On Coresets for Clustering in Small Dimensional Euclidean spaces

Lingxiao Huang [*1] **Ruiyuan Huang** [*2] **Zengfeng Huang** [*2] **Xuan Wu** [*3]

Abstract

We consider the problem of constructing small coresets for k-MEDIAN in Euclidean spaces. Given a large set of data points $P \subset \mathbb{R}^d$, a coreset is a much smaller set $S \subset \mathbb{R}^d$, so that the k-MEDIAN costs of any k centers w.r.t. P and S are close. Existing literature mainly focuses on the high-dimension case and there has been a great success in obtaining dimension-independent bounds, whereas the case for small d is largely unexplored. Considering many applications of Euclidean clustering algorithms are in small dimensions and the lack of systematic studies in the current literature, this paper investigates coresets for k-MEDIAN in small dimensions. For small d, a natural question is whether existing near-optimal dimension-independent bounds can be significantly improved. We provide affirmative answers to this question for a range of parameters. Moreover, new lower bound results are also proved, which are the highest for small d. In particular, we completely settle the coreset size bound for 1-d k-MEDIAN (up to log factors). Interestingly, our results imply a strong separation between 1-d 1-MEDIAN and 1-d 2-MEDIAN. As far as we know, this is the first such separation between $k = 1$ and $k = 2$ in any dimension.

1. Introduction

Processing huge datasets is always computationally challenging. In this paper, we consider the coreset paradigm, which is an effective data-reduction tool to alleviate the computation burden on big data. Roughly speaking, given a large dataset, the goal is to construct a much smaller dataset, called *coreset*, so that vital properties of the original dataset are preserved. Coresets for various problems have been extensively studied (Har-Peled & Mazumdar, 2004; Feldman & Langberg, 2011; Feldman et al., 2013; Cohen-Addad et al., 2022; Braverman et al., 2022). In this paper, we investigate coreset construction for k-MEDIAN in Euclidean spaces.

Coreset construction for Euclidean k-MEDIAN has been studied for nearly two decades (Har-Peled & Mazumdar, 2004; Feldman & Langberg, 2011; Huang et al., 2018; Cohen-Addad et al., 2021; 2022). For this particular problem, an ε-coreset is a (weighted) point set in the same Euclidean space that satisfies: given any set of k centers, the k-MEDIAN costs of the centers w.r.t. the original point set and the coreset are within a factor of $1 + \varepsilon$. The most important task in theoretical research here is to characterize the minimum size of ε-coresets. Recently, there has been great progress in closing the gap between upper and lower bounds in high-dimensional spaces. However, researches on the coreset size in small dimensional spaces are rare. There are still large gaps between upper and lower bounds even for 1-d 1-MEDIAN.

Clustering in small dimensional Euclidean spaces is of both theoretical and practical importance. In practice, many applications involve clustering points in small dimensional spaces. A typical example is clustering objects in $\mathbb{R}^2$ or $\mathbb{R}^3$ based on their spatial coordinates (Wheeler, 2007; Fonseca-Rodríguez et al., 2021). Another example is spectral clustering for graph and social network analysis (Von Luxburg, 2007; Kunegis et al., 2010; Zhang et al., 2014; Narantsatsralt & Kang, 2017). In spectral clustering, nodes are first embedded into a small dimensional Euclidean space using spectral methods and then Euclidean clustering algorithms are applied in the embedding space. Even the simplest 1-d k-MEDIAN has numerous practical applications (Arnaboldi et al., 2012; Jeske et al., 2013; Pennacchioli et al., 2014).

On the theory side, existing techniques for coresets in high dimensions may not be sufficient to obtain optimal coresets in small dimensions. For example, much smaller size is achievable in $\mathbb{R}^1$ by using geometric methods, while the sampling methods for strong coresets in high dimension (Langberg & Schulman, 2010; Cohen-Addad et al.,

[*]Equal contribution [1]State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing, China [2]School of Data Science, Fudan University, Shanghai, China [3]Huawei TCS Lab, Shanghai, China. Correspondence to: Lingxiao Huang <huanglingxiao1990@126.com>, Zengfeng Huang <huangzf@fudan.edu.cn>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

2021; Huang et al., 2022) seem not viable to obtain such bounds in low dimensions. This suggests that optimal coreset construction in small dimensions may require new techniques, which provides a partial explanation of why 1-d 1-MEDIAN is still open after two decades of research. That being said, the coreset problem for clustering in small dimensional spaces is of great theoretical interest and practical value. Yet it is largely unexplored in the literature. This paper aims to fill the gap and study the following question:

Question 1. *What is the tight coreset size for Euclidean k-MEDIAN problem in $\mathbb{R}^d$ for small d?*

1.1. Problem Definitions and Previous Results

Euclidean k-MEDIAN. In the Euclidean k-MEDIAN problem, we are given a dataset $P \subset \mathbb{R}^d$ ($d \geq 1$) of n points and an integer $k \geq 1$; and the goal is to find a k-center set $C \subset \mathbb{R}^d$ that minimizes the objective function

$$\operatorname{cost}(P, C) := \sum_{p \in P} d(p, C) = \sum_{p \in P} \min_{c \in C} d(p, c), \quad (1)$$

where $d(p, c)$ represents the Euclidean distance between p and c. It has many application domains including approximation algorithms, unsupervised learning, and computational geometry (Lloyd, 1982; Tan et al., 2006; Arthur & Vassilvitskii, 2007; Coates & Ng, 2012).

Coresets. Let $\mathcal{C}$ denote the collection of all k-center sets, i.e., $\mathcal{C} := \{C \subset \mathbb{R}^d \ : \ |C| = k\}$.

Definition 1.1 **(ε-Coreset for Euclidean k-MEDIAN (Har-Peled & Mazumdar, 2004)).** *Given a dataset $P \subset \mathbb{R}^d$ of n points, an integer $k \geq 1$ and $\varepsilon \in (0, 1)$, an ε-coreset for Euclidean k-MEDIAN is a subset $S \subseteq P$ with weight $w : S \to \mathbb{R}_{\geq 0}$, such that*

$$\forall C \in \mathcal{C}, \qquad \sum_{p \in S} w(p) \cdot d(p, C) \in (1 \pm \varepsilon) \cdot \operatorname{cost}(P, C).$$

For Euclidean k-MEDIAN, the best known upper bound on ε-coreset size is $\tilde{O}(\min\left\{\frac{k^{4/3}}{\varepsilon^2}, \frac{k}{\varepsilon^3}\right\})$ (Huang et al., 2022; Cohen-Addad et al., 2022) and $\Omega(\frac{k}{\varepsilon^2})$ is the best existing lower bound (Cohen-Addad et al., 2022). The upper bound is dimension-independent, since using dimensionality reduction techniques such as Johnson–Lindenstrauss transform, the dimension can be reduced to $\tilde{\Theta}(\frac{1}{\varepsilon^2})$. Thus, most previous work essentially only focus on $d = \tilde{\Theta}(\frac{1}{\varepsilon^2})$, whereas the case for $d < \frac{1}{\varepsilon^2}$ is largely unexplored. The lower bound requires $d = \Omega(\frac{k}{\varepsilon^2})$, as the hard instance for the lower bound is an orthonormal basis of size $\Omega(\frac{k}{\varepsilon^2})$. For constant k and large enough d, the upper and lower bounds match up to a polylog factor.

On the contrary, for $d \ll \Theta(\frac{1}{\varepsilon^2})$, tight coreset sizes for k-MEDIAN are far from well-understood, even when $k = 1$. Specifically, for constant d, the current best upper bound is $\tilde{O}(\frac{k}{\varepsilon^3}, \frac{kd}{\varepsilon^2})$ (Feldman & Langberg, 2011; Cohen-Addad et al., 2022), and the best lower bound is $\Omega(\frac{k}{\sqrt{\varepsilon}})$ (Baker et al., 2020). Thus, there is a still large gap between the upper and lower bounds for small d. Perhaps surprisingly, this is the case even for $d = 1$: Har-Peled & Kushal (2005) present a coreset of size $\tilde{O}(\frac{k}{\varepsilon})$ in $\mathbb{R}$ while the best known lower bound is $\Omega(\frac{k}{\sqrt{\varepsilon}})$.

1.2. Our Results

We provide a complete characterization of the coreset size (up to a logarithm factor) for $d = 1$ and partially answer Question 1 for $1 < d < \Theta(\frac{1}{\varepsilon^2})$. Our results are summarized in Table 1.

For $d = 1$, we construct coresets with size $\tilde{O}(\frac{1}{\sqrt{\varepsilon}})$ for 1-MEDIAN (Theorem 2.1) and prove that the coreset size lower bound is $\Omega(\frac{k}{\varepsilon})$ for $k \geq 2$ (Theorem 2.10). Previous work has shown coresets with size $\tilde{O}(\frac{k}{\varepsilon})$ exist for k-MEDIAN (Har-Peled & Kushal, 2005) in 1-d, and thus our lower bound nearly matches this upper bound. On the other hand, it was proved that the coreset size of 1-MEDIAN in 1-d is $\Omega(\frac{1}{\sqrt{\varepsilon}})$ (Baker et al., 2020), which shows our upper bound result for 1-MEDIAN is nearly tight.

For $d > 1$, we provide a discrepancy-based method that constructs deterministic coresets of size $\tilde{O}(\frac{\sqrt{d}}{\varepsilon})$ for 1-MEDIAN (Theorem 3.2). Our result improves over the existing $\tilde{O}(\frac{1}{\varepsilon^2})$ upper bound (Cohen-Addad et al., 2021) for $1 < d < \Theta(\frac{1}{\varepsilon^2})$ and matches the $\Omega(\frac{1}{\varepsilon^2})$ lower bound (Cohen-Addad et al., 2022) for $d = \Theta(\frac{1}{\varepsilon^2})$. We further prove a lower bound of $\Omega(kd)$ for k-MEDIAN in $\mathbb{R}^d$ (Theorem D.3). Combining with our 1-d lower bound $\Omega(\frac{k}{\varepsilon})$, this improves over the existing $\Omega(\frac{k}{\sqrt{\varepsilon}} + d)$ lower bound (Baker et al., 2020; Cohen-Addad et al., 2022).

1.3. Technical Overview

We first discuss the 1-d k-MEDIAN problem and show that the framework of (Har-Peled & Kushal, 2005) is optimal with significant improvement for $k = 1$. Then we briefly summarize our approaches for $2 \leq d \leq \varepsilon^{-2}$.

The Bucket-Partitioning Framework for 1-d k-MEDIAN in (Har-Peled & Kushal, 2005). Our main results in 1-d are based on the classic bucket-partitioning framework, developed in (Har-Peled & Kushal, 2005), which we briefly review now. They greedily partition a dataset $P \subset \mathbb{R}$ into $O(k\varepsilon^{-1})$ consecutive buckets B's and collect the mean point $\mu(B)$ together with weight $|B|$ as their coreset S. Their construction requires that the cumulative error $\delta(B) = \sum_{p \in B} |p - \mu(B)| \leq \varepsilon \cdot \mathsf{OPT}/k$ holds for every bucket B, where OPT is the optimal k-MEDIAN cost of P. Their important geometric observation is that

Table 1. Comparison of coreset sizes for k-MEDIAN in $\mathbb{R}^d$. We use following abbreviations: [1] for (Har-Peled & Kushal, 2005), [2] for (Feldman & Langberg, 2011), [3] for (Baker et al., 2020), [4] for (Cohen-Addad et al., 2021), [5] for (Cohen-Addad et al., 2022) and [6] for (Huang et al., 2022). The symbol † represents that the results can be generalized to (k, z)-CLUSTERING (Definition 3.1).

Paremeters d, k		Best Known Upper Bound	Best Known Lower Bound	Our Results
$d = 1$	$k = 1$	$\tilde{O}(\varepsilon^{-1})$ [1]	$\Omega(\varepsilon^{-1/2})$ [3]	$\tilde{O}(\varepsilon^{-1/2})$ (Thm. 2.1)
	$k > 1$	$O(k\varepsilon^{-1})$ [1]	$\Omega(k\varepsilon^{-1/2})$ [3]	$\Omega(k\varepsilon^{-1})$ (Thm. 2.10)
$1 < d < \Theta(\varepsilon^{-2})$	$k = 1$	$\tilde{O}(\varepsilon^{-2})$ [4]	$\Omega(\varepsilon^{-1/2})$ [3]	$\tilde{O}(\sqrt{d}\varepsilon^{-1})^\dagger$ (Thm. 3.2)
	$k > 1$	$\tilde{O}(\min\left\{\frac{kd}{\varepsilon^2}, \frac{k}{\varepsilon^3}, \frac{k^{4/3}}{\varepsilon^2}\right\})$ [2,5, 6]	$\Omega(k\varepsilon^{-1/2})$ [3]	$\Omega(kd + k\varepsilon^{-1})^\dagger$ (Thm. D.3)
$d = \Omega(\varepsilon^{-2})$	$k \geq 1$	$\tilde{O}(\min\left\{\frac{k}{\varepsilon^3}, \frac{k^{4/3}}{\varepsilon^2}\right\})$ [5, 6]	$\Omega(k\varepsilon^{-2})$ [5]	

the induced error $|\mathrm{cost}(B, C) - |B| \cdot d(\mu(B), C)|$ of every bucket B is at most $\delta(B)$, and even is 0 when all points in B assign to the same center. Consequently, only $O(k)$ buckets induce a non-zero error for every center set C and the total induced error is at most $\varepsilon \cdot \mathsf{OPT}$, which concludes that S is a coreset of size $O(k\varepsilon^{-1})$.

Reducing the Number of Buckets for 1-d 1-MEDIAN via Adaptive Cumulative Errors. In the case of $k = 1$ where there is only one center $c \in \mathbb{R}$, we improve the result in (Har-Peled & Kushal, 2005) (Theorem 2.1) through the following observation: $\mathrm{cost}(P, c)$ can be much larger than OPT when center c is close to either of the endpoints of P, and consequently, can allow a larger induced error of coreset than $\varepsilon \cdot \mathsf{OPT}$. This observation motivates us to adaptively select cumulative errors for different buckets according to their locations. Inspired by this motivation, our algorithm (Algorithm 1) first partitions dataset P into blocks B_i according to clustering cost, i.e., $\mathrm{cost}(P, c) \approx 2^i \cdot \mathsf{OPT}$ for all $c \in B_i$, and then further partition each block B_i into buckets $B_{i,j}$ with a carefully selected cumulative error bound $\delta(B_{i,j}) \leq \varepsilon \cdot 2^i \cdot \mathsf{OPT}$. Intuitively, our selection of cumulative errors is proportional to the minimum clustering cost of buckets, which results in a coreset.

For the coreset size, we first observe that there are only $O(\log \varepsilon^{-1})$ non-empty blocks B_i (Lemma 2.8) since we can "safely ignore" the leftmost and the rightmost εn points and the remaining points $p \in P$ satisfy $\mathrm{cost}(P, p) \leq \varepsilon^{-1}\mathsf{OPT}$. The most technical part is that we show the number m of buckets in each B_i is at most $O(\varepsilon^{-1/2})$ (Lemma 2.9), which results in our improved coreset size $\tilde{O}(\varepsilon^{-1/2})$. The basic idea is surprisingly simple: the clustering cost of a bucket is proportional to its distance to center c, and hence, the clustering cost of m consecutive buckets is proportional to m^2 instead of m. According to this idea, we find that $m^2 \cdot \delta(B_{i,j}) \leq 2^i \cdot \mathsf{OPT}$ for every B_i, which implies a desired bound $m = O(\varepsilon^{-1/2})$ by our selection of $\delta(B_{i,j}) \approx \varepsilon \cdot 2^i \cdot \mathsf{OPT}$.

Hardness Result for 1-d 2-MEDIAN: Cumulative Error is Unavoidable. We take $k = 2$ as an example here and show the tightness of the $O(\varepsilon^{-1})$ bound by (Har-Peled & Mazumdar, 2004). The extension to $k > 2$ is standard via an idea of (Baker et al., 2020).

We construct the following worst-case instance $P \subset \mathbb{R}$ of size ε^{-1}: We construct $m = \varepsilon^{-1}$ consecutive buckets $B_1, B_2, \ldots, B_m$ such that the length of buckets exponentially increases while the number of points in buckets exponentially decreases. We fix a center at the leftmost point of P (assuming to be 0 w. l. o. g.) and move the other center c along the axis. Such dataset P satisfies the following:

- the clustering cost is stable: for all c, $f_P(c) := \mathrm{cost}(P, \{0, c\}) \approx \varepsilon^{-1}$ up to a constant factor;
- the cumulative error for every bucket B_i is $\delta(B_i) \approx 1$;
- for every B_i, $\mathrm{cost}(B_i, \{0, c\})$ is a quadratic function that first decreases and then increases as c moves from left to right within B_i, and the gap between the maximum and the minimum values is $\Omega(\delta(B_i))$.

Suppose $S \subseteq P$ is of size $o(\varepsilon^{-1})$. Then there must exist a bucket B such that $S \cap B = \emptyset$. We find that function $f_S(c) := \mathrm{cost}(S, \{0, c\})$ is an affine linear function when c is located within B_i (Lemma 2.12). Consequently, the maximum induced error $\max_{c \in B_i} |f_P(c) - f_S(c)|$ is at least $\Omega(\delta(B_i))$ since the estimation error of an affine linear function f_S to a quadratic function f_P is up to certain "cumulative curvature" of f_P (Lemma 2.11), which is $\Omega(\delta(B_i))$ due to our construction. Hence, S is not a coreset since $f_P(c) \approx \varepsilon^{-1}$ always holds.

We remind the readers that the above cost function f_P is actually a piecewise quadratic function with $O(\varepsilon^{-1})$ pieces instead of a quadratic one, which ensures the stability of f_P. This is the main difference from $k = 1$, which leads to a gap of $\varepsilon^{-1/2}$ on the coreset size between $k = 1$ and $k = 2$. As far as we know, this is the first such separation in any dimension.

Our Approaches when $2 \le d \le \varepsilon^{-2}$. For 1-MEDIAN, our upper bound result (Theorem 3.2) combines a recent hierarchical decomposition coreset framework in (Braverman et al., 2022), that reduces the instance to a hierarchical ring structure (Theorem A.2), and the discrepancy approaches (Theorem A.4) developed by (Karnin & Liberty, 2019). The main idea is to extend the analytic analysis of (Karnin & Liberty, 2019) to handle multiplicative errors in a scalable way.

For k-MEDIAN, our lower bound result (Theorem D.3) extends recently developed approaches in (Cohen-Addad et al., 2022). Their hard instance is an orthonormal basis in $\mathbb{R}^d$, the size of which is at most d, and hence cannot obtain a lower bound higher than $\Omega(d)$. We improve the results by embedding $\Theta(k)$ copies of their hard instance in $\mathbb{R}^d$, each of which lies in a different affine subspace. We argue that the errors from all subspaces add up. However, the error analysis from (Cohen-Addad et al., 2022) cannot be directly used; we need to overcome several technical challenges. For instance, points in the coreset are not necessary in any affine subspace, so the error in each subspace is not a corollary of their result. Moreover, errors from different subspaces may cancel each other.

1.4. Other Related Work

Coresets for Clustering in Metric Spaces Recent works (Cohen-Addad et al., 2022; Cohen-Addad et al., 2022; Huang et al., 2023a) show that Euclidean (k, z)-CLUSTERING admits ε-coresets of size $\tilde{O}(k\varepsilon^{-2} \cdot \min\{\varepsilon^{-z}, k^{\frac{z}{z+2}}\})$ and a nearly tight bound $\tilde{O}(\varepsilon^{-2})$ is known when $k = 1$ (Cohen-Addad et al., 2021). Apart from the Euclidean metric, the research community also studies coresets for clustering in general metric spaces a lot. For example, Feldman & Langberg (2011) construct coresets of size $\tilde{O}(k\varepsilon^{-2}\log n)$ for general discrete metric. Baker et al. (2020) show that the previous $\log n$ factor is unavoidable. There are also works on other specific metrics spaces: doubling metrics (Huang et al., 2018) and graphs with shortest path metrics (Baker et al., 2020; Braverman et al., 2021; Cohen-Addad et al., 2021), to name a few.

Coresets for Variants of Clustering Coresets for variants of clustering problems are also of great interest. For example, Braverman et al. (2022) construct coresets of size $\tilde{O}(k^3\varepsilon^{-6})$ for capacitated k-MEDIAN, which is improved to $\tilde{O}(k^3\varepsilon^{-5})$ by (Huang et al., 2023a). Other important variants of clustering include ordered clustering (Braverman et al., 2019), robust clustering (Huang et al., 2023b), and time-series clustering (Huang et al., 2021).

2. Tight Coreset Sizes for 1-d k-MEDIAN

2.1. Near Optimal Coreset for 1-d 1-MEDIAN

We have the following theorem.

Theorem 2.1 (Improved Coreset for one-dimensional 1-MEDIAN). *There is a polynomial time algorithm, such that given an input data set $P \subset \mathbb{R}$, it outputs an ε-coreset of P for 1-MEDIAN with size $\tilde{O}(\varepsilon^{-\frac{1}{2}})$.*

Remark 2.2 (Discussion of our coreset result). *We note that there are fast algorithms for 1-MEDIAN and 1-d k-MEDIAN (Cohen et al., 2016; Grønlund et al., 2018) and PTAS for general k-MEDIAN (Har-Peled & Kushal, 2005). Even with these algorithms, coresets are also of great importance because coreset constructions have numerous uses beyond computing the optimal solution of a problem. For example, one can answer any query by computing on the coreset rather than computing on the full dataset, which greatly reduces the time complexity of answering queries. Furthermore, for any problem that admits a coreset construction, we can convert an arbitrary offline algorithm to a simultaneously parallel and streaming algorithm via a black-box reduction (Munteanu & Schwiegelshohn, 2018). Moreover, a small coreset can automatically accelerate a PTAS by replacing the original dataset with the coreset and running the PTAS on the coreset (Har-Peled & Kushal, 2005).*

In the case of $d = 1$, the black-box reduction converts any off-line algorithm to a single-pass streaming algorithm with only $O(m\, polylog(n))$ spaces, where n is the stream length and m is the coreset size. Specifically, for the 1-d 1-MEDIAN problem (computing the median) our paper immediately provides an ε-approximated streaming algorithm with $O(\varepsilon^{-1/2} polylog(n))$ spaces. Furthermore, as a coreset maintains the answer to all queries, we can compute the sum of the distance to a set of centers in time independent of the point numbers.

Useful Notations and Facts. Throughout this section, we use $P = \{p_1, \cdots, p_n\} \subset \mathbb{R}$ with $p_1 < p_2 < \cdots < p_n$. Let $c^\star = p_{\lfloor \frac{n}{2} \rfloor}$, we have the following simple observations for $\mathrm{cost}(P, c)$.

Observation 2.3. *$\mathrm{cost}(P, c)$ is a convex piecewise affine linear function of c and $\mathsf{OPT} = \mathrm{cost}(P, c^\star)$ is the optimal 1-MEDIAN cost on P.*

The following notions, proposed by (Har-Peled & Mazumdar, 2004), are useful for our coreset construction.

Definition 2.4 (Bucket). *A bucket B is a continuous subset $\{p_l, p_{l+1} \dots, p_r\}$ of P for some $1 \le l \le r \le n$.*

Definition 2.5 (Mean and cumulative error (Har-Peled & Kushal, 2005)). *Given a bucket $B = \{p_l, \dots, p_r\}$ for some $1 \le l \le r \le n$, denote $N(B) := r - l + 1$ to be* *the number of points within B and $L(B) := p_r - p_l$ to be the length of B. We define the* mean *of B to be $\mu(B) := \frac{1}{N(B)}\sum_{p\in B} p$, and define the* cumulative error *of B to be $\delta(B) := \sum_{p\in B}|p-\mu(B)|$.*

Note that $\mu(B) \in [p_l, p_r]$ always holds, which implies the following fact.

Fact 2.6. $\delta(B) \le N(B)\cdot L(B)$.

The following lemma shows that for each bucket B, the coreset error on B is no more than $\delta(B)$.

Lemma 2.7 (Cumulative error controls coreset error (Har-Peled & Kushal, 2005)). *Let $B = \{p_l, \ldots, p_r\} \subseteq P$ for $1 \le l \le r \le n$ be a bucket and $c \in \mathbb{R}$ be a center. We have*

1. *if $c \in (p_l, p_r)$, $|\mathrm{cost}(B,c) - N(B)d(\mu(B),c)| \le \delta(B)$;*
2. *if $c \notin (p_l, p_r)$, $|\mathrm{cost}(B,c) - N(B)d(\mu(B),c)| = 0$.*

Algorithm for Theorem 2.1. Our algorithm is summarized in Algorithm 1. We improve the framework in (Har-Peled & Kushal, 2005), which partitions P into multiple buckets so that the cumulative errors in different buckets are the same and collects their means as a coreset. Our main idea is to carefully select an adaptive cumulative error for different buckets. In Lines 2-3, we take the leftmost εn points and the rightmost εn points, and add their weighted means to our coreset S. In Lines 4 (and 7), we divide the remaining points into disjoint blocks B_i (B_i') such that for every $p \in B_i$, $\mathrm{cost}(P,p) \approx 2^i\cdot\mathsf{OPT}$, and then greedily divide each B_i into disjoint buckets $B_{i,j}$ with a cumulative error roughly $\varepsilon\cdot 2^i\cdot\mathsf{OPT}$ in Line 5. We remind the readers that the cumulative error in (Har-Peled & Kushal, 2005) is always $\varepsilon\cdot\mathsf{OPT}$.

We define function $f_P : \mathbb{R} \to \mathbb{R}_{\ge 0}$ such that $f_P(c) = \mathrm{cost}(P,c)$ for every $c \in \mathbb{R}$ and define $f_S : \mathbb{R} \to \mathbb{R}_{\ge 0}$ such that $f_S(c) = \mathrm{cost}(S,c)$ for every $c \in \mathbb{R}$. By Observation 2.3, $f_P(c)$ is decreasing on $(-\infty, c^*]$ and increasing on $[c^*, \infty)$. As a result, each $B_i(B_i')$ consists of consecutive points in P. The following lemma shows that the number of blocks $B_i(B_i')$ is $O(\log\frac{1}{\varepsilon})$.

Lemma 2.8 (Number of blocks). *There are at most $O(\log(\frac{1}{\varepsilon}))$ non-empty blocks B_i or B_i'.*

Proof: We prove Algorithm 1 divides $\{p_{L+1}, \ldots, p_{\lfloor\frac{n}{2}\rfloor}\}$ into at most $O(\log(\frac{1}{\varepsilon}))$ non-empty blocks B_i. Argument for $\{p_{\lfloor\frac{n}{2}\rfloor+1}, \ldots, p_R\}$ is entirely symmetric.

If B_i is non-empty for some $i \ge 0$, we must have $f_P(p) \ge 2^i\cdot\mathsf{OPT}$ for $p \in B_i$. We also have $p > p_L$ since $p \in B_i \subset \{p_{L+1}, \ldots, p_{\lfloor\frac{n}{2}\rfloor}\}$. Since f_P is convex, we have $2^i\cdot\mathsf{OPT} \le f_P(p) \le f_P(p_L)$. If we show that $f_P(p_L) \le (1+\varepsilon^{-1})\cdot\mathsf{OPT} = (1+\varepsilon^{-1})\cdot f_P(c^\star)$ then we have $2^i \le (1+\varepsilon^{-1})$ thus $i \le O(\log(\frac{1}{\varepsilon}))$.

To prove $f_P(p_L) \le (1+\varepsilon^{-1})\cdot f_P(c^\star)$, we use triangle inequality to obtain that

$$\begin{aligned} f_P(p_L) &= \sum_{i=1}^{n}|p_i - p_L| \\ &\le \sum_{i=1}^{n}(|p_i - c^\star| + |c^\star - p_L|) \\ &= f_P(c^\star) + n\cdot|c^\star - p_L|. \end{aligned}$$

Moreover, we note that by the choice of p_L, $|c^\star - p_L| \le \frac{1}{L}\cdot\sum_{i=1}^{L}|c^\star - p_i| \le \frac{f_P(c^\star)}{\varepsilon n}$. Thus we have,

$$f_P(p_L) \le f_P(c^\star) + n\cdot\frac{f_P(c^\star)}{\varepsilon n} = (1+\varepsilon^{-1})\cdot f_P(c^\star).$$

$\square$

Algorithm 1 Coreset1d(P, ε)

Input: Dataset $P = \{p_1, \cdots, p_n\} \subset \mathbb{R}$ with $p_1 < \cdots < p_n$, and $\varepsilon \in (0,1)$.
Output: An ε-coreset S of P for 1-d 1-MEDIAN

1: Set $S \leftarrow \emptyset$.
2: Set $L \leftarrow \lfloor\varepsilon n\rfloor$ and $R \leftarrow n - \lfloor\varepsilon n\rfloor$. Set $B_- \leftarrow \{p_1, \ldots, p_L\}$ and $B_+ \leftarrow \{p_{R+1}, \ldots, p_n\}$.
3: Add $\mu(B_-)$ with weight $N(B_-)$ and $\mu(B_+)$ with weight $N(B_+)$ into S.
4: Divide $\{p_{L+1}, \ldots, p_{\lfloor\frac{n}{2}\rfloor}\}$ into disjoint blocks $\{B_i\}_{i\ge 0}$ where $B_i := \{p \in \{p_{L+1}, \ldots p_{\lfloor\frac{n}{2}\rfloor}\} : 2^i\cdot\mathsf{OPT} \le \mathrm{cost}(P,p) < 2^{i+1}\cdot\mathsf{OPT}\}$.
5: For each non-empty block B_i ($i \ge 0$), consider the points within B_i from left to right and group them into buckets $\{B_{i,j}\}_{j\ge 0}$ in a greedy way: each bucket $B_{i,j}$ is a maximal set with $\delta(B_{i,j}) \le \varepsilon\cdot 2^i\cdot\mathsf{OPT}$.
6: For every bucket $B_{i,j}$, add $\mu(B_{i,j})$ with weight $N(B_{i,j})$ into S.
7: Symmetrically divide $\{p_{\lfloor\frac{n}{2}\rfloor+1}, \ldots, p_R\}$ into disjoint buckets $\{B_{i,j}'\}_{i,j\ge 0}$ and add $\mu(B_{i,j}')$ with weight $N(B_{i,j}')$ into S for every bucket $B_{i,j}'$.
8: Return S.

We next give a key lemma that we use to obtain an improved coreset size.

Lemma 2.9 (Number of buckets). *Each non-empty block B_i or B_i' is divided into $O(\varepsilon^{-1/2})$ buckets.*

Proof: We prove that each block $B_i \subset \{p_{L+1}, \ldots, p_{\lfloor\frac{n}{2}\rfloor}\}$ is divided into at most $O(\varepsilon^{-1/2})$

buckets $B_{i,j}$. Argument for $B'_i \subset \{p_{\lfloor \frac{n}{2} \rfloor+1}, \ldots, p_R\}$ is entirely symmetric.

Suppose $B_i = \{p_{l_i}, \ldots, p_{r_i}\}$ and we divide B_i into t buckets $\{B_{i,j}\}_{j=0}^{t-1}$. Since each $B_{i,j}$ is the maximal bucket with $\delta(B_{i,j}) \leq \varepsilon \cdot 2^i \cdot \mathsf{OPT}$, we have $\delta(B_{i,2j} \cup B_{i,2j+1}) > \varepsilon \cdot 2^i \cdot \mathsf{OPT}$ for $2j+1 < t$. Denote $B_{i,2j} \cup B_{i,2j+1}$ by C_j for $j \in \{0, \ldots, \lfloor \frac{t-2}{2} \rfloor\}$, we have

$$\begin{aligned}
4 \cdot 2^i \cdot \mathsf{OPT} &\geq f_P(p_{l_i}) + f_P(p_{r_i}) \\
&\geq \sum_{p \in B_i} (|p - p_{l_i}| + |p - p_{r_i}|) \\
&= N(B_i)(p_{r_i} - p_{l_i}) \\
&\geq \Big(\sum_{j=1}^{\lfloor \frac{t-2}{2} \rfloor} N(C_j)\Big) \cdot \Big(\sum_{j=1}^{\lfloor \frac{t-2}{2} \rfloor} L(C_j)\Big) \\
&\geq \Big(\sum_{j=1}^{\lfloor \frac{t-2}{2} \rfloor} N(C_j)^{\frac{1}{2}} L(C_j)^{\frac{1}{2}}\Big)^2 \qquad (2) \\
&\geq \Big(\sum_{j=1}^{\lfloor \frac{t-2}{2} \rfloor} \delta(C_j)^{\frac{1}{2}}\Big)^2 \quad \text{by Fact 2.6} \\
&> (\lfloor \frac{t-2}{2} \rfloor)^2 \cdot \varepsilon \cdot 2^i \cdot \mathsf{OPT}.
\end{aligned}$$

Here (2) is from Cauchy-Schwarz inequality. So we have $(\lfloor \frac{t-2}{2} \rfloor)^2 \cdot \varepsilon \cdot 2^i \cdot \mathsf{OPT} < 4 \cdot 2^i \cdot \mathsf{OPT}$, which implies $t \leq O(\varepsilon^{-\frac{1}{2}})$. □

Now we are ready to prove Theorem 2.1.

Proof: [of Theorem 2.1] We first verify that the set S is an $O(\varepsilon)$-coreset. Our goal is to prove that for every $c \in \mathbb{R}$, $f_S(c) \in (1 \pm \varepsilon) \cdot f_P(c)$. We prove this for any $c \in (-\infty, c^\star]$. The argument for $c \in (c^\star, +\infty)$ is entirely symmetric.

For any $c \in (-\infty, c^\star]$, we have

$$f_P(c) - f_S(c) = \sum_B \mathrm{cost}(B, c) - N(B) \cdot d(\mu(B), c)$$

where B takes over all buckets. We then separately analyze the $c \in (-\infty, p_L]$ case and the $c \in (p_L, c^*]$ case.

When $c \in (-\infty, p_L]$, we note that $f_P(p_L) = f_S(p_L)$(Lemma 2.7). By elementary calculus, both $\frac{df_P(c)}{dc}$ and $\frac{df_S(c)}{dc}$ are within $[-n, -(1-2\varepsilon)n]$; hence differ by at most a multiplicative factor of $1+\varepsilon$. Thus, $|f_P(c) - f_S(c)| \leq O(\varepsilon) \cdot f_P(c)$.

When $c \in (p_L, c^*]$, there is at most one bucket $B = \{p_l, \ldots, p_r\}$ such that $c \in (p_l, p_r)$ since these buckets are disjoint. If such a bucket B does not exist, we have $f_P(c) = f_S(c)$. Now suppose such a bucket B exists. Since $c > p_L$, we have $B \subset B_i$ for some block B_i. Thus, by Lemma 2.7 and the construction of buckets:

$$|f_P(c) - f_S(c)| \leq \delta(B) \leq \varepsilon \cdot 2^i \cdot \mathsf{OPT}.$$

We have $f_P(p_l) \geq 2^i \cdot \mathsf{OPT}$ and $f_P(p_r) \geq 2^i \cdot \mathsf{OPT}$. Since f_P is convex (thus decreasing on $(-\infty, c^*]$) and $c \in (p_l, p_r)$, we also have $f_P(c) \geq 2^i \cdot \mathsf{OPT}$. This implies $|f_P(c) - f_S(c)| \leq \varepsilon \cdot f_P(c)$.

It remains to show that the size of S, which is the total number of buckets, is $\tilde{O}(\varepsilon^{-1/2})$. However, by Lemma 2.8, there are $O(\log(1/\varepsilon))$ blocks, and by Lemma 2.9, each block contains $O(\varepsilon^{-1/2})$ buckets. Thus, there are at most $\tilde{O}(\varepsilon^{-1/2})$ buckets. □

2.2. Tight Lower Bound on Coreset Size for 1-d k-MEDIAN when $k \geq 2$

In this subsection, we prove that the size lower bound of ε-coreset for k-MEDIAN problem in $\mathbb{R}^1$ is $\Omega(\frac{k}{\varepsilon})$. This lower bound matches the upper bound in (Har-Peled & Kushal, 2005).

Theorem 2.10 (Coreset lower bound for 1-d k-MEDIAN when $k \geq 2$). *For a given integer $k \geq 2$ and $\varepsilon \in (0,1)$, there exists a dataset $P \subset \mathbb{R}$ such that any ε-coreset S must have size $|S| \geq \Omega(k\varepsilon^{-1})$.*

For ease of exposition, we only prove the lower bound for 2-MEDIAN here. The generalization to k-MEDIAN is straightforward and can be found in Appendix B.

We first prove a technical lemma, which shows that a quadratic function cannot be approximated well by an affine linear function in a long enough interval. We note that similar technical lemmas appear in coresets lower bound of other related clustering problems (Braverman et al., 2019) (Baker et al., 2020). The lemma in (Braverman et al., 2019) shows that the function $\sqrt{x}$ cannot be approximated well by an affine linear function while our lemma is about approximating a quadratic function. The lemma in (Baker et al., 2020) shows that a quadratic function cannot be approximated well by an affine linear function on a bounded interval, a situation slightly different from ours.

Lemma 2.11 (Quadratic function cannot be approximated well by affine linear functions). *Let $[a,b]$ be an interval, $f(c)$ be a quadratic function on interval $[a,b]$, $\alpha > 0$ and $\beta > 0$ be two constants, and $0 \leq \varepsilon < \frac{1}{32}\frac{\beta}{\alpha}$ be a non-negative real number. If $|f(c)| \leq \alpha$ and $(b-a)^2 f''(c) \geq \beta$ for all $c \in [a,b]$, then there is no affine linear function g such that $|g(c) - f(c)| \leq \varepsilon f(c)$ for all $c \in [a,b]$.*

Proof: Assume there is an affine linear function $g(c)$ that satisfies $|g(c) - f(c)| \leq \varepsilon f(c)$ for all $c \in [a,b]$. We denote the error function by $r(c) = f(c) - g(c)$, which has two properties. First, its l_∞ norm $\|r\|_\infty = \sup_{c\in[a,b]} |r(c)| \le \varepsilon\alpha$. Second, it is quadratic and satisfies $r''(c) = f''(c)$, thus $(b-a)^2 r''(c) \ge \beta$ for all $c \in [a,b]$.

Define $L = b - a$. By the mean value theorem, there is a point $c_{1/4} \in [a, \frac{a+b}{2}]$ such that $|r'(c_{1/4})| = |\frac{1}{L/2}[r(\frac{a+b}{2}) - r(a)]| \le \frac{4}{L}\|r\|_\infty$. Similarly there is a point $c_{3/4} \in [\frac{a+b}{2}, b]$ such that $|r'(c_{3/4})| \le \frac{4}{L}\|r\|_\infty$. Since r is a quadratic function, its derivative is monotonic and $|r'(\frac{a+b}{2})| \le \max(|r'(c_{1/4})|, |r'(c_{3/4})|) \le \frac{4}{L}\|r\|_\infty$. Thus we have

$$\begin{aligned} r(b) - r(\frac{a+b}{2}) &= \int_{\frac{a+b}{2}}^{b} r'(c)\mathrm{d}c \\ &= \int_{\frac{a+b}{2}}^{b} r'(\frac{a+b}{2}) + \int_{\frac{a+b}{2}}^{c} r''(t)\mathrm{d}t\mathrm{d}c \\ &= \frac{L}{2} r'(\frac{a+b}{2}) + \int_{\frac{a+b}{2}}^{b}\int_{\frac{a+b}{2}}^{c} r''(t)\mathrm{d}t\mathrm{d}c \\ &\ge -\frac{L}{2}\frac{4}{L}\|r\|_\infty + \frac{1}{8}(b-a)^2 r''(c) \\ &\ge -2\varepsilon\alpha + \frac{1}{8}\beta. \end{aligned}$$

On the other hand $r(b) - r(\frac{a+b}{2}) \le 2\|r\|_\infty \le 2\varepsilon\alpha$. We have $2\varepsilon\alpha \ge -2\varepsilon\alpha + \frac{1}{8}\beta$. Thus $\varepsilon \ge \frac{1}{32}\frac{\beta}{\alpha}$. □

For any dataset P, with a slight abuse of notations, we denote the cost function for 2-MEDIAN with one query point fixed in 0 by $f_P(c) = \mathrm{cost}(P, \{0, c\})$. The following lemma shows that $f_P(c)$ is a piecewise affine linear function and all the transition points are $P \cup \{2p \mid p \in P\}$.

Lemma 2.12 (The function $f_P(c)$ is piecewise affine linear). *Let $P \subset \mathbb{R}$ be a weighted dataset. The function $f_P(c)$ is a piecewise affine linear function. All the transition points between two affine pieces are $P \cup \{2p \mid p \in P\}$.*

Proof: We denote the weight of point p by $w(p)$ and denote the midpoint between any point c and 0 by $\mathrm{mid} = \frac{c}{2}$. Now assume $c \ge 0$ and both c and $\frac{c}{2}$ are not in the dataset P. The clustering cost of a single point p is

$$\mathrm{cost}(p, \{0, c\}) = \begin{cases} w(p)p & \text{for } p \in [0, \mathrm{mid}], \\ w(p)(c-p) & \text{for } p \in [\mathrm{mid}, c], \\ w(p)(p-c) & \text{for } p \in [c, +\infty). \end{cases}$$

If c changes to $c + \mathrm{d}c$ we have

$$\begin{aligned} &\mathrm{cost}(p, \{0, c+\mathrm{d}c\}) - \mathrm{cost}(p, \{0, c\}) \\ =&\begin{cases} 0 & \text{for } p \in [0, \mathrm{mid}], \\ w(p)\mathrm{d}c & \text{for } p \in [\mathrm{mid} + \frac{1}{2}\mathrm{d}c, c], \\ -w(p)\mathrm{d}c & \text{for } p \in [c + \mathrm{d}c, +\infty). \end{cases} \end{aligned}$$

Assume $|\mathrm{d}c|$ is small enough, then there are no data points in $[\mathrm{mid}, \mathrm{mid} + \frac{1}{2}\mathrm{d}c]$ and $[c, c + \mathrm{d}c]$. We have

$$\begin{aligned} &f_P(c + \mathrm{d}c) - f_P(c) \\ =& \sum_{p \in P\cap[\mathrm{mid}, c]} w(p)\mathrm{d}c - \sum_{p\in P\cap[c,+\infty)} w(p)\mathrm{d}c, \end{aligned}$$

thus

$$f'_P(c) = \sum_{p \in P\cap[\mathrm{mid}, c]} w(p) - \sum_{p\in P\cap[c,+\infty)} w(p).$$

Consider c moves in $\mathbb{R}$ from left to right, the derivative $f'_P(c)$ changes only when c or $\mathrm{mid} = \frac{c}{2}$ pass a data point in P. The same conclusion also holds for $c < 0$ by a symmetric argument. This is exactly what we want. □

Proof: [2-MEDIAN case of Theorem 2.10] We first construct the dataset P. The dataset P is a union of $\frac{1}{\varepsilon}$ disjoint intervals $\{I_i\}_{i=1}^{\frac{1}{\varepsilon}}$. Denote the left endpoint and right endpoint of I_i by l_i and r_i respectively. We recursively define $l_i = r_{i-1}$ for $i \ge 2$, $r_i = l_i + 4^{i-1}$ for $i \ge 1$, and $l_1 = 0$. Thus $r_i = l_{i+1} = \frac{1}{3}(4^i - 1)$. The weight of points is specified by a measure λ on P. The measure is absolutely continuous with respect to Lebesgue measure m such that its density on the ith interval is $\frac{\mathrm{d}\lambda}{\mathrm{d}m} = (\frac{1}{16})^{i-1}$. We denote the density on the ith interval by μ_i and the density at point p by $\mu(p)$. Note that P is a hard case with an infinite number of points. We can derive hard cases with a finite number of points by discretizing P, see the end of this proof.

The cost function $f_P(c)$ has following two features:

1. the function value $f_P(c) \in [0, \frac{2}{\varepsilon}]$ for any $c \in \mathbb{R}$,
2. the function is quadratic on the interval $[l_i + \frac{1}{3}(r_i - l_i), r_i]$ and satisfies $[\frac{2}{3}(r_i - l_i)]^2 f''_P(c) = \frac{2}{3}$ for each i.

We show how to prove theorem 2.10 from these features and defer verification of these features later. Note that feature 2 does not contradict lemma 2.12 since the dataset contains infinite points.

Assume that S is an $\frac{\varepsilon}{300}$-coreset of P. We prove $|S| \ge \frac{1}{2\varepsilon}$ by contradiction. If $|S| < \frac{1}{2\varepsilon}$, then there is an interval $I_i = [l_i, r_i]$ such that $(l_i, r_i) \cap S = \varnothing$ by the pigeonhole's principle. Consider function $f_S(c)$ on interval $[l_i + \frac{1}{3}(r_i - l_i), r_i]$. When $c \in [l_i + \frac{1}{3}(r_i - l_i), r_i]$, we have $\frac{c}{2} \in [l_i, r_i]$. Thus both c and $\frac{c}{2}$ do not pass points in S when c moves from $l_i + \frac{1}{3}(r_i - l_i)$ to r_i. By lemma 2.12, function $f_S(c)$ is affine linear on interval $[l_i + \frac{1}{3}(r_i - l_i), r_i]$. Since S is an $\frac{\varepsilon}{300}$-coreset of P, we have $|f_S(c) - f_P(c)| \le \frac{\varepsilon}{300} f_P(c)$ on interval $[l_i + \frac{1}{3}(r_i - l_i), r_i]$. However, by applying lemma 2.11 to $f_P(c)$ and $f_S(c)$ on interval $[l_+ \frac{1}{3}(r_i - l_i), r_i]$ with $\alpha = \frac{2}{\varepsilon}$

and $\beta = \frac{2}{3}$, we obtain that $\frac{\varepsilon}{300} \geq \frac{1}{32} \times \frac{2}{3} \times \frac{\varepsilon}{2} > \frac{\varepsilon}{300}$. This is a contradiction.

We now verify the two features of $f_P(c)$. We verify feature 1 by direct computations. For any point c, the function satisfies

$$\begin{aligned}0 \leq f_P(c) \leq \mathrm{cost}(P, \{0,0\}) &= \int_P p\mu(p)\mathrm{d}p \\ &\leq \sum_{i=1}^{\frac{1}{\varepsilon}} \lambda(I_i) r_i \leq \sum_{i=1}^{\frac{1}{\varepsilon}} (\frac{1}{4})^{i-1} \times 2 \times 4^{i-1} \\ &= \frac{2}{\varepsilon}.\end{aligned}$$

To verify feature 2, we compute the first order derivative by computing the change of the function value $f_P(c+\mathrm{d}c) - f_P(c)$ up to the first order term when c increases an infinitesimal number $\mathrm{d}c$. The unweighted clustering cost of a single point p is

$$\mathrm{cost}(p, \{0, c\}) = \begin{cases} p & \text{for } p \in [0, \mathrm{mid}], \\ c - p & \text{for } p \in [\mathrm{mid}, c], \\ p - c & \text{for } p \in [c, +\infty). \end{cases}$$

As c increases to $c+\mathrm{d}c$, the clustering cost of a single point changes

$$\begin{aligned}&\mathrm{cost}(p, \{0, c+\mathrm{d}c\}) - \mathrm{cost}(p, \{0, c\})] \\ =& \begin{cases} 0 & \text{for } p \in [0, \mathrm{mid}], \\ O(\mathrm{d}c) & \text{for } p \in [\mathrm{mid}, \mathrm{mid} + \frac{1}{2}\mathrm{d}c], \\ \mathrm{d}c & \text{for } p \in [\mathrm{mid} + \frac{1}{2}\mathrm{d}c, c], \\ O(\mathrm{d}c) & \text{for } p \in [c, c+\mathrm{d}c], \\ -\mathrm{d}c & \text{for } p \in [c+\mathrm{d}c, +\infty). \end{cases}\end{aligned}$$

The cumulative clustering cost changes

$$\begin{aligned}& f_P(c+\mathrm{d}c) - f_P(c) \\ =& \int_0^{+\infty} \mathrm{cost}(p, \{0, c+\mathrm{d}c\}) - \mathrm{cost}(p, \{0, c\}) \mathrm{d}\lambda \\ =& \int_0^{\mathrm{mid}} 0\mathrm{d}\lambda + \int_{\mathrm{mid}}^{\mathrm{mid}+\frac{1}{2}\mathrm{d}c} O(\mathrm{d}c)\mathrm{d}\lambda + \int_{\mathrm{mid}+\frac{1}{2}\mathrm{d}c}^{c} \mathrm{d}c\mathrm{d}\lambda \\ &+ \int_c^{c+\mathrm{d}c} O(\mathrm{d}c)\mathrm{d}\lambda + \int_{c+\mathrm{d}c}^{+\infty} -\mathrm{d}c\mathrm{d}\lambda \\ =& \lambda([\mathrm{mid}, c])\mathrm{d}c - \lambda([c, +\infty))\mathrm{d}c + O(\mathrm{d}c)^2.\end{aligned}$$

Thus the first order derivative $f'_P(c) = \lambda([\frac{c}{2}, c]) - \lambda([c, +\infty))$ and the second order derivative

$$\begin{aligned}f''_P(c) &= \frac{\mathrm{d}}{\mathrm{d}c}\big(\lambda([\frac{c}{2}, c]) - \lambda([c, +\infty))\big), \\ &= 2\mu(c) - \frac{1}{2}\mu(\frac{c}{2}).\end{aligned}$$

For $c \in [l_i + \frac{1}{3}(r_i - l_i), r_i]$, the two points c and $\frac{c}{2}$ both lie in interval $[l_i, r_i]$. We have $\mu(c) = \mu(\frac{c}{2}) = \mu_i$ and $f''_P(c) = \frac{3}{2}\mu_i$. Thus the function $f_P(c)$ is quadratic on $[l_i + \frac{1}{3}(r_i - l_i), r_i]$ and $[\frac{2}{3}(r_i - l_i)]^2 f''_P(c) = \frac{2}{3}$.

The only remaining part is constructing hard cases with a finite number of points from P. We construct P_n from P for each n by creating a bucket B_i of $(\frac{1}{4})^{i-1}n$ equally spaced points in each interval I_i and assigning weight $\frac{1}{n}$ to every point. We show that P_n is a hard case for n large enough. We assume that each P_n has a 300ε-coreset of size less than $\frac{1}{4\varepsilon}$ and prove that it leads to a contradiction. The key is the following proposition: $\lim_{n\to+\infty} \|f_{P_n} - f_P\|_\infty = 0$.

To see this, denote the point $l_i + \frac{j}{(1/4)^{i-1}n}(r_i - l_i)$ by $p_{i,j}$ for $j \in \{0, \ldots, (1/4)^{i-1}n - 1\}$. We have $I_i = \coprod_{j=0}^{(1/4)^{i-1}n}[p_{i,j}, p_{i,j+1})$ and $B_i = \cup_{j=0}^{(1/4)^{i-1}n} p_{i,j}$. For any center $c \in \mathbb{R}$, we have

$$\begin{aligned}f_P(c) - f_{P_n}(c) = & \sum_{i=1}^{\frac{1}{\varepsilon}} \sum_{j=1}^{(\frac{1}{4})^{i-1}n} \mathrm{cost}([p_{i,j-1}, p_{i,j}), \{0, c\}) \\ & - \sum_{i=1}^{\frac{1}{\varepsilon}} \sum_{j=1}^{(\frac{1}{4})^{i-1}n} \mathrm{cost}(p_{i,j-1}, \{0, c\}).\end{aligned}$$

Since $\lambda([p_{i,j-1}, p_{i,j})) = 1/n = w(p_{i,j-1})$, we have

$$\begin{aligned}& |\mathrm{cost}([p_{i,j-1}, p_{i,j}), \{0, c\}) - \mathrm{cost}(p_{i,j-1}, \{0, c\}) \\ \leq & \frac{1}{n}\frac{r_i - l_i}{(1/4)^{i-1}n}.\end{aligned}$$

Thus we have

$$\begin{aligned}&\sup_c |f_P(c) - f_{P_n}(c)| \\ \leq & \sum_{i=1}^{\frac{1}{\varepsilon}} \sum_{j=1}^{(\frac{1}{4})^{i-1}n} \frac{1}{n}\frac{r_i - l_i}{(1/4)^{i-1}n} = \frac{1}{n}\sum_{i=1}^{\frac{1}{\varepsilon}}(r_i - l_i).\end{aligned}$$

The proposition then follows from $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{\frac{1}{\varepsilon}}(r_i - l_i) = 0$.

Now, for each P_n, denote its 300ε-coreset of size less than $\frac{1}{4\varepsilon}$ by S_n. By the definition of coreset, we have $\sup_c[|f_{P_n}(c) - f_{S_n}(c)| - 300\varepsilon f_{P_n}(c)] \leq 0$. Using the proposition we have $\limsup_{n\to+\infty}\sup_c[|f_P(c) - f_{S_n}(c)| - 300\varepsilon f_P(c)] \leq 0$. Take a large enough n, we have $|f_P(c) - f_{S_n}(c)| \leq 301\varepsilon f_P(c)$ for any c and S_n is a 301ε-coreset of P. However, the set S_n contains less than $\frac{1}{4\varepsilon}$ points, which is impossible since we have proved that any 300ε-coreset of P contains at least $\frac{1}{2\varepsilon}$ points. This leads to a contradiction. $\square$

3. Improve Coreset Sizes when $2 \le d \le \varepsilon^{-2}$

In this section, we consider the case of constant d, $2 \le d \le \varepsilon^{-2}$, and provide several improved coreset bounds for a general problem of Euclidean k-MEDIAN, called Euclidean (k, z)-CLUSTERING. The only difference from k-MEDIAN is that the goal is to find a k-center set $C \subset \mathbb{R}^d$ that minimizes the objective function

$$\mathrm{cost}_z(P, C) := \sum_{p \in P} d^z(p, C) = \sum_{p \in P} \min_{c \in C} d^z(p, c), \quad (3)$$

where d^z represents the z-th power of the Euclidean distance. The coreset notion is as follows.

Definition 3.1 (**ε-Coreset for Euclidean (k, z)-CLUSTERING** (Har-Peled & Mazumdar, 2004)). *Given a dataset $P \subset \mathbb{R}^d$ of n points, an integer $k \ge 1$, constant $z \ge 1$ and $\varepsilon \in (0, 1)$, an ε-coreset for Euclidean (k, z)-CLUSTERING is a subset $S \subseteq P$ with weight $w : S \to \mathbb{R}_{\ge 0}$, such that*

$$\forall C \in \mathcal{C}, \qquad \sum_{p \in S} w(p) \cdot d^z(p, C) \in (1 \pm \varepsilon) \cdot \mathrm{cost}_z(P, C).$$

3.1. Improved Coreset Size in $\mathbb{R}^d$ when $k = 1$

We prove the following main theorem for $k = 1$ whose center is a point $c \in \mathbb{R}^d$.

Theorem 3.2 (**Coreset for Euclidean $(1, z)$-CLUSTERING**). *Let integer $d \ge 1$, constant $z \ge 1$ and $\varepsilon \in (0, 1)$. There exists a randomized polynomial time algorithm that given a dataset $P \subset \mathbb{R}^d$, outputs an ε-coreset for Euclidean $(1, z)$-CLUSTERING of size at most $z^{O(z)} \sqrt{d} \varepsilon^{-1} \log \varepsilon^{-1}$.*

The proof can be found in Appendix A. The above theorem is powerful and leads to the following results for $z = O(1)$.

1. By dimension reduction as in (Huang & Vishnoi, 2020; Cohen-Addad et al., 2021; 2022), we can assume $d = O(\varepsilon^{-2} \log \varepsilon^{-1})$. Consequently, our coreset size is upper bounded by $\tilde{O}(\varepsilon^{-2})$, which matches the nearly tight bound in (Cohen-Addad et al., 2022).

2. For $d = O(1)$, our coreset size is $O(\varepsilon^{-1})$, which is the first known result in small dimensional space. Specifically, the prior known coreset size in $\mathbb{R}^2$ is $\tilde{O}(\varepsilon^{-3/2})$ (Braverman et al., 2022), and our result improves it by a factor of $\varepsilon^{-1/2}$.

3. All results previously stated in Remark 2.2 hold. We can still convert an off-line algorithm to a streaming algorithm in a black-box manner. We can also answer any query in time independent of the point numbers. We further emphasize that a small coreset can automatically accelerate a PTAS. Suppose the time complexity of the PTAS is $O(f(n)g(\varepsilon))$, replacing the original dataset with our coreset immediately improves the time complexity of the PTAS to $O(f(\tilde{O}(\sqrt{d}\varepsilon^{-1}))g(\varepsilon))$. The time required for the coreset construction is usually negligible compared to the running time of the PTAS.

We conjecture that our coreset size is almost tight, i.e., there exists a coreset lower bound $\Omega(\sqrt{d}\varepsilon^{-1})$ for constant $2 \le d \le \varepsilon^{-2}$, which leaves as an interesting open problem.

3.2. Improved Coreset Lower Bound in $\mathbb{R}^d$ when $k \ge 2$

We present a lower bound for the coreset size in small dimensional spaces.

Theorem 3.3 (**Coreset lower bound in small dimensional spaces**). *Given an integer $k \ge 1$, constant $z \ge 1$ and a real number $\varepsilon \in (0, 1)$, for any integer $d \le \frac{1}{100\varepsilon^2}$, there is a dataset $P \subset \mathbb{R}^{d+1}$ such that its ε-coreset for (k, z)-CLUSTERING must contain at least $\frac{dk}{10z^4}$ points.*

When $d = \Theta(\frac{1}{\varepsilon^2})$, Theorem 3.3 gives the well known lower bound $\frac{k}{\varepsilon^2}$. When $d \ll \Theta(\frac{1}{\varepsilon^2})$, the theorem is non-trivial.

4. Conclusion

This work studies coresets for k-MEDIAN problem in small dimensional Euclidean spaces. We give tight size bounds for k-MEDIAN in $\mathbb{R}$ and show that the framework in (Har-Peled & Kushal, 2005), with significant improvement, is optimal. For $d \ge 2$, we improve existing coreset upper bounds for 1-MEDIAN and prove new lower bounds.

Our work leaves several interesting problems for future research. One of which is to close the gap between upper bounds and lower bounds for $d \ge 2$. Another one is to generalize our results to (k, z)-CLUSTERING for general z. Note that the generalization is non-trivial even for $d = 1$ since the cost function is piece-wise linear for k-MEDIAN while piece-wise polynomial of order z for general (k, z)-CLUSTERING.

Acknowledgement

Zengfeng Huang is supported by National Natural Science Foundation of China (No. 62276066, No. U2241212).

References

Arnaboldi, V., Conti, M., Passarella, A., and Pezzoni, F. Analysis of ego network structure in online social networks. *2012 International Conference on Privacy, Secu-*

rity, Risk and Trust and 2012 International Confernece on Social Computing, pp. 31–40, 2012.

Arthur, D. and Vassilvitskii, S. k-means++: the advantages of careful seeding. In *SODA*, pp. 1027–1035, 2007.

Baker, D. N., Braverman, V., Huang, L., Jiang, S. H., Krauthgamer, R., and Wu, X. Coresets for clustering in graphs of bounded treewidth. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 569–579. PMLR, 2020.

Banaszczyk, W. Balancing vectors and gaussian measures of n-dimensional convex bodies. *Random Struct. Algorithms*, 12(4):351–360, 1998.

Bansal, N., Dadush, D., Garg, S., and Lovett, S. The gram-schmidt walk: A cure for the banaszczyk blues. *Theory Comput.*, 15:1–27, 2019.

Braverman, V., Jiang, S. H.-C., Krauthgamer, R., and Wu, X. Coresets for ordered weighted clustering. In *International Conference on Machine Learning*, 2019.

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

Braverman, V., Cohen-Addad, V., Jiang, S., Krauthgamer, R., Schwiegelshohn, C., Toftrup, M. B., and Wu, X. The power of uniform sampling for coresets. In *62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2022*. IEEE Computer Society, 2022.

Coates, A. and Ng, A. Y. Learning feature representations with k-means. In *Neural Networks: Tricks of the Trade - Second Edition*, pp. 561–580. 2012.

Cohen, M. B., Lee, Y. T., Miller, G. L., Pachocki, J. W., and Sidford, A. Geometric median in nearly linear time. *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, 2016.

Cohen-Addad, V., Saulpic, D., and Schwiegelshohn, C. Improved coresets and sublinear algorithms for power means in Euclidean spaces. In *Neural Information Processing Systems*, 2021.

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

Cohen-Addad, V., Larsen, K. G., Saulpic, D., and Schwiegelshohn, C. Towards optimal lower bounds for k-median and k-means coresets. In *Proceedings of the firty-fourth annual ACM symposium on Theory of computing*, 2022.

Cohen-Addad, V., Larsen, K. G., Saulpic, D., Schwiegelshohn, C., and Sheikh-Omar, O. A. Improved coresets for Euclidean k-means. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.

Feldman, D. and Langberg, M. A unified framework for approximating and clustering data. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 569–578. ACM, 2011.

Feldman, D., Schmidt, M., and Sohler, C. 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*, pp. 1434–1453. SIAM, 2013.

Fonseca-Rodríguez, O., Gustafsson, P. E., Sebastiån, M. S., and Connolly, A.-M. F. Spatial clustering and contextual factors associated with hospitalisation and deaths due to COVID-19 in Sweden: a geospatial nationwide ecological study. *BMJ Global Health*, 6, 2021.

Grønlund, A., Larsen, K. G., Mathiasen, A., Nielsen, J. S., Schneider, S., and Song, M. Fast exact k-means, k-medians and bregman divergence clustering in 1d, 2018.

Har-Peled, S. and Kushal, A. Smaller coresets for k-median and k-means clustering. *Discrete & Computational Geometry*, 37:3–19, 2005.

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

Huang, L. and Vishnoi, N. K. Coresets for clustering in Euclidean spaces: importance sampling is nearly optimal. In Makarychev, K., Makarychev, Y., Tulsiani, M., Kamath, G., and Chuzhoy, J. (eds.), *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, pp. 1416–1429. ACM, 2020.

Huang, L., Jiang, S. H.-C., Li, J., and Wu, X. Epsilon-coresets for clustering (with outliers) in doubling metrics. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pp. 814–825. IEEE Computer Society, 2018.

Huang, L., Sudhir, K., and Vishnoi, N. K. Coresets for time series clustering. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 22849–22862, 2021.

Huang, L., Li, J., and Wu, X. On optimal coreset construction for euclidean (k, z)-clustering. *ArXiv*, abs/2211.11923, 2022.

Huang, L., Jiang, S., Li, J., and Wu, X. Coresets for clustering with general assignment constraints. *CoRR*, abs/2301.08460, 2023a.

Huang, L., Jiang, S. H.-C., Lou, J., and Wu, X. Near-optimal coresets for robust clustering. In *The Eleventh International Conference on Learning Representations*, 2023b.

Jeske, O., Jogler, M., Petersen, J., Sikorski, J., and Jogler, C. From genome mining to phenotypic microarrays: Planctomycetes as source for novel bioactive molecules. *Antonie van Leeuwenhoek*, 104:551–567, 2013.

Karnin, Z. S. and Liberty, E. Discrepancy, coresets, and sketches in machine learning. In Beygelzimer, A. and Hsu, D. (eds.), *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1975–1993. PMLR, 2019.

Kunegis, J., Schmidt, S., Lommatzsch, A., Lerner, J., Luca, E. W. D., and Albayrak, S. Spectral analysis of signed graphs for clustering, prediction and visualization. In *SDM*, 2010.

Langberg, M. and Schulman, L. J. Universal ε-approximators for integrals. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pp. 598–607. SIAM, 2010.

Lloyd, S. P. Least squares quantization in PCM. *IEEE Trans. Information Theory*, 28(2):129–136, 1982.

Munteanu, A. and Schwiegelshohn, C. Coresets-methods and history: A theoreticians design pattern for approximation and streaming algorithms. *KI-Künstliche Intelligenz*, 32:37–53, 2018.

Narantsatsralt, U.-U. and Kang, S. Social network community detection using agglomerative spectral clustering. *Complex.*, 2017:3719428:1–3719428:10, 2017.

Pennacchioli, D., Coscia, M., Rinzivillo, S., Giannotti, F., and Pedreschi, D. The retail market as a complex system. *EPJ Data Science*, 3:1–27, 2014.

Phillips, J. M. and Tai, W. M. Improved coresets for kernel density estimates. In Czumaj, A. (ed.), *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018*, pp. 2718–2727. SIAM, 2018a.

Phillips, J. M. and Tai, W. M. Near-optimal coresets of kernel density estimates. In Speckmann, B. and Tóth, C. D. (eds.), *34th International Symposium on Computational Geometry, SoCG 2018, June 11-14, 2018, Budapest, Hungary*, volume 99 of *LIPIcs*, pp. 66:1–66:13. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018b.

Tai, W. M. Optimal coreset for Gaussian kernel density estimation. In Goaoc, X. and Kerber, M. (eds.), *38th International Symposium on Computational Geometry, SoCG 2022, June 7-10, 2022, Berlin, Germany*, volume 224 of *LIPIcs*, pp. 63:1–63:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022.

Tan, P.-N., Steinbach, M., Kumar, V., et al. Cluster analysis: basic concepts and algorithms. *Introduction to data mining*, 8:487–568, 2006.

Von Luxburg, U. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Wheeler, D. C. A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio, 1996 – 2003. *International Journal of Health Geographics*, 6:13 – 13, 2007.

Zhang, Y., Levina, E., and Zhu, J. Detecting overlapping communities in networks using spectral methods. *SIAM J. Math. Data Sci.*, 2:265–283, 2014.

A. Proof of Theorem 3.2

A.1. Useful Notations and Facts

For preparation, we first propose a notion of mixed coreset (Definition A.1), and then introduce some known discrepancy results.

Reduction to mixed coreset. Let $B(a, r)$ denote the ℓ_2-ball in $\mathbb{R}^d$ that centers at $a \in \mathbb{R}^d$ with radius $r \geq 0$. Specifically, $B(0, 1)$ is the unit ball centered at the original point.

Definition A.1 (Mixed coreset for Euclidean $(1, z)$-CLUSTERING). *Given a dataset $P \subset B(0, 1)$ and $\varepsilon \in (0, 1)$, an ε-mixed-coreset for Euclidean $(1, z)$-CLUSTERING is a subset $S \subseteq P$ with weight $w : S \to \mathbb{R}_{\geq 0}$, such that $\forall c \in \mathbb{R}^d$,*

$$\sum_{p \in S} w(p) \cdot d^z(p, c) \in \mathrm{cost}_z(P, c) \pm \varepsilon \max\{1, \|c\|_2\}^z \cdot |P|. \tag{4}$$

Actually, prior work (Cohen-Addad et al., 2021; 2022; Braverman et al., 2022) usually consider the following form: $\forall c \in \mathbb{R}^d$,

$$\sum_{p \in S} w(p) \cdot d^z(p, c) \in (1 \pm \varepsilon) \cdot \mathrm{cost}_z(P, c) \pm \varepsilon |P|. \tag{5}$$

Compared to Definition 1.1, the above inequality allows both a multiplicative error $\varepsilon \cdot \mathrm{cost}_z(P, c)$ and an additional additive error $\varepsilon|P|$. Note that for a small $r = O(1)$, the additive error $\varepsilon|P|$ dominates the total error; while for a large $r \gg \Omega(1)$, the multiplicative error $\varepsilon \cdot \mathrm{cost}_z(P, c) \approx \varepsilon \|c\|_2 \cdot |P|$ dominates the total error. Hence, it is not hard to check that Inequality (5) is an equivalent form of Inequality (4) (up to an $2^{O(z)}$-scale). This is also the reason that we call Definition A.1 mixed coreset. We have the following useful reduction.

Theorem A.2 (Reduction from coreset to mixed coreset (Braverman et al., 2022)). *Let $\varepsilon \in (0, 1)$. Suppose there exists a polynomial time algorithm A that constructs an ε-mixed coreset for Euclidean $(1, z)$-CLUSTERING of size Γ. Then there exists a polynomial time algorithm A' that constructs an ε-coreset for Euclidean $(1, z)$-CLUSTERING of size $O(\Gamma \log \varepsilon^{-1})$.*

Thus, it suffices to prove that an ε-mixed coreset is of size $z^{O(z)}\sqrt{d}\varepsilon^{-1}$, which implies Theorem 3.2.

Class discrepancy. For preparation, we introduce the notion of class discrepancy introduced by (Karnin & Liberty, 2019). The idea of combining discrepancy and coreset construction has been studied in the literature, specifically for kernel density estimation (Phillips & Tai, 2018a;b; Karnin & Liberty, 2019; Tai, 2022). We propose the following definition.

Definition A.3 (Class discrepancy (Karnin & Liberty, 2019)). *Let $m \geq 1$ be an integer. Let $f : \mathcal{X}, \mathcal{C} \to \mathbb{R}$ and $P \subseteq \mathcal{X}$ with $|P| = m$. The class discrepancy of of P w.r.t. $(f, \mathcal{C})$ is*

$$\begin{aligned} D_P^{(\mathcal{C})}(f) &:= \min_{\sigma \in \{-1,1\}^P} D_P^{(\mathcal{C})}(f, \sigma) \\ &= \min_{\sigma \in \{-1,1\}^P} \max_{c \in \mathcal{C}} \frac{1}{m} \left| \sum_{p \in P} \sigma_p \cdot f(p, c) \right|. \end{aligned}$$

Moreover, we define $D_m^{(\mathcal{X},\mathcal{C})}(f) := \max_{P \subseteq \mathcal{X} : |P| = m} D_P^{(\mathcal{C})}(f)$ to be the class discrepancy w.r.t. $(f, \mathcal{X}, \mathcal{C})$.

Here, $\mathcal{X}$ is the instance space and $\mathcal{C}$ is the parameter space. Specifically, for Euclidean $(1, z)$-CLUSTERING, we let $\mathcal{X}, \mathcal{C} \subseteq \mathbb{R}^d$ and f be the Euclidean distance. The class discrepancy $D_m^{(\mathcal{X},\mathcal{C})}(f)$ measures the capacity of $\mathcal{C}$. Intuitively, if the capacity of $\mathcal{C}$ is large and leads to a complicated geometric structure of vector $(f(p, c))_{p \in P}$ for $c \in \mathcal{C}$, $D_m^{(\mathcal{X},\mathcal{C})}(f)$ tends to be large.

Useful discrepancy results. For a vector $p \in \mathbb{R}^d$ and integer $l \geq 1$, let $p^{\otimes l}$ present the l-dimensional tensor obtained from the outer product of p with itself l times. For a l-dimensional tensor X with d^l entries, we consider the measure $\|X\|_{T_l} := \max_{c \in \mathbb{R}^d : \|q\|=1} |\langle X, q^{\otimes l} \rangle|$. Next, we provide some known results about the class discrepancy.

Theorem A.4 (An upper bound for class discrepancy (restatement of Theorem 18 of Karnin & Liberty (2019))). *Let $\mathcal{X} = B(0,1)$ in $\mathbb{R}^d$. Let $f : \mathbb{R} \to \mathbb{R}$ be analytic satisfying that for any integer $l \geq 1$, $|\frac{d^l f}{dx^l}(x)| \leq \gamma_1 C^l l!$ for some constant $\gamma_1, C > 0$. Let $\mathcal{C} = B(0, \frac{1}{2C})$ and $m \geq 1$ be an integer. The class discrepancy w.r.t. $(f = f(\langle p, c\rangle), \mathcal{X}, \mathcal{C})$ is at most $D_m^{(\mathcal{X},\mathcal{C})}(f) \leq \gamma_2\gamma_1\sqrt{d}/m$ for some constant $\gamma_2 > 0$.*

$\cdot$ *Moreover, for any dataset $P \subset \mathcal{X}$ of size m, there exists a randomized polynomial time algorithm that constructs $\sigma \in \{-1,1\}^P$ satisfying that for any integer $l \geq 1$, we have*

$$\|\sum_{p\in P} \sigma_p \cdot p^{\otimes l}\|_{T_l} = O(\sqrt{dl \log^3 l}).$$

This σ satisfies $D_P^{(\mathcal{C})}(f,\sigma) \leq \gamma_2\gamma_1\sqrt{d}/m$.

Note that the above theorem is a constructive result instead of an existential result in Theorem 18 of (Karnin & Liberty, 2019). This is because Theorem 18 of (Karnin & Liberty, 2019) applies the existential version of Banaszczyk's theorem (Banaszczyk, 1998), which has been proven to be constructive recently (Bansal et al., 2019). Also, note that the construction of σ only depends on P and does not depend on the selection of $\mathcal{C}$. This observation is important for the construction of mixed coresets via discrepancy.

A.2. Proof of Theorem 3.2

We are ready to prove Theorem 3.2. The main lemma is as follows.

Lemma A.5 (Class discrepancy for Euclidean $(1,z)$-CLUSTERING). *Let $m \geq 1$ be an integer. Let $f = d^z$ and $\mathcal{X} = B(0,1)$. For a given dataset $P \subset \mathcal{X}$ of size m, there exists a vector $\sigma \in \{-1,1\}^P$ such that for any $r > 0$,*

$$D_P^{(B(0,r))}(f,\sigma) \leq z^{O(z)} \max\{1,r\}^z \cdot \sqrt{d}/m.$$

The above lemma indicates that the class discrepancy for Euclidean $(1,z)$-CLUSTERING linearly depends on the radius r of the parameter space $\mathcal{C}$. Note that the lemma finds a vector σ that satisfies all levels of parameter spaces $\mathcal{C} = B(0,r)$ simultaneously. This requirement is slightly different from Definition A.3 that considers a fixed parameter space. Observe that the term $\max\{1,r\}$ is similar to $\max\{1, \|c\|_2\}$ in Definition A.1, which is the key of reduction from Lemma A.5 to Theorem 3.2. The proof idea is similar to that of Fact 6 of (Karnin & Liberty, 2019).

Proof: [of Theorem 3.2] Let $P \subset B(0,1)$ be a dataset of size n and $\Lambda = z^{O(z)}\sqrt{d}\varepsilon^{-1}$. By the same argument as in Fact 6 of (Karnin & Liberty, 2019), we can iteratively applying Lemma A.5 to construct a subset $S \subseteq P$ of size $m = \Theta(\Lambda)$ together with weights $w(p) = \frac{n}{|S|}$ for $p \in S$ and a vector $\sigma \in \{-1,1\}^S$, and (S,σ) satisfies that for any $c \in \mathbb{R}^d$,

$$\begin{aligned} & \left|\sum_{p\in S} w(p)\cdot d(p,c) - \mathrm{cost}_z(P,c)\right| \\ \leq \quad & n \cdot D_S^{(B(0,\|c\|_2))}(f,\sigma) \\ \leq \quad & \varepsilon \max\{1,\|c\|_2\} \cdot n. \end{aligned}$$

This implies that S is an $O(\varepsilon)$-mixed coreset for Euclidean $(1,z)$-CLUSTERING of size at most $\Lambda = z^{O(z)}\sqrt{d}\varepsilon^{-1}$, which completes the proof of Theorem 3.2. □

It remains to prove Lemma A.5.

Proof: [of Lemma A.5] Let $P \subset B(0,1)$ be a dataset of size m. We first construct a vector $\sigma \in \{-1,1\}^P$ by the following way:

1. For each $p \in P$, construct a point $\phi(p) = (\frac{1}{2}\|p\|_2^2, \frac{\sqrt{2}}{2}p, \frac{1}{2}) \in \mathbb{R}^{d+2}$.

2. By Theorem A.4, construct $\sigma \in \{-1,1\}^P$ such that for any integer $l \geq 1$,

$$\|\sum_{p\in P} \sigma_p \cdot \phi(p)^{\otimes l}\|_{T_l} = O(\sqrt{(d+2)l\log^3 l}).$$

Let $\phi(P)$ be the collection of all $\phi(p)$s. Note that $\|\phi(p)\|_2 \leq 1$ by construction, which implies that $\phi(P) \subset B(0,1) \subset \mathbb{R}^{d+2}$. In the following, we show that σ satisfies Lemma A.5.

Fix $r \geq 1$ and let $\mathcal{C} = B(0,r)$. We construct another dataset $P' = \left\{p' = \frac{p}{4r} : p \in P\right\}$. For any $c \in \mathcal{C} = B(0,r)$, we let $c' = \frac{c}{4r} \in B(0,\frac{1}{4})$. By definition, we have for any $p \in \mathcal{X}$ and $c \in \mathcal{C}$,

$$\frac{1}{m}\left|\sum_{p\in P} \sigma_p \cdot f(p,c)\right| = \frac{(4r)^z}{m}\left|\sum_{p'\in P'} \sigma_p \cdot f(p',c')\right|,$$

which implies that

$$D_P^{(\mathcal{C})}(f,\sigma) = (4r)^z \cdot D_{P'}^{(B(0,\frac{1}{4}))}(f,\sigma).$$

Thus, it suffices to prove that

$$D_{P'}^{(B(0,\frac{1}{4}))}(f,\sigma) \leq z^{O(z)}\sqrt{d}/m, \tag{6}$$

which implies the lemma. The proof idea of Inequality (6) is similar to that of Theorem 22 of (Karnin & Liberty, 2019).[1] For each $p' \in P'$ and $c' \in B(0,\frac{1}{4})$, let $\psi(c') = (\frac{1}{8r^2}, -\frac{\sqrt{2}}{2r}c', 2\|c'\|_2^2) \in \mathbb{R}^{d+2}$ and we can rewrite $f(p',c')$ as follows:

$$f(p',c') = \|p'-c'\|_2^z = (\langle \phi(p), \psi(c')\rangle)^{z/2}.$$

We note that $\phi(p) \in B(0,1)$ and $\psi(c') \in B(0,\frac{1}{3})$ since $c' \in B(0,\frac{1}{4})$. Construct another function $g : P \times B(0,\frac{1}{3})$ as follows: for each $p \in P$ and $c \in B(0,\frac{1}{3})$,

1. If for any $p' \in P$, $\langle p', c\rangle \geq 0$, let $g(p,c) = g(\langle p,c\rangle) = (\langle p,c\rangle)^{z/2}$;
2. Otherwise, let $g(p,c) = 0$.

We have $|\frac{d^l g}{dx^l}(x)| \leq z^{O(z)} l!$ for any integer $l \geq 1$. By the construction of σ and Theorem A.4, we have that

$$D_{\phi(P)}^{(B(0,\frac{1}{3}))}(g,\sigma) \leq z^{O(z)}\sqrt{d}/m,$$

which implies Inequality (6) since $D_{P'}^{(B(0,\frac{1}{4}))}(f,\sigma) \leq D_{\phi(P)}^{(B(0,\frac{1}{3}))}(g,\sigma)$ due to the fact that $\psi(c') \in B(0,\frac{1}{3})$.

Overall, we complete the proof. □

B. Coreset Lower Bound for General k-MEDIAN in $\mathbb{R}$

We prove the general case of Theorem 2.10 here.

Proof: [the general case of Theorem 2.10]

We first construct the hard instance P. Let P_1 denote the hard instance we have constructed in the proof of Theorem 2.10. We take a large enough constant $L > 0$, take $P_i = (i-1)L + P_1$, and take $P = \cup_{i=1}^{\frac{k}{2}} P_i$. Here $(i-1)L + P_1$ means $\{(i-1)L + p | p \in P_1\}$.

[1]Note that the proof of Theorem 22 of (Karnin & Liberty, 2019) is actually incorrect. Applying Theorem 18 of (Karnin & Liberty, 2019) may lead to an upper bound $\|\tilde{q}\|_2 < 1$, which makes R in Theorem 22 of (Karnin & Liberty, 2019) not exist.

The dataset P is a unification of $\frac{k}{2}$ copies of P_1. These copies are far from each other. Thus k-MEDIAN problem on P can be decomposed to 2-MEDIAN problem on each copy. We prove the k-MEDIAN lower bound by applying the argument for the 2-MEDIAN lower bound on every single copy and combining them together.

We denote $P_1 = \cup_{j=1}^{\frac{1}{\varepsilon}} I_{1,j}$, where $I_{1,j}$ is the j-th interval we constructed in the proof of the 2-MEDIAN case of Theorem 2.10. We denote $I_{i,j} = (i-1)L + I_{1,j}$, denote the left endpoint and right endpoint of $I_{i,j}$ by $l_{i,j}$ and $r_{i,j}$ respectively. We have $P_i = \cup_{j=1}^{\frac{1}{\varepsilon}} I_{i,j}$.

Now, assume that S is an $\frac{\varepsilon}{300}$ coreset of P such that $|S| < \frac{k}{4\varepsilon}$. We prove that there must be a contradiction. Since $|S| < \frac{k}{4\varepsilon}$, there must be at least half of i such that $(l_{i,j_i}, r_{i,j_i}) \cap S = \varnothing$ for some j_i. We assume that these indexes are $1, 2, \ldots, \frac{k}{4}$, without loss of generality. We define a parametrized query family as $Q(t) = \cup_{i=1}^{\frac{k}{2}} Q_i(t)$, where $t \in [\frac{1}{3}, 1]$ and

$$Q_i(t) = \begin{cases} \{l_{i,1}, l_{i,j_i} + t(r_{i,j_i} - l_{i,j_i}), r_{i,j_i}\} & \text{for } i \le \frac{k}{4}, \\ \{l_{i,1}\} & \text{otherwise.} \end{cases}$$

Consider $\mathrm{cost}(P, Q(t))$, a function of t. Since L is large enough, we have $\mathrm{cost}(P, Q(t)) = \sum_{i=1}^{\frac{k}{2}} \mathrm{cost}(P_i, Q_i(t))$. The computation we have done in the proof of the 2-MEDIAN case of Theorem 2.10 implies that $\mathrm{cost}(P_i, Q_i(t)) \le \frac{2}{\varepsilon}$ for each i and

$$(1 - \frac{1}{3})^2 \frac{\mathrm{d}^2}{\mathrm{d}t^2} \mathrm{cost}(P_i, Q_i(t)) = \begin{cases} \frac{4}{9} & \text{for } i \le \frac{k}{4}, \\ 0 & \text{otherwise.} \end{cases}$$

Thus we have $\mathrm{cost}(P, Q(t)) \le \frac{k}{\varepsilon}$ and $(1 - \frac{1}{3})^2 \frac{\mathrm{d}^2}{\mathrm{d}t^2} \mathrm{cost}(P, Q(t)) = \frac{k}{9}$.

It's easy to see that $\mathrm{cost}(S, Q(t))$ is affine linear since $(l_{i,j_i}, r_{i,j_i}) \cap S = \varnothing$ for $i \le \frac{k}{4}$. Since S is an $\frac{\varepsilon}{300}$ coreset, we have $|\mathrm{cost}(S, Q(t)) - \mathrm{cost}(P, Q(t))| \le \frac{\varepsilon}{300} \mathrm{cost}(P, Q(t))$. By Lemma 2.11, we must have $\frac{\varepsilon}{300} \ge \frac{1}{32} \frac{\varepsilon}{k} \frac{k}{9} > \frac{\varepsilon}{300}$, which leads to a contradiction. □

C. Proof of Theorem 3.3 for k-MEANS

Theorem C.1 (**Restatement of Theorem 3.3**). *Given an integer $k \ge 1$, constant $z \ge 1$ and a real number $\varepsilon \in (0, 1)$, for any integer $d \le \frac{1}{100\varepsilon^2}$, there is a dataset $P \subset \mathbb{R}^{d+1}$ such that its ε-coreset for (k, z)-CLUSTERING must contain at least $\frac{dk}{10z^4}$ points.*

We first prove Theorem C.1 for $z = 2$, and then show how to extend to general $z \ge 1$ in Section E

Proof: [of Theorem C.1 for $z = 2$] We work on $\mathbb{R}^{d+1}$ instead of working on $\mathbb{R}^d$ for technical reasons. We will construct k d-dimensional affine subspaces in $\mathbb{R}^{d+1}$, each of them is far away from others. Then we consider the standard basis in each subspace and show that a coreset of the data set must contain at least $\Omega(kd)$ points.

Denote the standard basis in $\mathbb{R}^{d+1}$ by $e_0, \ldots, e_d$. For each $j \in [k]$, we consider the data set $P_j \triangleq jLe_0 + \{e_1, e_2, \ldots, e_d\}$ where L is a positive number large enough. We take the full data set P as $P = \cup_{j\in[k]} P_j$. Each P_j lies in a d-dimensional affine subspace $jLe_0 + \mathrm{span}\langle e_1, \ldots, e_d\rangle$. These affine subspaces are far away from each other since L is large enough, and this seperation property assures that we can analyze these affine subspaces independently.

Denote the coreset of P as C. For each $j \in [k]$, denote $C_j \triangleq C \cap P_j$. If $jLe_0 + e_i \in C_j$, we denote $i \in C_j$ for the sake of convenience, we also denote its coreset weight as w_i. Denote $v_j \triangleq \sum_{i\in C_j} w_i e_i - \sum_{i=1}^d e_i$, we consider the query $Q = \{q_1, \ldots, q_k\}$ where $q_j \triangleq \frac{v_j}{\|v_j\|_2} + jLe_0$.

Note that each $p \in P_j$ has q_j as its closet query point since L is large enough, thus $\mathrm{cost}(P, Q) = \sum_{j=1}^k \mathrm{cost}(P_j, Q) = \sum_{j=1}^k \mathrm{cost}(P_j, q_j)$ and $\mathrm{cost}(C, Q) = \sum_{j=1}^k \mathrm{cost}(C_j, q_j)$. We compute $\mathrm{cost}(P, Q)$ and $\mathrm{cost}(C, Q)$ by computing $\mathrm{cost}(P_j, q_j)$ and $\mathrm{cost}(C_j, q_j)$

We compute $\mathrm{cost}(P_j, q_j)$ first, that is

$$\mathrm{cost}(P_j, q_j) = \sum_{i=1}^{d} \|jLe_0 + e_i - q_j\|^2 = \sum_{i=1}^{d} \|e_i - \frac{v_j}{|v_j|}\|^2 = 2d - 2\sum_{i=1}^{d} \left\langle e_i, \frac{v_j}{|v_j|} \right\rangle .$$

Similarly we have $\mathrm{cost}(C_j, q_j) = 2\sum_{i\in C_j} w_i - 2\sum_{i\in C_j} w_i \left\langle e_i, \frac{v_j}{|v_j|} \right\rangle$.

Combining them all, we have

$$\begin{aligned} \mathrm{cost}(P,Q) - \mathrm{cost}(C,Q) = &\quad 2kd - 2\sum_{j=1}^{k}\sum_{i\in C_j} w_i + 2\sum_{j=1}^{k}\left\langle \sum_{i\in C_j} w_i e_i - \sum_{i=1}^{d} e_i, \frac{v_j}{|v_j|} \right\rangle \\ = &\quad 2kd - 2\sum_{j=1}^{k}\sum_{i\in C_j} w_i + 2\sum_{j=1}^{k} \|v_j\|. \end{aligned}$$

The coreset property implies that $|\mathrm{cost}(P,Q) - \mathrm{cost}(C,Q)| \le \varepsilon \mathrm{cost}(P,Q) \le 4\varepsilon kd$, thus we have

$$2kd - 2\sum_{j=1}^{k}\sum_{i\in C_j} w_i + 2\sum_{j=1}^{k} \|v_j\| \le 4\varepsilon kd. \tag{7}$$

Taking $\tilde{Q} = \{Le_0, 2Le_0, 3Le_0, \dots, kLe_0\}$, we have $\mathrm{cost}(P, \tilde{Q}) = kd$ and $\mathrm{cost}(C, \tilde{Q}) = \sum_{j=1}^{K}\sum_{i\in C_j} w_i$, the coreset property then gives that $|kd - \sum_{j=1}^{K}\sum_{i\in C_j} w_i| \le \varepsilon kd$. Substitute this inequality to inequality 7, we get

$$\sum_{j=1}^{k} \|v_j\| \le 3\varepsilon kd. \tag{8}$$

For each $j \in [k]$, we have that $\|v_j\| = \|\sum_{i\in C_j} w_i e_i - \sum_{i=1}^{d} e_i\| \ge \|\sum_{i\notin C_j} e_i\| = \sqrt{|P_j| - |C_j|}$. Substitute this inequality to inequality 8, we have

$$\sum_{j=1}^{k} \sqrt{|P_j| - |C_j|} \le 3\varepsilon kd. \tag{9}$$

Our goal is to show that $|C| \ge \frac{1}{10}kd = \frac{1}{10}|P|$, we prove it by contradiction. We will show that if $|C| \ge \frac{1}{10}|P|$ then the dimension d is larger than $\frac{1}{45}\frac{1}{\varepsilon^2}$, which contradicts to the assumption on d.

Assume that $|C| \le \frac{1}{10}|P| = \frac{1}{10}kd$, then for at least half of P_j we have $|C_j| \le \frac{1}{5}|P_j|$ and thus $|P_j| - |C_j| \ge \frac{4}{5}|P_j| = \frac{4}{5}d$. Summing over these P_j we have $\sum_{j=1}^{k} \sqrt{|P_j| - |C_j|} \ge \frac{k}{2}\sqrt{\frac{4}{5}d} = k\sqrt{\frac{1}{5}d}$. By inequality 9 we have $k\sqrt{\frac{1}{5}d} \le 3\varepsilon kd$, thus $\frac{1}{45}\frac{1}{\varepsilon^2} \le d$. This leads to a contradiction. □

Remark C.2. *The proof assumes that the coreset is a subset of the original data set, and the proof holds for coreset with offset.*

D. Generalized lower bound for k-MEANS clustering with general S.

The lower bound proved above relies on the assumption that the coreset S is a subset of the original dataset. Next we generalize the result by allowing arbitrary S in $\mathbb{R}^{d+1}$.

D.1. Preparation

Additional notation Let $e_0, \cdots, e_d$ be the standard basis vectors of $\mathbb{R}^{d+1}$, and $H_1, \cdots, H_{k/2}$ be $k/2$ d-dimensional affine subspaces, where $H_j := jLe_0 + \mathrm{span}\{e_1, \dots, e_d\}$ for a sufficiently large constant L. For any $p \in \mathbb{R}^{d+1}$, we use $\tilde{p}$ to denote the d-dimensional vector $p_{1:d}$ (i.e., discard the 0-th coordinate of p).

Hard instance The hard instance is the same as in Section C, except that now there are $k/2$ affine subspaces and in each affine subspace H_j, we only put $d/2$ points, which are $jLe_0 + e_1, \cdots, jLe_0 + e_{d/2}$. Similarly, we use P_j to denote the data points in H_j ($j = 1, \cdots, k/2$) and let P be the union of all P_j. Thus, $|P| = kd/4$. In our proof, we always put two centers in each H_j; for large enough L, all $p \in P_j$ must be assigned to centers in H_j.

We will use the following two technical lemmas from (Cohen-Addad et al., 2022).

Lemma D.1. *For any $k \geq 1$, let $\{c_1, \cdots, c_k\}$ be arbitrary k unit vectors in $\mathbb{R}^d$, we have*

$$\sum_{i=1}^{d/2} \min_{\ell=1}^{k} \|e_i - c_\ell\|^2 \geq d - \sqrt{dk/2}.$$

Lemma D.2. *Let S be a set of points in $\mathbb{R}^d$ of size t and $w : S \to \mathbb{R}^+$ be their weights. There exist 2 unit vectors v_1, v_2, such that*

$$\sum_{p\in S} w(p) \min_{\ell=1,2} \|p - v_\ell\|^2 \leq \sum_{s\in P} w(p)(\|p\|^2 + 1) - \frac{2\sum_{p\in S} w(p)\|p\|}{\sqrt{t}}.$$

D.2. Proof of the Lower Bound

Next, we present the lower bound result and its proof.

Theorem D.3 (**Same coreset lower bound when S can be arbitrary).** *Given an integer k, a real number $\varepsilon \in (0, 1)$, and integer $d \leq \frac{1}{100\varepsilon^2}$, let $P \subset \mathbb{R}^{d+1}$ be the point set described above. For any $S \subset \mathbb{R}^{d+1}$, if it is a ε-coreset of P, then we must have $|S| = \Omega(dk)$.*

Proof: Note that points in S might not be in any H_j. We first map each point $p \in S$ to an index $j_p \in [k/2]$ such that H_{j_p} is the nearest subspace of p. The mapping is quite simple:

$$j_p = \arg\min_{j\in[k/2]} |p_0 - jL|,$$

where p_0 is the 0-th coordinate of p. Let $\Delta_p = p_0 - j_p L$, which is the distance of p to the closest affine subspace. Let $S_j := \{p \in S : j_p = j\}$ be the set of points in P, whose closest affine subspace is H_j. Define $I := \{j \in [k/2] : |S_j| \leq d/4\}$. Consider any k-center set C such that $H_j \bigcap C \neq \emptyset$. Then $\mathrm{cost}(P, C) \ll 0.1L$ for sufficiently large L. On the other hand, $\mathrm{cost}(S, C) \geq \sum_{p\in S} \Delta_p^2$. Since S is a coreset, $\Delta_p^2 \ll L$ for all $p \in S$. [2] Therefore each $p \in S$ must be very close to its closest affine subspace; in particular, we can assume that p must be assigned to some center in H_{j_p} (if there exists one).

In the proof follows, we consider three different set of k centers C_1, C_2 and C_3 and compare the costs $\mathrm{cost}(P, C_i)$ and $\mathrm{cost}(S, C_i)$ for $i = 1, 2, 3$. In each C_i, there are two centers in each H_j. As we have discussed above, for large enough L, the total cost for both P and S can be decomposed into the sum of costs over all affine subspaces.

For each $j \in \bar{I}$, the corresponding centers in H_j are the same across C_1, C_2, C_3. Let c_j be any point in H_j such that $c_j - jLe_0$ has unit norm and is orthogonal to $e_1, \cdots, e_{d/2}$; in other words, $\|\tilde{c}_j\| = 1$ and the first $d/2$ coordinates of $\tilde{c}_j = 1$ are all zero. Specifically, we set $c_j = jLe_0 + e_{d/2+1}$ and the two centers in H_j are two copies of c_j for $j \in \bar{I}$.

We first consider the following k centers denoted by C_1. As we have specified the centers for $j \in \bar{I}$, we only describe the centers for each $j \in I$. Since by definition, $|S_j| \leq d/4$, we can find a vector $c_j \in \mathbb{R}^{d+1}$ in H_j such that $c_j - jLe_0$ has unit norm and is orthogonal to $e_1, \cdots, e_{d/2}$ and all vectors in S_j. Let C_1 be the set of k points with each point in $\{c_1, \cdots, c_{k/2}\}$ copied twice. We evaluate the cost of C_1 with respect to P and S.

Lemma D.4. *For C_1 constructed above, we have $\mathrm{cost}(P, C_1) = \frac{kd}{2}$ and*

$$\mathrm{cost}(S, C_1) = \sum_{p\in S} w(p)(\Delta_p^2 + \|\tilde{p}\|^2 + 1) - 2\sum_{j\in\bar{I}} \sum_{p\in S_j} w(p)\langle p - jLe_0, jLe_0 - c_j\rangle.$$

[2] Here we do not allow offsets to simplify the proof, but our technique can be extended to handle offsets.

Proof: Since e_i is orthogonal to $c_j - jLe_0$ and $c_j - jLe_0$ has unit norm for all i, j, it follows that

$$\begin{aligned}\mathrm{cost}(P, C_1) &= \sum_{j=1}^{k/2}\sum_{i=1}^{d/2}\min_{c\in C_1}\|jLe_0 + e_i - c\|^2 = \sum_{j=1}^{k/2}\sum_{i=1}^{d/2}\|jLe_0 + e_i - c_j\|^2 \\ &= \sum_{j=1}^{k/2}\sum_{i=1}^{d/2}(\|e_i\|^2 + \|c_j - jLe_0\|^2 - 2\langle e_i, c_j - jLe_0\rangle) \\ &= \frac{kd}{2}. \end{aligned} \tag{10}$$

On the other hand, the cost of C w.r.t. S_j is

$$\begin{aligned}\sum_{p\in S_j}\min_{c\in C_1} w(p)\|p-c\|^2 &= \sum_{p\in S_j} w(p)\|p - c_j\|^2 = \sum_{p\in S_j} w(p)\|p - jLe_0 + jLe_0 - c_j\|^2 \\ &= \sum_{p\in S_j} w(p)\left(\|p - jLe_0\|^2 + 1 - 2\langle p - jLe_0, jLe_0 - c_j\rangle\right) \\ &= \sum_{p\in S_j} w(p)(\Delta_p^2 + \|\tilde{p}\|^2 + 1) - 2w(p)\langle p - jLe_0, jLe_0 - c_j\rangle. \end{aligned} \tag{11}$$

Recall $\tilde{p} \in \mathbb{R}^d$ is $p_{1:d}$. For $j \in I$, the inner product is 0, and thus the total cost w.r.t. S is

$$\mathrm{cost}(S, C_1) = \sum_{p\in S} w(p)(\Delta_p^2 + \|\tilde{p}\|^2 + 1) - 2\sum_{j\in\bar{I}}\sum_{p\in S_j} w(p)\langle p - jLe_0, jLe_0 - c_j\rangle,$$

which finishes the proof. □

For notational convenience, we define $\kappa := 2\sum_{j\in\bar{I}}\sum_{p\in S_j} w(p)\langle p - jLe_0, jLe_0 - c_j\rangle$. Since S is an ε-coreset of P, we have

$$dk/2 - \varepsilon dk/2 \le \sum_{p\in S} w(p)(\Delta_p^2 + \|p'\|^2 + 1) - \kappa \le dk/2 + \varepsilon dk/2. \tag{12}$$

Next we consider a different set of k centers denoted by C_2. By Lemma D.2, there exists unit vectors $v_1^j, v_2^j \in \mathbb{R}^d$ such that

$$\sum_{p\in S_j} w(p)(\min_{\ell=1,2}\|\tilde{p} - v_\ell^j\|^2 + \Delta_p^2) \le \sum_{p\in S_j} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2) - \frac{2\sum_{p\in S_j} w(p)\|\tilde{p}\|}{\sqrt{|S_j|}}. \tag{13}$$

Applying this to all $j \in I$ and get corresponding v_1^j, v_2^j for all $j \in I$. Let $C_2 = \{u_1^1, u_2^2, \cdots, u_1^{k/2}, u_2^{k/2}\}$ be a set of k centers in $\mathbb{R}^{d+1}$ defined as follows: if $j \in I$, u_ℓ^j is v_ℓ^j with an additional 0th coordinate with value jL, making them lie in H_j; for $j \in \bar{I}$, we use the same centers as in C_1, i.e., $u_1^j = u_2^j = c_j$.

Lemma D.5. *For C_2 constructed above, we have*

$$\mathrm{cost}(P, C_2) \ge \frac{kd}{2} - \sqrt{d}|I| \text{ and}$$

$$\mathrm{cost}(S, C_2) \le \sum_{p\in S} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2) - \sum_{j\in I}\frac{2\sum_{p\in S_j} w(p)\|\tilde{p}\|}{\sqrt{|S_j|}} - \kappa.$$

Proof: By (13),

$$\begin{aligned}\mathrm{cost}(S, C_2) &= \sum_{j=1}^{k/2}\sum_{p\in S_j} w(p)\min_{c\in C_2}\|p - c\|^2 = \sum_{j\in I}\sum_{p\in S_j} w(p)\min_{\ell=1,2}(\|\tilde{p} - v_\ell^j\|^2 + \Delta_p^2) + \sum_{j\in\bar{I}}\sum_{p\in S_j} w(p)\|p - c_j\|^2 \\ &\le \sum_{p\in S} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2) - \sum_{j\in I}\frac{2\sum_{p\in S_j} w(p)\|\tilde{p}\|}{\sqrt{|S_j|}} - \kappa. \end{aligned}$$

By Lemma D.1 (with $k = 2$), we have

$$\sum_{i=1}^{d/2} \min_{\ell=1,2} \|e_i - v_\ell^j\|^2 \geq d - \sqrt{d}.$$

It follows that

$$\begin{aligned}\mathrm{cost}(P, C_2) &= \sum_{j=1}^{k/2} \sum_{i=1}^{d/2} \min_{c \in C_2} \|jLe_0 + e_i - c\|^2 = \sum_{j \in I} \sum_{i=1}^{d/2} \min_{\ell=1,2} \|e_i - v_\ell^j\|^2 + \sum_{j \in \bar{I}} \sum_{i=1}^{d/2} \|jLe_0 + e_i - c\|^2 \\ &\geq \frac{kd}{2} - \sqrt{d}|I|,\end{aligned}$$

where in the inequality, we also used the orthogonality between e_i and $c_j - jLe_0$. □

Since S is an ε-coreset of P, we have

$$\frac{dk}{2} - |I|\sqrt{d} - \frac{\varepsilon dk}{2} \leq (\frac{dk}{2} - |I|\sqrt{d})(1-\varepsilon) \leq \sum_{p \in S} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2) - \sum_{j \in I} \frac{2\sum_{p \in S_j} w(p)\|\tilde{p}\|}{\sqrt{|S_j|}} - \kappa,$$

which implies

$$\begin{aligned}\sum_{j \in I} \frac{2\sum_{p \in S_j} w(p)\|\tilde{p}\|}{\sqrt{|S_j|}} &\leq \sum_{p \in S} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2) - \frac{dk - 2|I|\sqrt{d} - \varepsilon kd}{2} - \kappa \\ &\leq \frac{dk + \varepsilon dk}{2} - \frac{dk - 2|I|\sqrt{d} - \varepsilon kd}{2} \quad \text{by (12)} \\ &= |I|\sqrt{d} + \varepsilon kd.\end{aligned}$$

By definition, $|S_j| \leq d/4$, so

$$\sum_{j \in I} \frac{2\sum_{p \in S_j} w(p)\|\tilde{p}\|}{\sqrt{d/4}} \leq \sum_{j \in I} \frac{2\sum_{p \in S_j} w(p)\|\tilde{p}\|}{\sqrt{|S_j|}},$$

and it follows that

$$\frac{\sum_{j \in I}\sum_{p \in S_j} w(p)\|\tilde{p}\|}{\sqrt{d}} \leq \frac{|I|\sqrt{d} + \varepsilon kd}{4}. \tag{14}$$

Finally we consider a third set of k centers C_3. Similarly, there are two centers per group. We set m be a power of 2 in $[d/2, d]$. Let $h_1, \cdots, h_m$ be the m-dimensional Hadamard basis vectors. So all h_ℓ's are $\{-\frac{1}{\sqrt{m}}, \frac{1}{\sqrt{m}}\}$ vectors and $h_1 = (\frac{1}{\sqrt{m}}, \cdots, \frac{1}{\sqrt{m}})$. We slightly abuse notation and treat each h_ℓ as a d-dimensional vector by concatenating zeros in the end. For each h_ℓ construct a set of k centers as follows. For each $j \in \bar{I}$, we still use two copies of c_j. For $j \in I$, the 0th coordinate of the two centers is jL, then we concatenate h_ℓ and $-h_\ell$ respectively to the first and the second centers.

Lemma D.6. *Suppose C_3 is constructed based on h_ℓ. Then for all $\ell \in [m]$, we have*

$$\mathrm{cost}(P, C_3) = \frac{kd}{2} - \frac{d|I|}{\sqrt{m}} \text{ and}$$

$$\mathrm{cost}(S, C_3) = \sum_{p \in S} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2) - 2\sum_{j \in I}\sum_{p \in S_j} \langle w(p)\tilde{p}, h_\ell^p \rangle - \kappa.$$

Proof: For $j \in I$, the cost of the two centers w.r.t. P_j is

$$\mathrm{cost}(P_j, C_3) = \sum_{i=1}^{d/2} \min_{s=-1,+1} \|e_i - s \cdot h_\ell\|^2 = \sum_{i=1}^{d/2}(2 - 2 \max_{s=-1,+1} \langle h_\ell, e_i\rangle) = \sum_{i=1}^{d/2}(2 - \frac{2}{\sqrt{m}}) = d - \frac{d}{\sqrt{m}}.$$

For $j \in \bar{I}$, the cost w.r.t. P_j is d by (10). Thus, the total cost over all subspaces is

$$\mathrm{cost}(P, C_3) = (d - \frac{d}{\sqrt{m}})|I| + \left(\frac{k}{2} - |I|\right) d = \frac{kd}{2} - \frac{d|I|}{\sqrt{m}}.$$

On the other hand, for $j \in I$, the cost w.r.t. S_j is

$$\begin{aligned}\sum_{p \in S_j} w(p)(\Delta_p^2 + \min_{s=\{-1,+1\}} \|\tilde{p} - s \cdot h_\ell\|^2) &= \sum_{p \in S_j} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2 - 2 \max_{s=\{-1,+1\}} \langle \tilde{p}, s \cdot h_\ell\rangle) \\ &= \sum_{p \in S_j} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2 - 2\langle \tilde{p}, h_\ell^p\rangle).\end{aligned}$$

Here $h_\ell^p = s^p \cdot h_\ell$, where $s^p = \arg\max_{s=\{-1,+1\}} \langle \tilde{p}, s \cdot h_\ell\rangle$. For $j \in \bar{I}$, the cost w.r.t. S_j is $\sum_{p \in S_j} w(p)(\Delta_p^2 + \|\tilde{p}\|^2 + 1) - 2\langle p - jLe_0, jLe_0 - c_j\rangle)$ by (11). Thus, the total cost w.r.t. S is

$$\mathrm{cost}(S, C_3) = \sum_{p \in S} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2) - 2\sum_{j \in I}\sum_{p \in S_j} \langle w(p)\tilde{p}, h_\ell^p\rangle - \kappa.$$

This finishes the proof. □

Corollary D.7. *Let S be a ε-coreset of P, and $I = \{j : |S_j| \le d/4\}$. Then*

$$\sum_{j \in I}\sum_{p \in S_j} w(p)\|\tilde{p}\| \ge \frac{d|I| - \varepsilon kd\sqrt{d}}{2}.$$

Proof: Since S is an ε-coreset, we have by Lemma D.6

$$\begin{aligned}2\sum_{j \in I}\sum_{p \in S_j} \langle w(p)\tilde{p}, h_\ell^p\rangle &\ge \sum_{p \in S} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2) - \kappa - (\frac{kd}{2} - \frac{d|I|}{\sqrt{m}})(1+\varepsilon) \\ &\ge \sum_{p \in S} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2) - \kappa - \frac{kd}{2} + \frac{d|I|}{\sqrt{m}} - \frac{\varepsilon kd}{2} \\ &\ge \frac{dk - \varepsilon dk}{2} - \frac{kd}{2} + \frac{d|I|}{\sqrt{m}} - \frac{\varepsilon kd}{2} \quad \text{by (12)} \\ &= \frac{d|I|}{\sqrt{m}} - \varepsilon kd.\end{aligned}$$

Note that the above inequality holds for all $\ell \in [m]$, then

$$2\sum_{\ell=1}^{m}\sum_{j \in I}\sum_{p \in S_j} \langle w(p)\tilde{p}, h_\ell^p\rangle \ge d|I|\sqrt{m} - \varepsilon kdm.$$

By the Cauchy-Schwartz inequality,

$$\sum_{\ell=1}^{m}\sum_{j \in I}\sum_{p \in S_j} \langle w(p)\tilde{p}, h_\ell^p\rangle = \sum_{j \in I}\sum_{p \in S_j} \langle w(p)\tilde{p}, \sum_{\ell=1}^{m} h_\ell^p\rangle \le \sum_{j \in I}\sum_{p \in S_j} w(p)\|\tilde{p}\|\|\sum_{\ell=1}^{m} h_\ell^p\| = \sqrt{m}\sum_{j \in I}\sum_{p \in S_j} w(p)\|\tilde{p}\|.$$

Therefore, we have

$$\sum_{j\in I}\sum_{p\in S_j} w(p)\|\tilde{p}\| \geq \frac{d|I| - \varepsilon kd\sqrt{m}}{2} \geq \frac{d|I| - \varepsilon kd\sqrt{d}}{2}.$$

□

Combining the above corollary with (14), we have

$$\frac{\sqrt{d}|I| - \varepsilon kd}{2} \leq \frac{|I|\sqrt{d} + \varepsilon kd}{4} \implies |I| \leq 3\varepsilon k\sqrt{d}.$$

By the assumption $d \leq \frac{1}{100\varepsilon^2}$, it holds that $|I| \leq \frac{3k}{10}$ or $|\bar{I}| \geq \frac{k}{2} - \frac{3k}{10} = \frac{k}{5}$. Moreover, since $|S_j| > \frac{d}{4}$ for each $j \in \bar{I}$, we have $|S| > \frac{d}{4} \cdot \frac{k}{5} = \frac{kd}{20}$. □

E. Proof of Theorem 3.3 for general $z \geq 1$

Using similar ideas from (Cohen-Addad et al., 2022), our proof of the lower bound for $z = 2$ can be extended to arbitrary z. First, we provide two lemmas analogous to Lemma D.1 and D.2 for general $z \geq 1$. Their proofs can be found in Appendix A in (Cohen-Addad et al., 2022).

Lemma E.1. *For any even number $k \geq 1$, let $\{c_1, \cdots, c_k\}$ be arbitrary k unit vectors in $\mathbb{R}^d$ such that for each i there exist some j satisfying $c_i = -c_j$. We have*

$$\sum_{i=1}^{d/2} \min_{\ell=1}^{k} \|e_i - c_\ell\|^z \geq 2^{z/2-1}d - 2^{z/2}\max\{1, z/2\}\sqrt{\frac{kd}{2}}.$$

Lemma E.2. *Let S be a set of points in $\mathbb{R}^d$ of size t and $w: S \to \mathbb{R}^+$ be their weights. For arbitrary Δ_p for each p, there exist 2 unit vectors v_1, v_2 satisfying $v_1 = -v_2$, such that*

$$\sum_{p\in S} w(p) \min_{\ell=1,2} \left(\|p - v_\ell\|^2 + \Delta_p^2\right)^{z/2} \leq \sum_{s\in P} w(p)(\|p\|^2 + 1 + \Delta_p^2)^{z/2} - \min\{1, z/2\} \cdot \frac{2\sum_{p\in S} w(p)(\|p\|^2 + 1 + \Delta_p^2)^{z/2-1}\|p\|}{\sqrt{t}}.$$

In this proof, the original point set P and three sets of k-centers, namely C_1, C_2, C_3, are the same as for the case $z = 2$. The difference is that now $I = \{j : |S_j| \leq \frac{d}{2^z}\}$ and when constructing C_2, we use Lemma E.2 in place of Lemma D.2. Again, we compare the cost of P and S w.r.t. C_1, C_2, C_3 and get the following lemmas.

Lemma E.3. *For C_1 constructed above, we have $\mathrm{cost}(P, C_1) = \frac{kd}{4} \cdot 2^{z/2}$ and*

$$\mathrm{cost}(S, C_1) = \sum_{j\in I}\sum_{p\in S_j} w(p)(\Delta_p^2 + \|\tilde{p}\|^2 + 1)^{z/2} + \sum_{j\in \bar{I}}\sum_{p\in S_j} w(p)\|p - c_j\|^z.$$

Proof: Since e_i is orthogonal to $c_j - jLe_0$ and $c_j - jLe_0$ has unit norm for all i, j, it follows that

$$\begin{aligned}\mathrm{cost}(P, C_1) &= \sum_{j=1}^{k/2}\sum_{i=1}^{d/2} \min_{c\in C_1} \|jLe_0 + e_i - c\|^{2\cdot z/2} = \sum_{j=1}^{k/2}\sum_{i=1}^{d/2} \|jLe_0 + e_i - c_j\|^{2\cdot z/2} \\ &= \sum_{j=1}^{k/2}\sum_{i=1}^{d/2} (\|e_i\|^2 + \|c_j - jLe_0\|^2 - 2\langle e_i, c_j - jLe_0\rangle)^{z/2} \\ &= \frac{kd}{4} \cdot 2^{z/2}. \end{aligned} \tag{15}$$

On the other hand, the cost of C_1 w.r.t. S_j is

$$\begin{aligned}\sum_{p\in S_j}\min_{c\in C_1} w(p)\|p-c\|^{2\cdot z/2} &= \sum_{p\in S_j} w(p)\|p-c_j\|^{2\cdot z/2} = \sum_{p\in S_j} w(p)\|p-jLe_0+jLe_0-c_j\|^{2\cdot z/2} \\ &= \sum_{p\in S_j} w(p)\left(\|p-jLe_0\|^2+1-2\langle p-jLe_0, jLe_0-c_j\rangle\right)^{z/2}. \end{aligned} \tag{16}$$

For $j \in I$, the inner product is 0, and thus the total cost w.r.t. S is

$$\mathrm{cost}(S, C_1) = \sum_{j\in I}\sum_{p\in S_j} w(p)(\Delta_p^2+\|\tilde{p}\|^2+1)^{z/2} + \sum_{j\in \bar{I}}\sum_{p\in S_j} w(p)\|p-c_j\|^z,$$

which finishes the proof. □

For notational convenience, we define $\kappa := \sum_{j\in\bar{I}}\sum_{p\in S_j} w(p)\|p-c_j\|^z$. Since S is an ε-coreset of P, we have

$$\frac{kd}{4}\cdot 2^{z/2} - \frac{\varepsilon kd}{4}\cdot 2^{z/2} \le \sum_{j\in I}\sum_{p\in S_j} w(p)(\Delta_p^2+\|\tilde{p}\|^2+1)^{z/2} + \kappa \le \frac{kd}{4}\cdot 2^{z/2} + \frac{\varepsilon kd}{4} 2^{z/2}. \tag{17}$$

Next we consider a different set of k centers denoted by C_2. By Lemma E.2, there exists unit vectors $v_1^j, v_2^j \in \mathbb{R}^d$ satisfying $v_1^j = -v_2^j$ such that

$$\begin{aligned}\sum_{p\in S_j} w(p)(\min_{\ell=1,2}\left(\|\tilde{p}-v_\ell^j\|^2+\Delta_p^2\right)^{z/2}) &\le \sum_{p\in S_j} w(p)(\|\tilde{p}\|^2+1+\Delta_p^2)^{z/2} \\ &\quad - \min\{1, z/2\}\frac{2\sum_{p\in S_j} w(p)(\|\tilde{p}\|^2+1+\Delta_p^2)^{z/2-1}\|\tilde{p}\|}{\sqrt{|S_j|}}.\end{aligned} \tag{18}$$

Applying this to all $j \in I$ and get corresponding v_1^j, v_2^j for all $j \in I$. Let $C_2 = \{u_1^1, u_2^2, \cdots, u_1^{k/2}, u_2^{k/2}\}$ be a set of k centers in $\mathbb{R}^{d+1}$ defined as follows: if $j \in I$, u_ℓ^j is v_ℓ^j with an additional 0th coordinate with value jL, making them lie in H_j; for $j \in \bar{I}$, we use the same centers as in C_1, i.e., $u_1^j = u_2^j = c_j$.

Lemma E.4. *For C_2 constructed above, we have*

$$\mathrm{cost}(P, C_2) \ge 2^{z/2}\left(\frac{kd}{4} - \max\{1, z/2\}\sqrt{d}|I|\right), \textit{ and}$$

$$\begin{aligned}\mathrm{cost}(S, C_2) \le& \sum_{j\in I}\sum_{p\in S_j} w(p)(\|\tilde{p}\|^2+1+\Delta_p^2)^{z/2} \\ &- \min\{1, z/2\}\sum_{j\in I}\frac{2\sum_{p\in S_j} w(p)(\|\tilde{p}\|^2+1+\Delta_p^2)^{z/2-1}\|\tilde{p}\|}{\sqrt{|S_j|}} + \kappa.\end{aligned}$$

Proof: By (18),

$$\begin{aligned}\mathrm{cost}(S, C_2) &= \sum_{j=1}^{k/2}\sum_{p\in S_j} w(p)\min_{c\in C_2}\|p-c\|^{2\cdot z/2} = \sum_{j\in I}\sum_{p\in S_j} w(p)\min_{\ell=1,2}(\|\tilde{p}-v_\ell^j\|^2+\Delta_p^2)^{z/2} + \kappa \\ &\le \sum_{j\in I}\sum_{p\in S_j} w(p)(\|\tilde{p}\|^2+1+\Delta_p^2)^{z/2} \\ &- \min\{1, z/2\}\sum_{j\in I}\frac{2\sum_{p\in S_j} w(p)(\|\tilde{p}\|^2+1+\Delta_p^2)^{z/2-1}\|\tilde{p}\|}{\sqrt{|S_j|}} + \kappa.\end{aligned}$$

By Lemma E.1 (with $k=2$), we have

$$\sum_{i=1}^{d/2} \min_{\ell=1,2} \|e_i - v_\ell^j\|^z \geq 2^{z/2-1}d - 2^{z/2}\max\{1, z/2\}\sqrt{d}.$$

It follows that

$$\begin{aligned}\mathrm{cost}(P, C_2) &= \sum_{j=1}^{k/2}\sum_{i=1}^{d/2} \min_{c\in C_2} \|jLe_0 + e_i - c\|^z = \sum_{j\in I}\sum_{i=1}^{d/2} \min_{\ell=1,2} \|e_i - v_\ell^j\|^{2\cdot z/2} + \sum_{j\in \bar{I}}\sum_{i=1}^{d/2} \|jLe_0 + e_i - c_j\|^{2\cdot z/2} \\ &\geq \left(2^{z/2-1}d - 2^{z/2}\max\{1, z/2\}\sqrt{d}\right)|I| + |\bar{I}|\frac{d}{2}\cdot 2^{z/2} = \frac{kd}{4}2^{z/2} - 2^{z/2}\max\{1, z/2\}\sqrt{d}|I|,\end{aligned}$$

where in the inequality, we also used the orthogonality between e_i and $c_j - jLe_0$. □

Since S is an ε-coreset of P, we have

$$\begin{aligned}&2^{z/2}\left(\frac{dk}{4} - \max\{1, z/2\}|I|\sqrt{d} - \frac{\varepsilon dk}{4}\right) \leq 2^{z/2}\left(\frac{kd}{4} - \max\{1, z/2\}\sqrt{d}|I|\right)(1-\varepsilon) \\ &\leq \sum_{j\in I}\sum_{p\in S_j} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2)^{z/2} - \min\{1, z/2\}\sum_{j\in I}\frac{2\sum_{p\in S_j} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2)^{z/2-1}\|\tilde{p}\|}{\sqrt{|S_j|}} + \kappa,\end{aligned}$$

which implies

$$\begin{aligned}&\min\{1, z/2\}\sum_{j\in I}\frac{2\sum_{p\in S_j} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2)^{z/2-1}\|\tilde{p}\|}{\sqrt{|S_j|}} \\ &\leq \sum_{j\in I}\sum_{p\in S_j} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2)^{z/2} - 2^{z/2}\left(\frac{dk}{4} - \max\{1, z/2\}|I|\sqrt{d} - \frac{\varepsilon dk}{4}\right) + \kappa \\ &\leq \frac{kd}{4}\cdot 2^{z/2} + \frac{\varepsilon kd}{4}2^{z/2} - 2^{z/2}\left(\frac{dk}{4} - \max\{1, z/2\}|I|\sqrt{d} - \frac{\varepsilon dk}{4}\right) \quad \text{by (17)} \\ &= \max\{1, z/2\}|I|\sqrt{d}2^{z/2} + \frac{\varepsilon kd}{2}2^{z/2}.\end{aligned}$$

By definition, $|S_j| \leq d/t^2$, so

$$\min\{1, \frac{z}{2}\}\sum_{j\in I}\frac{2\sum_{p\in S_j} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2)^{z/2-1}\|\tilde{p}\|}{\sqrt{d/t^2}} \leq \min\{1, \frac{z}{2}\}\sum_{j\in I}\frac{2\sum_{p\in S_j} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2)^{z/2-1}\|\tilde{p}\|}{\sqrt{|S_j|}},$$

and it follows that

$$\min\{1, \frac{z}{2}\}\sum_{j\in I}\frac{\sum_{p\in S_j} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2)^{z/2-1}\|\tilde{p}\|}{\sqrt{d}} \leq \frac{\max\{1, z/2\}|I|\sqrt{d}2^{z/2} + \frac{\varepsilon kd}{2}2^{z/2}}{2t}. \tag{19}$$

Finally we consider a third set of k centers C_3. Similarly, there are two centers per group. We set m be a power of 2 in $[d/2, d]$. Let $h_1, \cdots, h_m$ be the m-dimensional Hadamard basis vectors. So all h_ℓ's are $\{-\frac{1}{\sqrt{m}}, \frac{1}{\sqrt{m}}\}$ vectors and $h_1 = (\frac{1}{\sqrt{m}}, \cdots, \frac{1}{\sqrt{m}})$. We slightly abuse notation and treat each h_ℓ as a d-dimensional vector by concatenating zeros in the end. For each h_ℓ construct a set of k centers as follows. For each $j \in \bar{I}$, we still use two copies of c_j. For $j \in I$, the 0th coordinate of the two centers is jL, then we concatenate h_ℓ and $-h_\ell$ respectively to the first and the second centers.

Lemma E.5. *Suppose C_3 is constructed based on h_ℓ. Then for all $\ell \in [m]$, we have*

$$\mathrm{cost}(P, C_3) \leq 2^{z/2}\left(\frac{kd}{4} - \frac{d|I|}{2}\cdot\frac{\min\{1, z/2\}}{\sqrt{m}}\right), \text{ and}$$

$$\mathrm{cost}(S, C_3) \geq \sum_{j\in I}\sum_{p\in S_j} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2)^{\frac{z}{2}} - 2\max\{1, \frac{z}{2}\}\sum_{j\in I}\sum_{p\in S_j} w(p)\langle\tilde{p}, h_\ell^p\rangle(\|\tilde{p}\|^2 + 1 + \Delta_p^2)^{\frac{z}{2}-1} + \kappa.$$

Proof: For $j \in I$, the cost of the two centers w.r.t. P_j is

$$\begin{aligned}\mathrm{cost}(P_j, C_3) &= \sum_{i=1}^{d/2} \min_{s=-1,+1} \|e_i - s \cdot h_\ell\|^z = \sum_{i=1}^{d/2} (2 - 2 \max_{s=-1,+1} \langle h_\ell, e_i \rangle)^{z/2} = \frac{d}{2}(2 - \frac{2}{\sqrt{m}})^{z/2} \\ &\leq \frac{d}{2} \cdot 2^{z/2} \left(1 - \frac{\min\{1, z/2\}}{\sqrt{m}}\right).\end{aligned}$$

For $j \in \bar{I}$, the cost w.r.t. P_j is $\frac{d}{2} \cdot 2^{z/2}$ by (15). Thus, the total cost over all subspaces is

$$\mathrm{cost}(P, C_3) \leq \frac{d}{2} \cdot 2^{z/2} \left(1 - \frac{\min\{1, z/2\}}{\sqrt{m}}\right) |I| + \left(\frac{k}{2} - |I|\right) \frac{d}{2} \cdot 2^{z/2} = 2^{z/2} \left(\frac{kd}{4} - \frac{d|I|}{2} \cdot \frac{\min\{1, z/2\}}{\sqrt{m}}\right).$$

On the other hand, for $j \in I$, the cost w.r.t. S_j is

$$\begin{aligned}&\sum_{p \in S_j} w(p)(\Delta_p^2 + \min_{s=\{-1,+1\}} \|\tilde{p} - s \cdot h_\ell\|^2)^{z/2} = \sum_{p \in S_j} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2 - 2 \max_{s=\{-1,+1\}} \langle \tilde{p}, s \cdot h_\ell \rangle)^{z/2} \\ &= \sum_{p \in S_j} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2 - 2\langle \tilde{p}, h_\ell^p \rangle)^{z/2} \\ &\geq \sum_{p \in S_j} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2)^{\frac{z}{2}} - 2\max\{1, \frac{z}{2}\} \sum_{p \in S_j} w(p) \langle \tilde{p}, h_\ell^p \rangle (\|\tilde{p}\|^2 + 1 + \Delta_p^2)^{\frac{z}{2}-1}.\end{aligned}$$

Here $h_\ell^p = s^p \cdot h_\ell$, where $s^p = \arg\max_{s=\{-1,+1\}} \langle \tilde{p}, s \cdot h_\ell \rangle$. For $j \in \bar{I}$, the total cost w.r.t. S_j is κ. Thus, the total cost w.r.t. S is

$$\mathrm{cost}(S, C_3) \geq \sum_{j \in I} \sum_{p \in S_j} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2)^{\frac{z}{2}} - 2\max\{1, \frac{z}{2}\} \sum_{j \in I} \sum_{p \in S_j} w(p) \langle \tilde{p}, h_\ell^p \rangle (\|\tilde{p}\|^2 + 1 + \Delta_p^2)^{\frac{z}{2}-1} + \kappa.$$

This finishes the proof. □

Corollary E.6. *Let S be a ε-coreset of P, and $I = \{j : |S_j| \leq d/4\}$. Then*

$$2\max\{1, \frac{z}{2}\} \sum_{j \in I} \sum_{p \in S_j} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2)^{\frac{z}{2}-1} \|\tilde{p}\| \geq 2^{z/2} \cdot \left(\frac{d|I|}{2} \cdot \min\{1, z/2\} - \frac{\varepsilon k d \sqrt{d}}{2}\right).$$

Proof: Since S is an ε-coreset, we have by Lemma E.5

$$\begin{aligned}2\max\{1, \frac{z}{2}\} \sum_{j \in I} \sum_{p \in S_j} w(p) \langle \tilde{p}, h_\ell^p \rangle (\|\tilde{p}\|^2 + 1 + \Delta_p^2)^{\frac{z}{2}-1} &\geq \sum_{j \in I} \sum_{p \in S_j} w(p)(\|\tilde{p}\|^2 + 1 + \Delta_p^2)^{\frac{z}{2}} + \kappa \\ &\quad - 2^{z/2} \left(\frac{kd}{4} - \frac{d|I|}{2} \cdot \frac{\min\{1, z/2\}}{\sqrt{m}}\right)(1 + \varepsilon) \\ &\geq \frac{kd}{4} \cdot 2^{z/2} - \frac{\varepsilon kd}{4} \cdot 2^{z/2} - 2^{z/2} \left(\frac{kd}{4} - \frac{d|I|}{2} \cdot \frac{\min\{1, z/2\}}{\sqrt{m}} + \frac{\varepsilon kd}{4}\right) \quad \text{by (17)} \\ &= 2^{z/2} \cdot \frac{d|I|}{2} \cdot \frac{\min\{1, z/2\}}{\sqrt{m}} - \frac{\varepsilon kd}{2} \cdot 2^{z/2}.\end{aligned}$$

Note that the above inequality holds for all $\ell \in [m]$, then

$$2\max\{1, \frac{z}{2}\} \sum_{\ell=1}^{m} \sum_{j \in I} \sum_{p \in S_j} w(p) \langle \tilde{p}, h_\ell^p \rangle (\|\tilde{p}\|^2 + 1 + \Delta_p^2)^{\frac{z}{2}-1} \geq 2^{z/2} \cdot \left(\frac{d|I|\sqrt{m}}{2} \cdot \min\{1, z/2\} - \frac{\varepsilon kdm}{2}\right).$$

By the Cauchy-Schwartz inequality,

$$\begin{aligned}\sum_{\ell=1}^{m}\sum_{j\in I}\sum_{p\in S_j} w(p)\langle\tilde{p},h_\ell^p\rangle(\|\tilde{p}\|^2+1+\Delta_p^2)^{\frac{z}{2}-1} &= \sum_{j\in I}\sum_{p\in S_j} w(p)\langle\tilde{p},\sum_{\ell=1}^{m}h_\ell^p\rangle(\|\tilde{p}\|^2+1+\Delta_p^2)^{\frac{z}{2}-1}\\ &\leq \sum_{j\in I}\sum_{p\in S_j} w(p)(\|\tilde{p}\|^2+1+\Delta_p^2)^{\frac{z}{2}-1}\|\tilde{p}\|\cdot\|\sum_{\ell=1}^{m}h_\ell^p\|\\ &= \sqrt{m}\sum_{j\in I}\sum_{p\in S_j} w(p)(\|\tilde{p}\|^2+1+\Delta_p^2)^{\frac{z}{2}-1}\|\tilde{p}\|.\end{aligned}$$

Therefore, we have

$$\begin{aligned}2\max\{1,\frac{z}{2}\}\sum_{j\in I}\sum_{p\in S_j} w(p)(\|\tilde{p}\|^2+1+\Delta_p^2)^{\frac{z}{2}-1}\|\tilde{p}\| &\geq 2^{z/2}\cdot\left(\frac{d|I|}{2}\cdot\min\{1,z/2\}-\frac{\varepsilon kd\sqrt{m}}{2}\right)\\ &\geq 2^{z/2}\cdot\left(\frac{d|I|}{2}\cdot\min\{1,z/2\}-\frac{\varepsilon kd\sqrt{d}}{2}\right).\end{aligned}$$

□

Combining the above corollary with (19), we have

$$\frac{\min\{1,z/2\}}{2\max\{1,z/2\}}2^{z/2}\cdot\left(\frac{\sqrt{d}|I|}{2}\cdot\min\{1,z/2\}-\frac{\varepsilon kd}{2}\right)\leq\frac{\left(\max\{1,z/2\}|I|\sqrt{d}+\frac{\varepsilon kd}{2}\right)2^{z/2}}{2t},$$

which implies that

$$\left(\frac{\min\{1,(z/2)^2\}}{4\max\{1,(z/2)\}}-\frac{\max\{1,z/2\}}{2t}\right)|I|\leq\frac{\min\{1,(z/2)\}\varepsilon kd}{4\max\{1,(z/2)\}}+\frac{\varepsilon k\sqrt{d}}{4t}.$$

So if we set $t=\frac{4\max\{1,(z/2)^2\}}{\min\{1,(z/2)^2\}}$, then

$$\frac{\min\{1,(z/2)^2\}}{8\max\{1,(z/2)\}}|I|\leq\frac{\min\{1,(z/2)\}\varepsilon k\sqrt{d}}{2\max\{1,(z/2)\}}\implies|I|\leq\frac{4\varepsilon k\sqrt{d}}{\min\{1,z/2\}}.$$

By the assumption $d\leq\frac{\min\{1,(z/2)^2\}}{100\varepsilon^2}$, it holds that $|I|\leq\frac{2k}{5}$ or $|\bar{I}|\geq\frac{k}{2}-\frac{2k}{5}=\frac{k}{10}$. Moreover, since $|S_j|>\frac{d}{t^2}$ for each $j\in\bar{I}$, we have $|S|>\frac{d}{t^2}\cdot\frac{k}{5}=\frac{kd\min\{1,(z/2)^4\}}{\max\{1,(z/2)^4\}}$.