result in [33]. Moreover, the coreset size can be further decreased by applying terminal embedding similar to Theorem 5.3.

5.6 Generalization of Theorem 5.1 to ℓ_p-Metrics

Given $p \ge 1$, ℓ_p-metric is induced by distance function $d_p : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\ge 0}$, where for any two points $x,y \in \mathbb{R}^d$,

$$d_p(x,y) := \left(\sum_{i\in[d]} |x_i - y_i|^p\right)^{1/p}. \qquad (26)$$

The formulation captures classic distances, including Manhattan distance (where $p = 1$), Euclidean distance (where $p = 2$) and Chebyshev distance (where $p = \infty$). With ℓ_p-metric, the (k,z)-Clustering objective with respect to some $C \in \mathcal{C}$ is defined as follows

$$\mathrm{cost}_{p,z}(X,C) := \sum_{x\in X} d_p^z(x,C),$$

where, throughout, d_p^z denotes ℓ_p^d-distance raised to power $z \ge 1$, and

$$d_p(x,C) := \min\left\{d_p(x,c) : c\in C\right\}.$$

We can generalize Definition 1.1 to ℓ_p-metrics.

Definition 5.17 (**Coresets for (k,z)-Clustering with ℓ_p-metric in $\mathbb{R}^d$**). Given a collection $X \subseteq \mathbb{R}^d$ of n weighted points and $\varepsilon \in (0,1)$, an ε-coreset for (k,z)-Clustering in ℓ_p^d metric spaces is a subset $S \subseteq \mathbb{R}^d$ with weights $w : S \to \mathbb{R}_{\ge 0}$ such that for any k-center set $C \in \mathcal{C}$, the (k,z)-Clustering objective with respect to C is ε-approximately preserved, i.e.,

$$\sum_{x\in S} w(x)\cdot d_p^z(x,C) \in (1\pm\varepsilon)\cdot \mathrm{cost}_{p,z}(X,C).$$

Note that Theorem 5.1 considers the Euclidean distance where $p = 2$ and we want to generalize Theorem 5.1 to all $p \ge 1$. In this section, we show that Theorem 5.1 can be generalized to ℓ_p-metrics for $1 \le p \le 2$; see the following corollary. The main idea is that for $1 \le p < 2$, there exists an isometric embedding from ℓ_p to ℓ_2 square [24]. By this idea, we can reduce the problem of constructing an ε-coreset for (k,z)-Clustering with ℓ_p-metric to constructing an $O(\varepsilon)$-coreset for $(k,2z)$-Clustering with ℓ_2-metric.

Corollary 5.18 (**Coresets for (k,z)-Clustering with ℓ_p-metrics ($1 \le p < 2$)**). *There exists a randomized algorithm that, for a given dataset X of n points in $\mathbb{R}^d$, integer $k \ge 1$, $1 \le p < 2$, constant $z \ge 2$ and $\varepsilon \in (0,0.5)$, with probability at least $1-\delta$, constructs an ε-coreset for (k,z)-Clustering with ℓ_p-metric of size*

$$O\left(\min\left\{\varepsilon^{-4z-2}, 2^{4z}\varepsilon^{-4}k\right\} k\log k \log\frac{k}{\varepsilon\delta}\right)$$

and runs in time

$$O\left(ndk + nd\log(n/\delta) + k^2\log^2 n + \log^2(1/\delta)\log^2 n\right).$$

The proof can be found in the full version of this paper.

6 PROOF OF [18, THEOREM 15.6]

The proof of Theorem 15.6 in [18] has some typos for proving (85) by (84), where (84) does not satisfy the condition of [18, Lemma 14.2]. To fix the typo, it suffices to prove the following lemma.

Lemma 6.1. *Let $a, b, c \geq 0$, $\varepsilon > 0$ and $z \geq 1$. If $|a - b| \leq c$ and $|a^z - b^z| > \frac{zc^z}{\varepsilon^{z-1}}$, then we have $|a^z - b^z| \leq z\varepsilon \cdot (\max\{a, b\})^z$.*

Proof. Without loss of generality, assume that $a > b > 0$. By scalability, we can also assume that $b = 1$. Then we have

$$a^z - 1 > \frac{zc^z}{\varepsilon^{z-1}} \geq \frac{z(a-1)^z}{\varepsilon^{z-1}}.$$

Moreover, we claim that

$$a^z - 1 \leq (a-1)z \cdot a^{z-1}.$$

This is because that considering function $f(a) = (a-1)z \cdot a^{z-1} - (a^z - 1)$, we have $\nabla_a f(a) = (z-1)z(a^{z-1} - a^{z-2}) \geq 0$ when $a \geq 1$ and, hence, $f(a) \geq f(1) = 0$. Combining the above inequalities, we have $\frac{a-1}{a} \leq \varepsilon$. Then we have

$$a^z - 1 \leq (a-1)z \cdot a^{z-1} \overset{\frac{a-1}{a} \leq \varepsilon}{\leq} z\varepsilon \cdot a^z,$$

which completes the proof. □

Let $a = \text{dist}(p, x)$, $b = \text{dist}(p', x)$, $c = \text{dist}(p, p')$ and $\varepsilon' = \varepsilon/z$ in Lemma 6.1, we complete the proof of Theorem 15.6 in [18] from (84) to (85).

ACKNOWLEDGMENTS

This research was supported in part by NSF CCF-1908347.

REFERENCES

[1] Pankaj K Agarwal and Cecilia Magdalena Procopiuc. 2002. Exact and approximation algorithms for clustering. *Algorithmica* 33, 2 (2002), 201–226.

[2] David Arthur and Sergei Vassilvitskii. 2007. K-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 1027–1035.

[3] Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. 2004. Local search heuristics for k-median and facility location problems. *SIAM Journal on computing* 33, 3 (2004), 544–562.

[4] Frank Ban, Vijay Bhattiprolu, Karl Bringmann, Pavel Kolev, Euiwoong Lee, and David P Woodruff. 2019. A PTAS for l_p-Low Rank Approximation. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 747–766.

[5] Sayan Bandyapadhyay and Kasturi Varadarajan. 2016. On Variants of k-means Clustering. In *32nd International Symposium on Computational Geometry (SoCG 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[6] Luca Becchetti, Marc Bury, Vincent Cohen-Addad, Fabrizio Grandoni, and Chris Schwiegelshohn. 2019. Oblivious dimension reduction for k-means: beyond subspaces and the Johnson-Lindenstrauss lemma. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 1039–1050.

[7] Vladimir Braverman, Dan Feldman, and Harry Lang. 2016. New Frameworks for Offline and Streaming Coreset Constructions. *CoRR* abs/1612.00889 (2016).

[8] Vladimir Braverman, Lingxiao Huang, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. 2019. Coresets for Clustering in Graphs of Bounded Treewidth. *CoRR* abs/1907.04733 (2019).

[9] Vladimir Braverman, Shaofeng Jiang, Robert Krauthgamer, and Xuan Wu. 2019. Coresets for Ordered Weighted Clustering. In *ICML 2019 : Thirty-sixth International Conference on Machine Learning*. 744–753.

[10] Ke Chen. 2009. On Coresets for k-Median and k-Means Clustering in Metric and Euclidean Spaces and Their Applications. *SIAM J. Comput.* 39, 3 (Aug. 2009), 923–947.

[11] Flavio Chierichetti, Sreenivas Gollapudi, Ravi Kumar, Silvio Lattanzi, Rina Panigrahy, and David P Woodruff. 2017. Algorithms for l_p-Low-Rank Approximation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 806–814.

[12] Adam Coates and Andrew Y Ng. 2012. Learning feature representations with k-means. In *Neural networks: Tricks of the trade*. Springer, 561–580.

[13] Michael B Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. 2015. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. ACM, 163–172.

[14] Michael B Cohen and Richard Peng. 2015. L_p-row sampling by Lewis weights. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. ACM, 183–192.

[15] Vincent Cohen-Addad, Philip N Klein, and Claire Mathieu. 2019. Local search yields approximation schemes for k-means and k-median in Euclidean and minor-free metrics. *SIAM J. Comput.* 48, 2 (2019), 644–667.

[16] Amit Deshpande, Madhur Tulsiani, and Nisheeth K Vishnoi. 2011. Algorithms and hardness for subspace approximation. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 482–496.

[17] Amit Deshpande and Kasturi Varadarajan. 2007. Sampling-based dimension reduction for subspace approximation. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. ACM, 641–650.

[18] Dan Feldman and Michael Langberg. 2011. A unified framework for approximating and clustering data. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*. ACM, 569–578.

[19] Dan Feldman, Melanie Schmidt, and Christian Sohler. 2013. Turning big data into tiny data: Constant-size coresets for k-means, PCA and projective clustering. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 1434–1453.

[20] Sariel Har-Peled. 2004. Clustering motion. *Discrete & Computational Geometry* 31, 4 (2004), 545–565.

[21] Sariel Har-Peled and Soham Mazumdar. 2004. On coresets for k-means and k-median clustering. In *36th Annual ACM Symposium on Theory of Computing,*. 291–300.

[22] Lingxiao Huang, Shaofeng Jiang, Jian Li, and Xuan Wu. 2018. Epsilon-Coresets for Clustering (with Outliers) in Doubling Metrics. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 814–825.

[23] Lingxiao Huang, Shaofeng H.-C. Jiang, and Nisheeth Vishnoi. 2019. Coresets for Clustering with Fairness Constraints. In *NeurIPS 2019 : Thirty-third Conference on Neural Information Processing Systems*. 7589–7600.

[24] Csam Kahane and Jean-Pierre Kahane. 1993. *Some random series of functions*. Vol. 5. Cambridge University Press.

[25] Michael Langberg and Leonard J Schulman. 2010. Universal ε-approximators for integrals. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*. SIAM, 598–607.

[26] Stuart P. Lloyd. 1982. Least squares quantization in PCM. *IEEE Trans. Information Theory* 28, 2 (1982), 129–136.

[27] Konstantin Makarychev, Yury Makarychev, and Ilya Razenshteyn. 2019. Performance of Johnson-Lindenstrauss transform for k-means and k-medians clustering. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 1027–1038.

[28] Ramgopal R Mettu and C Greg Plaxton. 2004. Optimal time bounds for approximate clustering. *Machine Learning* 56, 1-3 (2004), 35–60.

[29] Sakib A Mondal. 2018. An improved approximation algorithm for hierarchical clustering. *Pattern Recognition Letters* 104 (2018), 23–28.

[30] Shyam Narayanan and Jelani Nelson. 2019. Optimal terminal dimensionality reduction in Euclidean space. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 1064–1069.

[31] Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. 2019. Fair Coresets and Streaming Algorithms for Fair k-means. In *International Workshop on Approximation and Online Algorithms*. 232–251.

[32] Nariankadu D Shyamalkumar and Kasturi Varadarajan. 2007. Efficient subspace approximation algorithms. In *SODA*, Vol. 7. 532–540.

[33] Christian Sohler and David P. Woodruff. 2018. Strong Coresets for k-Median and Subspace Approximation: Goodbye Dimension. In *FOCS*. IEEE Computer Society, 802–813.

[34] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, et al. 2006. Cluster analysis: basic concepts and algorithms. *Introduction to data mining* 8 (2006), 487–568.

[35] Kasturi Varadarajan and Xin Xiao. 2012. On the Sensitivity of Shape Fitting Problems. In *IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2012)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.