Towards Optimal Lower Bounds for k-Median and k-Means Coresets

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

David Saulpic
david.saulpic@lip6.fr
Sorbonne Université, CNRS, LIP6
Paris, France

Kasper Green Larsen
larsen@cs.au.dk
Aarhus University
Aarhus, Denmark

Chris Schwiegelshohn
cschwiegelshohn@gmail.com
Aarhus University
Aarhus, Denmark

ABSTRACT
The \((k, z)\)-clustering problem consists of finding a set of \(k\) points called centers, such that the sum of distances raised to the power of \(z\) of every data point to its closest center is minimized. Among the most commonly encountered special cases are \(k\)-median problem \((z = 1)\) and \(k\)-means problem \((z = 2)\). The \(k\)-median and \(k\)-means problems are at the heart of modern data analysis and massive data applications have given rise to the notion of coreset: a small (weighted) subset of the input point set preserving the cost of any solution to the problem up to a multiplicative \((1 \pm \varepsilon)\) factor, hence reducing from large to small scale the input to the problem.

While there has been an intensive effort to understand what is the best coreset size possible for both problems in various metric spaces, there is still a significant gap between the state-of-the-art upper and lower bounds. In this paper, we make progress on both upper and lower bounds, obtaining tight bounds for several cases, namely:

- In finite \(n\) point general metrics, any coreset must consist of \(\Omega(k \log n/e^2)\) points. This improves on the \(\Omega(k \log n/e)\) lower bound of Braverman, Jiang, Krauthgamer, and Wu [ICML’19] and matches the upper bounds proposed for \(k\)-median by Feldman and Langberg [STOC’11] and \(k\)-means by Cohen-Addad, Saulpic and Schwiegelshohn [STOC’21] up to polylog factors.
- For doubling metrics with doubling constant \(D\), any coreset must consist of \(\Omega(k/D^2)\) points. This matches the \(k\)-median and \(k\)-means upper bounds by Cohen-Addad, Saulpic, and Schwiegelshohn [STOC’21] up to polylog factors.
- In \(d\)-dimensional Euclidean space, any coreset for \((k, z)\) clustering requires \(\Omega(k/e^2)\) points. This improves on the \(\Omega(k/\sqrt{\varepsilon})\) lower bound of Baker, Braverman, Huang, Jiang, Krauthgamer, and Wu [ICML’20] for \(k\)-median and complements the \(\Omega(k \min(d, 2z/20))\) lower bound of Huang and Vishnoi [STOC’20].
- We complement our lower bound for \(d\)-dimensional Euclidean space with the construction of a coreset of size \(\tilde{O}(k^z/z^2)\) points. This improves over the \(\tilde{O}(k^z/e^2)\) upper bound for general power of \(z\) proposed by Braverman Jiang, Krauthgamer, and Wu [SODA’21] and over the \(\tilde{O}(k/e^4)\) upper bound for \(k\)-median by Huang and Vishnoi [STOC’20]. In fact, ours is the first construction breaking through the \(e^{-2} \cdot \min(d, e^{-2})\) barrier inherent in all previous coreset constructions. To do this, we employ a novel chaining based analysis that may be of independent interest. Together our upper and lower bounds for \(k\)-median in Euclidean spaces are tight up to a factor \(O(e^{-1}\text{polylog} k/e)\).

CCS CONCEPTS:
- Theory of computation → Facility location and clustering.

KEYWORDS:
Clustering; coreset; lower-bound; k-means; k-median; sketch

ACM Reference Format:
Vincent Cohen-Addad, Kasper Green Larsen, David Saulpic, and Chris Schwiegelshohn. 2022. Towards Optimal Lower Bounds for k-Median and k-Means Coresets. In Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing (STOC ’22), June 20–24, 2022, Rome, Italy. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3519935.3519946

1 INTRODUCTION
A clustering is a partition of a data set \(P\) such that data points in the same cluster are similar and points in different clusters are dissimilar. Various clustering problems have become important cornerstones in combinatorial optimization and machine learning problems. Among these, center-based clustering problems are arguably the most widely studied and used. Here, the data elements lie in a metric space, every cluster is associated with a center point and the cost of a data point is some function of the distance between data point and its assigned cluster. The \((k, z)\)-Clustering problem captures this and other important objectives via the cost function

\[
\text{cost}(P, S) := \sum_{p \in P} \min_{s \in S} d(p, s)^z,
\]

where \(z\) is a positive integer, \(|S| = k\) and \(d(\cdot, \cdot)\) denotes the distance function. For \(z = 1\), this is the \(k\)-median problem and for \(z = 2\), this is the equally intensely studied \(k\)-means problem.

The classical algorithmic way of thinking assumes that whole dataset can easily fit in memory. In most application however, this is out of the question: datasets may be distributed, change dynamically over time, or contain hundreds of millions of points. In those cases,
it is necessary to find a compact representation of the input data in order to simply store it, and to find a clustering of it. This leads to a tradeoff: the smaller the dataset, the less storage we need and the faster we can run an algorithm on the data set, but conversely, the smaller the data set the more information about the original data will be lost. Coresets formalize and study this tradeoff. Specifically, given a precision parameter $\varepsilon$, $k$ and $z$, an $(\varepsilon, k, z)$-coreset $\Omega$ is a subset of $P$ with weights $w : \Omega \to \mathbb{R}$ that approximates the cost of $P$ for any candidate solution $S$ up to a $(1 \pm \varepsilon)$ factor, namely

$$\forall S, (1 - \varepsilon)\text{cost}(P, S) \leq \sum_{p \in \Omega} w(p)\text{cost}(p, S) \leq (1 + \varepsilon)\text{cost}(P, S).$$

A small $(\varepsilon, k, z)$-coreset is therefore a good compression of the initial dataset, since it preserves the cost of any possible solution. Instead of storing the full dataset, one can simply store the coreset, saving on memory footprint and speeding up performances. We pay a particular attention to Euclidean Spaces. For those, we formalize and study this tradeoff. Specifically, in that case, the coreset cost of solution $S$ is $(1 \pm \varepsilon)\text{cost}(P, S) + \Delta$. In the case where the input space is infinite (e.g., Euclidean space), the coreset points may be chosen from the whole space, and are not restricted to be part of the input.

Although numerous great work focused on improving the size of coreset constructions, our understanding of coreset lower bounds is comparatively limited, and there is a significant gap between the best upper and lower bounds on the possible coreset size. For example, even for Euclidean $k$-means, nothing beyond the trivial $\Omega(k)$ lower bound is known. In this work, we attempt to systematically obtain lower bounds for these problems.

We pay a particular attention to Euclidean Spaces. For those, we complement our lower bound with a new coreset construction that has an optimal dependency in $1/\varepsilon$.

1.1 Our Results

We settle the complexity of the problem for several cases. First, for finite $n$-point metrics, we prove the following theorem.

**Theorem 1.** For any $0 < \varepsilon < 1/2$, $k$ and $n \geq \varepsilon^{-5}$ such that $\log k = O(\log n)$, there exists a finite $n$ point metric such that any $(\varepsilon, k, z)$-coreset using offset $\Delta$ consists of at least $\Omega\left(\frac{k}{\varepsilon^2 \log n}\right)$ points.

Our result improves over the $\Omega(k\varepsilon^{-1} \log n)$ lower bound of Baker, Braverman, Huang, Jiang, Krauthgamer, and Wu [5]. For the $k$-median and $k$-means objective matches the upper bounds proposed in Feldman and Langberg [42] and Cohen-Addad, Saulpic, and Schwiegelshohn [36] up to polylog($1/\varepsilon$) factors.

For metric space with doubling dimension $D$, we present a lower bound similar to that of Theorem 1:

**Corollary 2.** For any $\varepsilon, k, D$ such that $D \geq 5 \log 1/\varepsilon$ and $k = O(D)$, there exists a graph with doubling dimension $D$ on which any $(\varepsilon, k, z)$-coreset using offset $\Delta$ must have size $\Omega\left(\frac{kd}{\varepsilon^2}\right)$.

This matches up to polylog($1/\varepsilon$) factors the upper bound from [36] for $k$-median and $k$-means.

We also study Euclidean spaces more specifically. Here, the difficulty is that centers can be placed arbitrarily in the space, and not only at input points. Note however that a dependency in the dimension $d$ is not necessary, as it is possible to use dimension reduction to reduce $d$ to $O(\varepsilon^{-2} \log k)$. Our main results for Euclidean spaces is the following.

**Theorem 3** (See Theorem 5 for the exact statement). For any $0 < \varepsilon < 1/2$ and any $k$, there exists a point set such that any $(\varepsilon, k, z)$-coreset using offset $\Delta$ consists of at least $\Omega\left(\frac{1}{\varepsilon^3 \max\{kz\}}\right)$ points.

This lower bound holds for any selection of points (i.e. the coreset may use non-input points), and for any additive offset, which is a generalization initially proposed by Feldman, Schmidt, and Sohler [44] and which has since been used in a number of other papers, see Cohen, Elder, Musco, Musco, and Persu [24], Sohler and Woodruff [82] and Cohen-Addad, Saulpic and Schwiegelshohn [35]. The only previously known results are the $\Omega(k/\sqrt{d})$ bound for $k$-median by Baker, Braverman, Jiang, Krauthgamer, and Wu [3], and the $\Omega(k \cdot \min(d, 2^d/20))$ bound by Huang and Vishnoi [56]. Thus, we obtain the first non-trivial lower bound for Euclidean $k$-means.

We complement the lower bound with the following theorem.

**Theorem 4.** Given a set of points $P$ in $d$-dimensional Euclidean space and any $\varepsilon > 0$, there exists an $(\varepsilon, k, z)$-coreset of size $\tilde{O}(k \cdot \varepsilon^{-2} \cdot d^2(\varepsilon \log 2) \cdot \min(\varepsilon^{-2}, k))$.

This is the first coreset construction with an optimal dependency on $\varepsilon$, at the cost of a quadratic dependency on $d$. Previously, all upper bounds either had a dependency of at least $\varepsilon^{-4}$ [17, 36, 56] or a dependency on $d$ [22, 42].

We note that for the special case of Euclidean $k$-median, we improve the best coreset size from $O(k \cdot \varepsilon^{-4})$ to $O(k \cdot \varepsilon^{-3})$, taking a step to reduce the gap with the lower bound.

A complete overview of previous coreset bounds for Euclidean spaces and finite metrics is given in Table 2. For further related work, we refer to Section 2.

1.2 Overview of Our Techniques

Our results for the Euclidean setting require several important new technical insights and we thus review them first. We later review our approach for our lower bound for general metrics.

**Euclidean Lower Bounds.** The lower bound proof consists of three separate steps which combined proves that any coreset for the point set $P = \{e_1, \ldots, e_d\} \subset \mathbb{R}^d$ (i.e., the standard basis of $\mathbb{R}^d$) must have size $\Omega(k \cdot \varepsilon^{-2})$ (in this proof overview, we focus on $z = 2$) when $d = \Theta(k \cdot \varepsilon^{-2})$. The basic approach is to show that any clustering of $P$ with $k$ centers has large cost, while at the same time, for any coreset $\Omega$ using $o(d)$ weighted points, there is a low cost clustering. Combining the two yields the lower bound. We carry out this proof in three steps. In the first step, we show that any clustering of $P$ using unit norm centers has cost at least $2d - O(\sqrt{d\varepsilon})$. In the next step, we show that for any coreset $\Omega$ consisting of $t$ points and a weighing $w : \Omega \to \mathbb{R}^+$, there is a low-cost clustering using unit norm centers that has cost $2d - \sum_{p \in \Omega} w(p)\|p\|_2$. Combining this with step one implies $\sum_{p \in \Omega} w(p)\|p\|_2 = O(\sqrt{d\varepsilon})$. In the final step, we show that any coreset $\Omega$ must have $\sum_{p \in \Omega} w(p)\|p\|_2 = \Omega(d)$ when $d = \Theta(k \cdot \varepsilon^{-2})$. Combining this with the previous two steps finally yields $\sqrt{d\varepsilon} = \Omega(d) \Rightarrow t = \Omega(d) \Rightarrow t = \Omega(k \cdot \varepsilon^{-2})$. In the following, we elaborate on the high level ideas needed for each of the steps:
Table 1: Comparison between the state-of-the-art bounds and our results. Results marked with * are tight for $k$-median and $k$-means.

| Metric Space | Best upper bound | Best lower bound | Our result |
|--------------|------------------|------------------|------------|
| Discrete Metrics | $O(ke^{-\max(2,\varepsilon)} \log n)$ [36] | $\Omega(ke^{-1} \log n)$ [5] | $\Omega(ke^{-2} \log n)^*$ |
| with doubling dimension $D$ | $O(ke^{-\max(2,\varepsilon)} D)$ [36] | - | $\Omega(ke^{-2} D)^*$ |
| Euclidean $k$-median | $O(ke^{-4})$ [56] | $\Omega(ke^{-1/2})$ [5] | $O(ke^{-3}) \Omega(ke^{-2}D)$ |
| Euclidean $k$-means | $O(ke^{-4})$ [36] | - | $\Omega(ke^{-2})$ |
| Euclidean | $O(ke^{-2} \min(\varepsilon,k))$ [16, 36] | $\Omega(k\varepsilon^{3/100})$ [56] | $O(ke^{-2} \cdot \min(\varepsilon^{-2}, k)) \Omega(ke^{-2})$ |

1. First, we show that any clustering of $P$ using $k$ cluster centers $c_1, \ldots, c_k$ of unit norm, must have cost at least $2d - O(\sqrt{dk})$. To see this, notice that if $e_i$ is assigned to cluster center $c_j$, then the cost of $e_i$ is $\|e_i - c_j\|^2 = \|e_i\|^2 + \|c_j\|^2 - 2\langle e_i, c_j \rangle = 2 - 2c_{ij}$, where $c_{ij}$ denotes the $i$th coordinate of $c_j$. Any cluster center $c_j$ can thus at most reduce the cost of the clustering below $2d$ by an additive $2\sum_{i=1}^k |c_{ij}| \leq 2\|c_j\|$. Moreover, it is only "wasteful" to assign a value different from 0 to $c_{ij}$ if $e_i$ is not assigned to center $c_j$ (wasteful since $c_j$ is required to have unit norm). Thus the k centers can be thought of as having disjoint supports. Thus on average, they only have $d/k$ coordinates available. By Cauchy-Schwarz (i.e. the maximum ratio between $\|c_j\|$ and $\|c_j\|_2$), we can argue that $\sum_j |c_j| \leq \sqrt{dk} \sum_j |c_j|_2 = \sqrt{dk}$ and the conclusion follows.

2. Next, we argue that for any coreset $\Omega$ consisting of $t$ points and a weighing $w : \Omega \rightarrow \mathbb{R}^\ast$, we can find a low-cost clustering in terms of $\sum_{p \in \Omega} w(p)\|p\|_2$ using unit norm centers. This is achieved by partitioning the points of the coreset into $k$ groups of $t = t/k$ points each and using one center for each group. For a group of $t$ points $r_1, \ldots, r_t$, we choose the center as something that resembles the mean scaled to have unit norm. More precisely, we consider a random vector $u = \sum_{j=1}^t \sigma_j w(r_j) r_j$ for uniform random and independent signs $\sigma_j$. We can then argue that there is a existing of the signs, such that if $u$ is scaled to have unit norm and this is repeated for all $k$ groups, the resulting cluster cost is at most $2d - O(\sqrt{t\|w\|_2})$.

3. In the last step, we need to argue that any coreset $\Omega$ and weighing $w : \Omega \rightarrow \mathbb{R}^\ast$ must have $\sum_{p \in \Omega} w(p)\|p\|_2 = \Omega(d)$ when $d = \Theta(k \cdot \varepsilon^{-2})$. This is the technically most challenging part of the proof. The basic idea for arguing this, is to exploit that $\Omega$ must be a coreset for many different clusterings of $P = \{e_1, \ldots, e_q\}$. In particular, we consider the Hadamard basis over $q = d/k$ coordinates. The Hadamard basis consists of $q$ orthogonal vectors with coordinates in $\{-1/\sqrt{q}, 1/\sqrt{q}\}$, all having at least half of the coordinates equal to $1/\sqrt{q}$. For each point $v$ in the basis, we consider a clustering where we use $k$ centers $c_1, \ldots, c_k$ that are all copies of $v$ shifted to take up either the first $q$ coordinates in $\mathbb{R}^d$, the next $q$ coordinates and so on. Since half of the coordinates of any $v$ are $1/\sqrt{q}$, the cost of this clustering on $P$ is $2d - \Omega(d/\sqrt{q})$ (if $e_i$ is assigned to a center with the $i$th coordinate equal to $1/\sqrt{q}$ then the cost of $e_i$ is $2 - 2/\sqrt{q}$). Thus intuitively, the points $r_1, \ldots, r_t$ in any coreset $\Omega$ also must have $\sum_{i=1}^t \max_{j=1}^k \langle r_i, c_j \rangle = \Omega(d/\sqrt{q})$. This means that on average over all $r_i$, we have $\max_{j=1}^k \langle r_i, c_j \rangle = \Omega(d/(t\sqrt{q}))$. The crucial observation is that we can repeat this argument for every $\varepsilon$ in the basis. There are $q$ such $\varepsilon$'s. Moreover, for any point $r_j$ in the coreset, the set of $q$ centers $c_{ij}^1, \ldots, c_{ij}^q$ it is assigned to in these $q$ different clusterings are all orthogonal vectors. Thus by Cauchy-Schwarz, we must have $\sqrt{q/d} t \leq \sum_{j=1}^q (\langle w(r_j), c_{ij} \rangle^2)$, that is $\langle w(r_j), c_{ij} \rangle^2 \geq \sum_{j=1}^q \langle c_{ij}, c_{ij} \rangle / q = \Omega(d/\sqrt{q})$. Summing over all $r_i$ completes the proof. Finally, let us remark where the requirement $d = \Theta(k \cdot \varepsilon^{-2})$ enters the picture. We argued that the cost of clustering $P$ using the Hadamard basis was $2d - \Omega(d/\sqrt{q})$. In the coreset, the clustering is allowed to be a factor $(1 + \varepsilon)$ larger. Thus the overall cost of $(2d - \Omega(d/\sqrt{q})) (1 + \varepsilon) \leq 2d - \Omega(d/\sqrt{q})$, which is satisfied when $d = \Omega(d/\sqrt{q}) \Leftrightarrow q = \Omega(\varepsilon^{-2})$. But $q = d/k$ and thus this translates into $d = \Omega(k \cdot \varepsilon^{-2})$.

Upper Bounds. Our main technical contribution is an application of chaining techniques used to analyse Gaussian processes for coreset construction, see Talagrand for an extensive introduction [83]. To the best of our knowledge, we are not aware of any prior attempts of using chaining to improve coreset bounds directly. For readers that may not be familiar with the technique, we now highlight how it allows us to improve over previous constructions. For every candidate solution $S$, we say that $\sigma^S$ is the cost vector associated with $S$, where $\sigma^S_p$ is simply the cost of point $p$ in $S$. A sampling based coreset now picks rows of $\sigma^S$ according to some distribution and approximates $\|\sigma^S\|_2 = \sum_{p \in S} \sigma^S_p$ as the weighted average of the costs of the picked points. To show that this weighted average is concentrated, we require two ingredients. First, we bound the variance for approximating any $\|\sigma^S\|_2$. Suppose we make the simplifying assumption that all points less than 1 and that we are aiming for an additive error of at most $\varepsilon \cdot n$. In this case, the variance is constant, upon which applying a Chernoff bound requires only $\text{Var} \cdot \varepsilon^2$ samples to approximate any single $\|\sigma^S\|_2$. Second, we have to apply a union bound over all $\sigma^S$. In Euclidean spaces, a naive union bound is useless, as there are infinitely many candidate solutions. To discretize $S$, previous work, either implicitly or explicitly, showed that there exists a small set of vectors $N^\varepsilon$, henceforth called a net, such that for every $\sigma^S$ there exists $\sigma^{N^\varepsilon} \in N^\varepsilon$ with $\|\sigma^S - \sigma^{N^\varepsilon}\|_2 \leq \varepsilon$. Thus, an accurate estimation of $\|\sigma^S\|_2$ for all $\sigma^S \in N^\varepsilon$ is sufficient to achieve an estimation for all $\sigma^S$. Unfortunately, the only known bounds of $N^\varepsilon$ are of the order $\exp(k \min(d, \varepsilon^{-2}))$, which combined with bound of the variance leads to $\log |N^\varepsilon| \cdot \text{Var} \cdot \varepsilon^{-2} = k \cdot \varepsilon^{-2} \cdot \min(\varepsilon^{-2}, d)$ many samples.
To improve upon this idea, we use nets at different scales, i.e. we have nets $\mathbb{N}^1, \mathbb{N}^{1/2}, \mathbb{N}^{1/4}$ and so on. These nets allow us to write every $\nu^S$ as a telescoping sum of net vectors at different scales, that is

$$\nu^S = \sum_{h=0}^{\infty} \nu^S_{2^{-h}(h+1)} - \nu^S_{2^{-h}} ,$$

where $\nu^S_{2^{-h}}$ is an element of $B^{2^{-h}}$. Instead of applying the union bound for all vectors in $B^\varepsilon$ at once, we apply the union bound for all difference vectors at various scales, i.e. we show that for all difference vectors $\nu^S_{2^{-h}(h+1)} - \nu^S_{2^{-h}}$

$$\mathbb{P}\left[ |\nu^S_{2^{-h}(h+1)} - \nu^S_{2^{-h}} - \mathbb{E}[\nu^S_{2^{-h}(h+1)} - \nu^S_{2^{-h}}] | \geq \varepsilon \cdot n \right]$$

is small.

The reason why this improves over the naive discretization is that as the nets get finer, the difference also gets smaller, i.e. $|\nu^S_{2^{-h}(h+1)} - \nu^S_{2^{-h}}| \leq 2 \cdot 2^{-h}$. This difference directly affects the bound on the variance, which decreases from a constant to roughly $2 \cdot 2^{-h} \cdot O(1)$. Since there are only $|B^{2^{-h}(h+1)}| \cdot |B^{2^{-h}}| \in \exp(\cdot 2^{-2h}) \cdot O(1)$ many difference vectors, we can compensate the increase in net size by a decrease in variance, i.e. we require only

$$\log(|B^{2^{-h}(h+1)}| \cdot |B^{2^{-h}}|) \cdot \text{Var} \cdot \varepsilon^{-2} \approx k \cdot 2^{-2h} \cdot O(1) \cdot 2^{-2h} \cdot \varepsilon^{-2} = k \cdot \varepsilon^{-2} \cdot O(1)$$

many samples. Applying this idea to every successive summand of the telescoping sum (or rather to every link of the chain of net vectors), leads to an overall number of samples of the order $k \cdot \varepsilon^{-2}$, ignoring polylog factors.

Unfortunately, improving the analysis from an additive approximation to a multiplicative approximation leads to several difficulties. Without using the assumption that all points cost less than 1, the variance increases. Indeed, contrasting to the previous work [35] that used a chaining-based analysis to obtain coreset bounds for a single center and previous work [36] that used a chaining-inspired variance reduction technique, both of which managed to obtain constant variance, bounding the variance in this setting is highly non-trivial and requires a number of new ideas. The lowest variance we could show for estimating $\|\nu^S\|_1$ is only of the order $\min(\varepsilon^{-2}, k)$, leading to the (likely suboptimal) bound of $O(k \cdot \varepsilon^{-2} \cdot \min(\varepsilon^{-2}, k))$ and moreover this bound on the variance is tight. Further ideas will be necessary to reach the (conjectured) optimal bound of $O(k \cdot \varepsilon^{-2})$.

**Lower Bound for discrete metric spaces.** The general idea behind our lower bound is to use the tight concentration and anti-concentration bounds on the sum of random variables.

We first build an instance for $k = 1$, and combines several copies of it to obtain a lower bound for any arbitrary $k$. Our instance for $k = 1$ is such that: (1) when $|\Omega| \leq \varepsilon^{-2} \log |C|$ there exists a center with cost $\Omega(c) > (1 + 100\varepsilon)c$ and (2) for any $|\Omega| > \varepsilon^{-2} \log |C|$ there exists a center $c$ with $\text{cost}(\Omega, c) \in (1 \pm \varepsilon)c\text{cost}(c)$.

To show the existence of such an instance, we consider a complete bipartite graph with nodes $P \cup C$ where there is an edge between each point of $P$ and each point of $C$, with length 1 with probability 1/4 and 2 otherwise. The set of clients is $P$. For simplicity, we will assume here that the coreset weights are uniform.
1.3 Roadmap

We start by giving an extensive overview of related work. The proof of the Euclidean lower bound for $k$-Means is given in Section 4. The proof for general powers is given in in the full version of the paper [30]. The lower bounds for finite metrics and doubling metrics are given in Section 5. The proof of the upper bound is given in the full version of the paper.

2 RELATED WORK

For the most part, related work on coresets for $k$ clustering in Euclidean spaces are given in Table 2. A closely related line of research focuses on dimension reduction for $k$-clustering objectives, particularly $k$-means. Starting with [38], a series of results [7, 9–11, 24, 37, 44, 45, 65, 73, 82] explored the possibility of using dimension reduction methods for $k$-clustering, with a particular focus on principal component analysis (PCA) and random projections. The problem of dimension reduction, at least with respect to these techniques has been mostly resolved by now: Cohen, Elder, Musco, Musco, and Persu [24] proved tight bounds of $[k/\epsilon]$ for PCA and Makarychev, Makarychev and Rakhmanov [73] gave a bound of $O(k^2 \log k/\epsilon)$ for random projections, which nearly matches the lower bound by Larsen and Nelson [67]. The arguably most important technique for combining dimension reduction with coresets is the recent work on terminal embeddings, see [23, 39, 72]. Notably, Narayanan and Nelson [79] gave an optimal bound of $O((e^\epsilon \log n)$.

While Euclidean spaces are doubtlessly the most intensively studied metric, a number of further metrics have also been considered, including finite metrics [22, 26, 42], doubling metrics [36, 52], and graph metrics [5, 17, 36]. Coresets also feature prominently in streaming literature; see [12, 13, 19, 46, 47] for results with a special focus on various streaming models. Other related work considers generalizations of $k$-median and $k$-means by either adding capacity constraints [6, 31, 53, 81], generalizing the notion of centers to subspaces [18, 42, 43], time series [55] or sets [62] or considering more general objective functions [4, 14]. Coresets have also been studied for many other problems: we cite non-comprehensively decision trees [61], kernel methods [60, 63, 80], determinant maximization [58], diversity maximization [59], shape fitting problems [1, 21], linear regression [8, 54, 84], logistic regression [57, 78], Gaussian mixtures [76], dependency networks [76], or low-rank approximation [71]. The interested reader is referred to [2, 41, 77] and similar surveys for more pointers to coreset literature.

In terms of approximation guarantee, the best known approximation ratio for general metrics is $2.67$ due to Byrka et al. [20], improving over the result of 2.71 of Li and Svensson [69] while computing a better than $1 + 2/e$-approximation has been shown to be NP-hard by Guha and Khuller [49]. In Euclidean spaces of arbitrary dimension, the best known approximation is 2.408 and 5.957 for $k$-median and $k$-means, respectively, due to a recent result of Cohen-Addad et al. [26] who improved over the work of Grandoni et al. [48] and Ahmadian et al. [3]. The best known hardness of approximation is 1.73 and 1.27 for $k$-means and $k$-median assuming the Johnson–Coverage Hypothesis or 1.17 and 1.07 respectively assuming $P \neq NP$ [28] (see also [33, 34, 68]). For graphs excluding a fixed-minor, the problem is NP-Hard [74] and a PTAS is known [29, 32]. For doubling metrics, the problem is NP-Hard.

### Table 2: Comparison of coreset sizes for $(k, z)$-Clustering in Euclidean spaces.

| Reference | Size (Number of Points) |
|-----------|--------------------------|
| **Lower Bounds** | |
| Baker, Braverman, Huang, Jiang, Krauthgamer, Wu (ICML’19) [14] | $\Omega(k \cdot e^{-1/2})$ |
| Huang, Vishnoi (STOC’20) [56] | $\Omega(k \cdot \min(d, 22^z/20))$ |
| This paper | $\Omega(k \cdot e^{-2 \cdot z^2})$ |
| **Upper Bounds** | |
| Har-Peled, Mazumdar (STOC’04) [51] | $O(k \cdot e^{-d \cdot \log n})$ |
| Har-Peled, Kushal (DCG’07) [50] | $O(k^3 \cdot e^{-\log d})$ |
| Chen (Sicomp’09) [22] | $O(k^2 \cdot d \cdot e^{-2 \cdot z^2})$ |
| Langberg, Schulman (SODA’10) [66] | $O(k^3 \cdot d \cdot e^{-2 \cdot z^2})$ |
| Feldman, Langberg (STOC’11) [42] | $O(k \cdot d \cdot e^{-2 \cdot z^2})$ |
| Feldman, Schmidt, Sohler (Sicomp’20) [14] | $O(k^3 \cdot e^{-4})$ |
| Sohler, Woodruff (FOCS’18) [82] | $O(k^2 \cdot e^{-\log (z+1)})$ |
| Becchetti, Bury, Cohen-Addad, Grandoni, Schwegelshohn (STOC’19) [7] | $O(k \cdot e^{-8})$ |
| Huang, Vishnoi (STOC’20) [56] | $O(k \cdot e^{-2 \cdot z^2})$ |
| Bravermann, Jiang, Krautgamer, Wu (SODA’21) [16] | $O(k^2 \cdot e^{-4})$ |
| Cohen-Addad, Saulpic, Schwegelshohn (STOC’21) [36] | $\tilde{O}(k \cdot e^{-2 \cdot \max(2z, 1)})$ |
| This paper | $O(k \cdot e^{-2 \cdot \min(e^{z^2}, k)})$ |

**General $n$-point metrics,**

$D$ denotes the doubling dimension.

| **Lower Bounds** | |
| Braverman, Jiang, Krautgamer, Wu (ICML’19) [15] | $\Omega(k \cdot e^{-1 \cdot \log n})$ |
| This paper | $\Omega(k \cdot e^{-2 \cdot \log n})$ |
| This paper | $\Omega(k \cdot e^{-2 \cdot D})$ |
| **Upper Bounds** | |
| Chen (Sicomp’09) [22] | $O(k^2 \cdot e^{-2 \cdot z^2 \cdot \log^2 n})$ |
| Feldman, Langberg (STOC’11) [42] | $O(k \cdot e^{-2z} \cdot \log n)$ |
| Huang, Jiang, Li, Wu (FOCS’18) [52] | $O(k^3 \cdot e^{-2 \cdot D})$ |
| Cohen-Addad, Saulpic, Schwegelshohn (STOC’21) [36] | $\tilde{O}(k \cdot e^{-\max(2z,D)})$ |
| Cohen-Addad, Saulpic, Schwegelshohn (STOC’21) [36] | $\tilde{O}(k \cdot e^{-\max(2z, 1) \cdot \log n})$ |
We first prove the bound for \( \beta \) and a linear-time approximation scheme when the dimension is considered constant is known [25, 27, 64].

3 PRELIMINARIES

General Preliminaries. Given two points \( p \) and \( c \) in some metric space with distance function \( d \), the \( k \)-clustering cost of \( p \) to \( c \) is \( \text{cost}(p, c) = \text{dist}^2(p, c) \). The \( \ell_2 \) norm of a \( d \) dimensional vector \( x \) is defined as \( \|x\|_p = \sqrt{\sum_{i=1}^{d} |x_i|^p} \). If the value of \( p \) is unspecified, it is meant to be the Euclidean norm \( |c| = 2 \). Given a set of points \( P \) with weights \( w : P \to \mathbb{R}^+ \) on a metric space \( I \) and a solution \( S \), we define \( \text{cost}_p(P, S) := \sum_{p \in P} w(p) \text{cost}(p, S) \).

Definition 1. Let \((X, \text{dist})\) be a metric space, let \( P \subset X \) be a set of clients and let \( \Omega \) be a set of points with weights \( w : \Omega \to \mathbb{R}^+ \) and a constant \( \Delta \). \( \Omega \) is an \((\varepsilon, k, z)\)-coreset using offset \( \Delta \) if for any set \( S \subset X \), \( |S| = k \),

\[
\left| \sum_{p \in P} \text{cost}(p, S) - \left( \Delta + \sum_{p \in \Omega} w(p) \text{cost}(p, S) \right) \right| \leq \varepsilon \sum_{p \in P} \text{cost}(p, S)
\]

\( \Omega \) is a \((\varepsilon, k, z)\)-coreset using offset \( \Delta \) with additive error \( E \) if for any set \( S \subset X \), \( |S| = k \),

\[
\left| \sum_{p \in P} \text{cost}(p, S) - \left( \Delta + \sum_{p \in \Omega} w(p) \text{cost}(p, S) \right) \right| \leq \varepsilon \sum_{p \in P} \text{cost}(p, S) + E.
\]

The offset \( \Delta \) is often 0 for most coreset constructions, with a few exceptions [24, 44, 82]. In our algorithm, \( \Delta = 0 \). The lower bounds hold for any choice of \( \Delta \).

4 LOWER BOUNDS IN EUCLIDEAN SPACES FOR \( k \)-MEANS

We first prove the bound for \( k \)-means, i.e. for \( z = 2 \). The generalization to arbitrary powers is available in the full version of the paper: the proof idea is exactly alike, but a few new technicalities arise.

As mentioned in the proof outline in Section 1.2, we proceed in three steps. First we show that any clustering of \( e_1, \ldots, e_d \) using \( k \) cluster centers of unit norm must have cost at least \( 2d - O(\sqrt{dk}) \). Next, we show that for any coreset \( \Omega \) of \( t \) points and weights \( w : \Omega \to \mathbb{R}^+ \), there is a clustering that has cost at most \( 2d - O(\sqrt{kt} \cdot \sum_{p \in \Omega} w(p) \|p\|_2) \). Combined with step one, this shows that \( \sum_{p \in \Omega} w(p) \|p\|_2 = O(\sqrt{kt} \cdot \sum_{p \in \Omega} w(p) \|p\|_2) = O(\sqrt{td}) \). Finally we show that \( \Omega \) must satisfy \( \sum_{p \in \Omega} w(p) \|p\|_2 = \Omega(d) \) when \( d = \Theta(k \cdot \varepsilon^{-2}) \). Combining all of these implies \( \sqrt{td} = \Omega(d) \Rightarrow t = \Omega(d) = \Omega(k \cdot \varepsilon^{-2}) \).

For technical reasons, we consider the point set \( e_1, \ldots, e_d \) as residing in \( \mathbb{R}^{2d} \) and not \( \mathbb{R}^d \). The reason for this, is that we need to be able to find a vector that is orthogonal to all \( e_i \) and all points in a coreset (see proof of Lemma 4). If the size of the coreset is \( t < d \), then such a vector exists in \( \mathbb{R}^{2d} \).

Step One. We start by showing that any clustering of \( e_1, \ldots, e_d \) using \( k \) centers of unit norm must have large cost:

Lemmas 1. For any \( d \), consider the point set \( P = \{e_1, \ldots, e_d\} \) in \( \mathbb{R}^{2d} \). For any set of \( k \) centers \( c_1, \ldots, c_k \in \mathbb{R}^{2d} \) with unit norm, it holds that \( \sum_{i=1}^{k} \text{min}_{j \neq i} \|e_i - c_j\|_2^2 \geq 2d - 2\sqrt{dk} \).

Proof. We see that

\[
\sum_{i=1}^{d} \min_{j=1}^{k} \|e_i - c_j\|_2^2 = \sum_{i=1}^{d} \min_{j=1}^{k} \|e_i\|_2^2 + \|c_j\|_2^2 - 2(e_i, c_j) \\
= 2d - 2 \sum_{i=1}^{d} \max_{j=1}^{k} (e_i, c_j) \\
= 2d - 2 \sum_{j=1}^{k} \sum_{i,j \neq \text{argmax}_k (e_i, c_j)} (e_i, c_j).
\]

Now, for each \( c_j \), define \( \hat{c}_j \) to equal \( c_j \), except that we set the \( i \)th coordinate to 0 if \( j \neq \text{argmax}_k (e_i, c_j) \). Then:

\[
2d - 2 \sum_{j=1}^{k} \sum_{i,j \neq \text{argmax}_k (e_i, c_j)} (e_i, \hat{c}_j) = 2d - 2 \sum_{i=1}^{d} \sum_{j=1}^{k} (e_i, \hat{c}_j) \\
\geq 2d - 2|k| \sum_{j=1}^{k} \|\hat{c}_j\|_1.
\]

By Cauchy-Schwartz, we have \( \| \sum_{j=1}^{k} \hat{c}_j \|_1 \leq \| \sum_{j=1}^{k} \hat{c}_j \|_2 \cdot \sqrt{k} \). Since \( \hat{c}_j \)’s are orthogonal and have norm at most 1, we have \( \| \sum_{j=1}^{k} \hat{c}_j \|_2 \leq \sqrt{k} \). Thus we conclude \( \sum_{i=1}^{d} \min_{j \neq \text{argmax}_k (e_i, c_j)} \|e_i - c_j\|_2^2 \geq 2d - 2\sqrt{dk} \).

Step Two. Next, we show that for any coreset \( \Omega \) of \( t \) points and weights \( w : \Omega \to \mathbb{R}^+ \), there is a clustering that has cost at most \( 2d - \Omega(\sqrt{kt} \cdot \sum_{p \in \Omega} w(p) \|p\|_2) \). To prove this, we start by considering the case of using a single cluster center to cluster \( t \) weighted points:

Lemma 2. Let \( r_1, \ldots, r_t \in \mathbb{R}^{2d} \) and let \( w_1, \ldots, w_t \in \mathbb{R}^+ \). There exists a unit vector \( v \) such that \( \sum_{i=1}^{t} w_i \|r_i - v\|_2^2 \geq \frac{\sum_{i=1}^{t} w_i \|r_i\|_2^2}{\sqrt{t}} \).

Proof. Consider the random vector \( u = \sum_{i=1}^{t} w_i \sigma_i r_i \) where the \( \sigma_i \) are i.i.d. uniform Rademachers \((-1 + 1) / 2\) with probability 1/2). We see that

\[
\sum_{i=1}^{t} w_i \|r_i - u\|_2^2 = \sum_{i=1}^{t} w_i \sum_{j=1}^{t} w_j \sigma_j (r_i, r_j) \\
= \sum_{i=1}^{t} w_i \sum_{j=1}^{t} w_j \sigma_j (r_i, r_j) \\
\geq \sum_{i=1}^{t} w_i \sum_{j=1}^{t} w_j \sigma_j (r_i, r_j) \\
= \|u\|_2^2.
\]

We may then define the unit vector \( v = u / \|u\|_2 \) (with \( u = 0 \) when \( u = 0 \)) and conclude that

\[
\sum_{i=1}^{t} w_i \|r_i - v\|_2^2 \geq \|u\|_2^2.
\]
Towards Optimal Lower Bounds for k-Median and k-Means Coresets

$\Omega(\sqrt{kd})$}.

**Lemma 4.** For any $d$, consider the point set $P = \{e_1, \ldots, e_d\} \in \mathbb{R}^{2d}$. Let $r_1, \ldots, r_t \in \mathbb{R}^{2d}$ and let $w_1, \ldots, w_t \in \mathbb{R}^+$ be an $\epsilon$-coreset for $P$, using offset $\Delta$ and with $t < d$. Then we must have $\Delta + \sum_{i=1}^{t} w_i (\|r_j\|^2 + 1) \in (1 \pm 2\epsilon)2d$.

**Proof.** Since $t + \epsilon < 2d$ there exists a unit vector $v$ that is orthogonal to all $r_j$ and $e_j$. Consider placing all $k$ centers at $v$. Then the cost of clustering $P$ with these centers is $2d$. It therefore must hold that $\Delta + \sum_{i=1}^{t} w_i (\|r_j\|^2 + 1 - 2(r_j, v)) = \Delta + \sum_{i=1}^{t} w_i (\|r_j\|^2 + 1) \in (1 \pm 2\epsilon)2d$.

**Lemma 5.** For any $d$ and any $k > 1$, let $P = \{e_1, \ldots, e_d\} \in \mathbb{R}^{2d}$. Let $r_1, \ldots, r_t \in \mathbb{R}^{2d}$ and let $w_1, \ldots, w_t \in \mathbb{R}^+$ be an $\epsilon$-coreset for $P$ with $t < d$, using offset $\Delta$. Then

$$\sum_{i=1}^{t} w_i \|r_i\|_2 \leq 4 \epsilon d + 2 \sqrt{d \Delta}.$$ 

**Proof.** By Lemma 3, we can find $k$ unit vectors $v_1, \ldots, v_k$ such that $\sum_{i=1}^{t} -2w_i \max_{j=1}^{k} (r_i, v_j) \leq -\sqrt{2kd}/t \cdot \sum_{i=1}^{t} w_i \|r_i\|_2$ and moreover, for all $i$ we have $\max_{j=1}^{k} (r_i, v_j) \leq 0$.

We can now extend this to using $k$ centers of unit norm to cluster $t$ weighted points:

**Lemma 3.** Let $r_1, \ldots, r_t \in \mathbb{R}^{2d}$ and let $w_1, \ldots, w_t \in \mathbb{R}^+$ for any positive even integer $d$, there exists a set of $k$ unit vectors $v_1, \ldots, v_k$ such that $\sum_{i=1}^{t} -2w_i \max_{j=1}^{k} (r_i, v_j) \leq -\sqrt{2kd}/t \cdot \sum_{i=1}^{t} w_i \|r_i\|_2$ and moreover, for all $i$ we have $\max_{j=1}^{k} (r_i, v_j) \leq 0$.

**Proof.** Partition $r_1, \ldots, r_t$ into $k/2$ disjoint groups $G_1, \ldots, G_{k/2}$ of at most $2k/t$ vectors each. For each group $G_j$, apply Lemma 2 to find a unit vector $u_j$ with $\sum_{r_i \in G_j} w_i (r_i, u_j) \geq \sum_{r_i \in G_j} w_i \|r_i\|_2 / \sqrt{2/t}$. Let $v_{2j-1} = u_j$ and $v_{2j} = -u_j$. Since we always add both $u_j$ and $-u_j$, it holds for all $r_j$ that $\max_{j=1}^{k} (r_i, v_j) = \max_{j=1}^{k} (r_i, u_j)$. We therefore conclude (notice the $\leq$ rather than $\geq$ due to the negation):

$$\sum_{i=1}^{t} -2w_i \max_{j=1}^{k} (r_i, u_j) = \sum_{i=1}^{t} -2w_i \max_{j=1}^{k} (r_i, u_j) \leq \sum_{j=1}^{k/2} \sum_{r_i \in G_j} -2w_i (r_i, u_j) \leq -2 \sum_{j=1}^{k/2} \sum_{r_i \in G_j} w_i \|r_i\|_2 / \sqrt{2/t} \leq -\sqrt{2} \sum_{i=1}^{t} w_i \|r_i\|_2 / \sqrt{t/k}.$$ 

With this established, we now combine this with step one to show that for any coreset $\Omega$ with $t$ points, we must have the equality $\sum_{p \in \Omega} w(p) \|p\|_2 = O(\sqrt{kd} \sqrt{d}) = O(\sqrt{kd})$. This is established in two smaller steps:

**Lemma 4.** For any $d$, consider the point set $P = \{e_1, \ldots, e_d\} \in \mathbb{R}^{2d}$. Then $\sum_{p \in \Omega} w(p) \|p\|_2 = \Omega(d)$ when $d = O(k \cdot \epsilon^{-2})$.

**Proof.** Consider the Hadamard basis $h_1, \ldots, h_d$ on $q = 1/(36d^2)$ coordinates, i.e. the set of rows in the normalized Hadamard matrix.
This is a set of $q$ orthogonal unit vectors with all coordinates in \((-1/\sqrt{q}, 1/\sqrt{q})\). All $h_i$ except $h_1$ have equally many coordinates that are $-1/\sqrt{q}$ and $1/\sqrt{q}$ and $h_1$ have all coordinates $1/\sqrt{q}$. Now partition the first $d$ coordinates into $k$ groups $G_1, \ldots, G_k$ of $q$ coordinates each. For any $h_i$, consider the $k$ centers $v'_1, \ldots, v'_k$ obtained as follows: For each group $G_j$ of $q$ coordinates, copy $h_i$ into those coordinates to obtain the vector $v'_j$. We must have that 
$$\Sigma_{h=1}^d \min_{j=1}^k \|v_h - v'_j\|^2 = \Sigma_{h=1}^d \min_{j=1}^k \|v_h\|^2 + \|v'_j\|^2 - 2(v_h, v'_j).$$
Since $k > 1$, there is always a $j$ such that $(v_h, v'_j) = 0$. Moreover, for $i = 1$, we have max $\max_{j=1}^k (v_h, v'_j) = 1/\sqrt{q}$ and for $i \neq 1$, it holds that precisely half of all $v_h$ have $\max_{j=1}^k (v_h, v'_j) = 1/\sqrt{q}$. Thus we have 
$$\Sigma_{h=1}^d \min_{j=1}^k \|v_h - v'_j\|^2 \leq (d/2)^2 + (d/2)(2 - 2/\sqrt{q}) = 2d - d/\sqrt{q}.$$ Thus:

$$(1 + \epsilon)(2d - d/\sqrt{q}) \geq \sum_{h=1}^t w_h (\|r_h\|_2^2 + 1 - 2\max_{j=1}^k (r_h, v'_j))$$

By Lemma 4, this is at least

$$(1 - \epsilon)2d - 2 \sum_{h=1}^t w_h k \max_{j=1}^k (r_h, v'_j).$$

We have thus shown

$$(1 + \epsilon)(2d - d/\sqrt{q}) \geq (1 - \epsilon)2d - 2 \sum_{h=1}^t w_h k \max_{j=1}^k (r_h, v'_j) \Rightarrow$$

$$4\epsilon d - (1 + \epsilon)d/\sqrt{q} \geq -2 \sum_{h=1}^t w_h k \max_{j=1}^k (r_h, v'_j) \Rightarrow$$

$$\sum_{h=1}^t w_h k \max_{j=1}^k (r_h, v'_j) \geq (1 + \epsilon)d/(2\sqrt{q}) - 2\epsilon d \Rightarrow$$

$$\sum_{h=1}^t w_h k \max_{j=1}^k (r_h, v'_j) \geq d/(2\sqrt{q}) - 2\epsilon d.$$

Now consider any $r_h$ with weight $w_h$. Collect the vectors $v'_h$, such that $u'_h = v'_j$ with $j^* = \arg\max_j (r_h, v'_j)$. By construction, all these $q$ vectors are orthogonal (either disjoint support or distinct vectors from the Hadamard basis). By Cauchy-Schwarz, we then have 
$$\langle w_h r_h; \sum_{i=1}^q u'_h \rangle \leq w_h \|r_h\|_2 \|\sum_{i=1}^q u'_h\|_2 = w_h \|r_h\|_2 \sqrt{q}.$$ We then see that 
$$dq/(2\sqrt{q}) - 2d\epsilon q \leq \sum_{i=1}^t \sum_{h=1}^t w_h k \max_{j=1}^k (r_h, v'_j) \Rightarrow$$

$$\sum_{h=1}^t \sum_{i=1}^q w_h k \|r_h\|_2 \leq 2 \sum_{h=1}^t w_h \|r_h\|_2 \sqrt{q}.$$ We have thus shown

$$\sum_{h=1}^t w_h \|r_h\|_2 \geq d/2 - 2d\epsilon d/\sqrt{q} = d/2 - 2d/(6\epsilon) = d/2 - d/3 = d/6. \square$$

**Combining it All.**

**Theorem 5.** For any $0 < \epsilon < 1/2$ and any positive even integer $k$, let $d = k/(36\epsilon^2)$ and let $P = \{v_1, \ldots, v_d\} \in \mathbb{R}^{2d}$. Let $u_1, \ldots, u_t \in \mathbb{R}^{2d}$ and let $w_1, \ldots, w_t \in \mathbb{R}^t$ be an $\epsilon$-coreset for $P$, using offset $\Delta$. Then $t \geq \epsilon^{-5}k/180$.

**Proof.** If $t \geq d$, then we are done. Otherwise, we combine Lemma 5 and Lemma 6, to get:

$$d/6 \leq \sum_{i=1}^t w_i \|v_i\|_2$$

$$\leq \frac{4\epsilon d}{\sqrt{2k/t}}$$

$$= \frac{4\epsilon d}{\sqrt{2k/t}}$$

$$= \frac{10\epsilon d}{\sqrt{2k/t}}.$$ This finally implies:

$$t \geq \epsilon^{-5}k/180. \square$$

**5 LOWER BOUNDS FOR DISCRETE METRICS**

We show in this section Theorem 1, that we recall here for convenience:

**Theorem 1.** For any $0 < \epsilon < 1/2$, and $n > \epsilon^{-5}$ such that

$$\log k = O(\log n),$$

there exists a finite $n$ point metric such that any $(\epsilon, k, z)$ coreset using offset $\Delta$ consists of at least $\Omega(\frac{k}{\epsilon^2} \log n)$ points.

To prove the theorem, we create a subinstance that implies a lower bound for the case $k = 1$. The general lower bound for arbitrary $k$ then naturally combines several copies of the subinstance. The key technical part of our proof is the use of some Azuma-Hoeffding type concentration inequality, but where the concentration probability is lower bounded. The results we use are developed in Section 5.1. We present the subinstance in Section 5.2, and the general lower bound in Section 5.3.

**5.1 Technical Lemmas**

Our proof relies on Lemma 8, which we prove using the following result from [40].

**Lemma 7 (Equation 2.11 in [40]).** Let $\xi_1, \ldots, \xi_m$ be independent centered random variables, and $\bar{\epsilon}$ such that

$$\forall i, k \geq 3 \left[ E[\xi_i^2] \right] \leq \frac{1}{2} k! \epsilon^k - 2E[\xi_i^2].$$

Let $\sigma^2 = \sum E[\xi_i^2]$, and $S_m = \sum_{i=1}^m \xi_i$. Then, for all $0 \leq x \leq 0.1\sigma$,

$$\Pr[S_m \geq \sigma x] \geq \left(1 - \Phi\left(x(1 - cx^2/\sigma)\right)\right) \cdot \left(1 - c(1 + x)\frac{\bar{\epsilon}}{\sigma}\right),$$

where $c$ is an absolute positive constant and $\Phi$ is the standard normal distribution function.
We now turn to proving a lower bound for the case where \( k \) and so by the assumptions of the lemma \( \tilde{\xi}_i \) verifies the condition of Lemma 7. They are independent and centered, and:

\[
\mathbb{E}[\tilde{\xi}^2_i] = p(w_i - pw_i)^2 + (1-p)(pw_i)^2
\]

\[
= w_i^2 \left( p - 2p^2 + p^3 - 2p^3 + p^4 \right)
\]

\[
= w_i^2 (p - 2p^3 + p^4) \geq \frac{w_i^2p}{2},
\]

using \( p \leq 1/4 \). The \( k \)-th moment verifies:

\[
\mathbb{E}[\tilde{\xi}^k_i] = w_i^k \cdot \left( (1-p)^k + (1-p) \cdot (p)^k \right) \leq w_i^k p.
\]

hence \( \tilde{\xi}_i \) verifies the condition of Lemma 7 with \( \tilde{\xi} = \max_i w_i \). We want to apply that lemma to \( x \) of the order \( \tilde{\xi}^2 \); therefore, we need to bound that quantity. Note that

\[
\sigma^2 \geq \frac{p}{2} \cdot \sum w_i^2 \geq \frac{p}{2} \cdot \left( \sum w_i^2 / m \right).
\]

(1)

and so by the assumptions of the lemma \( \frac{\sigma}{\tilde{\xi}} \geq \frac{\sqrt{mp}}{10y^2} \). Furthermore,

\[
\frac{\mu}{\sigma} \leq \frac{p \cdot \sum w_i}{\sqrt{\frac{p}{2m} \cdot \sum w_i}} \leq \sqrt{2mp}
\]

(2)

Now, let \( x := \frac{\tilde{\xi}}{10y \cdot \sqrt{2mp}} \). Thus, \( x \) verifies \( x \leq \frac{\sqrt{mp}}{10y^2} \). and so applying Lemma 7 we obtain:

\[
\Pr\left[ \sum w_iX_i - \mu > \epsilon \mu \right] \geq \left( 1 - \Phi \left(\frac{1 - c + 1 + x \cdot \tilde{\xi}}{\sigma}\right) \right) \cdot \left( 1 - c \cdot (1+x) \cdot \tilde{\xi} / \sigma \right)
\]

\[
\geq (1 - \Phi(0.9x)) \cdot 0.9
\]

\[
= 0.9 \cdot \Pr\left[ N(0,1) \geq 0.9x \right]
\]

\[
\geq 0.9 \cdot \frac{1}{2} \left( 1 - \sqrt{1 - e^{-0.9x^2}} \right)
\]

\[
\geq \exp\left( -\frac{\beta}{y^2} \cdot \epsilon \cdot 2mp \right)
\]

\[
\geq \exp\left( -\frac{\beta}{y^2} \cdot \epsilon \cdot 2mp \right),
\]

where \( \beta \) is some absolute constant, and where the last line uses Eq. (2). \( \square \)

### 5.2 A Subinstance for the Case \( k = 1 \)

We now turn to proving a lower bound for the case where \( k = 1 \). This is going to be our building block in the next subsection where we generalize the result to arbitrary \( k \). Let \( \delta = 1/4 \) be a parameter.

**Definition 2.** A subinstance \( U_\delta \) is defined as follows. Let \( C \) be a set of \( n \) candidate centers and \( P \) a set of \( n_U \) clients. The metric on the ground set \( P \cup C \) is defined according to the following probability distribution.

**Lemma 8.** Let \( X_1, \ldots, X_m \) be independent Bernoulli random variables with expectation \( p \leq 1/4, \epsilon > 0 \) and \( w_1, \ldots, w_m \) be some positive weights, such that max \( w_i \leq \gamma \cdot \frac{\sqrt{w_i}}{\sum w_i} \), for some \( \gamma \). Let \( \mu = p \cdot \sum w_i \). Then, there exists a constant \( \beta \) such that

\[
\Pr\left[ \sum w_iX_i - \mu > \epsilon \mu \right] \geq \exp\left( -\frac{\beta}{y^2} \cdot \epsilon \cdot 2mp \right)
\]
• \( \Omega_3 := \{ x \in \Omega : w_x \in [1, \epsilon^{-1}] \} \)
• \( \Omega_4 := \{ x \in \Omega : w_x \in [\epsilon^{-1}, 10 \log(1/\delta) \cdot \epsilon^{-2}] \} \)
• \( \Omega_5 := \{ x \in \Omega : w_x \geq 20 \log(1/\delta) \cdot \epsilon^{-2} \} \)

We will show that, \( \forall i \in \{2, \ldots, 5\} \), \( w_i(c, \Omega_i) \) exceeds its expectation by a factor \((1 + 205\epsilon)\) with large probability, and that \( w_i(c, \Omega_i) \) is negligible.

First, note that since \( \sum_{x \in \Omega} w_x \leq (1 + 1/2)n_U = 3/2 \epsilon^{-2} \log |C| \), it must be that
\[
|\Omega_5| \leq \frac{\log |C|}{10 \log(1/\delta)}.
\]
Hence, \( c \) is connected with length 1 to all points of \( \Omega_5 \) with probability \( \delta^{|\Omega_5|} \geq \exp(\log |C|/10) = |C|^{-1/10} \).

Now, on each group \( \Omega_2, \Omega_3, \Omega_4, \Omega_5 \), the maximum weight cannot be more than \( 20 \log(1/\delta) \cdot \epsilon^{-1} \) times the average.

For \( i \in \{2, 3, 4\} \), \( w_i(c, \Omega_i) \) is the sum of \( m = |\Omega_i| \leq \epsilon^{-2} \eta \log |C| \) random variables \( X_{i,x} \), for \( x \in \Omega_i \), with \( X_x = 0 \) with probability \((1 - \delta)\) and \( X_x = w_x \) with probability \( \delta = 1/4 \). Hence, Lemma 8 gives that:

\[
\Pr[w_i(c, \Omega_i) \geq (1 + 205\epsilon) \cdot E[w_i(c, \Omega_i)]] > \exp\left(-\frac{\beta}{\log(1/\delta)} \epsilon^{-2} \delta \Omega_i\right) \geq \exp(-\log |C|/10),
\]
for some absolute constant \( \beta \) given by Lemma 8 and \( \eta \leq \frac{\log(1/\delta)^2}{\log(\epsilon)} \).

Finally, to deal with \( \Omega_1 \), we note that \( E[w_i(c, \Omega_1)] \leq \epsilon \delta n_U \).

Hence, \( \sum_{i=2}^{5} \mathbb{E}[w_i(c, \Omega_i)] \geq \mathbb{E}[w_i(c, \Omega_i)] - c \delta n_U \), and
\[
\sum_{i=2}^{5} \mathbb{E}[w_i(c, \Omega_i)] \geq (1 + 205\epsilon) \sum_{i=2}^{5} \mathbb{E}[w_i(c, \Omega_i)]
\]
\[
\Rightarrow w_1(c, \Omega) \geq (1 + 200\epsilon) \delta n_U.
\]

Since all groups are disjoint, the variables \( w_i(c, \Omega_i) \) are independent and we can combine the previous equations to get:
\[
\Pr\left[ \sum_{x \in \Omega : \text{dist}(x, \tilde{c}) = 1} w_x \geq (1 + 200\epsilon) \cdot \delta n_U \right] > |C|^{-3/10}.
\]

Since the length of the edges are chosen independently, the probability that there exists no center \( \tilde{c} \) with \( \sum_{x \in \Omega : \text{dist}(x, \tilde{c}) = 1} w_x \geq (1 + 200\epsilon) \cdot \delta n_U \) is at most
\[
\left(1 - |C|^{-3/10}\right)^{|C|} = \exp\left(|C| \log(1 - |C|^{-3/10})\right) 
\]
\[
\leq \exp(-|C|^{-7/10}).
\]

And hence with probability at least \( 1 - \exp(-|C|^{-7/10}) \) there is a center \( \tilde{c} \) with \( \sum_{x \in \Omega : \text{dist}(x, \tilde{c}) = 1} w_x \geq (1 + 200\epsilon) \cdot \delta n_U \).

To conclude the proof of the first bullet, it remains to do a union-bound over all possible weighted subset \( \Omega \). Such an \( \Omega \) consists of at most \( n_U \) different points, with \( \epsilon/2 \)-rounded weights in \([0, (1 + 1/2) n_U]\). Hence, there are at most \( \frac{1}{2} n_U \) many different weights.

Therefore, there are \( \frac{1}{2} n_U \cdot n_U \) many possible weighted subset \( \Omega \) with \( \epsilon/2 \)-rounded weights, i.e.,
\[
\exp\left(\epsilon^{-2} \log |C| \cdot \log\left(2 \epsilon^{-3} \log |C|\right)\right).
\]

We can conclude that there exists a center \( \tilde{c} \in C \) with \( \sum_{x \in \Omega} w_x = (1 + 200\epsilon) \cdot \delta n_U \) with probability at least
\[
1 - \exp\left(\epsilon^{-2} \log |C| \cdot \log\left(2 \epsilon^{-3} \log |C|\right)\right) \cdot \exp(-|C|^{-7/10}) \geq \frac{99}{100}
\]
by our choice of \( \delta \). Furthermore, \( \Pr[|x \in P : \text{dist}(x, \tilde{c}) = 1| \geq \delta n_U] \geq 1/2 \), because \( |x \in P : \text{dist}(x, \tilde{c}) = 1| \) follows a binomial law with mean \( \delta n_U \). This concludes the proof of the first bullet.

We now turn to the second bullet of the claim, for which the proof is a more standard application of Azuma inequality. Fix some coreset \( \Omega \) of size at least \( \epsilon^{-2} \eta \log |C| \), and a center \( c \). We have,
\[
\Pr[w_i(c, \Omega) \notin (1 \pm \epsilon) \cdot \delta n_U] \leq \exp(-2\epsilon^2 \delta^2 n_U^2) \cdot \exp(-1/2 \cdot \delta^2 \epsilon^2 |C|),
\]
where the second inequality uses \( n_U^2 \geq 1/4 (\sum w_j) \geq 1/4 \cdot \sum w_i^2 \).

Since those events are independent for different centers \( c \), the probability that there exists no center \( c \in C \) with \( w_i(c, \Omega) \in (1 \pm \epsilon) \cdot \delta n_U \) is at most \( \exp(-1/2 \cdot \delta^2 \epsilon^2 |C|) \).

Hence, a union-bound over the \( \frac{n_U}{2} \)-many possible weighted subset \( \Omega \) ensures that the following holds with probability at most
\[
1 - \left(\frac{n_U}{2}\epsilon\right) \cdot \exp(-1/2 \cdot \delta^2 \epsilon^2 |C|) \geq 99/100:
\]
For any \( \Omega \) there exists a center \( c \) with \( w_i(c, \Omega) \in (1 \pm \epsilon) \cdot \delta |\Omega| \) as desired. \( \square \)

### 5.3 Combining Subinstances

![Figure 2: Illustration of a full instance, in the case \( z = 1 \). The subinstance are inside squares, and there is an edge from a node to a square when the node is linked to every point of the subinstance, with the distance written on the edge. \( D_\infty \) is set to be \( 4^{1/\epsilon} \cdot \frac{n_U \cdot k}{\epsilon} \). The node \( e^\epsilon \) is not represented.](image-url)

We now conclude the proof of the lower bound for the \((c, k, z)\) coreset using offset \( \Delta \). We consider \( k \) copies of the subinstance given by Lemma 9, \( U_1^\delta, \ldots, U_k^\delta \), where the set of clients in each subinstance has size \( n_U = 10\epsilon^{-2} \log n \), and the set of candidate centers has size \( |C| \) such that \( |C| \geq \epsilon^{-5} \) and \( (|C| \cdot k) = O(\log |C|) \). In total, there are \( k|C| \) many candidate centers, and \( kn_U \) many different clients. The subinstances are numbered from 1 to \( k \), and connected together in a star-graph metric centered at an arbitrary point \( e^0 \), where all points are at distance \( \frac{n_U \cdot k}{\epsilon} \) of \( e^0 \). There is some additional...
candidate centers: $c_i$, at distance $4 \cdot \frac{n_{U_i}}{\epsilon}$ of every client, and for subinstance $i$ there is a center $c_i', at distance 21/\gamma$ from every client of the subinstance. Fig. 2 illustrates that construction.

We can now turn to the proof of the theorem. For this, we start with three claims: The first one shows that the total weight of the coreset must be very close to the number of point in the instance. The second shows that the offset $\Delta$ must be negligible, and the third that the coreset weight in each subinstance is close to $n_{U_i}$, the number of point in a subinstance.

Claim 6. If $\Omega$ is an $\epsilon$-coreset with offset $\Delta$ for the instance, then the total weight verifies $w(\Omega) \in (1 \pm 2\epsilon)kn_{U}$. 

Proof. Consider the solution consisting only of one center located at $c^{*}$. Let $D_{c^{*}} = \frac{n_{U}}{\epsilon}$, $\epsilon$. This solution has cost $\text{cost}(c^{*}) = kn_{U} \cdot D_{c^{*}}^z$, and $\text{cost}(\Omega, c^{*}) = w(\Omega) \cdot D_{c^{*}}^z$. Hence, 

$$\Delta + w(\Omega) \cdot D_{c^{*}}^z \in (1 \pm 2\epsilon)kn_{U} \cdot D_{c^{*}}^z.$$ 

Similarly, considering the solution that places only one center at $c^{4-\epsilon}$ gives 

$$\Delta + w(\Omega) \cdot D_{c^{4-\epsilon}}^z \in (1 \pm \epsilon)kn_{U} \cdot D_{c^{4-\epsilon}}^z.$$ 

Subtracting those two equations yields: 

$$(4^2 - 1)w(\Omega) \cdot D_{c^{4-\epsilon}}^z = ((4^2 - 1) \pm (4^2 + 1)\epsilon)kn_{U} \cdot D_{c^{4-\epsilon}}^z,$$

and so $w(\Omega) \in (1 \pm 2\epsilon)kn_{U}$. □

Claim 7. If $\Omega$ is an $\epsilon$-coreset with offset $\Delta$ for the instance, then $|\Delta| \leq 3\epsilon k \cdot n_{U}$. 

Proof. Consider the solution $S^2 = \{c_i', \forall i\}$. We have cost($S^2$) = $2kn_{U}$ and cost($\Omega, S^2$) = $2w(\Omega) \in (1 \pm 2\epsilon)\text{cost}(S^2)$, using Claim 6. Since $|\Delta + \text{cost}(\Omega, S^2) - \text{cost}(S^2)| \leq \epsilon \text{cost}(S^2)$, it must be that $|\Delta| \leq 3\epsilon \text{cost}(S^2) = 3\epsilon kn_{U}$. □

Claim 8. If $\Omega$ is an $\epsilon$-coreset with offset $\Delta$ for the instance, then in every subinstance, the sum of the coreset weights is in $(1 \pm 1/2)kn_{U}$. 

Proof. Assume towards contradiction that, in some subinstance, say subinstance $i$, the coreset mass is not in $(1 \pm 1/2)kn_{U}$, and consider a solution $S$ that places one center in each subinstance but subinstance $i$. Suppose w.l.o.g. that the subinstance is overweighted: the coreset places a total weight larger than $3/2 \cdot n_{U}$. in it. The cost of the solution is a most 

$$\text{cost}(S) \leq 2(k-1)kn_{U} + n_{U} \cdot \left(kn_{U} \epsilon^{-1}\right)^z \leq (1 + \epsilon) \left(k \cdot n_{U} \epsilon^{-1}\right)^z,$$

while the cost in the coreset verifies 

$$\Delta + \text{cost}(\Omega, S) > 3\epsilon kn_{U} + 3/2 \cdot n_{U} \cdot \left(kn_{U} \epsilon^{-1}\right)^z \geq (1 + \epsilon) \left(k \cdot n_{U} \epsilon^{-1}\right)^z.$$

(Using Claim 7 and keeping only the cost of the overweighted subinstance) 

$$> (1 + \epsilon) \left(k \cdot n_{U} \epsilon^{-1}\right)^z \geq (1 + \epsilon) \text{cost}(S),$$

hence contradicting the fact that $\Omega$ is an $\epsilon$-coreset with offset $\Delta$. 

The proof of the case where some subinstance is underweighted is done exactly alike. □

We can now turn to the proof of the theorem.

**Proof of Theorem 1.** Assume toward contradiction that there exists an $\epsilon$-coreset with offset $\Delta$ of size smaller than $n_{U} \cdot k \epsilon^{-2} \log |C|$ where $\eta$ is the constant of Lemma 9.

First, this implies the existence of a $2\epsilon$-coreset with $\epsilon$-rounded weights, simply by rounding each weight to the closest multiple of $\epsilon$.

Using Claim 8, we can apply Lemma 9 on each subinstance. The total coreset size is $n_{U} \cdot k \epsilon^{-2} \log |C|$; that means that there are at least $k/10$ subinstances for which the coreset contains no more than $n_{U} \epsilon^{-2} \log |C|$ many different points. We refer to these subinstances as the bad subinstances. Using Lemma 9, we construct a solution $S$ by taking the center given by bullet 1 for the bad subinstances, i.e.: center $\tilde{c}$ as per the notation of Lemma 9, and bullet 2 for the others, i.e.: center $c^{*}$ as per the notation of Lemma 9. The cost of that solution is $n_{1} + 2(kn_{U} - n_{1}) = 2kn_{U} - n_{1}$, where $n_{1}$ the number of edges of length 1 from the clients to $S$. Similarly, the cost of $S$ for the coreset is $2 \cdot w(\Omega) = w_{1}(S, \Omega)$, where $w(\Omega)$ is the total coreset weight and $w_{1}(S, \Omega)$ the weighted number of length 1 edges from $\Omega$ to $S$. By construction of $S$, $w_{1}(S, \Omega)$ verifies 

$$w_{1}(\Omega, \Omega) \geq k/10 \cdot (1 + 200\epsilon)\eta n_{U} + 9k/10 \cdot (1 - \epsilon)\eta n_{U}$$

$$> (1 + 19\epsilon) \delta \cdot \eta n_{U}.$$ 

Furthermore, using properties of Lemma 9, $n_{1} \leq \delta kn_{U}$. Hence, the cost of $S$ in the coreset satisfies 

$$\Delta + 2 \cdot w(\Omega) - w_{1}(S, \Omega) < 3\delta kn_{U} + 2 \cdot (1 + 2\epsilon)kn_{U} - (1 + 19\epsilon) \delta \cdot \eta n_{U}$$

$$\leq (2kn_{U} - n_{1}) + \epsilon n_{U} \cdot (7 - 38\delta)$$

$$< (1 - \epsilon) (2k|S| - n_{1}),$$

where the last inequality uses $\delta = 1/4$, so that $(38\delta - 7)kn_{U} \geq 2kn_{U}$.

Therefore the cost of the coreset for $S$ is smaller than a $(1 - \epsilon)$ factor times the cost of $P$ for $S$, a contradiction that concludes the proof. □

A simple corollary of that proof is a lower bound for metric with bounded doubling dimension. Since any $n$ points metric has doubling dimension $O(\log n)$, the metric constructed has doubling dimension $D = O(\log n) = O(\log |C|)$, which implies Corollary 2.

**ACKNOWLEDGEMENTS**

Kasper Green Larsen is supported by Independent Research Fund Denmark (DFF) Sapere Aude Research Leader grant No 9064-00068B. 

David Saulpic was partially supported by the grant ANR-19-CE48-0016 from the French National Research Agency (ANR). 

Chris Schwiegelshohn is supported by a DFF Sapere Aude Research Leader Grant.

**REFERENCES**

[1] Pankaj K. Agarwal, Sariel Har-Peled, and Kasturi R. Varadarajan. 2004. Approximating extent measures of points. *J. ACM 51*, 4 (2004), 606–635. https://doi.org/10.1145/1008751.1008736
Alejandro Molina, Alexander Munteanu, and Kristian Kersting. 2018. Core Dependency Networks. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (IAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 3820–3827. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16847

Alexander Munteanu and Chris Schwiegelshohn. 2018. Coresets-Methods and History: A Theoreticians Design Pattern for Approximation and Streaming Algorithms. Künstliche Intell. 32, 1 (2018), 37–53. https://doi.org/10.1007/s13218-017-0519-3

Christian Sohler and David P. Woodruff. 2018. Strong Coresets for k-Median and Subspace Approximation: Goodbye Dimension. In 59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018. 802–813. https://doi.org/10.1109/FOCS.2018.00081

Michel Talagrand. 1996. Majorizing measures: the generic chaining. The Annals of Probability 24, 3 (1996), 1049 – 1103. https://doi.org/10.1214/aop/1065725175

Murali Tukac, Alaa Maalouf, and Dan Feldman. 2020. Coresets for Near-Convex Functions. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, Hugo Larochelle, Marc Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/0afe095e81a6ac76f3fe9975cb3e7ae-Abstract.html