HAL open science

Towards Optimal Lower Bounds for k-median and k-means Coresets

Vincent Cohen-Addad, Kasper Green Larsen, David Saulpic, Chris Schwiegelshohn

To cite this version:

Vincent Cohen-Addad, Kasper Green Larsen, David Saulpic, Chris Schwiegelshohn. Towards Optimal Lower Bounds for k-median and k-means Coresets. Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing, Jun 2022, Rome, Italy. pp.1038-1051, 10.1145/3519935.3519946 . hal-03944755

HAL Id: hal-03944755

https://hal.sorbonne-universite.fr/hal-03944755

Submitted on 18 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

arXiv:2202.12793v1 [cs.DS] 25 Feb 2022

Towards Optimal Lower Bounds for k-median and k-means Coresets

Vincent Cohen-Addad* Kasper Green Larsen† David Saulpic‡
Chris Schwiegelshohn†

Abstract

Given a set of points in a metric space, 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. Special cases include the famous 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 raise 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 $\Omega(k \log n/\varepsilon^2)$ points. This improves on the $\Omega(k \log n/\varepsilon)$ 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 $\Omega(kD/\varepsilon^2)$ 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 $\Omega(k/\varepsilon^2)$ 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 $\Omega(k \min(d, 2^{z/20}))$ 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 $\tilde{O}(k/\varepsilon^2 \cdot \min(\varepsilon^{-z}, k))$. This improves over the $\tilde{O}(k^2\varepsilon^{-4})$ upper bound for general power of z proposed by Braverman Jiang, Krauthgamer, and Wu [SODA'21] and over the $\tilde{O}(k/\varepsilon^4)$ upper bound for k-median by Huang and Vishnoi [STOC'20]. In fact, ours is the first construction breaking through the $\varepsilon^{-2} \cdot \min(d, \varepsilon^{-2})$ 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 $O(\varepsilon^{-1}\text{polylog } k/\epsilon)$.

*Google Research, Zurich.
†Aarhus University
‡Sorbonne Université, Paris

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) problem captures this and other important objectives via the cost function

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

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

Datasets used in practice are often huge, containing hundred of millions of points, distributed, or evolving over time. Hence, in these settings classical heuristics (such as Lloyd or k-means++) are lapsed; the size of the dataset forbids multiple passes over the input data and finding a "compact representation" of the input data is of primary importance. 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 orginal data will be lost. *Coresets* formalize and study this tradeoff. Specifically, given a precision parameter ε, k and z, an (ε, k, z) coreset Ω is a subset of P with weights $w : \Omega \to \mathbb{R}$ that approximates the cost of P for any candidate solution $\mathcal{S}$ up to a $(1 \pm \varepsilon)$ factor, namely

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

A small (ε, 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 note that in some definitions, an offset Δ is added to the coreset: in that case, the coreset cost of solution $\mathcal{S}$ is $(1 \pm \varepsilon)\text{cost}(P, \mathcal{S}) + \Delta$. 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.

Metric Space	Best upper bound	Best lower bound	Our result
Discrete Metrics	$O(k\varepsilon^{-\max(2,z)}\log n)$ [35]	$\Omega(k\varepsilon^{-1}\log n)$ [6]	$\Omega(k\varepsilon^{-2}\log n)$*
with doubling dimension D	$O(k\varepsilon^{-\max(2,z)}D)$ [35]	-	$\Omega(k\varepsilon^{-2}D)$*
Euclidean k-median	$\widetilde{O}(k\varepsilon^{-4})$ [55]	$\Omega(k\varepsilon^{-1/2})$ [6]	$\widetilde{O}(k\varepsilon^{-3})$ $\Omega(k\varepsilon^{-2})$
Euclidean k-means	$\widetilde{O}(k\varepsilon^{-4})$ [35]	-	$\Omega(k\varepsilon^{-2})$
Euclidean	$\widetilde{O}(k\varepsilon^{-2-\max(2,z)})$ [35] $\widetilde{O}(k^2\varepsilon^{-4})$ [17]	$\Omega(k2^{z/100})$ [55]	$\widetilde{O}_z(k\varepsilon^{-2}\cdot\min(\varepsilon^{-z},k))$ $\Omega(k\varepsilon^{-2})$

Figure 1: Comparison between the state-of-the-art bounds and our results. Results marked with * are tight for k-median and k-means.

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 (ε,k,z) coreset using offset Δ 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 [6]. For the k-median and k-means objective matches the upper bounds proposed in Feldman and Langberg [41] and Cohen-Addad, Saulpic, and Schwiegelshohn [35] 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 ε,k,D such that $D\geq 5\log 1/\varepsilon$ and $\log k=O(D)$, there exists a graph with doubling dimension D on which any (ε,k,z)-coreset using offset Δ must have size $\Omega\left(\frac{kD}{\varepsilon^2}\right)$.*

This matches up to polylog$(1/\varepsilon)$ factors the upper bound from [35] 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. Our main results for Euclidean spaces is the following.

Theorem 3 (See Theorem 11 for the exact statement)**.** *For any $0<\varepsilon<1/2$ and any k, there exists a point set such that any (ε,k,z) coreset using offset Δ consists of at least $\Omega\left(\frac{k}{\varepsilon^2\max\{1,z^4\}}\right)$ 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 [43] and which has since been used in a number of other papers, see Cohen, Elder, Musco, Musco, and Persu [25], Sohler and Woodruff [86] and Cohen-Addad, Saulpic and Schwiegelshohn [34]. The only previously known results are the $\Omega(k/\sqrt{\varepsilon})$ bound for k-median by Baker, Braverman, Jiang, Krauthgamer, and Wu [6], and the $\Omega(k\cdot\min(d,2^{z/20}))$ bound by Huang and Vishnoi [55]. 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 (ε, k, z) *coreset of size* $\tilde{O}(k \cdot \varepsilon^{-2} \cdot 2^{O(z \log z)} \cdot \min(\varepsilon^{-z}, k))$.

This is the first coreset construction with an optimal dependency on ε, at the cost of a quadratic dependency on k. Previously, all upper bounds either had a dependency of at least ε^{-4} [18, 35, 55] or a dependency on d [23, 41].

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 1. 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 $P = \{e_1, \ldots, e_d\}$ in $\mathbb{R}^d$ (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 Ω 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 $2d - O(\sqrt{dk})$. In the next step, we show that for any coreset Ω 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 $2d - \Omega(\sqrt{k/t} \cdot \sum_{p \in \Omega} w(p)\|p\|_2)$. Combining this with step one implies $\sum_{p \in \Omega} w(p)\|p\|_2 = O(\sqrt{td})$. In the final step, we show that any coreset Ω 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 $\sqrt{td} = \Omega(d) \Rightarrow t = \Omega(d) \Rightarrow t = \Omega(k \cdot \varepsilon^{-2})$. In the following, we elaborate on the high level ideas needed for each of the steps:

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 $\|e_i - c_j\|_2^2 = \|e_i\|_2^2 + \|c_j\|_2^2 - 2\langle e_i, c_j \rangle = 2 - 2c_{j,i}$, where $c_{j,i}$ 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 $2\sum_i c_{j,i} \leq 2\|c_j\|_1$. Moreover, it is only "wasteful" to assign a value different from 0 to $c_{j,i}$ 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-Schwartz (i.e. the maximum ratio between $\|c_j\|_1$ and $\|c_j\|_2$), we can argue that $\sum_j \|c_j\|_1 \leq \sqrt{d/k} \sum_j \|c_j\|_2 = \sqrt{dk}$ and the conclusion follows.

2. Next, we argue that for any coreset Ω consisting of t points and a weighing $w : \Omega \to \mathbb{R}^+$, 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 $\ell = t/k$ points each and using one center for each group. For a group of ℓ points $r_1, \ldots, r_\ell$, we choose the center as something that resembles the mean scaled to have unit norm. More precisely, we consider a random vector $u = \sum_{i=1}^{\ell} \sigma_i w(r_i) r_i$ for uniform random and independent signs σ_i. We can then argue that there is a fixing 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 $2d - \Omega(\sqrt{k/t}\sum_{p\in\Omega} w(p)\|p\|_2)$.

3. In the last step, we need to argue that any coreset Ω and weighing $w : \Omega \to \mathbb{R}^+$ 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 Ω must be a coreset for many different clusterings of $P = \{e_1, \ldots, e_d\}$. 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 vector 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 is 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 Ω 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 $\max_{k=1}^{k} w(r_i)\langle r_i, c_j\rangle = \Omega(d/(t\sqrt{q}))$. The crucial observation is that we can repeat this argument for every v in the basis. There are q such v's. Moreover, for any point r_i in the coreset, the set of q centers $c^1_{i_1}, \ldots, c^q_{i_q}$ it is assigned to in these q different clusterings are all orthogonal vectors. Thus by Cauchy-Schwartz, we must have $\sqrt{q}d/t \leq \sum_{j=1}^{q}\langle w(r_i)r_i, c^j_{i,j}\rangle = \langle w(r_i)r_i, \sum_{j=1}^{q} c^j_{i_j}\rangle \leq \|w(r_i)r_i\|_2\|\sum_{j=1}^{q} c^j_{i,j}\|_2 = w(r_i)\|r_i\|_2\sqrt{q}$. That is, $w(r_i)\|r_i\|_2 = \Omega(d/t)$. 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. We thus require that $(2d - \Omega(d/\sqrt{q}))(1+\varepsilon) \leq 2d - \Omega(d/\sqrt{q})$, which is satisfied when $d\varepsilon = O(d/\sqrt{q}) \Leftrightarrow q = O(\varepsilon^{-2})$. But $q = d/k$ and thus this translates into $d = O(k \cdot \varepsilon^{-2})$.

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 [87]. 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 $\mathcal{S}$, we say that $v^{\mathcal{S}}$ is the cost vector associated with $\mathcal{S}$, where $v^{\mathcal{S}}_p$ is simply the cost of point p in $\mathcal{S}$. A sampling based coreset now picks rows of $v^{\mathcal{S}}$ according to some distribution and approximates $\|v^{\mathcal{S}}\|_1 = \sum v^{\mathcal{S}}_p$ 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 $\|v^{\mathcal{S}}\|_1$. 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 $\mathbf{Var} \cdot \varepsilon^{-2}$ samples to approximate any single $\|v^{\mathcal{S}}\|_1$.

Second, we have to apply a union bound over all $v^{\mathcal{S}}$. In Euclidean spaces, a naive union bound is useless, as there are infinitely many candidate solutions. To discretize $\mathcal{S}$, previous work, either implicitly or explicitly, showed that there exists a small set of vectors $\mathbb{N}^{\varepsilon}$, henceforth called a net, such that for every $v^{\mathcal{S}}$ there exists $v^{\mathcal{S},\varepsilon} \in \mathbb{N}^{\varepsilon}$ with $|v^{\mathcal{S},\varepsilon}_p - v^{\mathcal{S}}_p| \leq \varepsilon$. Thus, an accurate estimation of $\|v\|_1$ for all $v \in \mathbb{N}^{\varepsilon}$ is sufficient to achieve an estimation for all $v^{\mathcal{S}}$. Unfortunately, the only known bounds of $\mathbb{N}^{\varepsilon}$ are of the order $\exp(k \min(d, \varepsilon^{-2}))$, which combined with bound of the variance leads

to $\log|\mathbb{N}^\varepsilon| \cdot \mathbf{Var} \cdot \varepsilon^{-2} = k \cdot \varepsilon^{-2} \cdot \min(\varepsilon^{-2}, d)$ 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 $v^{\mathcal{S}}$ as a telescoping sum of net vectors at different scales, that is

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

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

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

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. $|v_p^{\mathcal{S},2^{-(h+1)}} - v_p^{\mathcal{S},2^{-h}}| \leq 2 \cdot 2^{-h}$. This difference directly affects the bound on the variance, which decreases from a constant to roughly $2^{-2h} \cdot O(1)$. Since there are only $|\mathbb{N}^{2^{-(h+1)}}| \cdot |\mathbb{N}^{2^{-h}}| \in \exp(k \cdot 2^{-2h} \cdot O(1))$ many difference vectors, we can compensate the increase in net size by a decrease in variance, i.e. we require only

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

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 [34] that used a chaining-based analysis to obtain coreset bounds for a single center and previous work [35] 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 $\|v^{\mathcal{S}}\|_1$ is only of the order $\min(\varepsilon^{-z}, k)$, leading to the (likely suboptimal) bound of $\tilde{O}(k \cdot \varepsilon^{-2} \cdot \min(\varepsilon^{-z}, k))$ and moreover this bound on the variance is tight. Further ideas will be necessary to reach the (conjectured) optimal bound of $\Theta(k \cdot \varepsilon^{-2})$.

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 $\text{cost}(\Omega, c) > (1 + 100\varepsilon)\text{cost}(c)$, and (2): for any $|\Omega| > \varepsilon^{-2} \log|C|$ there exists a center c with $\text{cost}(\Omega, c) \in (1 \pm \varepsilon)\text{cost}(c)$.

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. Making the idea work for non-uniform weights requires several other technical ingredients.

In that instance for $k = 1$, the cost of a solution (with a single center, c) is fully determined by $n_1(c)$, the number of length 1 edges to c. Indeed, $\text{cost}(c) = 2(|P| - n_1(c)) + n_1 - c) = 2|P| - n_1(c)$. Let us further assume that $n_1(c)$ is equal to its expectation, $\delta|P|$. For a fixed subset of points Ω, the cost of the solution for Ω with uniform weights $\frac{|P|}{|\Omega|}$ verifies the same equation: it is $2|P| - n_1(\Omega, c) \cdot \frac{|P|}{|\Omega|}$, where $n_1(\Omega, c)$ the number of length 1 edges from Ω to c. Note that $\mathbb{E}[n_1(\Omega, c)] = \delta|\Omega|$.

Using anti-concentration inequalities, we show that $n_1(\Omega, c) > (1+200\varepsilon)\mathbb{E}[n_1(\Omega, c)]$ with probability at least $\exp(-\alpha\varepsilon^2|\Omega|)$, for some constant α. When this event happens, then Ω does not preserve the cost of solution c: indeed,

$$\begin{aligned} 2|P| - n_1(\Omega, c) \cdot \frac{|P|}{|\Omega|} &> 2|P| - (1 + 200\varepsilon)\delta|\Omega| \cdot \frac{|P|}{|\Omega|} \\ &= 2|P| - \delta|P| + 200\varepsilon\delta|P| > (1 + 100\varepsilon)(2|P| - n_1(c)). \end{aligned}$$

Since the edges are drawn independently, the coreset cost for all possible centers c is independent. Hence, there exists one center with $n_1(\Omega, c) > (1 + 200\varepsilon)\delta|\Omega|$ with probability at least $1 - (1 - \exp(-\alpha\varepsilon^2|\Omega|))^{|C|}$. By doing a union-bound over all possible subsets Ω, one can show the following: with positive (close to 1) probability, for any $|\Omega| \leq \varepsilon^{-2} \log |C|$ there exists a center with $\text{cost}(\Omega, c) > (1 + 100\varepsilon)\text{cost}(c)$.

Using standard concentration inequality, one can show that with probability close to 1, for any $|\Omega| > \varepsilon^{-2} \log |C|$, there exists a center c with $\text{cost}(\Omega, c) \in (1 \pm \varepsilon)\text{cost}(c)$. Since the probabilities are taken on the edges randomness, those two result ensure the existence of a graph that verifies properties (1) and (2) desired for the $k = 1$ instance.

Now, the full instance is made of k distinct copies $X_1, ..., X_k$ of the $k = 1$ instance, placed at infinite distance from each other. Let P_i be the set of clients of X_i: the clients for the full instance are $\cup P_i$. Let Ω be a set of at most $1/100 \cdot k\varepsilon^{-2} \log n$ points: we show that Ω cannot be a coreset. By Markov's inequality, there are at least $99/100k$ copies that contain less than $\varepsilon^{-2} \log n$ points of Ω. We say those copies are *bad*, the others are *good*. Consider now the solution $\mathcal{S}$ defined as follows: from each X_i, take the center such that $\text{cost}(\Omega \cap P_i, c) > (1 + 100\varepsilon)\text{cost}(P_i, c)$ when X_i is bad, and the center such that $\text{cost}(\Omega \cap P_i, c) \in (1 \pm \varepsilon)\text{cost}(P_i, c)$ when X_i is good. Observe also that by construction of the instance for $k = 1$, the cost in each copy must lie in $[|P|, 2|P|]$. For that solution, we have:

$$\begin{aligned} \text{cost}(\Omega, \mathcal{S}) = \sum \text{cost}(\Omega \cap P_i, s_i) &= \sum_{i \text{ bad}} \text{cost}(\Omega \cap P_i, s_i) + \sum_{i \text{ good}} \text{cost}(\Omega \cap P_i, s_i) \\ &> \sum_{i \text{ bad}} (1 + 100\varepsilon)\text{cost}(P_i, s_i) + \sum_{i \text{ good}} (1 - \varepsilon)\text{cost}(P_i, s_i) \\ &> \text{cost}(\mathcal{S}) + \frac{99k}{100} \cdot 100\varepsilon|P| - \frac{k}{100} \cdot \varepsilon 2|P| > \text{cost}(\mathcal{S}) + 98k\varepsilon|P| > (1 + \varepsilon)\text{cost}(\mathcal{S}). \end{aligned}$$

Hence, any Ω with $|\Omega| \leq 1/100 \cdot k\varepsilon^{-2} \log n$ cannot be a coreset for our instance, which concludes the proof.

2 Related Work

Reference	Size (Number of Points)
Coreset Bounds in Euclidean Spaces	
Lower Bounds	
Baker, Braverman, Huang, Jiang, Krauthgamer, Wu (ICML'19) [15]	$\Omega(k \cdot \varepsilon^{-1/2})$
Huang, Vishnoi (STOC'20) [55]	$\Omega(k \cdot \min(d, 2^{z/20}))$
This paper	$\Omega(k \cdot \varepsilon^{-2}/z^4)$
Upper Bounds	
Har-Peled, Mazumdar (STOC'04) [50]	$O(k \cdot \varepsilon^{-d} \cdot \log n)$
Har-Peled, Kushal (DCG'07) [49]	$O(k^3 \cdot \varepsilon^{-(d+1)})$
Chen (Sicomp'09) [23]	$O(k^2 \cdot d \cdot \varepsilon^{-2} \cdot \log n)$
Langberg, Schulman (SODA'10) [65]	$O(k^3 \cdot d^2 \cdot \varepsilon^{-2})$
Feldman, Langberg (STOC'11) [41]	$O(k \cdot d \cdot \varepsilon^{-2z})$
Feldman, Schmidt, Sohler (Sicomp'20) [43]	$O(k^3 \cdot \varepsilon^{-4})$
Sohler, Woodruff (FOCS'18) [86]	$O(k^2 \cdot \varepsilon^{-O(z)})$
Becchetti, Bury, Cohen-Addad, Grandoni, Schwiegelshohn (STOC'19) [8]	$O(k \cdot \varepsilon^{-8})$
Huang, Vishnoi (STOC'20) [55]	$O(k \cdot \varepsilon^{-2-2z})$
Bravermann, Jiang, Krautgamer, Wu (SODA'21) [17]	$O(k^2 \cdot \varepsilon^{-4})$
Cohen-Addad, Saulpic, Schwiegelshohn (STOC'21) [35]	$\tilde{O}(k \cdot \varepsilon^{-2-\max(2,z)})$
This paper	$\tilde{O}(k \cdot \varepsilon^{-2} \cdot \min(\varepsilon^{-z}, k))$
General n-point metrics, D denotes the doubling dimension	
Lower Bounds	
Braverman, Jiang, Krauthgamer, Wu (ICML'19) [16]	$\Omega(k \cdot \varepsilon^{-1} \cdot \log n)$
This paper	$\Omega(k \cdot \varepsilon^{-2} \cdot \log n)$
This paper	$\Omega(k \cdot \varepsilon^{-2} \cdot D)$
Upper Bounds	
Chen (Sicomp'09) [23]	$O(k^2 \cdot \varepsilon^{-2} \cdot \log^2 n)$
Feldman, Langberg (STOC'11) [41]	$O(k \cdot \varepsilon^{-2z} \cdot \log n)$
Huang, Jiang, Li, Wu (FOCS'18) [51]	$O(k^3 \cdot \varepsilon^{-2} \cdot D)$
Cohen-Addad, Saulpic, Schwiegelshohn (STOC'21) [35]	$\tilde{O}(k \cdot \varepsilon^{-\max(2,z)} \cdot D)$
Cohen-Addad, Saulpic, Schwiegelshohn (STOC'21) [35]	$\tilde{O}(k \cdot \varepsilon^{-\max(2,z)} \cdot \log n)$

Table 1: Comparison of coreset sizes for (k, z)-Clustering in Euclidean spaces. [15] only applies to k-median, [49, 50] only applies to k-means and k-median, and[8, 43] only applies to k-means. [86] runs in exponential time, which has been addressed by Feng, Kacham, and Woodruff [44]. Aside from [49, 50], the algorithms are randomized and succeed with constant probability. Any dependency on $2^{O(z \log z)}$, as well as polylog factors have been omitted in the upper bounds.

For the most part, related work on coresets for k clustering in Euclidean spaces are given in Table 1. A closely related line of research focusses on dimension reduction for k-clustering objectives, particularly k-means. Starting with [37], a series of results [8, 10, 11, 12, 25, 36, 43, 44, 64, 73, 86] 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 [25] proved tight bounds of $\lceil k/\varepsilon \rceil$ for PCA and Makarychev, Makarychev and Razenshteyn [73] gave a bound of $O(\varepsilon^{-2} \log k/\varepsilon)$ for random projections, which nearly matches the lower bound by Larsen and Nelson [66]. The arguably most important technique for combining dimension reduction with coresets is the recent work on terminal embeddings, see [24, 38, 72]. Notably, Narayanan and Nelson [82] gave an optimal bound of $O(\varepsilon^{-2} \log n)$. We will discuss specifics on terminal embeddings in Section 6.4.

While Euclidean spaces are doubtlessly the most intensively studied metric, a number of further metrics have also been considered, including finite metrics [23, 35, 41], doubling metrics [35, 51], and graph metrics [6, 18, 35]. Coresets also feature prominently in streaming literature, see [13, 14, 20, 45, 46] 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 [7, 29, 52, 85], generalizing the notion of centers to subspaces [19, 41, 42], time series [54] or sets [61] or considering more general objective functions [5, 15]. Coresets have also been studied for many other problems: we cite non-comprehensively decision trees [60], kernel methods [59, 62, 83], determinant maximization [57], diversity maximization [58], shape fitting problems [2, 22], linear regression [9, 53, 88], logistic regression [56, 81], Gaussian mixtures [70], dependency networks [79], or low-rank approximation [71]. The interested reader is referred to [3, 40, 80] 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. [21], 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 [48]. 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. [1] who improved over the work of Grandoni et al. [47] and Ahmadian et al. [4]. 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 $\mathrm{P} \neq \mathrm{NP}$ [32] (see also [31, 33, 68]). For graphs excluding a fixed-minor, the problem is NP-Hard [75] and a PTAS is known [28, 30]. For doubling metrics, the problem is NP-Hard (even in the plane [77]) and a linear-time approximation scheme when the dimension is considered constant is known [27, 26, 63].

2.1 Roadmap

The proof of the Euclidean lower bound for k-Means is given in Section 4. The proof for general powers is given in Appendix A. The lower bounds for finite metrics and doubling metrics are given in Section 5. The proof of the upper bound is given in Section 6.

3 Preliminaries

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

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

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

Ω is a (ε, k, z)-coreset using offset Δ with additive error E if for any set $\mathcal{S} \subset X$, $|\mathcal{S}| = k$,

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

The offset Δ is often 0 for most coreset constructions, with a few exceptions [25, 43, 86]. In our algorithm, $\Delta = 0$. The lower bounds hold for any choice of Δ.

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 made in appendix: the proof idea is exactly alike, but a few new technicalities arise.

4.1 k-Means

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 Ω of t points and weights $w : \Omega \to \mathbb{R}^+$, there is a clustering that has cost at most $2d - \Omega(\sqrt{k/t} \cdot \sum_{p \in \Omega} w(p)\|p\|_2)$. Combined with step one, this shows that $\sum_{p \in \Omega} w(p)\|p\|_2 = O(\sqrt{t/k}\sqrt{dk}) = O(\sqrt{td})$. Finally we show that Ω 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 Ω (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:

Lemma 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 $\sum_{i=1}^{d} \min_{j=1}^{k} \|e_i - c_j\|_2^2 \geq 2d - 2\sqrt{dk}$.*

Proof. We see that

$$\begin{aligned}
\sum_{i=1}^{d} \min_{j=1}^{k} \|e_i - c_j\|_2^2 &= \sum_{i=1}^{d} \min_{j=1}^{k} \|e_i\|_2^2 + \|c_j\|_2^2 - 2\langle e_i, c_j \rangle \\
&= 2d - 2\sum_{i=1}^{d} \max_{j=1}^{k} \langle e_i, c_j \rangle \\
&= 2d - 2\sum_{j=1}^{k} \sum_{i: j = \mathrm{argmax}_h \langle e_i, c_h \rangle} \langle e_i, c_j \rangle.
\end{aligned}$$

Now, for each c_j, define $\hat{c}_j$ to equal c_j, except that we set the i'th coordinate to 0 if $j \neq \mathrm{argmax}_h \langle e_i, c_h \rangle$. Then:

$$\begin{aligned}
2d - 2\sum_{j=1}^{k} \sum_{i: j = \mathrm{argmax}_h \langle e_i, c_h \rangle} \langle e_i, c_j \rangle &= 2d - 2\sum_{i=1}^{d} \sum_{j=1}^{k} \langle e_i, \hat{c}_j \rangle \\
&= 2d - 2\sum_{i=1}^{d} \langle e_i, \sum_{j=1}^{k} \hat{c}_j \rangle \\
&\geq 2d - 2\| \sum_{j=1}^{k} \hat{c}_j \|_1.
\end{aligned}$$

By Cauchy-Schwartz, we have $\| \sum_{j=1}^{k} \hat{c}_j \|_1 \leq \| \sum_{j=1}^{k} \hat{c}_j \|_2 \cdot \sqrt{d}$. Since the $\hat{c}_j$'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 $\sum_{i=1}^{d} \min_{j=1}^{k} \|e_i - c_j\|_2^2 \geq 2d - 2\sqrt{dk}$. □

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

Lemma 2. *Let $r_1, \ldots, r_\ell \in \mathbb{R}^{2d}$ and let $w_1, \ldots, w_\ell \in \mathbb{R}^+$. There exists a unit vector v such that $\sum_{i=1}^{\ell} w_i |\langle r_i, v \rangle| \geq \frac{\sum_{i=1}^{t} w_i \|r_i\|_2}{\sqrt{\ell}}$.*

Proof. Consider the random vector $u = \sum_{i=1}^{\ell} w_i \sigma_i r_i$ where the σ_i are i.i.d. uniform Rademachers

(-1 and $+1$ with probability $1/2$). We see that

$$\begin{aligned}
\sum_{i=1}^{\ell} w_i |\langle r_i, u\rangle| &= \sum_{i=1}^{\ell} w_i \left| \sum_{j=1}^{\ell} w_j \sigma_j \langle r_i, r_j\rangle \right| \\
&= \sum_{i=1}^{\ell} w_i \left| \sum_{j=1}^{\ell} w_j \sigma_i \sigma_j \langle r_i, r_j\rangle \right| \\
&\geq \sum_{i=1}^{\ell} w_i \sum_{j=1}^{\ell} w_j \sigma_i \sigma_j \langle r_i, r_j\rangle \\
&= \|u\|_2^2.
\end{aligned}$$

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

$$\sum_{i=1}^{\ell} w_i |\langle r_i, v\rangle| \geq \|u\|_2.$$

Since $\mathbb{E}[\|u\|_2^2] = \sum_{i=1}^{\ell} w_i^2 \|r_i\|_2^2$ we conclude that there must exist a unit vector v with

$$\sum_{i=1}^{\ell} w_i |\langle r_i, v\rangle| \geq \sqrt{\sum_{i=1}^{\ell} w_i^2 \|r_i\|_2^2}.$$

By Cauchy-Schwartz, we have:

$$\sum_{i=1}^{\ell} |1 \cdot w_i \|r_i\|_2| \leq \sqrt{\sum_{i=1}^{\ell} w_i^2 \|r_i\|_2^2} \cdot \sqrt{\sum_{i=1}^{\ell} 1} = \sqrt{\sum_{i=1}^{\ell} w_i^2 \|r_i\|_2^2} \cdot \sqrt{\ell}$$

which finally implies

$$\sum_{i=1}^{\ell} w_i |\langle r_i, v\rangle| \geq \frac{\sum_{i=1}^{\ell} w_i \|r_i\|_2}{\sqrt{\ell}}.$$

□

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 k, there exists a set of k unit vectors $v_1, \ldots, v_k$ such that $\sum_{i=1}^{t} -2w_i \max_{j=1}^{k} \langle r_i, v_j\rangle \leq -\sqrt{2k/t} \cdot \sum_{i=1}^{t} w_i \|r_i\|_2$ and moreover, for all i we have $\max_{j=1}^{k} \langle r_i, v_j\rangle \geq 0$.*

Proof. Partition $r_1, \ldots, r_t$ arbitrarily into $k/2$ disjoint groups $G_1, \ldots, G_{k/2}$ of at most $2t/k$ vectors each. For each group G_j, apply Lemma 2 to find a unit vector u_j with $\sum_{r_i \in G_j} w_i |\langle r_i, u_j\rangle| \geq \frac{\sum_{r_i \in G_j} w_i \|r_i\|_2}{\sqrt{2t/k}}$. 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_i that $\max_{j=1}^k \langle r_i, v_j\rangle = \max_{j=1}^k |\langle r_i, v_j\rangle|$. We therefore conclude (notice the $\leq$ rather than $\geq$ due to the negation):

$$\begin{aligned}
\sum_{i=1}^{t} -2w_i \max_{j=1}^{k} \langle r_i, v_j\rangle &= \sum_{i=1}^{t} -2w_i \max_{j=1}^{k} |\langle r_i, v_j\rangle| \\
&\leq \sum_{j=1}^{k/2} \sum_{r_i \in G_j} -2w_i |\langle r_i, u_j\rangle| \\
&\leq -2 \sum_{j=1}^{k/2} \frac{\sum_{r_i \in G_j} w_i \|r_i\|_2}{\sqrt{2t/k}} \\
&= -\frac{\sqrt{2}\sum_{i=1}^{t} w_i \|r_i\|_2}{\sqrt{t/k}}.
\end{aligned}$$

□

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

Lemma 4. *For any d, consider the point set $P = \{e_1, \ldots, e_d\}$ in $\mathbb{R}^{2d}$. Let $r_1, \ldots, r_t \in \mathbb{R}^{2d}$ and let $w_1, \ldots, w_t \in \mathbb{R}^+$ be an ε-coreset for P, using offset Δ and with $t < d$. Then we must have $\Delta + \sum_{i=1}^t w_i(\|r_i\|_2^2 + 1) \in (1 \pm \varepsilon)2d$.*

Proof. Since $t + d < 2d$ there exists a unit vector v that is orthogonal to all r_i and all 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 $\Delta + \sum_{i=1}^t w_i(\|r_i\|_2^2 + \|v\|_2^2 - 2\langle r_i, v\rangle) = \Delta + \sum_{i=1}^t w_i(\|r_i\|_2^2 + 1) \in (1 \pm \varepsilon)2d$. □

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 ε-coreset for P with $t < d$, using offset Δ. Then*

$$\sum_{i=1}^{t} w_i \|r_i\|_2 \leq \frac{4\varepsilon d + 2\sqrt{dk}}{\sqrt{2k/t}}.$$

Proof. By Lemma 3, we can find k unit vectors $v_1, \ldots, v_k$ such that $\sum_{i=1}^t -2w_i \max_{j=1}^k \langle r_i, v_j\rangle \leq -\sqrt{2k/t} \cdot \sum_{i=1}^t w_i \|r_i\|_2$. By Lemma 1, it holds that $\sum_{p\in P} \min_{j=1}^k \|p - v_j\|_2^2 \geq 2d - 2\sqrt{dk}$. Since points $r_1, \ldots, r_t$ with respective weights $w_1, \ldots, w_t$ and offset Δ form an ε-coreset for P, we must

have

$$\begin{aligned}
(1-\varepsilon)(2d-2\sqrt{dk}) &\leq \Delta + \sum_{i=1}^{t} \min_{j=1}^{k} w_i \|r_i - v_j\|_2^2 \\
&= \Delta + \sum_{i=1}^{t} w_i(\|r_i\|_2^2 + \|v_j\|_2^2 - 2\max_{j=1}^{k}\langle r_i, v_j\rangle) \\
&= \Delta + \sum_{i=1}^{t} w_i(\|r_i\|_2^2 + 1) - 2\sum_{i=1}^{t} w_i \max_{j=1}^{k}\langle r_i, v_j\rangle \\
&\leq \Delta + \sum_{i=1}^{t} w_i(\|r_i\|_2^2 + 1) - \sqrt{2k/t}\cdot\sum_{i=1}^{t} w_i\|r_i\|_2.
\end{aligned}$$

By Lemma 4, this is at most

$$\leq (1+\varepsilon)2d - \sqrt{2k/t}\cdot\sum_{i=1}^{t} w_i\|r_i\|_2.$$

We have therefore shown that

$$\begin{aligned}
(1-\varepsilon)(2d-2\sqrt{dk}) &\leq (1+\varepsilon)2d - \sqrt{2k/t}\cdot\sum_{i=1}^{t} w_i\|r_i\|_2 \Rightarrow \\
\sqrt{2k/t}\cdot\sum_{i=1}^{t} w_i\|r_i\|_2 &\leq (1+\varepsilon)2d - (1-\varepsilon)(2d-2\sqrt{dk}) \Rightarrow \\
\sqrt{2k/t}\cdot\sum_{i=1}^{t} w_i\|r_i\|_2 &\leq 4\varepsilon d + (1-\varepsilon)2\sqrt{dk} \Rightarrow \\
\sum_{i=1}^{t} w_i\|r_i\|_2 &\leq \frac{4\varepsilon d + 2\sqrt{dk}}{\sqrt{2k/t}}.
\end{aligned}$$

□

Step Three. Finally we show that any coreset Ω must satisfy $\sum_{p\in\Omega} w(p)\|p\|_2 = \Omega(d)$ when $d = \Theta(k\cdot\varepsilon^{-2})$:

Lemma 6. *For any $0 < \varepsilon < 1/2$ and any positive even integer k, let $d = k/(36\varepsilon^2)$ and 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 ε-coreset for P with $t < d$, using offset Δ. Then $\sum_{i=1}^{t} w_i\|r_i\|_2 \geq d/6$.*

Proof. Consider the Hadamard basis $h_1, \ldots, h_q$ on $q = 1/(36\varepsilon^2)$ 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 $v_1^i, \ldots, v_k^i$ obtained as follows: For each group G_j of q coordinates, copy h_i into those coordinates to obtain the vector v_j^i. We must have that

$\sum_{h=1}^{d} \min_{j=1}^{k} \|e_h - v_j^i\|_2^2 = \sum_{h=1}^{d} \min_{j=1}^{k} \|e_h\|_2^2 + \|v_j^i\|_2^2 - 2\langle e_h, v_j^i\rangle$. Since $k > 1$, there is always a j such that $\langle e_h, v_j^i\rangle = 0$. Moreover, for $i = 1$, we have $\max_{j=1}^{k}\langle e_h, v_j^i\rangle = 1/\sqrt{q}$ and for $i \neq 1$, it holds that precisely half of all e_h have $\max_{j=1}^{k}\langle e_h, v_j^i\rangle = 1/\sqrt{q}$. Thus we have $\sum_{h=1}^{d} \min_{j=1}^{k} \|e_h - v_j^i\|_2^2 \leq (d/2)2 + (d/2)(2 - 2/\sqrt{q}) = 2d - d/\sqrt{q}$. Thus:

$$(1+\varepsilon)(2d - d/\sqrt{q}) \quad \geq \quad \Delta + \sum_{h=1}^{t} w_h(\|r_h\|_2^2 + 1 - 2\max_{j=1}^{k}\langle r_h, v_j^i\rangle)$$

By Lemma 4, this is at least

$$\geq \quad (1-\varepsilon)2d - 2\sum_{h=1}^{t} w_h \max_{j=1}^{k}\langle r_h, v_j^i\rangle.$$

We have thus shown

$$\begin{aligned}
(1+\varepsilon)(2d - d/\sqrt{q}) &\geq (1-\varepsilon)2d - 2\sum_{h=1}^{t} w_h \max_{j=1}^{k}\langle r_h, v_j^i\rangle \Rightarrow \\
4\varepsilon d - (1+\varepsilon)d/\sqrt{q} &\geq -2\sum_{h=1}^{t} w_h \max_{j=1}^{k}\langle r_h, v_j^i\rangle \Rightarrow \\
\sum_{h=1}^{t} w_h \max_{j=1}^{k}\langle r_h, v_j^i\rangle &\geq (1+\varepsilon)d/(2\sqrt{q}) - 2\varepsilon d \Rightarrow \\
\sum_{h=1}^{t} w_h \max_{j=1}^{k}\langle r_h, v_j^i\rangle &\geq d/(2\sqrt{q}) - 2\varepsilon d.
\end{aligned}$$

Now consider any r_h with weight w_h. Collect the vectors u_h^i such that $u_h^i = v_{j^*}^i$ with $j^* = \text{argmax}_j\langle r_h, v_j^i\rangle$. By construction, all these q vectors are orthogonal (either disjoint support or distinct vectors from the Hadamard basis). By Cauchy-Schwartz, we then have $\langle w_h r_h, \sum_{i=1}^{q} u_h^i\rangle \leq w_h\|r_h\|_2\|\sum_{i=1}^{q} u_h^i\|_2 = w_h\|r_h\|_2\sqrt{q}$. We then see that

$$\begin{aligned}
dq/(2\sqrt{q}) - 2\varepsilon dq &\leq \sum_{i=1}^{q}\sum_{h=1}^{t} w_h \max_{j=1}^{k}\langle r_h, v_j^i\rangle \\
&= \sum_{h=1}^{t}\sum_{i=1}^{q} w_h\langle r_h, u_h^i\rangle \\
&= \sum_{h=1}^{t}\langle w_h r_h, \sum_{i=1}^{q} u_h^i\rangle \\
&\leq \sum_{h=1}^{t} w_h\|r_h\|_2\sqrt{q}.
\end{aligned}$$

We have thus shown $\sum_{h=1}^{t} w_h\|r_h\|_2 \geq d/2 - 2\varepsilon d\sqrt{q} = d/2 - 2\varepsilon d/(6\varepsilon) = d/2 - d/3 = d/6$. □

Combining it All.

Theorem 5. *For any $0 < \varepsilon < 1/2$ and any positive even integer k, let $d = k/(36\varepsilon^2)$ and 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 ε-coreset for P, using offset Δ. Then $t \geq \varepsilon^{-2}k/180$.*

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

$$\begin{aligned} d/6 &\leq \sum_{i=1}^{t} w_i \|r_i\|_2 \\ &\leq \frac{4\varepsilon d + 2\sqrt{dk}}{\sqrt{2k/t}} \\ &= \frac{4\varepsilon d + 6\varepsilon d}{\sqrt{2k/t}} \\ &= \frac{10\varepsilon d}{\sqrt{2k/t}}. \end{aligned}$$

This finally implies:

$$t \geq \varepsilon^{-2}k/180.$$

□

5 Lower Bounds For Discrete Metrics

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

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 (ε, k, z) coreset using offset Δ consists of at least $\Omega\left(\frac{k}{\varepsilon^2}\log n\right)$ 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 [39].

Lemma 7 (Equation 2.11 in [39])**.** *Let $\xi_1, ..., \xi_m$ be independent centered random variables, and $\tilde{\varepsilon}$ such that*

$$\forall i, k \geq 3 \ |\mathbb{E}[\xi_i^k]| \leq \frac{1}{2}k!\tilde{\varepsilon}^{k-2}\mathbb{E}[\xi_i^2].$$

Let $\sigma^2 = \sum \mathbb{E}[\xi_i^2]$, and $S_m = \sum_{i=1}^m \xi_i$.

Then, for all $0 \leq x \leq 0.1\frac{\sigma}{\tilde{\varepsilon}}$,

$$\Pr[S_m \geq x\sigma] \geq \left(1 - \Phi\left(x(1 - cx\frac{\tilde{\varepsilon}}{\sigma}\right)\right) \cdot \left(1 - c(1+x)\frac{\tilde{\varepsilon}}{\sigma}\right),$$

where c is an absolute positive constant and Φ is the standard normal distribution function.

Lemma 8. *Let $X_1, ..., X_m$ be independent Bernouilli random variables with expectation $p \leq 1/4$, $\varepsilon > 0$ and $w_1, ..., w_m$ be some positive weights, such that $\max w_i \leq \gamma \cdot \frac{\sum w_i}{\varepsilon m}$, for some γ. Let $\mu = p \cdot \sum w_i$. Then, there exists a constant β such that*

$$\Pr\left[\sum w_i X_i - \mu > \varepsilon\mu\right] \geq \exp(-\frac{\beta}{\gamma^2}\varepsilon^2 mp)$$

Proof. Define $\xi_i = w_i X_i - pw_i$. We show that the variables ξ_i verify the conditions of Lemma 7. They are independent and centered, and:

$$\begin{aligned}\mathbb{E}[\xi_i^2] &= p(w_i - pw_i)^2 + (1-p)(pw_i)^2) \\ &= w_i^2\left(p - 2p^2 + p^3 + p^2 - 2p^3 + p^4\right) \\ &= w_i^2(p - p^2 - p^3 + p^4) \geq \frac{w_i^2 p}{2},\end{aligned}$$

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

$$\left|\mathbb{E}[\xi_i^k]\right| = w_i^k \cdot \left(p \cdot (1-p)^k + (1-p) \cdot (-p)^k\right) \leq w_i^k p,$$

hence ξ_i verifies the condition of Lemma 7 with $\tilde{\varepsilon} = \max_i w_i$. We want to apply that lemma to x of the order $\varepsilon\frac{\mu}{\sigma}$: therefore, we need to bound that quantity. Note that

$$\sigma^2 \geq \frac{p}{2}\sum w_i^2 \geq \frac{p}{2} \cdot \frac{(\sum w_i)^2}{m}, \tag{1}$$

and so by the assumptions of the lemma $\frac{\sigma}{\tilde{\varepsilon}} \geq \frac{\varepsilon\sqrt{mp}}{\gamma\sqrt{2}}$. Furthermore,

$$\frac{\mu}{\sigma} \leq \frac{p\sum w_i}{\sqrt{\frac{p}{2m}}\sum w_i} \leq \sqrt{2mp} \tag{2}$$

Now, let $x := \frac{\varepsilon}{10\gamma c\sqrt{2}} \cdot \frac{\mu}{\sigma}$. Thus, x verifies $x \leq \frac{\varepsilon}{10\gamma c\sqrt{2}} \cdot \sqrt{2pm} \leq 0.1\frac{\sigma}{c\tilde{\varepsilon}}$ and so applying Lemma 7 we obtain:

$$\begin{aligned}\Pr\left[\sum w_i X_i - \mu > \varepsilon\mu\right] &\geq \left(1 - \Phi\left(x(1 - cx\frac{\tilde{\varepsilon}}{\sigma})\right)\right) \cdot \left(1 - c(1+x)\frac{\tilde{\varepsilon}}{\sigma}\right) \\ &\geq (1 - \Phi(0.9x)) \cdot 0.9 \\ &= 0.9 \cdot \Pr[\mathcal{N}(0,1) \geq 0.9x] \\ &\geq 0.9 \cdot \frac{1}{2}\left(1 - \sqrt{1 - e^{-(0.9x)^2}}\right) \\ &\geq \exp(-\frac{\beta}{\gamma^2}\varepsilon^2\frac{\mu^2}{\sigma^2}) \\ &\geq \exp(-\frac{\beta}{\gamma^2}\varepsilon^2 mp),\end{aligned}$$

where β is some absolute constant, and where the last line uses Eq. (2). □

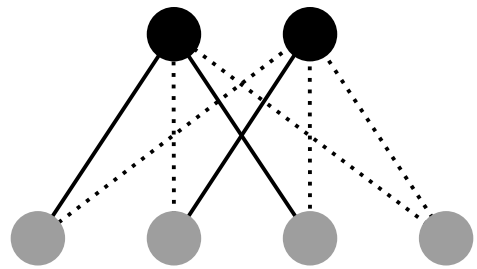

Figure 2: Illustration of an instance U_δ. Dashed edges have length $2^{1/z}$, black ones have length 1.

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_δ is defined as follows. Let C be a set of n* candidate centers *and P a set of of n_U* clients*. The metric on the ground set $P \cup C$ is defined according to the following probability distribution.*

For each pair $(p, c) \in P \times C$,

$$dist(p, c) = \begin{cases} 1 & \text{with probability } \delta \\ 2^{1/z} & \text{otherwise} \end{cases} \tag{3}$$

Distances between any pair of points $p, p' \in P$ or $c, c' \in C$ is set to $2^{1/z}$.

Fig. 2 illustrates the definition.

Since any complete graph with edge length only 1 or $\ell \leq 2$ defines a metric space, it immediately follows that $(P \cup C, \text{dist})$ is a metric space.

The important properties of the subinstance are summarized in the following lemma. We say that a set of weights is *ε-rounded* if all weights are multiples of ε.

Lemma 9. *There exists a constant η and an instance $U_\delta = (P, C, dist)$ with $|P| = n_U = \varepsilon^{-2} \log |C|$, $\delta \leq 1/4$ and $|C| \geq \varepsilon^{-5}$, the following holds. For any subset $\Omega \subseteq P$ with $\varepsilon/2$-rounded weights w_x being such that $\sum w_x \in (1 \pm 1/2)n_U$, we have:*

1. *If $|\Omega| < \varepsilon^{-2}\eta \log |C|$, there exists a center $\tilde{c} \in C$ such that*

$$\sum_{x \in \Omega : dist(x, \tilde{c}) = 1} w_x > (1 + 200\varepsilon)\delta n_U$$

   and $|x \in P : \ dist(x, \tilde{c}) = 1| \geq \delta n_U$

2. *If $|\Omega| \geq \varepsilon^{-2}\eta \log |C|$, there exists a center $c^* \in C$ such that*

$$\sum_{x \in \Omega : dist(x, c^*) = 1} w_x \geq (1 - \varepsilon)\delta n_U$$

   and $|x \in P : \ dist(x, c^*) = 1| \geq \delta n_U$.

Proof. We use the probabilistic method: we will show that, when U_δ is generated according to the process defined above, the two properties of the lemma hold with some positive probability. This is enough to ensure the existence of an instance U_δ verifying them.

We start by proving the first item. Fix some arbitrary subset of clients Ω of size at most $\varepsilon^{-2}\eta \log |C|$, with weight $w_x, \forall x \in \Omega$ and a candidate center $c \in C$. Let $w_1(c, \Omega) := \sum_{x\in\Omega:\mathrm{dist}(x,c)=1} w_x$ denote the (weighted) number of edges of length 1 from Ω to c. The expected value of $w_1(c, \Omega)$ over the random choice of edges is $\delta \cdot n_U$. We aim at applying Lemma 8 on the variable $w_1(c, \Omega)$. This cannot be done directly, as we have no control on $\max w_x$. Hence, we partition the points of Ω into five groups:

- $\Omega_1 := \{x \in \Omega : \ w_x < \varepsilon\}$
- $\Omega_2 := \{x \in \Omega : \ w_x \in [\varepsilon, 1)\}$
- $\Omega_3 := \{x \in \Omega : \ w_x \in [1, \varepsilon^{-1})\}$
- $\Omega_4 := \{x \in \Omega : \ w_x \in [\varepsilon^{-1}, 10 \log(1/\delta) \cdot \varepsilon^{-2})\}$
- $\Omega_5 := \{x \in \Omega : \ w_x \geq 20 \log(1/\delta) \cdot \varepsilon^{-2}\}$

We will show that, $\forall i \in \{2, ..., 5\}$, $w_1(c, \Omega_i)$ exceeds its expectation by a factor $(1+205\varepsilon)$ with large probability, and that $w_1(c, \Omega_1)$ is negligible.

First, note that since $\sum_{x\in\Omega} w_x \leq (1 + 1/2)n_U = \frac{3}{2}\varepsilon^{-2} \log |C|$, 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 Ω_5 with probability $\delta^{|\Omega_5|} \geq \exp(-\log |C|/10) = |C|^{-1/10}$.

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

For $i \in \{2, 3, 4\}$, $w_1(c, \Omega_i)$ is the sum of $m = |\Omega_i| \leq \varepsilon^{-2}\eta \log |C|$ random variables X_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:

$$\begin{aligned}\Pr\left[w_1(c, \Omega_i) \geq (1 + 205\varepsilon) \cdot \mathbb{E}[w_1(c, \Omega_i)]\right] &> \exp(-\frac{\beta}{\log(1/\delta)^2}\varepsilon^2\delta|\Omega_i|) \\ &\geq \exp(-\log |C|/10),\end{aligned}$$

for some absolute constant β given by Lemma 8 and $\eta \leq \frac{\log(1/\delta)^2}{10\beta\delta}$.

Finally, to deal with Ω_1, we note that $\mathbb{E}[w_1(c, \Omega_1)] \leq \varepsilon\delta n_U$. Hence, $\sum_{i=2}^{5} \mathbb{E}[w_1(c, \Omega_i)] \geq \mathbb{E}[w_1(c, \Omega)] - \varepsilon\delta n_U$, and

$$\sum_{i=2}^{5} w_1(c, \Omega_i) \geq (1 + 205\varepsilon) \sum_{i=2}^{5} \mathbb{E}[w_1(c, \Omega_i)] \quad \Rightarrow \quad w_1(c, \Omega) \geq (1 + 200\varepsilon)\delta n_U.$$

Since all groups are disjoint, the variables $w_1(c, \Omega_i)$ are independents and we can combine the previous equations to get:

$$\Pr\left[\sum_{x\in\Omega:\mathrm{dist}(x,\tilde{c})=1} w_x \geq (1+200\varepsilon)\cdot\delta n_U\right] > |C|^{-3/10}.$$

Since the length of the edges are chosen independently, the probability that there exists no center $\tilde{c}$ with $\sum_{x\in\Omega:\mathrm{dist}(x,\tilde{c})=1} w_x \geq (1+200\varepsilon)\cdot\delta n_U$ is at most

$$\begin{aligned}\left(1-|C|^{-3/10}\right)^{|C|} &= \exp\left(|C|\log(1-|C|^{-3/10})\right)\\ &\leq \exp(-|C|^{7/10}).\end{aligned}$$

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

To conclude the proof of the first bullet, it remains to do a union-bound over all possible weighted subset Ω. Such an Ω consists of at most n_U different points, with $\varepsilon/2$-rounded weights in $[0,(1+1/2)n_U]$. Hence, there are at most $\frac{4}{\varepsilon}n_U$ many different weights.

Therefore, there are $\left(\frac{4n_U}{\varepsilon}\right)^{n_U}$ many possible weighted subset Ω with $\varepsilon/2$-rounded weights, i.e.,

$$\exp\left(\varepsilon^{-2}\log|C|\cdot\log\left(2\varepsilon^{-3}\log|C|\right)\right).$$

We can conclude that there exists a center $\tilde{c}\in C$ with $\sum_{x\in\Omega:\mathrm{dist}(x,\tilde{c})=1} w_x \geq (1+200\varepsilon)\cdot\delta n_U$ with probability at least

$$1-\exp\left(\varepsilon^{-2}\log|C|\cdot\log\left(2\varepsilon^{-3}\log|C|\right)\right)\cdot\exp(-|C|^{7/10}) \geq \frac{99}{100}$$

by our choice of $|C|$. Furthermore, $\Pr[|x\in P:\ \mathrm{dist}(x,\tilde{c})=1| \geq \delta n_U] \geq 1/2$, because $|x\in P:\ \mathrm{dist}(x,\tilde{c})=1|$ follows a binomial law with mean δn_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 Ω of size at least $\varepsilon^{-2}\eta\log|C|$, and a center c. We have,

$$\begin{aligned}\Pr\left[w_1(c,\Omega)\notin(1\pm\varepsilon)\cdot\delta n_U\right] &\leq \exp(-2\varepsilon^2\delta^2\frac{n_U^2}{\sum w_i^2})\\ &\leq \exp(-2/4\cdot\delta^2\varepsilon^2)\\ &\leq \exp(-1/2\cdot\delta^2\varepsilon^2),\end{aligned}$$

where the second inequality uses $n_U^2 \geq 1/4\left(\sum w_i\right)^2 \geq 1/4\cdot\sum w_i^2$.

Since those events are independent for different centers c, the probability that there exists no center $c\in C$ with $w_1(c,\Omega)\in(1\pm\varepsilon)\cdot\delta n_U$ is at most $\exp(-1/2\cdot\delta^2\varepsilon^2|C|)$.

Hence, a union-bound over the $\left(\frac{4n_U}{\varepsilon}\right)^{n_U}$ many possible weighted subset Ω ensures that the following holds with probability at most $1-\left(\frac{4n_U}{\varepsilon}\right)^{n_U}\cdot\exp(-1/2\cdot\delta^2\varepsilon^2|C|) \geq 99/100$: For any Ω there exists a center c with $w_1(c,\Omega)\in(1\pm\varepsilon)\cdot\delta|\Omega|$ as desired. □

5.3 Combining the subinstances

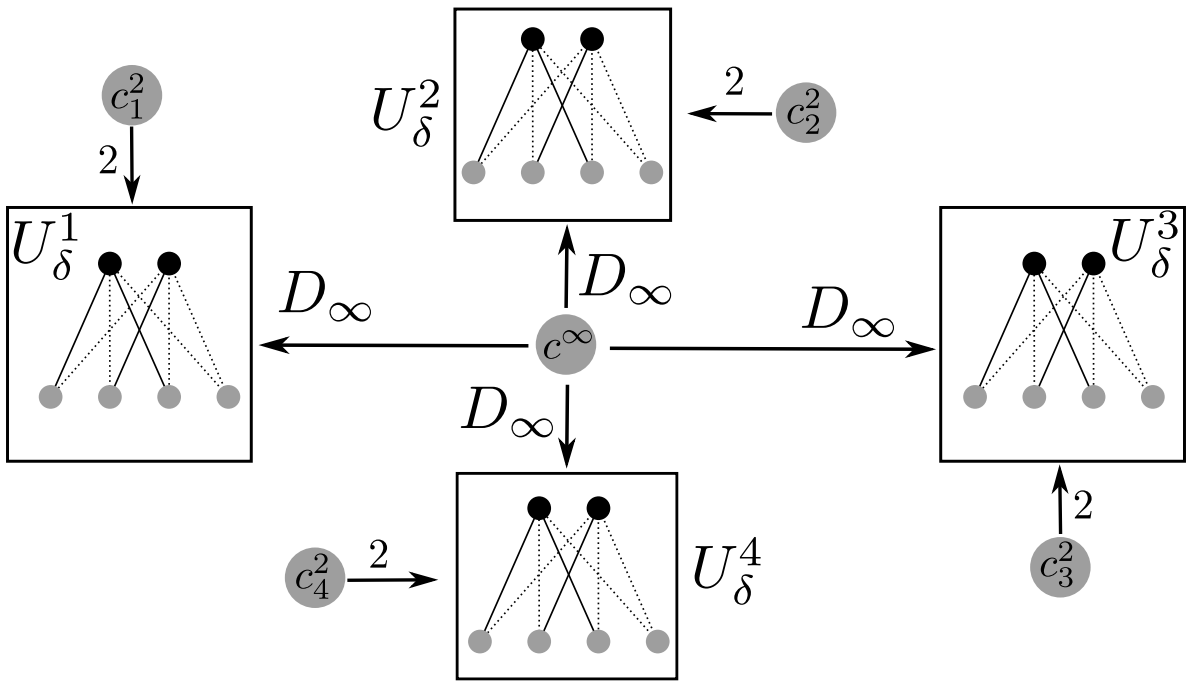

Figure 3: Illustration of a full instance, in the case $z = 1$. The subinstance are inside squares, and there is an edge from a node to a square when the node is linked to every point of the subinstance, with the distance written on the edge. D_∞ is set to be $4^i \cdot \frac{n_U \cdot k}{\varepsilon}$. The node $c^{4\cdot\infty}$ is not represented.

We now conclude the proof of the lower bound for the (ε, k, z) coreset using offset Δ. We consider k copies of the subinstance given by Lemma 9, $U_\delta^1, \ldots, U_\delta^k$, where the set of clients in each subinstance has size $n_U = 10\varepsilon^{-2} \log n$, and the set of candidate centers has size $|C|$ such that $|C| \geq \varepsilon^{-5}$ and $\log(|C| \cdot k) = O(\log |C|)$. 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 c^∞, where all points are at distance $\frac{n_U \cdot k}{\varepsilon}$ of c^∞. There is some additional candidate centers: $c^{4\cdot\infty}$, at distance $4 \cdot \frac{n_U \cdot k}{\varepsilon}$ of every client, and for subinstance i there is a center c_i^2, at distance $2^{1/z}$ from every client of the subinstance. Fig. 3 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 Δ must be negligible, and the third that the coreset weight in each subinstance is close to n_U, the number of point in a subinstance.

Claim 6. *If Ω is an ε-coreset with offset Δ for the instance, then the total weight verifies $w(\Omega) \in (1 \pm 2\varepsilon)kn_U$.*

Proof. Consider the solution consisting only of one center placed at c^∞. Let $D_\infty = \frac{n_U \cdot k}{\varepsilon}$. This solution has cost $\mathrm{cost}(c^\infty) = kn_U \cdot D_\infty^z$, and $\mathrm{cost}(\Omega, c^\infty) = w(\Omega) \cdot D_\infty^z$. Hence,

$$\Delta + w(\Omega) \cdot D_\infty^z \in (1 \pm \varepsilon)kn_U \cdot D_\infty^z.$$

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

$$\Delta + w(\Omega)4^z D_\infty^z \in (1 \pm \varepsilon)kn_U \cdot 4^z D_\infty^z.$$

Substracting those two equations yields:

$$(4^z - 1)w(\Omega) \cdot D_\infty^z \in ((4^z - 1) \pm (4^z + 1)\varepsilon)kn_U \cdot D_\infty^z,$$

and so $w(\Omega) \in (1 \pm 2\varepsilon)kn_U$. □

Claim 7. *If Ω is an ε-coreset with offset Δ for the instance, then $|\Delta| \leq 3\varepsilon k \cdot n_U$.*

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

Claim 8. *If Ω is an ε-coreset with offset Δ for the instance, then in every subinstance, the sum of the coreset weights is in $(1 \pm 1/2)n_U$.*

Proof. Assume towards contradiction that, in some subinstance, say subinstance i, the coreset mass is not in $(1 \pm 1/2)n_U$, and consider a solution $\mathcal{S}$ 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

$$\begin{aligned}\text{cost}(\mathcal{S}) &\leq \underbrace{2(k-1)n_U}_{\text{for subinstances that contain a center}} + \underbrace{n_U \cdot \left(kn_U\varepsilon^{-1}\right)^z}_{\text{for the overweighted subinstance}} \\ &\leq (1+\varepsilon)\left(k \cdot n_U^2\varepsilon^{-1}\right)^z,\end{aligned}$$

while the cost in the coreset verifies

$$\begin{aligned}\Delta + \text{cost}(\Omega, \mathcal{S}) &> -3\varepsilon kn_U + 3/2 \cdot n_U \cdot \left(kn_U\varepsilon^{-1}\right)^z \\ &\qquad \text{(using Claim 7 and keeping only the cost of the overweighted subinstance)} \\ &> (1+\varepsilon) \cdot (1+\varepsilon)\left(k \cdot n_U^2\varepsilon^{-1}\right)^z \\ &> (1+\varepsilon)\text{cost}(\mathcal{S}),\end{aligned}$$

hence contradicting the fact that Ω is an ε-coreset with offset Δ.

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 ε-coreset with offset Δ of size smaller than $\frac{\eta}{10} \cdot k\varepsilon^{-2}\log|C|$, where η is the constant of Lemma 9.

First, this implies the existence of an 2ε-coreset with ε-rounded weights, simply by rounding each weight to the closest multiple of ε.

Using Claim 8, we can apply Lemma 9 on each subinstance. The total coreset size is $\frac{\eta}{10} \cdot k\varepsilon^{-2} \log |C|$: that means that there are at least $k/10$ subinstances for which the coreset contains no more than $\eta\varepsilon^{-2} \log |C|$ many different points. We refer to these subinstances as the *bad* subinstances. Using Lemma 9, we construct a solution $\mathcal{S}$ by taking the center given by bullet 1 for the bad subinstances, i.e.: center $\hat{c}$ 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 $\mathcal{S}$. Similarly, the cost of $\mathcal{S}$ for the coreset is $2 \cdot w(\Omega) - w_1(\mathcal{S}, \Omega)$, where $w(\Omega)$ is the total coreset weight and $w_1(\mathcal{S}, \Omega)$ the weighted number of length 1 edges from Ω to $\mathcal{S}$. By construction of $\mathcal{S}$, $w_1(\mathcal{S}, \Omega)$ verifies

$$w_1(\mathcal{S}, \Omega) \geq k/10 \cdot (1 + 200\varepsilon)\delta n_U + 9k/10 \cdot (1 - \varepsilon)\delta n_U > (1 + 19\varepsilon)\delta \cdot kn_U$$

Furthermore, using properties of Lemma 9, $n_1 \leq \delta kn_U$. Hence, the cost of $\mathcal{S}$ in the coreset satisfies

$$\begin{aligned}\Delta + 2 \cdot w(\Omega) - w_1(\mathcal{S}, \Omega) &< 3\varepsilon kn_U + 2 \cdot (1 + 2\varepsilon)kn_U - (1 + 19\varepsilon)\delta \cdot kn_U \\ &\leq (2kn_U - n_1) + \varepsilon kn_U \cdot (7 - 38\delta) < (1 - \varepsilon)(2k|P| - n_1),\end{aligned}$$

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 $\mathcal{S}$ is smaller than a $(1-\varepsilon)$ factor times the cost of P for $\mathcal{S}$, 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.

6 Algorithm

Throughout this section, we use the following notation. We use $\|P\|_0$ to denote the distinct number of points in P. For a solution $\mathcal{S}$, we define the $|P|$ dimensional cost vector $v^{\mathcal{S}}$ induced by $\mathcal{S}$ as

$$v_p^{\mathcal{S}} = \text{cost}(p, \mathcal{S}).$$

Hence, $\|v^{\mathcal{S}}\|_1 = \text{cost}(P, \mathcal{S})$.

We will also make use the following lemma to have a weaker version of the triangle inequality for k-Means and more general powers of distances. See Appendix A from Makarychev, Makarychev, and Razenshteyn [73] for a proof.

Lemma 10 (Triangle Inequality for Powers)**.** *Let a, b, c be an arbitrary set of points in a metric space with distance function d and let z be a positive integer. Then for any $\varepsilon > 0$*

$$\begin{aligned}d(a, b)^z &\leq (1 + \varepsilon)^{z-1} d(a, c)^z + \left(\frac{1 + \varepsilon}{\varepsilon}\right)^{z-1} d(b, c)^z \\ |d(a, b)^z - d(a, c)^z| &\leq \varepsilon \cdot d(a, c)^z + \left(\frac{z + \varepsilon}{\varepsilon}\right)^{z-1} d(b, c)^z.\end{aligned}$$

We also require Bernstein's inequality:

Theorem 9 (Bernstein's Inequality)**.** *Let $X_1, \dots X_\delta$ be non-negative independent random variables. Let $S = \sum_{i=1}^{\delta} X_i$. If there exists an almost-sure upper bound $M \geq X_i$, then*

$$\mathbb{P}\left[|S - \mathbb{E}[S]| \geq t\right] \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^{\delta} Var[X_i] + \frac{2}{3} \cdot M \cdot t}\right).$$

6.1 Preprocessing and General Outline

We make the following three assumptions:

Assumption 1 The number of distinct points $\|P\|_0$ is in poly(k/ε).

Assumption 2 The dimension d of the points is in $O(\varepsilon^{-2} \log \|P\|_0) = O\left(\varepsilon^{-2} \log \frac{k}{\varepsilon} \cdot \text{poly } z\right)$.

Assumption 3 The point set is unweighted.

Assuming these simplifies the presentation significantly. The first assumption can be justified by computing a (potentially weighted) coreset in preprocessing. Coresets of size $\tilde{O}(k^2 \cdot \varepsilon^{-4} \cdot 2^{O(z)})$ are known to exists for all (k, z) clustering objectives [35], which is sufficient for our purposes.

The second assumption follows from a result on terminal embeddings due to Narayanan and Nelson [82]. We will discuss this result in more detail in Section 6.4. Suffice to say here is that there exists a coreset-preserving embedding from an arbitrary dimension to the desired target dimension.

The final assumption follows by scaling the weights and rounding them to integers. Each weight is then treated as a multiplicity of a point. Note that this does not increase the distinct number of points. For a proof of the validity of such an operation, we refer to Corollary 2.3 [35].

We now describe the algorithm. We first compute some constant factor approximation $\mathcal{A}$ for the entire instance.[1] Let C_i be the ith cluster induced by $\mathcal{A}$. The average cost of C_i is $\Delta_{C_i} = \frac{\text{cost}(C_i, \mathcal{A})}{|C_i|}$. For all i, j, the *ring* $R_{i,j}$ is the set of points $p \in C_i$ such that $2^j \Delta_{C_i} \leq \text{cost}(p, \mathcal{A}) \leq 2^{j+1} \Delta_{C_i}$. The *inner rings* $R_I(C_i) := \cup_{j \leq z \log(\varepsilon/z)} R_{i,j}$ (resp. *outer rings* $R_O(C_i) := \cup_{j > 2z \log(z/\varepsilon)} R_{i,j}$) of a cluster C_i consists of the points of C_i with cost at most $(\varepsilon/z)^z \, \Delta_{C_i}$ and resp. at least $(z/\varepsilon)^{2z} \, \Delta_{C_i}$. The *main rings* $R_M(C_i)$ consists of all the other points of C_i. For each j, R_j is defined to be $\cup_{i=1}^{k} R_{i,j}$. We then partition the input point set into the following groups.

- For each j, the rings $R_{i,j}$ are gathered into *groups* $G_{j,b}^M$:

$$G_{j,b}^M := \Big\{ p \mid \exists i,\ p \in R_{i,j} \text{ and } \left(\frac{\varepsilon}{4z}\right)^z \cdot \frac{\text{cost}(R_j, \mathcal{A})}{k} \cdot 2^b \leq \text{cost}(R_{i,j}, \mathcal{A}) \leq \left(\frac{\varepsilon}{4z}\right)^z \cdot 2^{b+1} \cdot \frac{\text{cost}(R_j, \mathcal{A})}{k} \Big\}.$$

- For any j, let $G_{j,min}^M := \cup_{b \leq 0} G_{j,b}^M$ be the union of the cheapest groups, and $G_{j,max}^M := \cup_{b \geq z \log \frac{4z}{\varepsilon}} G_{j,b}^M$ be the union of the most expensive ones. We define $G^M := \bigcup_j G_{j,max}^M \cup \bigcup_b G_{j,b}^M \setminus G_{j,min}^M$.

[1] A bicriteria approximation that uses $O(k)$ centers and yields a constant factor approximation would also be possible. See [74] for state of the art bounds on bicriteria approximations for k-median and k-means. For higher powers, see Mettu and Plaxton [78] for a $2^{O(z)}$ approximation.

Algorithm 1 Euclidean Coreset Construction

Compute a $O(2^z)$ approximation $\mathcal{A}$ to P.
Preprocess the instance such that Assumptions 1-3 hold.
Partition the points into groups $\mathcal{G} = \left(\bigcup_j G_{j,\max} \cup \bigcup_b G(j,b) \setminus G_{j,\min}\right) \cup \left(G^O_{\max} \cup \bigcup_b G^O_b \setminus G^O_{\min}\right)$.
for all Groups $G \in \mathcal{G}$ **do**
  Sample $\delta \in k \cdot \log \frac{k}{\varepsilon} \cdot \varepsilon^{-2} \cdot 2^{O(z\log(1+z))} \cdot \log^3 \varepsilon^{-1} \cdot \min(\varepsilon^{-z}, k)$ points Ω_G proportionate to $\frac{\mathrm{cost}(p,\mathcal{A})}{\mathrm{cost}(G,\mathcal{A})}$, and weighted by $\frac{\mathrm{cost}(G,\mathcal{A})}{\delta \cdot \mathrm{cost}(p,\mathcal{A})}$.
end for
for all $c_i \in \mathcal{A}$ **do**
  Weigh $c_i \in \mathcal{A}$ by the number of points not in $\mathcal{G} \cap C_i$
end for
Output $\Omega = \mathcal{A} \cup \bigcup_G \Omega_G$.

- The points in the outer rings are also partitioned into *outer groups*:

$$G^O_b = \Big\{p \;\mid\; \exists i,\ p \in C_i \text{ and } \left(\frac{\varepsilon}{4z}\right)^z \cdot \frac{\mathrm{cost}(R^{\mathcal{A}}_O, \mathcal{A})}{k} \cdot 2^b \le \mathrm{cost}(R_O(C_i), \mathcal{A}) \le \left(\frac{\varepsilon}{4z}\right)^z \cdot 2^{b+1} \cdot \frac{\mathrm{cost}(R^{\mathcal{A}}_O, \mathcal{A})}{k}\Big\}.$$

  We denote by $P^{G^O_b} := \{p \in P \mid p \in C \wedge C \cap G^O_b \neq \emptyset\}$ the set all points in clusters intersecting with G.

- We let as well $G^O_{min} = \cup_{b \le 0} G^O_b$ and $G^O_{max} = \cup_{b \ge z \log \frac{4z}{\varepsilon}} G^O_b$. We define $G^O := G^O_{max} \cup \bigcup_b G^O_b \setminus G^O_{min}$.

The set of all groups is denoted by $\mathcal{G} := G^M \cup G^O$. We sometimes abuse notation and also use $\mathcal{G}$ to denote the set of points in the groups $G^M \cup G^O$, i.e. $P \cap \mathcal{G} = \{p \in P \mid p \in G \in \mathcal{G}\}$ and $P \setminus \mathcal{G} = \{p \in P \mid p \in G \notin \mathcal{G}\}$. We summarize the group partitioning scheme with the following two facts.

Fact 1. *There exist at most $O(z^2 \log^2 z/\varepsilon)$ groups in $\mathcal{G}$.*

Fact 2. *All groups are pairwise disjoint. Moreover, every cluster C induced by $\mathcal{A}$ intersects with at most one group $G \in G^O$.*

The final algorithm now consists of sensitivity sampling for all groups $G \in \mathcal{G}$. Specifically, we pick a point $p \in G$ with probability $\frac{\mathrm{cost}(p,\mathcal{A})}{\mathrm{cost}(G,\mathcal{A})}$. We repeat this δ times, where δ is the size of the desired coreset. For each picked point p, we set the weight equal to $w_p := \frac{\mathrm{cost}(G,\mathcal{A})}{\delta \cdot \mathrm{cost}(p,\mathcal{A})}$. For every cluster C_i, we weigh the center c_i with the number of points in $C_i \cap \left(\bigcup_j G_{j,\min} \cup G^O_{\min}\right)$. The entire coreset construction then consists of steps required to satisfy the three initial assumptions followed by the sampling procedure, see Algorithm 1.

For every group $G \in \mathcal{G}$, we will prove that the sampling yields an (ε, E) coreset with $E = \varepsilon \cdot \mathrm{cost}(G, \mathcal{A})$.

Given a solution $\mathcal{S}$, the basic estimator for the error is

$$D_{\mathcal{S}}^{\Omega}(G) := \left| \sum_{p \in \Omega} w_p \cdot \mathrm{cost}(p, \mathcal{S}) - \mathrm{cost}(G, \mathcal{S}) \right|.$$

If for all solutions $\mathcal{S}$ we have a coreset of group $G \in \mathcal{G}$, we can compose the coresets of each group such that we have a coreset for P. Specifically, we will prove Theorem 4 by proving the following three lemmas.

The first lemma states that we can use the centers of $\mathcal{A}$ as proxies for all points not in $\mathcal{G}$. The second and third lemmas informally give the bounds such that sensitivity sampling for every group $G \in G^M$ and respectively $G \in G^O$ yield corsets.

Lemma 11. *Let P be a set of points and let $\mathcal{S}$ be an arbitrary solution. Then*

$$\left| cost(P \setminus \mathcal{G}, \mathcal{S}) - \sum_{C_i} |\mathcal{G} \cap C_i| \cdot cost(c_i, \mathcal{S}) \right| \leq \varepsilon \cdot \left(cost(P, \mathcal{S}) + cost(P, \mathcal{A}) \right).$$

Lemma 12. *Let P be a set of points and let $G \subset G^M$ be a group. Then there exist absolute constants $\gamma_1 > 0$ such that the sampling procedure of Algorithm 1 with $\delta \geq k \cdot \log \frac{k}{\varepsilon} \cdot \varepsilon^{-2} \cdot 2^{\gamma_1 \cdot z \log(1+z)} \cdot \log^3 \varepsilon^{-1} \cdot \min(\varepsilon^{-z}, k)$ yields*

$$\mathbb{E} \sup_{\mathcal{S}} \left[\frac{1}{cost(G, \mathcal{S}) + cost(G, \mathcal{A})} \cdot D_{\mathcal{S}}^{\Omega}(G) \right] \leq \varepsilon.$$

Lemma 13. *Let P be a set of points and let $G \subset \mathcal{G}^O$ be a group. Then there exist absolute constants $\gamma_2 > 0$ such that the sampling procedure of Algorithm 1 with $\delta \geq k \cdot \log \frac{k}{\varepsilon} \cdot \varepsilon^{-2} \cdot 2^{\gamma_2 \cdot z \log(1+z)} \cdot \log^3 \varepsilon^{-1}$ yields*

$$\mathbb{E} \sup_{\mathcal{S}} \left[\frac{1}{cost(P^G, \mathcal{S}) + cost(P^G, \mathcal{A})} \cdot D_{\mathcal{S}}^{\Omega}(G) \right] \leq \varepsilon.$$

First, we show that this lemma implies our main theorem.

Proof of Theorem 4. For every group $G \in \mathcal{G}$, let Ω_G be the set of points returned by the sampling

routine and let $\Omega_{\mathcal{G}}$ be the union of the output of all sampling routines. We consider

$$
\begin{aligned}
& \mathbb{E} \sup_{\mathcal{S}} \left[\frac{1}{\mathrm{cost}(P,\mathcal{S}) + \mathrm{cost}(P,\mathcal{A})} D_{\mathcal{S}}^{\Omega_{\mathcal{G}}}(P \cap \mathcal{G}) \right] \\
= \; & \mathbb{E} \sup_{\mathcal{S}} \left[\frac{1}{\mathrm{cost}(P,\mathcal{S}) + \mathrm{cost}(P,\mathcal{A})} \left| \sum_{G \in \mathcal{G}} \sum_{p \in \Omega_G} w_p \cdot \mathrm{cost}(p,\mathcal{S}) - \mathrm{cost}(G,\mathcal{S}) \right| \right] \\
\leq \; & \mathbb{E} \sup_{\mathcal{S}} \left[\frac{1}{\mathrm{cost}(P,\mathcal{S}) + \mathrm{cost}(P,\mathcal{A})} \sum_{G \in \mathcal{G}} D_{\mathcal{S}}^{\Omega_G}(G) \right] \\
\leq \; & \mathbb{E} \sup_{\mathcal{S}} \left[\sum_{G \in G^M} \frac{\mathrm{cost}(G,\mathcal{S}) + \mathrm{cost}(G,\mathcal{A})}{\mathrm{cost}(P,\mathcal{S}) + \mathrm{cost}(P,\mathcal{A})} \cdot \mathbb{E} \sup_{\mathcal{S}} \left[\frac{1}{\mathrm{cost}(G,\mathcal{S}) + \mathrm{cost}(G,\mathcal{A})} \cdot D_{\mathcal{S}}^{\Omega_G}(G) \right] \right. \\
& \left. + \sum_{G \in G^O} \frac{\mathrm{cost}(P^G,\mathcal{S}) + \mathrm{cost}(P^G,\mathcal{A})}{\mathrm{cost}(P,\mathcal{S}) + \mathrm{cost}(P,\mathcal{A})} \cdot \mathbb{E} \sup_{\mathcal{S}} \left[\frac{1}{\mathrm{cost}(P^G,\mathcal{S}) + \mathrm{cost}(P^G,\mathcal{A})} \cdot D_{\mathcal{S}}^{\Omega_G}(G) \right] \right] \\
(\textit{Lemma } 12) \quad \leq \; & \mathbb{E} \sup_{\mathcal{S}} \left[\sum_{G \in G^M} \frac{\mathrm{cost}(G,\mathcal{S}) + \mathrm{cost}(G,\mathcal{A})}{\mathrm{cost}(P,\mathcal{S}) + \mathrm{cost}(P,\mathcal{A})} \cdot \varepsilon \right. \\
(\textit{Lemma } 13) \quad & \left. + \sum_{G \in G^O} \frac{\mathrm{cost}(P^G,\mathcal{S}) + \mathrm{cost}(P^G,\mathcal{A})}{\mathrm{cost}(P,\mathcal{S}) + \mathrm{cost}(P,\mathcal{A})} \cdot \varepsilon \right] \\
\leq \; & \mathbb{E} \sup_{\mathcal{S}} [\varepsilon + \varepsilon] = 2\varepsilon
\end{aligned}
$$

Due to Markov's inequality, we have with probability at least 3/4 that $D_{\mathcal{S}}^{\Omega}(P \cap \mathcal{G}) \leq 8 \cdot \varepsilon \cdot (\mathrm{cost}(P,\mathcal{S}) + \mathrm{cost}(P,\mathcal{A}))$ for all $\mathcal{S}$. Combining this with Lemma 11, we then have for all $\mathcal{S}$

$$
\begin{aligned}
D_{\mathcal{S}}^{\Omega}(P) \; & \leq \; D_{\mathcal{S}}^{\Omega_{\mathcal{G}}}(P \cap \mathcal{G}) + \left| \mathrm{cost}(P \setminus \mathcal{G},\mathcal{S}) - \sum_{C_i} |\mathcal{G} \cap C_i| \cdot \mathrm{cost}(c_i,\mathcal{S}) \right| \\
& \leq \; 9 \cdot \varepsilon \cdot (\mathrm{cost}(P,\mathcal{S}) + \mathrm{cost}(P,\mathcal{A}))\,.
\end{aligned}
$$

Rescaling ε by a factor $9 \cdot \left(1 + \frac{\mathrm{cost}(P,\mathcal{A})}{\mathrm{OPT}}\right) \in 2^{O(z)}$ yields the desired accuracy. What is left is to prove the space bound. The maximum number of samples in any group required by Lemma 12 and Lemma 13 is in $O(k \cdot \log \frac{k}{\varepsilon} \cdot \varepsilon^{-2} \cdot \log^3 \varepsilon^{-1} \cdot 2^{O(z \log z)} \cdot \min(\varepsilon^{-z}, k))$. Due to Fact 1, the overall coreset therefore has size $O(k \cdot \log \frac{k}{\varepsilon} \cdot \varepsilon^{-2} \cdot \log^5 \varepsilon^{-1} \cdot 2^{O(z \log z)})$. □

The remainder of this section will now focus on the proofs of Lemma 12 and Lemma 13. Our main analysis tool will be a chaining argument. To do this, we require two things: (i) a reduction to a Gaussian process and (ii) controlling the variance of said Gaussian process. The proof of Lemma 11 is standard in this line of research and included in the appendix for completeness sake.

6.2 Setting up a Gaussian process

The chaining arguments we use for proving Lemma 12 and Lemma 13, while similar, are distinct enough that each lemma requires it's own notation and approach. We will focus on Lemma 12, as

it arguably the more interesting and important step. The differences for Lemma 13 are discussed at the end of this section.

Unless mentioned otherwise, we let the group G be in G^M. For proving Lemma 12, we need to have a handle on $\sum_{p\in G\cap\Omega} w_p \text{cost}(p,\mathcal{S})$, to show that $D^{\Omega}_{\mathcal{S}}(G)$ is concentrated around zero. We will not try to work directly with the basic cost estimator $\sum_{p\in P\cap\Omega} v_p^{\mathcal{S}} \cdot w_p$, since it has a too large variance. We denote the cost vector $v_p^{G,\mathcal{S}} = \begin{cases} v_p^{\mathcal{S}} & \text{if } p \in G \\ 0 & \text{else} \end{cases}$. We will split the cost vector $v^{G,\mathcal{S}}$ into two vectors for which we have separate estimators, for which we will be able show strong concentration.

To define those estimators, let us first characterize the clusters of the initial solution $\mathcal{A}$ as follow.

- We say that a cluster $C_i \cap G$ induced by $\mathcal{A}$ is *huge* if there exists a point $p \in C_i \cap G$ such that $\text{cost}(p,\mathcal{S}) \geq \left(\frac{4z}{\varepsilon}\right)^z \cdot \text{cost}(p,\mathcal{A})$. The set of huge clusters induced by $\mathcal{S}$ in G are denoted by $H_{G,\mathcal{S}}$.

Instead of estimating $\|v^{G,\mathcal{S}}\|_1$ directly, we now split $v^{G,\mathcal{S}}$ in two vectors for which we carry out the estimation separately. We, define the $|P|$-dimensional vector $u^{G,\mathcal{S}}$ with entries

$$u_p^{G,\mathcal{S}} := \begin{cases} \text{cost}(p,\mathcal{S}) & \text{if } p \in C \cap G \text{ and } C \in H_{G,\mathcal{S}} \\ 0 & \text{otherwise} \end{cases}. \tag{4}$$

Clearly $v^{G,\mathcal{S}} = v^{G,\mathcal{S}} - u^{G,\mathcal{S}} + u^{G,\mathcal{S}}$, but even more importantly

$$\|v^{G,\mathcal{S}}\|_1 = \|v^{G,\mathcal{S}} - u^{G,\mathcal{S}}\|_1 + \|u^{G,\mathcal{S}}\|_1 \tag{5}$$

as none of the entries of the considered vectors are negative.

For a group $G \in G^O$, we also characterize the clusters by a type.

- We say that a cluster $C_i \cap G$ induced by $\mathcal{A}$ is *far* if there exists a point $p \in C_i \cap G$ such that $\text{cost}(p,\mathcal{S}) \geq 4^z \cdot \text{cost}(p,\mathcal{A})$. The set of far clusters induced by $\mathcal{S}$ in G are denoted by $F_{G,\mathcal{S}}$.

Again, we split the cost vector $v^{G,\mathcal{S}}$ into two parts. Here we define

$$u_p^{G,\mathcal{S}} := \begin{cases} \text{cost}(p,\mathcal{S}) & \text{if } p \in C \cap G \text{ and } C \in F_{G,\mathcal{S}} \\ 0 & \text{otherwise} \end{cases}. \tag{6}$$

As above, Equation 5 holds for this definition of $u^{G,\mathcal{S}}$.

We will estimate $\|v^{G,\mathcal{S}} - u^{G,\mathcal{S}}\|_1$ in both cases by means of controlling a Gaussian process. Estimating $\|u^{G,\mathcal{S}}\|_1$ is done via more straightforward methods.

To show that $\sum_{p\in\Omega} w_p \cdot \left(v_p^{G,\mathcal{S}} - u_p^{G,\mathcal{S}}\right)$ is concentrated around its expectation $\|v^{G,\mathcal{S}} - u^{G,\mathcal{S}}\|_1$, we introduce a notion of nets for cost vectors defined as follows.

Definition 3. *Let I be a metric space, P a set of points and two positive integers k and z, and let $\alpha > 0$ be a precision parameter. Given some solution $\mathcal{A}$, suppose that G is a group of P. Let $\mathbb{C} \subset I^k$ be a (potentially infinite) set of candidate k-clusterings. We say that a set of cost vectors $\mathbb{N} \subset \mathbb{R}^{|P|}$ is an (α,k,z)-clustering net if for every $\mathcal{S} \in \mathbb{C}$ there exists a vector $v \in \mathbb{N}$ such that the following condition holds.*

For all $p \in C \cap G$ such that $C \cap G$ is not huge and not far,

$$|cost(p, \mathcal{S}) - v_p| \leq \alpha \cdot (cost(p, \mathcal{S}) + cost(p, \mathcal{A})).$$

For all $p \in C \cap G$ such that $C \cap G$ is either huge or far,

$$v_p = 0.$$

The existence of small clustering nets is given by Lemma 21 and 22 in Section 6.4 further below. Before we prove these lemmas, we first describe how this allows us to use a Gaussian process.

Consider a sequence of $|P|$ dimensional vectors $v^{\mathcal{S},1}, v^{\mathcal{S},2}, \ldots$ such that $v^{\mathcal{S},h}$ is the vector approximating the cost vector $v^{G,\mathcal{S}} - u^{G,\mathcal{S}}$ of $\mathcal{S}$ from a $(2^{-h}, k, z)$ clustering net $\mathbb{N}_h$. Let us now consider our estimator of $\|v^{G,\mathcal{S}} - u^{G,\mathcal{S}}\|_1$ defined as follows.

$$\begin{aligned} Y_{G,p,\mathcal{S}} &:= \left(\sum_{h=1}^{\infty} w_p \cdot \left(v_p^{\mathcal{S},h+1} - v_p^{\mathcal{S},h} \right) \right) + w_p \cdot v_p^{\mathcal{S},1} \\ Y_{G,\mathcal{S}} &:= \sum_{p \in \Omega} Y_{G,p} \end{aligned}$$

The following fact shows that this sum telescopes, and that the expectation of $Y_{G,\mathcal{S}}$ remains $\|v^{G,\mathcal{S}} - u^{G,\mathcal{S}}\|_1$.

Fact 3. $\mathbb{E}_\Omega [Y_{G,\mathcal{S}}] = \|v^{G,\mathcal{S}} - u^{G,\mathcal{S}}\|_1$.

Proof. For a fixed point p, it holds that $\lim_{h \to \infty} v_p^h = v_p^{G,\mathcal{S}} - u_p^{G,\mathcal{S}}$. Hence, the infinite sum is well defined and we have:

$$\begin{aligned} & \sum_{h=1}^{\infty} w_p \left(v_p^{\mathcal{S},h+1} - v_p^{\mathcal{S},h} \right) + w_p \cdot v_p^{\mathcal{S},1} \\ = \; & w_p v_p^{G,\mathcal{S}} - u^{G,\mathcal{S}} \text{ since the sum telescopes} \end{aligned}$$

Hence, summing over all points $p \in \Omega$ and taking the expecation concludes the lemma. □

Using this fact, we can estimate $\|v^{G,\mathcal{S}} - u^{G,\mathcal{S}}\|_1$ by $Y_{G,\mathcal{S}}$.

To prove Lemma 12, we in particular wish to show for $G \in G^M$

$$\mathbb{E}_\Omega \sup_{\mathcal{S}} \left| \frac{1}{\text{cost}(G, \mathcal{S}) + \text{cost}(G, \mathcal{A})} \cdot (Y_{G,\mathcal{S}} - \mathbb{E}[Y_{G,\mathcal{S}}]) \right| \leq \varepsilon.$$

Analogously, for Lemma 13, we wish to show for $G \in G^O$

$$\mathbb{E}_\Omega \sup_{\mathcal{S}} \left| \frac{1}{\text{cost}(P^G, \mathcal{S}) + \text{cost}(P^G, \mathcal{A})} \cdot (Y_{G,\mathcal{S}} - \mathbb{E}[Y_{G,\mathcal{S}}]) \right| \leq \varepsilon.$$

Unfortunately, it is difficult to apply the chaining framework with weighted Boolean variables. This is usually addressed using the following symmetrization argument. We pick δ independent standard normal Gaussian random variables $\xi_1, \ldots, \xi_\delta \sim \mathcal{N}(0,1)$ and analyse the following random variables for the respective cases $G \in G^M$ and $G \in G^O$

$$\begin{aligned} X_{G,\mathcal{S}} &:= \frac{\sum_{p\in\Omega}\left(\sum_{h=1}^{\infty} \xi_p \cdot w_p \cdot \left(v_p^{\mathcal{S},h+1} - v_p^{\mathcal{S},h}\right)\right) + \xi_p \cdot w_p \cdot v^{\mathcal{S},1}(p)}{\mathrm{cost}(G,\mathcal{S}) + \mathrm{cost}(G,\mathcal{A})}, \\ X_{G,\mathcal{S}} &:= \frac{\sum_{p\in\Omega}\left(\sum_{h=1}^{\infty} \xi_p \cdot w_p \cdot \left(v_p^{\mathcal{S},h+1} - v_p^{\mathcal{S},h}\right)\right) + \xi_p \cdot w_p \cdot v_p^{\mathcal{S},1}}{\mathrm{cost}(P^G,\mathcal{S}) + \mathrm{cost}(P^G,\mathcal{A})}. \end{aligned}$$

The following lemma is due to Rudra and Wootters [84], see also the book by Ledoux and Talagrand [67] for more general statements.

Lemma 14 (Appendix B.3 of [84])**.** *Let* $T = \frac{1}{cost(G,\mathcal{S})+cost(G,\mathcal{A})}$ *or* $T = \frac{1}{cost(P^G,\mathcal{S})+cost(P^G,\mathcal{A})}$*. Then* $\mathbb{E}_\Omega \sup_{\mathcal{S}} \left|\sum_{p\in\Omega}^{\delta} T \cdot (Y_{G,p,\mathcal{S}} - \mathbb{E}[Y_{G,p,\mathcal{S}}])\right| \leq \sqrt{2\pi} \cdot \mathbb{E}_\Omega \mathbb{E}_\xi \sup_{\mathcal{S}} |X_{G,\mathcal{S}}|$.

With these, we now prove the following lemmas.

Lemma 15. *Let* $G \in G^M$*. Suppose* $\delta = \gamma_3 \cdot k \cdot \log \frac{k}{\varepsilon} \cdot \varepsilon^{-2}$ *for some absolute constant* γ_5*. Then*

$$\mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\left|\frac{\sum_{p\in\Omega} w_p \cdot u_p^{G,\mathcal{S}} - \|u^{G,\mathcal{S}}\|_1}{cost(G,\mathcal{A}) + cost(G,\mathcal{S})}\right|\right] \leq \varepsilon.$$

Lemma 16. *Let* $G \in G^M$*. Suppose* $\delta = 2^{\gamma_4 \cdot z \log(1+z)} \cdot k \cdot \log \frac{k}{\varepsilon} \cdot \varepsilon^{-2} \cdot \log^3 \varepsilon^{-1} \cdot \min(\varepsilon^{-z}, k)$ *for absolute constants* γ_6 *and* γ_7*. Then*

$$\mathbb{E}_\Omega\ \mathbb{E}_\xi \sup_{\mathcal{S}} |X_{G,\mathcal{S}}| \leq \varepsilon.$$

Lemma 17. *Let* $G \in G^O$*. Suppose* $\delta = \gamma_5 \cdot k \cdot \log \frac{k}{\varepsilon} \cdot \varepsilon^{-2}$ *for some absolute constant* γ_9*. Then*

$$\mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\left|\frac{\sum_{p\in\Omega} w_p \cdot u_p^{G,\mathcal{S}} - \|u^{G,\mathcal{S}}\|_1}{cost(P^G,\mathcal{A}) + cost(P^G,\mathcal{S})}\right|\right] \leq \varepsilon.$$

Lemma 18. *Let* $G \in G^O$*. Suppose* $\delta = 2^{\gamma_6 \cdot z \log(1+z)} \cdot k \cdot \log \frac{k}{\varepsilon} \cdot \varepsilon^{-2} \cdot \log^3 \varepsilon^{-1}$ *for absolute constants* γ_8 *and* γ_9*. Then*

$$\mathbb{E}_\Omega\ \mathbb{E}_\xi \sup_{\mathcal{S}} |X_{G,\mathcal{S}}| \leq \varepsilon.$$

The proofs of Lemma 15 and Lemma 17 are in Section Section 6.6, the proofs of Lemma 16 and Lemma 18 is split into proving the existence of sufficiently small nets (Section 6.4) and analysing the variance of the Gaussian process. For now, we show why these lemmas imply Lemma 12 and Lemma 13.

Proof of Lemma 12. As mentioned in Equation 5, we have $\mathrm{cost}(G,\mathcal{S}) = \|v^{G,\mathcal{S}}\|_1 = \|v^{G,\mathcal{S}} - u^{G,\mathcal{S}}\|_1 + \|u^{G,\mathcal{S}}\|_1$. Due to Lemma 14, we have

$$\mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\frac{1}{\mathrm{cost}(G,\mathcal{S}) + \mathrm{cost}(G,\mathcal{A})} \cdot |Y_{G,\mathcal{S}} - \mathbb{E}[Y_{G,\mathcal{S}}]|\right] \leq \sqrt{2\pi} \cdot \mathbb{E}_\Omega\ \mathbb{E}_\xi \sup_{\mathcal{S}} |X_{G,\mathcal{S}}|.$$

Plugging in the bound from Lemma 16, we therefore have

$$\mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\frac{1}{\text{cost}(G,\mathcal{S})+\text{cost}(G,\mathcal{A})} \cdot |Y_{G,\mathcal{S}} - \mathbb{E}[Y_{G,\mathcal{S}}]|\right] \leq \sqrt{2\pi}\varepsilon.$$

Then

$$\begin{aligned}
& \mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\frac{1}{\text{cost}(G,\mathcal{S})+\text{cost}(G,\mathcal{A})} D^\Omega_{\mathcal{S}}(G)]\right] \\
= \; & \mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\left|\frac{\|v^{G,\mathcal{S}} - u^{G,\mathcal{S}}\|_1 + \|u^{G,\mathcal{S}}\|_1 - \sum_{p\in\Omega}\left(w_p \cdot \left(v_p^{G,\mathcal{S}} - u_p^{G,\mathcal{S}}\right) + w_p \cdot u_p^{G,\mathcal{S}}\right)}{\text{cost}(G,\mathcal{S})+\text{cost}(G,\mathcal{A})}\right|\right] \\
\leq \; & \mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\left|\frac{\sum_{p\in\Omega} w_p \cdot u_p^{G,\mathcal{S}} - \|u^{G,\mathcal{S}}\|_1}{\text{cost}(G,\mathcal{S})+\text{cost}(G,\mathcal{A})}\right|\right] \\
& + \mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\left|\frac{\sum_{p\in\Omega} w_p \cdot (v_p^{G,\mathcal{S}} - u_p^{G,\mathcal{S}}) - \|v^{G,\mathcal{S}} - u^{G,\mathcal{S}}\|_1}{\text{cost}(G,\mathcal{S})+\text{cost}(G,\mathcal{A})}\right|\right] \\
& \leq \mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\left|\frac{\sum_{p\in\Omega} w_p \cdot u_p^{G,\mathcal{S}} - \|u^{G,\mathcal{S}}\|_1}{\text{cost}(G,\mathcal{S})+\text{cost}(G,\mathcal{A})}\right|\right] + \mathbb{E}_\Omega \sup_{\mathcal{S}} \left|\sum_{p\in\Omega}^{\delta} \frac{Y_{G,p,\mathcal{S}} - \mathbb{E}[Y_{G,p,\mathcal{S}}]}{\text{cost}(G,\mathcal{S})+\text{cost}(G,\mathcal{A})}\right| \\
(\textit{Lemma } 15) \; \leq \; & \varepsilon + \sqrt{2\pi}\varepsilon.
\end{aligned}$$

Rescaling ε yields the claim. □

The proof of Lemma 13 is completely analogous. For completeness sake, we repeat the steps.

Proof of Lemma 13. As mentioned in Equation 5, we have $\text{cost}(G,\mathcal{S}) = \|v^{G,\mathcal{S}}\|_1 = \|v^{G,\mathcal{S}} - q^{G,\mathcal{S}}\|_1 + \|u^{G,\mathcal{S}}\|_1$. Due to Lemma 14, we have $\mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\frac{1}{\text{cost}(P^G,\mathcal{S})+\text{cost}(P^G,\mathcal{A})} \cdot |Y_{G,\mathcal{S}} - \mathbb{E}[Y_{G,\mathcal{S}}]|\right] \leq \sqrt{2\pi} \cdot \mathbb{E}_\Omega \, \mathbb{E}_\xi \sup_{\mathcal{S}} |X_{G,\mathcal{S}}|$. Plugging in the bound from Lemma 18, we therefore have $\mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\frac{1}{\text{cost}(P^G,\mathcal{S})+\text{cost}(P^G,\mathcal{A})} \cdot |Y_{G,\mathcal{S}} - \mathbb{E}[Y_{G,\mathcal{S}}]|\right] \leq \sqrt{2\pi}\varepsilon$. Then

$$\begin{aligned}
& \mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\frac{1}{\text{cost}(P^G,\mathcal{S})+\text{cost}(P^G,\mathcal{A})} D^\Omega_{\mathcal{S}}(G)]\right] \\
= \; & \mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\left|\frac{\|v^{G,\mathcal{S}} - u^{G,\mathcal{S}}\|_1 + \|u^{G,\mathcal{S}}\|_1 - \sum_{p\in\Omega} w_p \cdot \left(v_p^{G,\mathcal{S}} - u_p^{G,\mathcal{S}}\right) + w_p \cdot u_p^{G,\mathcal{S}}}{\text{cost}(P^G,\mathcal{S})+\text{cost}(P^G,\mathcal{A})}\right|\right] \\
\leq \; & \mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\left|\frac{\sum_{p\in\Omega} w_p \cdot u_p^{G,\mathcal{S}} - \|u^{G,\mathcal{S}}\|_1}{\text{cost}(P^G,\mathcal{S})+\text{cost}(P^G,\mathcal{A})}\right|\right] \\
& + \mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\left|\frac{\sum_{p\in\Omega} w_p \cdot (v_p^{G,\mathcal{S}} - u_p^{G,\mathcal{S}}) - \|v^{G,\mathcal{S}} - u^{G,\mathcal{S}}\|_1}{\text{cost}(P^G,\mathcal{S})+\text{cost}(P^G,\mathcal{A})}\right|\right] \\
\textit{Lemma } 17 \; \leq \; & \varepsilon + \sqrt{2\pi}\varepsilon.
\end{aligned}$$

Rescaling ε yields the claim. □

6.3 A Structural Lemma

We will use the property for G^M that we have a good estimator for the size of every cluster of $\mathcal{A}$. We will frequently use this property in subsequent sections. By definition of groups, we have for every point p of any cluster C with a non-empty intersection with $G \in G^M$

$$\text{cost}(G, \mathcal{A}) \leq 2k \cdot \text{cost}(C \cap G, \mathcal{A}) \leq 4k \cdot |C \cap G| \cdot \text{cost}(p, \mathcal{A}). \tag{7}$$

We first show that, given we sampled enough points, $|C \cap G|$ is well approximated for every cluster C. This lemma will also be used later for bounding the supremum of $X_{G,\mathcal{S}}$ in the proof of Lemma 16. We define event $\mathcal{E}_G$ to be for all clusters C,

$$\sum_{p \in C \cap G \cap \Omega} w_p = \sum_{p \in C \cap G \cap \Omega} \frac{\text{cost}(G, \mathcal{A})}{\delta \cdot \text{cost}(p, \mathcal{A})} = (1 \pm \varepsilon) \cdot |C \cap G|.$$

Lemma 19. *Let $G \in G^M$. We have that with probability at least $1 - k \cdot \exp\left(-\frac{\varepsilon^2}{9 \cdot k}\delta\right)$, event $\mathcal{E}_G$ happens.*

The proof is similar to the one used in Lemma 4.4 from [35]. The main difference is, due to using a slightly different sampling distribution, Hoeffding's inequality is insufficient and we have to rely on Bernstein's inequality.

Proof of Lemma 19. First, observe that $\mathbb{E}[\sum_{p \in C \cap G \cap \Omega} w_p] = |C \cap G|$. We will bound both the variance as well as M in order to apply Bernstein's inequality. Let Ω be the set of sampled points and let p_j be the jth point in the sample with respect to some arbitrary but fixed ordering. Consider the random variable $w_{p_j,C} = \begin{cases} w_p & \text{if } p_j = p \in C \cap G \\ 0 & \text{else} \end{cases}$. Then

$$\begin{aligned} \text{Var}[w_{p_j,C}] &\leq \mathbb{E}[w_{p_j,C}^2] = \sum_{p \in C \cap G} w_p^2 \cdot \mathbb{P}[p \in \Omega] = \sum_{p \in C \cap G} \frac{\text{cost}(G, \mathcal{A})}{\delta^2 \cdot \text{cost}(p, \mathcal{A})} \\ (Eq.\ 7) &\leq \sum_{p \in C \cap G} \frac{2k\text{cost}(C, \mathcal{A})}{\delta^2 \text{cost}(p, \mathcal{A})} \leq \sum_{p \in C \cap G} \frac{4k \cdot |C \cap G|}{\delta^2} \leq \frac{4k \cdot |C \cap G|^2}{\delta^2} \end{aligned} \tag{8}$$

For the maximum upper bound, we have again due to Equation 7

$$w_{p_j,C} = \frac{\text{cost}(P, \mathcal{A})}{\delta \cdot \text{cost}(p_j, \mathcal{A})} \leq \frac{4k \cdot |C \cap G|}{\delta} \tag{9}$$

Thus, combining Equation 8 and 9 with Bernstein's inequality, we have

$$\begin{aligned} \mathbb{P}\left[\left|\sum_{p \in C_i \cap \Omega} w_p - |C \cap G|\right| > \varepsilon \cdot |C \cap G|\right] &\leq \exp\left(-\frac{\varepsilon^2 \cdot |C \cap G|^2}{2\delta \cdot \frac{4k \cdot |C \cap G|^2}{\delta^2} + \frac{2}{3}\frac{4k \cdot |C \cap G|}{\delta} \cdot \varepsilon \cdot |C \cap G|}\right) \\ &\leq \exp\left(-\frac{\varepsilon^2}{9 \cdot k} \cdot \delta\right). \end{aligned}$$

The lemma now follows by taking a union bound over all clusters in $\mathcal{A}$. □

6.4 Existence of Small Clustering Nets

For a set of points P, a set of points $\mathcal{N}_\varepsilon$ is an *ε-net* of P if for every point $x \in P$ there exists some point $y \in \mathcal{N}_\varepsilon$ with $\|x - y\| \leq \varepsilon$. The existence of small nets in Euclidean spaces is given by the following statement.

Lemma 20 (Lemma 5.2 of [89])**.** *For the unit d-dimensional Euclidean ball centered around the origin, there exists an ε-net of cardinality $(1 + 2/\varepsilon)^d$.*

We further will crucially rely on terminal embeddings defined as follows. A terminal embedding of a set $P \in \mathbb{R}^d$ is a mapping $f : \mathbb{R}^d \to \mathbb{R}^m$ such that

$$\forall x \in P,\ \forall y \in \mathbb{R}^d,\ (1 - \varepsilon) \cdot \|x - y\|_2 \leq \|f(x) - f(y)\|_2 \leq (1 + \varepsilon) \cdot \|x - y\|_2.$$

The statement is closely related to the classic Johnson-Lindenstrauss lemma. The crucial generalization is that the pairwise distances between any point of $\mathbb{R}^d$ and any point of P, rather than just the pairwise distances of points in P, are preserved.

Theorem 10 (Theorem 1.1 of [82])**.** *For any point set P in $\mathbb{R}^d$, there exists a terminal embedding $f : \mathbb{R}^d \to \mathbb{R}^m$ with $m \in O(\varepsilon^2 \log \|P\|_0)$.*

The target dimension here is optimal for a wide range of parameters (see Larsen and Nelson for a matching lower bound [66]). Using both of these statements, we now show the existence of small clustering nets.

Lemma 21. *Let k, z be two positive integers, G be a group and $\mathcal{A}$ be a solution to (k, z)-clustering. Define $\mathbb{C}$ to be the set of possible candidate centers. For all $\alpha \leq 1/2$, there exists an (α, k, z)-clustering net $\mathbb{N}$ of $\mathbb{C}$ with*

$$|\mathbb{N}| \leq \exp\left(\gamma_7 \cdot z^2 \cdot k \cdot \log \|P\|_0 / \alpha^2 \cdot \log \frac{1}{\alpha \cdot \varepsilon})\right),$$

where γ_7 is an absolute constant.

Proof. Let f be a terminal embedding of P into $O((\frac{z}{\alpha})^2 \log \|P\|_0)$ dimensions given by Theorem 10. Given a solution $\mathcal{S}$, we then have for any $p \in P$

$$\mathrm{cost}(p, f(\mathcal{S})) = \left(\min_{s \in \mathcal{S}} \|f(p) - f(s)\|\right)^z = \left((1 \pm \alpha/O(z)) \cdot \min_{s \in \mathcal{S}} \|p - s\|\right)^z = (1 \pm \alpha) \cdot \mathrm{cost}(p, \mathcal{S})$$

Let B be an arbitrary subset of the clusters induced by $\mathcal{A}$. Here B is meant to contain the clusters that are not in $H_{G,\mathcal{S}}$ for a given candidate solution $\mathcal{S}$, but the exact interpretation of B is not important for the proof. We will show that for every B, there exists an (α, k, z) clustering net $\mathbb{N}_B$ of size

$$|\mathbb{N}_B| \in \exp\left(O\left(z^2/\alpha^2 k \cdot \log \|P\|_0 \cdot \log\left(\frac{z}{\alpha \cdot \varepsilon}\right)\right)\right). \tag{10}$$

Since there are at most 2^k subsets B, the overall size of the clustering net is then $\sum_B |\mathbb{N}_B| \leq 2^k \cdot \exp\left(O\left(z^2 \cdot k \cdot \log \|P\|_0 / \alpha^2 \cdot \log\left(\frac{z}{\alpha \cdot \varepsilon}\right)\right)\right) = \exp\left(O\left(z^2 \cdot k \cdot \log \|P\|_0 / \alpha^2 \cdot \log\left(\frac{z}{\alpha \cdot \varepsilon}\right)\right)\right)$.

We now justify Equation 10. We take an $\frac{\alpha}{z} \cdot \mathrm{dist}(p, \mathcal{A})$-net of the Euclidean ball centered around $p \in C \in B$ with radius $8 \cdot \left(\frac{4z}{\varepsilon}\right) \cdot \mathrm{dist}(p, \mathcal{A})$. Such a net has size at most

$$\exp\left(O\left(z^2 \log \|P\|_0/\alpha^2 \cdot \log\left(\frac{z}{\alpha \cdot \varepsilon}\right)\right)\right)$$

due to Lemma 20.

We now take the union of all $\frac{\alpha}{z} \cdot \mathrm{dist}(p, \mathcal{A})$-nets of all points $p \in C \in B$. This yields a total number of $\|P\|_0 \cdot \exp\left(K_z \log \|P\|_0 \cdot \alpha^{-2} \cdot \log \frac{1}{\alpha \cdot \varepsilon}\right)$ nets points. We set $\mathbb{N}_B$ to be set of all subsets of size k of the union of nets. Clearly, $|\mathbb{N}_B| = \exp\left(K_z \cdot k \cdot \log \|P\|_0/\alpha^2 \cdot \log \frac{1}{\alpha \cdot \varepsilon}\right)$ as desired in Equation 10. What is left to show is that $\mathbb{N}_B$ is an $(O(\alpha), k, z)$ clustering net. The lemma then follows by rescaling α.

Let $\mathcal{S} \in \mathbb{C}$ be a set of k centers. Consider the set N of k net points defined as follows : $N := \cup_{s \in \mathcal{S}} \{n_s : n_s \text{ is the closest net point to } f(s)\}$. Define the cost vector v^N such that $v^N(p) = \begin{cases} \mathrm{cost}(f(p), N) & \text{if } p \in C \notin H_{G,\mathcal{S}} \\ 0 & \text{else} \end{cases}$. Let p be a point from a cluster $C \notin H_{G,\mathcal{S}}$. By definition of $H_{G,\mathcal{S}}$, this implies

$$\mathrm{cost}(p, \mathcal{S}) \leq \left(\frac{4z}{\varepsilon}\right)^z \mathrm{cost}(p, \mathcal{A}). \tag{11}$$

We need to show that $|\mathrm{cost}(p, \mathcal{S}) - v_{N_{\mathcal{S}}}(p)| \leq \alpha \cdot (\mathrm{cost}(p, \mathcal{S}) + \mathrm{cost}(p, \mathcal{A}))$.

Let s be the center closest to p in $\mathcal{S}$, and c be p's closest center in $\mathcal{A}$. The terminal embedding ensures

$$\|f(p) - f(s)\|_2 \leq (1 + \alpha/O(z))\|p - s\|_2 \leq (1 + \alpha/O(z))^2 \cdot \|f(p) - f(s)\|_2.$$

Then,

$$\begin{aligned} \|f(c) - f(s)\|_2 &\leq \|f(c) - f(p)\|_2 + \|f(p) - f(s)\|_2 \\ &\leq (1 + \alpha/z) \cdot \|c - p\|_2 + (1 + \alpha/z) \cdot \|p - s\|_2 \\ &\leq (1 + \alpha/O(z)) \left(1 + \left(\frac{4z}{\varepsilon}\right)\right) \cdot \|p - c\|_2 \\ &\leq 4 \cdot \left(\frac{4z}{\varepsilon}\right) \|p - c\|_2 \end{aligned}$$

where third inequality holds due to Equation 11. Hence, $f(s)$ is in the ball of radius $8\left(\frac{4z}{\varepsilon}\right) \mathrm{dist}(p, c)$ centered around p which implies

$$|\mathrm{dist}(f(p), f(\mathcal{S})) - \mathrm{dist}(f(p), n_s)| \leq \left(\frac{2\alpha}{z}\right) \cdot \mathrm{dist}(f(p), f(\mathcal{A})).$$

We therefore have

$$
\begin{aligned}
& |\mathrm{cost}(f(p), f(\mathcal{S})) - \mathrm{cost}(f(p), n_s)| \\
= \; & |\mathrm{dist}(f(p), f(\mathcal{S})) - \mathrm{dist}(f(p), n_s)| \cdot \mathrm{dist}(f(p), f(\mathcal{S}))^{z-1} \cdot \sum_{i=0}^{z-1} \left(\frac{\mathrm{dist}(f(p), n_s)}{\mathrm{dist}(f(p), f(\mathcal{S}))} \right)^i \\
\leq \; & \left(\frac{2\alpha}{z} \right) \cdot \mathrm{dist}(f(p), f(\mathcal{A})) \cdot \mathrm{dist}(f(p), f(\mathcal{S}))^{z-1} \cdot z \cdot \left(1 + \left(\frac{2\alpha}{z} \right) \cdot \frac{\mathrm{dist}(f(p), f(\mathcal{A}))}{\mathrm{dist}(f(p), f(\mathcal{S}))} \right)^{z-1} \\
\leq \; & \left(\frac{2\alpha}{z} \right) \cdot \mathrm{dist}(f(p), f(\mathcal{A})) \cdot z \cdot \left(\mathrm{dist}(f(p), f(\mathcal{S})) + \left(\frac{2\alpha}{z} \right) \mathrm{dist}(f(p), f(\mathcal{A})) \right)^{z-1} \\
\leq \; & \left(\frac{2\alpha}{z} \right) \cdot \mathrm{dist}(f(p), f(\mathcal{A})) \cdot z \cdot \left(1 + \left(\frac{2\alpha}{z} \right) \right) \cdot \max\left(\mathrm{dist}(f(p), f(\mathcal{S})), \mathrm{dist}(f(p), f(\mathcal{A})) \right)^{z-1} \\
\leq \; & 4\alpha \cdot \left(\mathrm{cost}(f(p), f(\mathcal{A}) + \mathrm{cost}(f(p), f(\mathcal{S})) \right). \qquad (12)
\end{aligned}
$$

This implies that

$$
\begin{aligned}
|\mathrm{cost}(p, \mathcal{S}) - v_N(p)| &\leq |\mathrm{cost}(f(p), f(\mathcal{S})) - \mathrm{cost}(f(p), n_s)| + O(\alpha) \cdot \mathrm{cost}(p, \mathcal{S}) \\
(Eq.\ 12) &\leq 4\alpha \cdot (\mathrm{cost}(f(p), f(\mathcal{A}) + \mathrm{cost}(f(p), f(\mathcal{S})) + O(\alpha) \cdot \mathrm{cost}(p, \mathcal{S}) \\
&= O(\alpha) \cdot (\mathrm{cost}(p, \mathcal{S}) + \mathrm{cost}(p, \mathcal{A})).
\end{aligned}
$$

Thus, up to a rescaling of α by constant factors, we have the desired accuracy and thereby proving Equation 10. □

Lemma 22. *Let k, z be two positive integers, G be a group and $\mathcal{A}$ be a solution to (k, z)-clustering. Suppose the points of G lie in d dimensional Euclidean space. Define $\mathbb{C}$ to be the set of possible candidate centers. For all $\alpha \leq 1/2$, there exists an (α, k, z)-clustering net $\mathbb{N}$ of $\mathbb{C}$ with*

$$|\mathbb{N}| \leq \exp\left(\gamma_8 \cdot k \cdot d \cdot z \log(4z/(\alpha \cdot \varepsilon)) \right),$$

where γ_8 is an absolute constant.

Proof. The construction is essentially identical to that of Lemma 21. The main difference is that we now take nets of the d-dimensional Euclidean ball centered around every point p. These nets has size at most

$$\exp\left(O\left(d \cdot z \log(4z/(\alpha \cdot \varepsilon)) \right) \right) = \exp\left(\gamma_{11} \cdot d \cdot z \log(4z/(\alpha \cdot \varepsilon)) \right)$$

due to Lemma 20.

The remaining arguments from Lemma 21 are not affected by this change. □

Recall that we assumed that $d \in O(\varepsilon^{-2} \log \|P\|_0)$, which is a consequence of Theorem 10. We will describe the chaining procedure in more detail in Section 6.5. For those familiar with chaining: this assumption on d, combined with the bound of Lemma 22 will ensure that the chain converges after only a small number of steps.

6.5 Proofs of Lemma 16 and Lemma 18

We focus on the proof of Lemma 16. The proof of Lemma 18 follows along the same lines, but is far simpler. For completeness, we repeat the arguments at the end of the section.

Proof of Lemma 16. In the following, let $G \in G^M$. We recall the random variable

$$X_{G,\mathcal{S}} \quad := \quad \frac{1}{\text{cost}(G,\mathcal{A}) + \text{cost}(G,\mathcal{S})} \sum_{p \in \Omega} \left(\sum_{h=1}^{\infty} \xi_p \cdot w_p \cdot \left(v_p^{\mathcal{S},h+1} - v_p^{\mathcal{S},h} \right) \right) + \xi_p \cdot w_p \cdot v_p^{\mathcal{S},1}.$$

Define

$$\begin{aligned} X_{G,\mathcal{S},0} \quad &:= \quad \frac{1}{\text{cost}(G,\mathcal{A}) + \text{cost}(G,\mathcal{S})} \sum_{p \in \Omega} \xi_p \cdot w_p \cdot v_p^{\mathcal{S},1} \\ X_{G,\mathcal{S},h} \quad &:= \quad \frac{1}{\text{cost}(G,\mathcal{A}) + \text{cost}(G,\mathcal{S})} \sum_{p \in \Omega} \xi_p \cdot w_p \cdot \left(v_p^{\mathcal{S},h+1} - v_p^{\mathcal{S},h} \right). \end{aligned}$$

We have $\mathbb{E}_\Omega \mathbb{E}_\xi \sup_{\mathcal{S}} |X_{G,\mathcal{S}}| \leq \sum_{h=0}^{\infty} \mathbb{E}_\Omega \mathbb{E}_\xi \sup_{\mathcal{S}} |X_{G,\mathcal{S},h}|$. The number of vectors $v^{\mathcal{S},h}$ are bounded via Lemma 21 and 22. The primary remaining challenge is to control the variance of $X_{G,\mathcal{S},h}$.

Lemma 23. *Let $G \in G^M$. Fix a solution $\mathcal{S}$ and let $\beta_1, \beta_2 > 0$ be absolute constants. Then $X_{G,\mathcal{S},h}$ is Gaussian distributed with mean 0. The variance of $X_{G,\mathcal{S},h}$ is always at most*

$$\sum_{p \in \Omega} \left(\frac{w_p \cdot \left(v_p^{\mathcal{S},h+1} - v_p^{\mathcal{S},h} \right)}{cost(G,\mathcal{A}) + cost(G,\mathcal{S})} \right)^2 \in \delta^{-1} \cdot 2^{\beta_1 z \log(1+z)} 2^{-2h} \cdot \varepsilon^{-2z}.$$

Furthermore, conditioned on event $\mathcal{E}_G$, the variance of $X_{G,\mathcal{S},h}$ is at most

$$\sum_{p \in \Omega} \left(\frac{w_p \cdot \left(v_p^{\mathcal{S},h+1} - v_p^{\mathcal{S},h} \right)}{cost(G,\mathcal{A}) + cost(G,\mathcal{S})} \right)^2 \in \delta^{-1} \cdot 2^{\beta_2 z \log(1+z)} 2^{-2h} \cdot \min(\varepsilon^{-z}, k).$$

Proof. We recall the standard fact that if $\xi_p \sim \mathcal{N}(0,1)$, then $\sum_p a_p \cdot \xi_p$ is Gaussian distributed with mean 0 and variance $\sum_p a_p^2$.

For any $p \in C \notin H_{G,\mathcal{S}}$, we have $\text{cost}(p,\mathcal{S}) \leq \left(\frac{4z}{\varepsilon}\right)^z \text{cost}(p,\mathcal{A})$. A terminal embedding with target dimension $O(z^2 2^{2h} \log \|P\|_0)$ preserves the cost up to a factor (1 ± 2^{-h}), i.e. we have $(1 - 2^{-h}) \cdot$

$\mathrm{cost}(p,\mathcal{S}) \le v_p^{\mathcal{S},h} \le (1+2^{-h})\mathrm{cost}(p,\mathcal{S})$. Therefore

$$\begin{aligned}
&\sum_{p\in\Omega}\left(\frac{w_p\cdot\left(v_p^{\mathcal{S},h+1}-v_p^{\mathcal{S},h}\right)}{\mathrm{cost}(G,\mathcal{A})+\mathrm{cost}(G,\mathcal{S})}\right)^2\\
=&\sum_{p\in\Omega}\left(\frac{w_p\cdot\left(v_p^{\mathcal{S},h+1}-\mathrm{cost}(p,\mathcal{S})+\mathrm{cost}(p,\mathcal{S})-v_p^{\mathcal{S},h}\right)}{\mathrm{cost}(G,\mathcal{A})+\mathrm{cost}(G,\mathcal{S})}\right)^2\\
\le&\sum_{p\in\Omega}\left(\frac{w_p\cdot 2^{-h-1}\cdot\mathrm{cost}(p,\mathcal{S})}{\mathrm{cost}(G,\mathcal{A})+\mathrm{cost}(G,\mathcal{S})}\right)^2\\
\le&\sum_{p\in\Omega}2^{-2h+2}\left(\frac{\frac{\mathrm{cost}(G,\mathcal{A})}{\delta\mathrm{cost}(p,\mathcal{A})}\cdot\mathrm{cost}(p,\mathcal{S})}{\mathrm{cost}(G,\mathcal{A})+\mathrm{cost}(G,\mathcal{S})}\right)^2. \qquad (13)
\end{aligned}$$

To prove the first bound, we now merely recall since $p\in C\notin H_{G,\mathcal{S}}$, we have $\frac{\mathrm{cost}(p,\mathcal{S})}{\mathrm{cost}(p,\mathcal{A})}\le\left(\frac{4z}{\varepsilon}\right)^z$.

For the second bound, we first consider an arbitrary cluster C and let $q=\operatorname*{argmin}_{p\in C}\mathrm{cost}(p,\mathcal{S})$. Then we have for any solution $\mathcal{S}$

$$\mathrm{cost}(p,\mathcal{S})\le(\mathrm{dist}(q,\mathcal{S})+\mathrm{dist}(p,q))^z\le\frac{2^z\cdot\mathrm{cost}(C\cap G,\mathcal{S})+4^z\cdot\mathrm{cost}(C\cap G,\mathcal{A})}{|C|}. \qquad (14)$$

We first focus on the case $\varepsilon^z<k$. Continuing from Equation 13 and combining with Equation 14, we have

$$\begin{aligned}
&\sum_{p\in\Omega}2^{-2h+2}\left(\frac{\frac{\mathrm{cost}(G,\mathcal{A})}{\delta\mathrm{cost}(p,\mathcal{A})}\cdot\mathrm{cost}(p,\mathcal{S})}{\mathrm{cost}(G,\mathcal{A})+\mathrm{cost}(G,\mathcal{S})}\right)^2\\
\le&\frac{\mathrm{cost}(G,\mathcal{A})}{\delta\cdot(\mathrm{cost}(G,\mathcal{A})+\mathrm{cost}(G,\mathcal{S}))^2}\cdot 2^{-2h+2}\cdot\sum_{p\in\Omega}\frac{\mathrm{cost}(G,\mathcal{A})}{\delta\mathrm{cost}(p,\mathcal{A})}\cdot\mathrm{cost}(p,\mathcal{S})\cdot\left(\frac{4z}{\varepsilon}\right)^z\\
\le&\frac{\mathrm{cost}(G,\mathcal{A})\cdot 2^{-2h+2}\cdot\left(\frac{4z}{\varepsilon}\right)^z}{\delta\cdot(\mathrm{cost}(G,\mathcal{A})+\mathrm{cost}(G,\mathcal{S}))^2}\cdot\sum_C\frac{2^z\cdot\mathrm{cost}(C\cap G,\mathcal{S})+4^z\cdot\mathrm{cost}(C\cap G,\mathcal{A})}{|C|}\sum_{p\in C\cap\Omega}\frac{\mathrm{cost}(G,\mathcal{A})}{\delta\mathrm{cost}(p,\mathcal{A})}\\
\le&\frac{\mathrm{cost}(G,\mathcal{A})\cdot 2^{-2h+2}\cdot\left(\frac{4z}{\varepsilon}\right)^z}{\delta\cdot(\mathrm{cost}(G,\mathcal{A})+\mathrm{cost}(G,\mathcal{S}))^2}\cdot\sum_C 2^z\cdot\mathrm{cost}(C\cap G,\mathcal{S})+4^z\cdot\mathrm{cost}(C\cap G,\mathcal{A})\\
\le&\frac{\mathrm{cost}(G,\mathcal{A})\cdot 2^{-2h+2}\cdot\left(\frac{4z}{\varepsilon}\right)^z}{\delta\cdot(\mathrm{cost}(G,\mathcal{A})+\mathrm{cost}(G,\mathcal{S}))^2}\cdot 8^z\cdot(\mathrm{cost}(G,\mathcal{A})+\mathrm{cost}(G,\mathcal{S}))\\
\le&\frac{2^{-2h+2}\cdot\left(\frac{4z}{\varepsilon}\right)^z\cdot 8^z}{\delta}
\end{aligned}$$

where the third inequality uses event $\mathcal{E}_G$.

We now obtain a bound depending on k. Recall for any point $p\in C\cap G$, we have

$$\mathrm{cost}(G,\mathcal{A})\le 2k\cdot\mathrm{cost}(C\cap G,\mathcal{A})\le 4k\cdot\mathrm{cost}(p,\mathcal{A})\cdot|C| \qquad (15)$$

by definition of the groups. Then

$$
\begin{aligned}
& \sum_{p\in\Omega} 2^{-2h+2}\left(\frac{\frac{\mathrm{cost}(G,\mathcal{A})}{\delta\mathrm{cost}(p,\mathcal{A})}\cdot\mathrm{cost}(p,\mathcal{S})}{\mathrm{cost}(G,\mathcal{A})+\mathrm{cost}(G,\mathcal{S})}\right)^2 \\
\leq\ & \frac{1}{\delta\cdot(\mathrm{cost}(G,\mathcal{A})+\mathrm{cost}(G,\mathcal{S}))^2}\cdot 2^{-2h+2}\cdot\sum_{C}\sum_{p\in C\cap\Omega}\frac{4\cdot\mathrm{cost}(G,\mathcal{A})\cdot|C|\cdot k}{\delta\cdot\mathrm{cost}(p,\mathcal{A})}\cdot\mathrm{cost}^2(p,\mathcal{S}) \\
\leq\ & \frac{2^{-2h+4}\cdot k}{\delta\cdot(\mathrm{cost}(G,\mathcal{A})+\mathrm{cost}(G,\mathcal{S}))^2}\cdot\sum_{C}|C|\cdot\left(\frac{2^z\cdot\mathrm{cost}(C\cap G,\mathcal{S})+4^z\cdot\mathrm{cost}(C\cap G,\mathcal{A})}{|C|}\right)^2 \\
& \qquad\qquad\cdot\sum_{p\in C\cap\Omega}\frac{\mathrm{cost}(G,\mathcal{A})}{\delta\mathrm{cost}(p,\mathcal{A})} \\
\leq\ & \frac{2^{-2h+4}\cdot k}{\delta\cdot(\mathrm{cost}(G,\mathcal{A})+\mathrm{cost}(G,\mathcal{S}))^2}\cdot\sum_{C}\left(2^z\cdot\mathrm{cost}(C\cap G,\mathcal{S})+4^z\cdot\mathrm{cost}(C\cap G,\mathcal{A})\right)^2 \\
\leq\ & \frac{2^{-2h+4}\cdot k}{\delta\cdot(\mathrm{cost}(G,\mathcal{A})+\mathrm{cost}(G,\mathcal{S}))^2}\cdot\left(\sum_{C}2^z\cdot\mathrm{cost}(C\cap G,\mathcal{S})+4^z\cdot\mathrm{cost}(C\cap G,\mathcal{A})\right)^2 \\
\leq\ & \frac{2^{-2h+4}\cdot k\cdot 64^z}{\delta\cdot(\mathrm{cost}(G,\mathcal{A})+\mathrm{cost}(G,\mathcal{S}))^2}\cdot\left(\mathrm{cost}(G,\mathcal{A})+\mathrm{cost}(G,\mathcal{S})\right)^2 \\
\leq\ & \frac{2^{-2h+4}\cdot k\cdot 64^z}{\delta}
\end{aligned}
$$

where the first inequality uses Equation 15, the second inequality uses Equation 14 and the third inequality uses Equation 13. □

We now combine the bound on the variance with union bounds for $X_{G,\mathcal{S},h}$. Let δ be as given in the statement of Lemma 16.

We will use the following lemma for bounding the expected maximum of independent Gaussians.

Lemma 24 (Lemma 2.3 of Massart [76])**.** *Let $g_i \sim \mathcal{N}(0,\sigma_i^2)$, $i\in[n]$ be Gaussian random variables and suppose $\sigma_i \leq \sigma$ for all i. Then*

$$\mathbb{E}[\max_{i\in[n]}|g_i|] \leq 2\sigma\cdot\sqrt{2\ln n}.$$

We use the following fact:

Fact 4.

$$\mathbb{E}_{\Omega,\xi}[\sup_{\mathcal{S}} X_{G,\mathcal{S},h}] = \mathbb{P}_\Omega[\mathcal{E}_G]\cdot\mathbb{E}_\xi[\sup_{\mathcal{S}} X_{G,\mathcal{S},h}\mid\mathcal{E}_G] + \mathbb{P}_\Omega[\overline{\mathcal{E}_G}]\cdot\mathbb{E}_\xi[\sup_{\mathcal{S}} X_{G,\mathcal{S},h}\mid\overline{\mathcal{E}_G}].$$

$$\mathbb{E}_{\Omega,\xi}[\sup_{\mathcal{S}} X_{\mathcal{S}}] = \mathbb{P}_\Omega[\mathcal{E}_G]\cdot\mathbb{E}_{\Omega,\xi}[\sup_{\mathcal{S}} X_{G,\mathcal{S},h}\mid\mathcal{E}_G] + \mathbb{P}_\Omega[\overline{\mathcal{E}_G}]\cdot\mathbb{E}_{\Omega,\xi}[\sup_{\mathcal{S}} X_{G,\mathcal{S},h}\mid\overline{\mathcal{E}_G}].$$

Proof. Since $\mathcal{E}_G$ is independent of ξ, the law of total expectation gives

$$\begin{aligned}\mathbb{E}_{\Omega,\xi}[\sup_{\mathcal{S}} X_{G,\mathcal{S},h}] &= \mathbb{P}_{\Omega,\xi}[\mathcal{E}_G] \cdot \mathbb{E}_{\Omega,\xi}[\sup_{\mathcal{S}} X_{G,\mathcal{S},h} \mid \mathcal{E}_G] + \mathbb{P}_{\Omega,\xi}[\overline{\mathcal{E}}_G] \cdot \mathbb{E}_{\Omega,\xi}[\sup_{\mathcal{S}} X_{G,\mathcal{S},h} \mid \overline{\mathcal{E}}_G] \\ &= \mathbb{P}_{\Omega}[\mathcal{E}_G] \cdot \mathbb{E}_{\Omega,\xi}[\sup_{\mathcal{S}} X_{G,\mathcal{S},h} \mid \mathcal{E}_G] + \mathbb{P}_{\Omega}[\overline{\mathcal{E}}_G] \cdot \mathbb{E}_{\Omega,\xi}[\sup_{\mathcal{S}} X_{G,\mathcal{S},h} \mid \overline{\mathcal{E}}_G]\end{aligned}$$

□

We will bound the expectation, first assuming the (more likely case) that event $\mathcal{E}_G$ holds, then assuming the (more unlikely case) that event $\mathcal{E}_G$ does not hold.

Condition on $\mathcal{E}_G$. We simply upper bound $\mathbb{P}[\mathcal{E}_G]$ with 1. Assume that the points lie in dimension $\alpha \cdot \log \|P\|_0 \cdot \varepsilon^{-2}$ and let $t \in O(\log(\alpha \cdot \varepsilon^{-2}))$. We will show that the contribution of the $X_{G,\mathcal{S},h}$ with $h \leq t$ to the expectation is at most $\varepsilon/\log(1/\varepsilon)$, and then bound the expectation for all remaining $h \geq t$.

First, we recall that, conditioned on event $\mathcal{E}_G$ and due to Lemma 23 and by our choice of δ, we have

$$\mathbf{Var}[X_{h,\mathcal{S}} \mid \mathcal{E}_G] \leq \beta_3 \cdot \delta^{-1} \cdot 2^{\beta_4 z \log z} 2^{-2h} \cdot \min(\varepsilon^{-z}, k).$$

for absolute constants β_3 and β_4.

For the number of distinct $X_{G,\mathcal{S},h}$, we have an upper bound of $|\mathbb{N}_{h-1}| \cdot |\mathbb{N}_h| \leq |\mathbb{N}_h|^2$, where $|\mathbb{N}_h|$ is the size of an $(2^{-h}, k, z)$-clustering net. Due to Lemma 21, this is at most $\exp(2 \cdot \gamma_7 \cdot z^2 \cdot k \cdot \log \|P\|_0 \cdot 2^{2h} \cdot \log \frac{1}{2^{-h} \cdot \varepsilon})$, which by the upper bound on h is at most $\exp(2 \cdot \gamma_7 \cdot z^2 \cdot k \cdot \log \|P\|_0 \cdot 2^{2h} \cdot \log \frac{\alpha}{\varepsilon^{-3}})$. Therefore, using Lemma 24

$$\begin{aligned}& \mathbb{E}[\sup_{\mathcal{S}} |X_{G,\mathcal{S},h}| \mid \mathcal{E}_G] \\ \leq\ & 2\sqrt{\mathbf{Var}[X_{G,\mathcal{S},h} \mid \mathcal{E}_G]}\sqrt{2\ln(|\mathbb{N}_{h-1}| \cdot |\mathbb{N}_h|)} \\ \leq\ & 2\sqrt{\delta^{-1} \cdot 2^{\beta_1 z \log z} 2^{-2h} \cdot \min(\varepsilon^{-z}, k)} \cdot \sqrt{2 \cdot \gamma_7 \cdot z^2 \cdot k \cdot \log \|P\|_0 \cdot 2^{2h} \cdot \log\log \frac{\alpha}{\varepsilon^{-3}}} \\ =\ & 2 \cdot \left(\frac{2 \cdot \gamma_7 \cdot z^2 \cdot k \cdot \log \|P\|_0 \cdot 2^{2h} \cdot \log \frac{\alpha}{\varepsilon^{-3}} \cdot 2^{\beta_1 z \log z} 2^{-2h} \cdot \min(\varepsilon^{-z}, k)}{2^{\gamma_4 \cdot z \log(1+z)} \cdot k \cdot \log \frac{k}{\varepsilon} \cdot \varepsilon^{-2} \cdot \log^3 \varepsilon^{-1} \cdot \min(\varepsilon^{-z}, k))} \right)^{1/2} \\ \leq\ & \beta \cdot \frac{\varepsilon}{\log 1/\varepsilon}, \qquad (16)\end{aligned}$$

where the final inequality holds for a sufficiently large choice of the constants γ_4.

We now assume that $h \geq t$. This time, using Lemma 22, we have a net of size at most $|\mathbb{N}_h| \leq \exp\left(\gamma_8 \cdot k \cdot d \cdot z \log(4z/(2^{-h}\varepsilon))\right)$. Furthermore, using the assumption that the points lie in dimension $\log \|P\|_0 \cdot \varepsilon^{-2} \in O(\log k/\varepsilon) \cdot \varepsilon^{-2}$, we have

$$
\begin{aligned}
& \mathbb{E}[\sup_{\mathcal{S}} |X_{G,\mathcal{S},h}| \mid \mathcal{E}_G] \\
\leq\ & 2\sqrt{\mathbf{Var}[X_{G,\mathcal{S},h} \mid \mathcal{E}_G]}\sqrt{2\ln(|\mathbb{N}_{h-1}| \cdot |\mathbb{N}_h|)} && (17)\\
\leq\ & 2 \cdot \left(\frac{2 \cdot \gamma_8 \cdot k \cdot d \cdot z \log(4z/(2^{-h}\varepsilon)) \cdot \beta_3 \cdot 2^{\beta_4 z \log z} 2^{-2h} \cdot \min(\varepsilon^{-z}, k)}{2^{\gamma_4 \cdot z \log(1+z)} \cdot k \cdot \log \frac{k}{\varepsilon} \cdot \varepsilon^{-2} \cdot \log^3 \varepsilon^{-1} \cdot \min(\varepsilon^{-z}, k))} \right)^{1/2} \\
\leq\ & 2 \cdot \left(\frac{2 \cdot \gamma_8 \cdot k \cdot \alpha \cdot \|P\|_0 \varepsilon^{-2} \cdot z \log(4z/(2^{-h}\varepsilon)) \cdot 2^{\beta_2 z \log(1+z)} 2^{-2h} \cdot \min(\varepsilon^{-z}, k)}{2^{\gamma_4 \cdot z \log(1+z)} \cdot k \cdot \log \frac{k}{\varepsilon} \cdot \varepsilon^{-2} \cdot \log^3 \varepsilon^{-1} \cdot \min(\varepsilon^{-z}, k))} \right)^{1/2} \\
\leq\ & \beta \cdot \left(\frac{\log h}{2^{2h} \log^2 1/\varepsilon} \right)^{1/2}, && (18)
\end{aligned}
$$

where again the final inequality holds for a sufficiently large choice of the constant γ_4 and with the assumption $\|P\|_0 = \text{poly}(k/\varepsilon)$.

Summing up Equations 16 and 18 for all h, we then obtain

$$
\begin{aligned}
\mathbb{E}[\sup_{\mathcal{S}} |X_{\mathcal{S}}| \mid \mathcal{E}_G] \quad & \leq \quad \sum_{h=0}^{\infty} \mathbb{E}[\sup_{\mathcal{S}} |X_{h,\mathcal{S}}| \mid \mathcal{E}_G] \\
& = \quad \sum_{h=0}^{t-1} \mathbb{E}[\sup_{\mathcal{S}} |X_{h,\mathcal{S}}| \mid \mathcal{E}_G] + \sum_{h=t}^{\infty} \mathbb{E}[\sup_{\mathcal{S}} |X_{h,\mathcal{S}}| \mid \mathcal{E}_G] \\
\text{(Eq. 16 18)} \quad & \leq \quad \sum_{h=0}^{t-1} \beta \cdot \frac{\varepsilon}{\log 1/\varepsilon} + \sum_{h=t}^{\infty} \beta \cdot \left(\frac{\log h}{2^{2h} \log^2 1/\varepsilon} \right)^{1/2} \\
& \overset{t=\log(\alpha/\varepsilon^2)}{\leq} \quad \beta \cdot \varepsilon + \beta \cdot \varepsilon \sum_{h=0}^{\infty} \left(\frac{\log(h+t)}{\beta \cdot 2^{2h} \log^2 1/\varepsilon} \right)^{1/2} \in O(\varepsilon). \qquad (19)
\end{aligned}
$$

Now, we move onto the case $\overline{\mathcal{E}_G}$. Due to Lemma 19, we have

$$
\mathbb{P}[\overline{\mathcal{E}_G}] \leq k \cdot z^2 \log^2(z/\varepsilon) \exp(-O(1) \cdot \frac{\varepsilon^2}{k}\delta) \leq \varepsilon^{2z}, \qquad (20)
$$

where the second upper bound follows from our choice of sufficiently large constants in the definition of δ[2]

Using the worse variance bound from Lemma 23 for the variance in equations 16 and 18, we then have for $h \leq t$

$$
\mathbb{E}[\sup_{\mathcal{S}} |X_{h,\mathcal{S}}| \mid \overline{\mathcal{E}_G}] \leq \beta \frac{\varepsilon}{\log 1/\varepsilon} \cdot \varepsilon^{-z}. \qquad (21)
$$

[2]The bound is not close to tight and can be any power of ε. The stated bound happens to be sufficient here.

and for $h > t$

$$\mathbb{E}[\sup_{\mathcal{S}} |X_{h,\mathcal{S}}| \mid \overline{\mathcal{E}_G}] \leq \beta \cdot \left(\frac{\log h}{2^{2h} \log^2 1/\varepsilon} \right)^{1/2} \cdot \varepsilon^{-z}. \tag{22}$$

Using an analogous calculation to those in the derivation of Equation 19 using Equations 21 and 22, we now obtain

$$\mathbb{E}[\sup_{\mathcal{S}} |X_{\ell,\mathcal{S}}| \mid \overline{\mathcal{E}_G}] \in O(\varepsilon^{-z}). \tag{23}$$

Combining Equations 19, 20, and 23, we finally have

$$\mathbb{P}_\Omega[\mathcal{E}_G] \cdot \mathbb{E}_\xi[\sup_{\mathcal{S}} X_{\ell,\mathcal{S}} \mid \mathcal{E}_G] + \mathbb{P}_\Omega[\overline{\mathcal{E}_G}] \cdot \mathbb{E}_\xi[\sup_{\mathcal{S}} X_{\ell,\mathcal{S}} \mid \overline{\mathcal{E}_G}] \in O(\varepsilon + \varepsilon^{2z} \cdot \varepsilon^{-z}) = O(\varepsilon).$$

□

We now repeat these calculations for Lemma 18. The main difference is the variance bound, which improves over what we were able to prove for Lemma 16. The remaining arguments differ only in the calculations and are omitted.

Proof of Lemma 18. Let $G \in G^O$. As for Lemma 18, we bound the random variable

$$X_{G,\mathcal{S}} \quad := \quad \frac{1}{\text{cost}(G,\mathcal{A}) + \text{cost}(G,\mathcal{S})} \sum_{p \in \Omega} \left(\sum_{h=1}^{\infty} \xi_p \cdot w_p \cdot \left(v_p^{\mathcal{S},h+1} - v_p^{\mathcal{S},h} \right) \right) + \xi_p \cdot w_p \cdot v_p^{\mathcal{S},1}.$$

by bounding

$$\begin{aligned} X_{G,\mathcal{S},0} \quad &:= \quad \frac{1}{\text{cost}(G,\mathcal{A}) + \text{cost}(G,\mathcal{S})} \sum_{p \in \Omega} \xi_p \cdot w_p \cdot v_p^{\mathcal{S},1} \\ X_{G,\mathcal{S},h} \quad &:= \quad \frac{1}{\text{cost}(G,\mathcal{A}) + \text{cost}(G,\mathcal{S})} \sum_{p \in \Omega} \xi_p \cdot w_p \cdot \left(v_p^{\mathcal{S},h+1} - v_p^{\mathcal{S},h} \right). \end{aligned}$$

Lemma 25. *Let $G \in G^O$ and let $\beta_3 > 0$ be a constant. Fix a solution $\mathcal{S}$. Then $X_{G,\mathcal{S},h}$ is Gaussian distributed with mean 0. The variance of $X_{G,\mathcal{S},h}$ is always at most*

$$\sum_{p \in \Omega} \left(\frac{w_p \cdot \left(v_p^{\mathcal{S},h+1} - v_p^{\mathcal{S},h} \right)}{cost(G,\mathcal{A}) + cost(G,\mathcal{S})} \right)^2 \in \frac{16^z \cdot 2^{-2h+2}}{\delta}.$$

Proof. As in the proof of Lemma 23, we use that if $\xi_p \sim \mathcal{N}(0,1)$, then $\sum_p a_p \cdot \xi_p$ is Gaussian distributed with mean 0 and variance $\sum_p a_p^2$.

For any $p \in C \notin H_{G,\mathcal{S}}$, we have $\text{cost}(p,\mathcal{S}) \leq \left(\frac{4z}{\varepsilon}\right)^z \text{cost}(p\mathcal{A})$. A terminal embedding with target dimension $O(z^2 2^{2h} \log \|P\|_0)$ preserves the cost up to a factor (1 ± 2^{-h}), i.e. we have $(1 - 2^{-h}) \cdot$

$\text{cost}(p,\mathcal{S}) \leq v_p^{\mathcal{S},h} \leq (1+2^{-h})\text{cost}(p,\mathcal{S})$. Therefore

$$\begin{aligned}
&\sum_{p\in\Omega}\left(\frac{w_p\cdot\left(v_p^{\mathcal{S},h+1}-v_p^{\mathcal{S},h}\right)}{\text{cost}(G,\mathcal{A})+\text{cost}(G,\mathcal{S})}\right)^2\\
=&\sum_{p\in\Omega}\left(\frac{w_p\cdot\left(v_p^{\mathcal{S},h+1}-\text{cost}(p,\mathcal{S})+\text{cost}(p,\mathcal{S})-v_p^{\mathcal{S},h}\right)}{\text{cost}(G,\mathcal{A})+\text{cost}(G,\mathcal{S})}\right)^2\\
\leq&\sum_{p\in\Omega}\left(\frac{w_p\cdot 2^{-h-1}\cdot\text{cost}(p,\mathcal{S})}{\text{cost}(G,\mathcal{A})+\text{cost}(G,\mathcal{S})}\right)^2\\
\leq&\sum_{p\in\Omega}2^{-2h+2}\left(\frac{\frac{\text{cost}(G,\mathcal{A})}{\delta\text{cost}(p,\mathcal{A})}\cdot\text{cost}(p,\mathcal{S})}{\text{cost}(G,\mathcal{A})+\text{cost}(G,\mathcal{S})}\right)^2. \qquad (24)
\end{aligned}$$

By definition, none of the points with non-zero coordinates in the cost vector v are far, i.e. $\text{cost}(p,\mathcal{S}) \leq 4^z\cdot\text{cost}(p,\mathcal{A})$.

Therefore, Equation 24 becomes

$$\sum_{p\in\Omega}16^z\cdot 2^{-2h+2}\frac{1}{\delta^2}=\frac{16^z\cdot 2^{-2h+2}}{\delta}.$$

□

The remaining calculations are completely analogous to that of Lemma 16, albeit with a significantly better (and simpler) bound on the variance. □

6.6 Estimating $\|u^{G,\mathcal{S}}\|_1$ (Proofs of Lemma 15 and Lemma 17)

Lemma 26. *Let $G \in G^M$ be a group. Condition on event $\mathcal{E}_G$. Suppose $\varepsilon < 1/4$. Then, for any solution $\mathcal{S}$, and point $p \in C$ with $C \in H_{G,\mathcal{S}}$, we have:*

$$|\textit{cost}(C\cap G,\mathcal{S})-|C\cap G|\cdot\textit{cost}(p,\mathcal{S})|\leq 10\cdot\varepsilon\cdot\textit{cost}(C,\mathcal{S}).$$

Proof. Let $p,p' \in C \in H_{\mathcal{S}}$. First, we require an upper bound on $\text{cost}(p,p')$. We have due to Lemma 10 and since $\text{cost}(p,\mathcal{A}) \leq 2\cdot\text{cost}(p',\mathcal{A})$

$$\text{cost}(p,p')\leq 2^{z-1}\left(\text{cost}(p,\mathcal{A})+\text{cost}(p',\mathcal{A})\right)\leq 2^{z+1}\text{cost}(p',\mathcal{A}).$$

Let $p' \in C$ be a point such that $\text{cost}(p,\mathcal{S}) > \left(\frac{4z}{\varepsilon}\right)^z\text{cost}(p',\mathcal{A})$. We now give upper and lower bounds

for $\mathrm{cost}(p, \mathcal{S})$ in terms of $\mathrm{cost}(p', \mathcal{S})$, again using Lemma 10. For the upper bound:

$$\begin{aligned}
\mathrm{cost}(p, \mathcal{S}) &\leq (1+\varepsilon) \cdot \mathrm{cost}(p', \mathcal{S}) + \left(\frac{z+\varepsilon}{\varepsilon}\right)^{z-1} \mathrm{cost}(p, p') \\
&\leq (1+\varepsilon) \cdot \mathrm{cost}(p', \mathcal{S}) + \left(\frac{z+\varepsilon}{\varepsilon}\right)^{z-1} 2^{z+1} \cdot \mathrm{cost}(p', \mathcal{A}) \\
&\leq (1+\varepsilon) \cdot \mathrm{cost}(p', \mathcal{S}) + \left(\frac{z+\varepsilon}{\varepsilon}\right)^{z-1} 2^{z+1} \cdot \left(\frac{\varepsilon}{4z}\right)^{z} \mathrm{cost}(p', \mathcal{S}) \\
&\leq (1+2\varepsilon) \cdot \mathrm{cost}(p', \mathcal{S})
\end{aligned}$$

For the lower bound:

$$\begin{aligned}
\mathrm{cost}(p', \mathcal{S}) &\leq (1+\varepsilon) \cdot \mathrm{cost}(p, \mathcal{S}) + \left(\frac{z+\varepsilon}{\varepsilon}\right)^{z-1} \mathrm{cost}(p, p') \\
&\leq (1+\varepsilon) \cdot \mathrm{cost}(p, \mathcal{S}) + \left(\frac{z+\varepsilon}{\varepsilon}\right)^{z-1} 2^{z+1} \cdot \mathrm{cost}(p', \mathcal{A}) \\
&\leq (1+\varepsilon) \cdot \mathrm{cost}(p, \mathcal{S}) + \left(\frac{z+\varepsilon}{\varepsilon}\right)^{z-1} 2^{z+1} \cdot \left(\frac{\varepsilon}{4z}\right)^{z} \mathrm{cost}(p', \mathcal{S}) \\
&\leq (1+\varepsilon) \cdot \mathrm{cost}(p, \mathcal{S}) + \varepsilon \cdot \mathrm{cost}(p', \mathcal{S}) \\
\Rightarrow \mathrm{cost}(p, \mathcal{S}) &\geq \frac{1-\varepsilon}{1+\varepsilon} \cdot \mathrm{cost}(p', \mathcal{S}) \geq (1-2\varepsilon) \cdot \mathrm{cost}(p', \mathcal{S})
\end{aligned}$$

Thus we have $\mathrm{cost}(C_i, \mathcal{S}) = \sum_{p \in C_i} \mathrm{cost}(p, \mathcal{S}) = (1 \pm 2\varepsilon) \cdot |C_i| \cdot \mathrm{cost}(p', \mathcal{S})$. Conditioned on event $\mathcal{E}$, we now have $\sum_{p \in C_i \cap \Omega} w_p = (1 \pm \varepsilon) \cdot |C_i|$, hence

$$\begin{aligned}
\sum_{p \in C_i \cap \Omega} w_p \cdot \mathrm{cost}(p, \mathcal{S}) &\leq \sum_{p \in C_i \cap \Omega} w_p \cdot (1+2\varepsilon) \cdot \mathrm{cost}(p', \mathcal{S}) = (1+\varepsilon) \cdot (1+2\varepsilon) \cdot |C_i| \cdot \mathrm{cost}(p', \mathcal{S}) \\
&\leq \mathrm{cost}(C_i, \mathcal{S}) \cdot \frac{(1+\varepsilon) \cdot (1+2\varepsilon)}{1-2\varepsilon}
\end{aligned}$$

and analogously for the lower bound

$$\begin{aligned}
\sum_{p \in C_i \cap \Omega} w_p \cdot \mathrm{cost}(p, \mathcal{S}) &\geq \sum_{p \in C_i \cap \Omega} w_p \cdot (1-2\varepsilon) \cdot \mathrm{cost}(p', \mathcal{S}) = (1-\varepsilon) \cdot (1-2\varepsilon) \cdot |C_i| \cdot \mathrm{cost}(p', \mathcal{S}) \\
&\geq \mathrm{cost}(C_i, \mathcal{S}) \cdot \frac{(1-\varepsilon) \cdot (1-2\varepsilon)}{1+2\varepsilon}.
\end{aligned}$$

The final bound follows by observing for $\varepsilon < 1/4$, we have $\frac{(1+\varepsilon)\cdot(1+2\varepsilon)}{1-2\varepsilon} \leq 1 + 10\varepsilon$ and $\frac{(1-\varepsilon)\cdot(1-2\varepsilon)}{1+2\varepsilon} \geq 1 - 10\varepsilon$. □

Proof of Lemma 15. We have

$$\begin{aligned}
& \mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\left| \frac{\sum_{p\in\Omega} w_p \cdot u^{\mathcal{S}}(p) - \|u^{\mathcal{S}}\|_1}{\mathrm{cost}(G,\mathcal{A}) + \mathrm{cost}(G,\mathcal{S})} \right| \right] \\
\leq \; & \mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\left| \frac{\sum_{p\in\Omega} w_p \cdot u^{G,\mathcal{S}}(p) - \|u^{G,\mathcal{S}}\|_1}{\mathrm{cost}(G,\mathcal{A}) + \mathrm{cost}(G,\mathcal{S})} \right| \mid \mathcal{E}_G \right] \cdot \mathbb{P}_\Omega[\mathcal{E}_G] && (25) \\
& + \mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\left| \frac{\sum_{p\in\Omega} w_p \cdot u^{G,\mathcal{S}}(p) - \|u^{G,\mathcal{S}}\|_1}{\mathrm{cost}(G,\mathcal{A}) + \mathrm{cost}(G,\mathcal{S})} \right| \mid \overline{\mathcal{E}_G} \right] \cdot \mathbb{P}_\Omega[\overline{\mathcal{E}_G}] && (26)
\end{aligned}$$

We first consider the term 25. A trivial upper bound for $\mathbb{P}_\Omega[\mathcal{E}_G]$ is 1. Using Lemma 26 we have

$$\sum_{C\in H_{G,\mathcal{S}}} w_p \cdot u_p^{G,\mathcal{S}} = \sum_{C\in H_{G,\mathcal{S}}} w_p \cdot \mathrm{cost}(p,\mathcal{S}) = (1 \pm 10\varepsilon) \sum_{C\in H_{G,\mathcal{S}}} \mathrm{cost}(p,\mathcal{S}).$$

The remaining entries of $u^{G,\mathcal{S}}$ are 0. Since $\|u^{G,\mathcal{S}}\|_1 \leq \mathrm{cost}(G,\mathcal{S})$, we therefore have

$$\mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\left| \frac{\sum_{p\in\Omega} w_p \cdot u^{G,\mathcal{S}}(p) - \|u^{G,\mathcal{S}}\|_1}{\mathrm{cost}(G,\mathcal{A}) + \mathrm{cost}(G,\mathcal{S})} \right| \mid \mathcal{E}_G \right] \cdot \mathbb{P}_\Omega[\mathcal{E}_G] \leq 10\varepsilon. \tag{27}$$

We now focus on term 26. We distinguish between two cases. If $\sum_{p\in\Omega} w_p \cdot u^{\mathcal{S}}(p) \leq \|u^{\mathcal{S}}\|_1$ then we have

$$\left| \frac{\sum_{p\in\Omega} w_p \cdot u^{\mathcal{S}}(p) - \|u^{\mathcal{S}}\|_1}{\mathrm{cost}(G,\mathcal{A}) + \mathrm{cost}(G,\mathcal{S})} \right| \leq \frac{\|u^{\mathcal{S}}\|_1}{\mathrm{cost}(P,\mathcal{A}) + \mathrm{cost}(P,\mathcal{S})} \leq 1 \tag{28}$$

If $\sum_{p\in\Omega} w_p \cdot u^{\mathcal{S}}(p) \geq \|u^{\mathcal{S}}\|_1$ then we have

$$\begin{aligned}
\sum_{p\in\Omega} w_p \cdot u_p^{G,\mathcal{S}} &= \sum_{p\in\Omega} \frac{\mathrm{cost}(G,\mathcal{A})}{\delta \cdot \mathrm{cost}(p,\mathcal{A})} \cdot u_p^{G,\mathcal{S}} \\
(Eq.\ 7) \quad &\leq \sum_{C} \sum_{p\in\Omega\cap C\cap G} \frac{4k|C\cap G| \cdot \mathrm{cost}(G,\mathcal{A})}{\delta \cdot \mathrm{cost}(G,\mathcal{A})} \cdot u_p^{G,\mathcal{S}} \\
&\leq 4k \cdot \sum_{C} \sum_{p\in\Omega\cap C\cap G} \frac{|C\cap G|}{\delta} \cdot u_p^{G,\mathcal{S}} \\
&\leq 4k \cdot \|u^{\mathcal{S}}\|_1,
\end{aligned}$$

Therefore in this case

$$\left| \frac{\sum_{p\in\Omega} w_p \cdot u_p^{G,\mathcal{S}} - \|u^{G,\mathcal{S}}\|_1}{\mathrm{cost}(G,\mathcal{A}) + \mathrm{cost}(G,\mathcal{S})} \right| \leq \frac{4k \cdot \|u^{\mathcal{S}}\|_1}{\mathrm{cost}(G,\mathcal{A}) + \mathrm{cost}(G,\mathcal{S})} \leq 4k \tag{29}$$

Due to Lemma 19, $\mathbb{P}_\Omega[\overline{\mathcal{E}_G}] \leq k \cdot \exp\left(-\frac{\varepsilon^2}{9\cdot k}\delta\right)$. Hence, if we set $\delta \geq 9\varepsilon^{-2} k \log \frac{4k^2}{\varepsilon}$, we have $\mathbb{P}[\overline{\mathcal{E}_G}] \leq \frac{\varepsilon}{4k}$. This implies together with Equations 28 and 29

$$\mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\left| \frac{\sum_{p\in\Omega} w_p \cdot q^{G,\mathcal{S}}(p) - \|q^{G,\mathcal{S}}\|_1}{\mathrm{cost}(G,\mathcal{A}) + \mathrm{cost}(G,\mathcal{S})} \right| \mid \overline{\mathcal{E}_G} \right] \cdot \mathbb{P}_\Omega[\overline{\mathcal{E}_G}] \leq 4k \cdot \frac{\varepsilon}{4k} \leq \varepsilon. \tag{30}$$

The claim now follows by combining Equations 27 and 30 and rescaling ε. $\square$

We now turn our attention to Lemma 17. Henceforth, we let $G \in G^O$. We first require an analogue of event $\mathcal{E}_G$. We define event $\mathcal{E}_{far,G}$ that for all clusters C with

$$\sum_{p \in C \cap G \cap \Omega} \frac{\text{cost}(G, \mathcal{A})}{\delta \cdot \text{cost}(p, \mathcal{A})} \text{cost}(p, \mathcal{A}) = (1 \pm \varepsilon) \cdot \text{cost}(C \cap G, \mathcal{A}).$$

Furthermore, by definition of the groups, we have

$$\text{cost}(G, \mathcal{A}) \leq 2k \cdot \text{cost}(C \cap G, \mathcal{A}). \tag{31}$$

We start by bounding the probability that $\mathcal{E}_{far,G}$ fails to occur.

Lemma 27. *Event $\mathcal{E}_{far,G}$ happens with probability at least $1 - k \exp(\frac{\varepsilon^2}{5 \cdot k} \cdot \delta)$.*

Proof. Again, we aim to use Bernstein's Inequality. Let p_j be the jth point in the sample Ω with respect to arbitrary but fixed ordering. Consider the random variable $w_{p_j,C} = \begin{cases} w_p \cdot \text{cost}(p, \mathcal{A}) & \text{if } p_j = p \in C \cap G \\ 0 & \text{else} \end{cases}$. Then:

$$\begin{aligned} E[w_{p_j,C}^2] &= \sum_{p \in C \cap G} \left(\frac{\text{cost}(G, \mathcal{A})}{\delta \cdot \text{cost}(p, \mathcal{A})} \cdot \text{cost}(p, \mathcal{A}) \right)^2 \cdot \mathbb{P}[p \in \Omega] \\ &= \frac{\text{cost}(G, \mathcal{A})}{\delta^2} \cdot \sum_{p \in C \cap G} \text{cost}(p, \mathcal{A}) \\ &= \frac{\text{cost}(G, \mathcal{A})}{\delta^2} \text{cost}(C \cap G, \mathcal{A}) \\ (Eq.\ 31) \quad &\leq \frac{2k}{\delta^2} \cdot \text{cost}^2(C \cap G, \mathcal{A}) \end{aligned}$$

Furthermore, we have by the same argument the following upper bound for the maximum value any of the $w_{p_j,C}$:

$$M := \max_{p \in C \cap G} \frac{\text{cost}(G, \mathcal{A})}{\delta \cdot \text{cost}(p, \mathcal{A})} \cdot \text{cost}(p, \mathcal{A}) \leq \frac{2k}{\delta} \cdot \text{cost}(C \cap G, \mathcal{A}).$$

Combining both bounds with Bernstein's inequality now yields

$$\begin{aligned} &\mathbb{P}[|\text{cost}(C \cap G \cap \Omega, \mathcal{A}) - \text{cost}(C \cap G, \mathcal{A})| \leq \varepsilon \cdot \text{cost}(C \cap G, \mathcal{A})] \\ \leq \quad &\exp\left(-\frac{\varepsilon^2 \cdot \text{cost}^2(C \cap G, \mathcal{A})}{2 \sum_{i=1}^{\delta} Var[X_i] + \frac{1}{3} M \cdot \varepsilon \cdot \text{cost}(C \cap G, \mathcal{A})} \right) \leq \exp\left(-\frac{\varepsilon^2}{5 \cdot k} \cdot \delta \right) \end{aligned}$$

Reformulating, we now have

$$\sum_{p \in C \cap G \cap \Omega} \frac{\text{cost}(G, \mathcal{A})}{\delta \cdot \text{cost}(p, \mathcal{A})} \text{cost}(p, \mathcal{A}) = (1 \pm \varepsilon) \cdot \text{cost}(C \cap G, \mathcal{A}).$$

Taking a union bound over all clusters yields the claim. □

Lemma 28. *Condition on event $\mathcal{E}_{far,G}$. Suppose $C \in F_{G,\mathcal{S}}$. Then*

$$cost(C \cap G, \mathcal{S}) + \sum_{p \in C \cap G \cap \Omega} w_p \cdot cost(p, \mathcal{S}) \leq \varepsilon \cdot cost(C, \mathcal{S}).$$

Proof. First, we fix a cluster $C \in \mathcal{A}$, and show that points of $C \cap G_{far,\mathcal{S}}$ are very cheap compared to $\text{cost}(C, \mathcal{S})$, assuming that $C \in F_{G,\mathcal{S}}$. Let c be the center serving $p \in G_{far,\mathcal{S}} \cap C$ in $\mathcal{A}$. Let C_{close} be the points of C with cost at most $\left(\frac{2z}{\varepsilon}\right)^z \cdot \frac{\text{cost}(C,c)}{|C|}$. Consider an arbitrary point in $p' \in C_{close}$. Due to the triangle inequality and $\text{cost}(p, \mathcal{S}) > 4^z \cdot \text{cost}(p, c)$, we have $\text{dist}(c, \mathcal{S}) \geq \text{dist}(p, \mathcal{S}) - \text{dist}(p, c) \geq 4\text{dist}(p, c) - \text{dist}(p, c) \geq \text{dist}(p, c)$. Therefore $\text{cost}(c, \mathcal{S}) \geq \left(\frac{4z}{\varepsilon}\right)^{2z} \cdot \frac{\text{cost}(C,c)}{|C|}$. Using this and Lemma 10 we now have for any $p' \in C_{close}$

$$\begin{aligned}
\text{cost}(c, \mathcal{S}) &\leq (1+\varepsilon) \cdot \text{cost}(p', \mathcal{S}) + \left(\frac{2z+\varepsilon}{\varepsilon}\right)^{z-1} \cdot \text{cost}(p', c) \\
&\leq (1+\varepsilon) \cdot \text{cost}(p', \mathcal{S}) + \left(\frac{2z+\varepsilon}{\varepsilon}\right)^{z-1} \cdot \left(\frac{2z}{\varepsilon}\right)^{z} \cdot \frac{\text{cost}(C, c)}{|C|} \\
&\leq (1+\varepsilon) \cdot \text{cost}(p', \mathcal{S}) + \frac{\left(\frac{4z}{\varepsilon}\right)^{2z-1} \cdot \frac{\text{cost}(C,c)}{|C|}}{\text{cost}(p, c)} \cdot \text{cost}(p, c) \\
&\leq (1+\varepsilon) \cdot \text{cost}(p', \mathcal{S}) + \varepsilon \cdot \text{cost}(p, c) \qquad \text{since } p \in G \in G^O \\
&\leq (1+\varepsilon) \cdot \text{cost}(p', \mathcal{S}) + \varepsilon \cdot \text{cost}(c, \mathcal{S}) \\
\Rightarrow \text{cost}(p', \mathcal{S}) &\geq \frac{1-\varepsilon}{1+\varepsilon} \cdot \text{cost}(c, \mathcal{S}) \qquad (32)
\end{aligned}$$

We now bound $\text{cost}(C, \mathcal{S})$ in terms of $\text{cost}(C, c)$. We have due to Markov's inequality $|C \cap G| \leq \left(\frac{\varepsilon}{4z}\right)^{2z}$ and $|C_{close}| \geq (1-\varepsilon) \cdot |C|$ and therefore

$$\text{cost}(C, \mathcal{S}) \geq \text{cost}(C_{close}, \mathcal{S}) = \sum_{p' \in C_{close}} \text{cost}(p', \mathcal{S}) \geq |C_{close}| \cdot \frac{1-\varepsilon}{1+\varepsilon} \cdot cost(c, \mathcal{S}) \qquad (33)$$

$$\geq |C_{close}| \cdot \frac{1-\varepsilon}{1+\varepsilon} \cdot \left(\frac{4z}{\varepsilon}\right)^{2z} \cdot \frac{\text{cost}(C, c)}{|C|} \geq \left(\frac{4z}{\varepsilon}\right)^{2z-1} \cdot \text{cost}(C, c) \qquad (34)$$

which yields for any $C \in G_{far,\mathcal{S}}$.

$$
\begin{aligned}
& \mathrm{cost}(C \cap G, \mathcal{S}) = \sum_{C} \sum_{p \in C \cap G} \mathrm{cost}(p, \mathcal{S}) \\
(\textit{Lemma } 10) \quad & \leq \sum_{p \in C \cap G} (1+\varepsilon) \cdot \mathrm{cost}(c, \mathcal{S}) + \left(\frac{2z+\varepsilon}{\varepsilon}\right)^{z-1} \cdot \mathrm{cost}(p, c) \\
& \leq |C \cap G| \cdot (1+\varepsilon) \cdot \mathrm{cost}(c, \mathcal{S}) + \left(\frac{2z+\varepsilon}{\varepsilon}\right)^{z-1} \cdot \mathrm{cost}(C \cap G, c) \\
(\text{Markov}) \quad & \leq (1+\varepsilon) \cdot \left(\frac{\varepsilon}{2z}\right)^{2z} \cdot |C| \cdot \mathrm{cost}(c, \mathcal{S}) + \left(\frac{2z+\varepsilon}{\varepsilon}\right)^{z-1} \cdot \mathrm{cost}(C \cap G, c) \qquad (35) \\
(\text{Markov}) \quad & \leq \frac{1+\varepsilon}{1-\varepsilon} \cdot \left(\frac{\varepsilon}{2z}\right)^{2z} \cdot |C_{close}| \cdot \mathrm{cost}(c, \mathcal{S}) + \left(\frac{2z+\varepsilon}{\varepsilon}\right)^{z-1} \cdot \mathrm{cost}(C \cap G, c)
\end{aligned}
$$

$$
\begin{aligned}
(\textit{Eq. } 33) \quad & \leq \frac{(1+\varepsilon)^2}{(1-\varepsilon)^2} \cdot \left(\frac{\varepsilon}{2z}\right)^{2z} \cdot \mathrm{cost}(C, \mathcal{S}) + \left(\frac{2z+\varepsilon}{\varepsilon}\right)^{z-1} \cdot \mathrm{cost}(C \cap G, c) \\
(\textit{Eq. } 34) \quad & \leq \frac{(1+\varepsilon)^2}{(1-\varepsilon)^2} \cdot \left(\frac{\varepsilon}{2z}\right)^{2z} \cdot \mathrm{cost}(C, \mathcal{S}) + \left(\frac{2z+\varepsilon}{\varepsilon}\right)^{z-1} \cdot \left(\frac{\varepsilon}{4z}\right)^{2z-1} \cdot \mathrm{cost}(C, \mathcal{S}) \qquad (36) \\
& \leq \varepsilon \cdot \mathrm{cost}(C, \mathcal{S}) \qquad (37)
\end{aligned}
$$

What is left to show is that the weighted cost of the points in $G_{far,\mathcal{S}} \cap \Omega$ can be bounded similarly. For that, we use event $\mathcal{E}_{far,G}$ to show that $\sum_{p \in G_{far,\mathcal{S}} \cap C \cap \Omega} \frac{\mathrm{cost}(G, \mathcal{A}_0)}{\mathrm{cost}(p, \mathcal{A}_0)} \approx |G_{far,\mathcal{S}} \cap C|$. We have for all clusters C induced by $\mathcal{A}$

$$
\begin{aligned}
\sum_{p \in C \cap G \cap \Omega} \frac{\mathrm{cost}(G, \mathcal{A})}{\delta \cdot \mathrm{cost}(p, \mathcal{A})} \cdot \left(\frac{2z}{\varepsilon}\right)^{2z} \cdot \frac{\mathrm{cost}(C, \mathcal{A})}{|C|} & \leq \sum_{p \in C \cap G \cap \Omega} \frac{\mathrm{cost}(G, \mathcal{A})}{\delta \cdot \mathrm{cost}(p, \mathcal{A})} \mathrm{cost}(p, \mathcal{A}) \\
& \leq (1+\varepsilon) \cdot \mathrm{cost}(C \cap G, \mathcal{A}) \\
\Rightarrow \sum_{p \in C \cap G \cap \Omega} \frac{\mathrm{cost}(G_j, \mathcal{A})}{\delta \cdot \mathrm{cost}(p, \mathcal{A})} & \leq (1+\varepsilon) \cdot \left(\frac{\varepsilon}{2z}\right)^{2z} \cdot |C| \frac{\mathrm{cost}(C \cap G, \mathcal{A})}{\mathrm{cost}(C, \mathcal{A})} \\
& \leq (1+\varepsilon) \cdot \left(\frac{\varepsilon}{2z}\right)^{2z} \cdot |C| \qquad (38)
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
& \operatorname{cost}(G_{far,\mathcal{S}} \cap \Omega \cap C, \mathcal{S}) = \sum_{p \in G_{far,\mathcal{S}} \cap C} \frac{\operatorname{cost}(G, \mathcal{A})}{\delta \cdot \operatorname{cost}(p, \mathcal{A})} \cdot \operatorname{cost}(p, \mathcal{S}) \\
(\textit{Lemma } 10) \quad \leq \quad & \sum_{p \in G_{far,\mathcal{S}} \cap \Omega \cap C} \frac{\operatorname{cost}(G, \mathcal{A})}{\delta \cdot \operatorname{cost}(p, \mathcal{A})} \cdot \left((1+\varepsilon) \cdot \operatorname{cost}(c, \mathcal{S}) + \left(\frac{2z+\varepsilon}{\varepsilon} \right)^{z-1} \cdot \operatorname{cost}(p, c) \right) \\
\leq \quad & (1+\varepsilon) \cdot \operatorname{cost}(c, \mathcal{S}) \cdot \sum_{p \in G_{far,\mathcal{S}} \cap \Omega \cap C} \frac{\operatorname{cost}(G, \mathcal{A})}{\delta \cdot \operatorname{cost}(p, \mathcal{A})} \\
(\mathcal{E}_{far,G}) \quad & + \left(\frac{2z+\varepsilon}{\varepsilon} \right)^{z-1} \cdot (1+\varepsilon) \cdot \operatorname{cost}(C \cap G, \mathcal{A}) \\
(\textit{Eq. } 38) \quad \leq \quad & (1+\varepsilon)^2 \cdot \operatorname{cost}(c, \mathcal{S}) \cdot \left(\frac{\varepsilon}{2z} \right)^{2z} \cdot |C| + \left(\frac{2z+\varepsilon}{\varepsilon} \right)^{z-1} \cdot (1+\varepsilon) \cdot \operatorname{cost}(C \cap G, \mathcal{A}) \\
\leq \quad & (1+\varepsilon)^2 \cdot \left(\frac{\varepsilon}{2z} \right)^{2z} \cdot |C| \cdot \operatorname{cost}(c, \mathcal{S}) + \left(\frac{2z+\varepsilon}{\varepsilon} \right)^{z-1} \cdot \operatorname{cost}(C, c) \qquad (39) \\
\leq \quad & \varepsilon \cdot \operatorname{cost}(C, \mathcal{S})
\end{aligned}
$$

where the steps following Equation 39 are identical to those used to derive Equation 37 from Equation 35. Summing up Equations 37 and 39 and rescaling ε by a factor 2 yields the claim. □

Proof of Lemma 17. Similar to the proof of Lemma 15, we bound the expectation when conditioning on $\mathcal{E}_{far,G}$ and when $\mathcal{E}_{far,G}$ fails to hold:

$$
\begin{aligned}
& \mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\left| \frac{\sum_{p \in \Omega} w_p \cdot u_p^{G,\mathcal{S}} - \|u^{G,\mathcal{S}}\|_1}{\operatorname{cost}(P^G, \mathcal{A}) + \operatorname{cost}(P^G, \mathcal{S})} \right| \right] \\
= \quad & \mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\left| \frac{\sum_{p \in \Omega} w_p \cdot u_p^{G,\mathcal{S}} - \|u^{G,\mathcal{S}}\|_1}{\operatorname{cost}(P^G, \mathcal{A}) + \operatorname{cost}(P^G, \mathcal{S})} \right| \mathcal{E}_{far,G} \right] \cdot \mathbb{P}\left[\mathcal{E}_{far,G}\right] \qquad (40) \\
& + \mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\left| \frac{\sum_{p \in \Omega} w_p \cdot u_p^{G,\mathcal{S}} - \|u^{G,\mathcal{S}}\|_1}{\operatorname{cost}(P^G, \mathcal{A}) + \operatorname{cost}(P^G, \mathcal{S})} \right| \overline{\mathcal{E}_{far,G}} \right] \cdot \mathbb{P}\left[\overline{\mathcal{E}_{far,G}}\right] \qquad (41)
\end{aligned}
$$

For term 40, Lemma 28 states that

$$
\begin{aligned}
\sum_{p \in \Omega} w_p \cdot u_p^{G,\mathcal{S}} + \|u^{G,\mathcal{S}}\|_1 \quad &= \quad \sum_{C \in F_{G,\mathcal{S}}} \sum_{p \in \Omega \cap G} w_p \cdot u_p^{G,\mathcal{S}} + \sum_{p \in C \cap G} \operatorname{cost}(p, \mathcal{S}) \\
&\leq \quad \sum_{C \in F_{G,\mathcal{S}}} \varepsilon \cdot \operatorname{cost}(C \cap G, \mathcal{S}) = \varepsilon \cdot \operatorname{cost}(P^G, \mathcal{S}). \qquad (42)
\end{aligned}
$$

We now consider term 41. If $\|u^{G,S}\|_1 > \sum_{p \in \Omega} w_p \cdot u_p^{G,\mathcal{S}}$, we can bound $\frac{\sum_{p \in \Omega} w_p \cdot u_p^{G,\mathcal{S}} - \|u^{G,\mathcal{S}}\|_1}{\operatorname{cost}(P^G, \mathcal{A}) + \operatorname{cost}(P^G, \mathcal{S})}$ by 1. Otherwise, let $r_C = \max_{p \in C \cap G} \frac{\operatorname{cost}(p, \mathcal{S})}{\operatorname{cost}(p, \mathcal{A}} > 4^z$ and let $p' = \operatorname{argmax}_{p \in C \cap G} \frac{\operatorname{cost}(p, \mathcal{S})}{\operatorname{cost}(p, \mathcal{A}}$. We have $\operatorname{dist}(c, \mathcal{S}) \geq$

$\mathrm{dist}(p', \mathcal{S}) - \mathrm{dist}(p', c) \geq (r_C^{1/z} - 1) \cdot \mathrm{dist}(p', c)$, which implies $\frac{\mathrm{cost}(c,\mathcal{S})}{\mathrm{cost}(p',c)} \cdot 2^z \geq r_C$. Therefore

$$\begin{aligned}
\sum_{p\in\Omega} w_p u_p^{G,\mathcal{S}} &= \sum_C \sum_{p\in\Omega\cap C} \frac{\mathrm{cost}(G,\mathcal{A})}{\delta \cdot \mathrm{cost}(p,\mathcal{A})} \cdot \mathrm{cost}(p,\mathcal{S}) \\
(Eq.\ 31) \quad &\leq 4k \cdot \sum_C \max_{p\in C\cap G} \mathrm{cost}(C\cap G,\mathcal{A}) \cdot r_C \\
&\leq 4k \cdot \sum_C \max_{p\in C\cap G} \mathrm{cost}(C\cap G,\mathcal{A}) \cdot 2^z \cdot \frac{\mathrm{cost}(c,\mathcal{S})}{\mathrm{cost}(p',\mathcal{A})} \\
&\leq 2^{z+2}k \cdot \sum_C \max_{p\in C\cap G} \frac{\mathrm{cost}(C,\mathcal{A})}{\mathrm{cost}(p',\mathcal{A})} \cdot \mathrm{cost}(c,\mathcal{S}) \\
(Markov) \quad &\leq 2^{z+2}k \cdot \sum_C \left(\frac{\varepsilon}{4z}\right)^{2z} |C| \cdot \mathrm{cost}(c,\mathcal{S}) \\
(Lemma\ 10) \quad &\leq 2^{2z+2}k \cdot \sum_C (\mathrm{cost}(C,\mathcal{A}) + \mathrm{cost}(C,\mathcal{S})) \\
&\leq 2^{2z+2}k \cdot (\mathrm{cost}(P^G,\mathcal{A}) + \mathrm{cost}(P^G,\mathcal{S}))
\end{aligned}$$

With this, we may bound the ratio $\frac{\sum_{p\in\Omega} w_p \cdot u_p^{G,\mathcal{S}} - \|u^{G,\mathcal{S}}\|_1}{\mathrm{cost}(P^G,\mathcal{A})+\mathrm{cost}(P^G,\mathcal{S})}$ by $2^{2+2}k$. The probability of $\overline{\mathcal{E}_{far,G}}$ is at most $k \cdot \exp\left(-\frac{\varepsilon^2}{5\cdot k} \cdot \delta\right)$ due to Lemma 27. Therefore setting $\delta > 5k \cdot \log \frac{k^2}{2^{2z+2}\cdot\varepsilon}$ yields

$$\mathbb{E}_\Omega \sup_{\mathcal{S}} \left[\left| \frac{\sum_{p\in\Omega} w_p \cdot u_p^{G,\mathcal{S}} - \|u^{G,\mathcal{S}}\|_1}{\mathrm{cost}(P^G,\mathcal{A}) + \mathrm{cost}(P^G,\mathcal{S})} \right| \overline{\mathcal{E}_{far,G}} \right] \cdot \mathbb{P}\left[\overline{\mathcal{E}_{far,G}}\right] \leq 2^{2z+2}k \cdot \frac{\varepsilon}{2^{2z+2} \cdot k} \leq \varepsilon.$$

Summing this with Equation 42 and rescaling ε by a factor of 2 yields the claim. □

References

[1] Improved approximations for euclidean k-means and k-median, via nested quasi-independent sets.

[2] Pankaj K. Agarwal, Sariel Har-Peled, and Kasturi R. Varadarajan. Approximating extent measures of points. *J. ACM*, 51(4):606–635, 2004.

[3] Pankaj K Agarwal, Sariel Har-Peled, and Kasturi R Varadarajan. Geometric approximation via coresets. *Combinatorial and computational geometry*, 52:1–30, 2005.

[4] Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, and Justin Ward. Better guarantees for k-means and euclidean k-median by primal-dual algorithms. *SIAM J. Comput.*, 49(4), 2020.

[5] Olivier Bachem, Mario Lucic, and Silvio Lattanzi. One-shot coresets: The case of k-clustering. In Amos J. Storkey and Fernando Pérez-Cruz, editors, *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, volume 84 of *Proceedings of Machine Learning Research*, pages 784–792. PMLR, 2018.

[6] Daniel Baker, Vladimir Braverman, Lingxiao Huang, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. 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*, pages 569–579. PMLR, 2020.

[7] Sayan Bandyapadhyay, Fedor V. Fomin, and Kirill Simonov. On coresets for fair clustering in metric and euclidean spaces and their applications. *CoRR*, abs/2007.10137, 2020.

[8] Luca Becchetti, Marc Bury, Vincent Cohen-Addad, Fabrizio Grandoni, and Chris Schwiegelshohn. Oblivious dimension reduction for k-means: beyond subspaces and the johnson-lindenstrauss lemma. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019*, pages 1039–1050, 2019.

[9] Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near-optimal coresets for least-squares regression. *IEEE Trans. Inf. Theory*, 59(10):6880–6892, 2013.

[10] Christos Boutsidis, Michael W. Mahoney, and Petros Drineas. Unsupervised feature selection for the k-means clustering problem. In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada.*, pages 153–161, 2009.

[11] Christos Boutsidis, Anastasios Zouzias, and Petros Drineas. Random projections for k-means clustering. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada.*, pages 298–306, 2010.

[12] Christos Boutsidis, Anastasios Zouzias, Michael W. Mahoney, and Petros Drineas. Randomized dimensionality reduction for k-means clustering. *IEEE Trans. Information Theory*, 61(2):1045–1062, 2015.

[13] Vladimir Braverman, Dan Feldman, Harry Lang, and Daniela Rus. Streaming coreset constructions for m-estimators. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2019, September 20-22, 2019, Massachusetts Institute of Technology, Cambridge, MA, USA*, pages 62:1–62:15, 2019.

[14] Vladimir Braverman, Gereon Frahling, Harry Lang, Christian Sohler, and Lin F. Yang. Clustering high dimensional dynamic data streams. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 576–585, 2017.

[15] Vladimir Braverman, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for ordered weighted clustering. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 744–753, 2019.

[16] Vladimir Braverman, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for ordered weighted clustering. *CoRR*, abs/1903.04351, 2019.

[17] Vladimir Braverman, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for clustering in excluded-minor graphs and beyond. In Dániel Marx, editor, *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021, Virtual Conference, January 10 - 13, 2021*, pages 2679–2696. SIAM, 2021.

[18] Vladimir Braverman, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for clustering in excluded-minor graphs and beyond. In Dániel Marx, editor, *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021, Virtual Conference, January 10 - 13, 2021*, pages 2679–2696. SIAM, 2021.

[19] Vladimir Braverman, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for clustering with missing values. *CoRR*, abs/2106.16112, 2021.

[20] Vladimir Braverman, Harry Lang, Keith Levin, and Morteza Monemizadeh. Clustering problems on sliding windows. In Robert Krauthgamer, editor, *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pages 1374–1390. SIAM, 2016.

[21] Jaroslaw Byrka, Thomas Pensyl, Bartosz Rybicki, Aravind Srinivasan, and Khoa Trinh. An improved approximation for k-median, and positive correlation in budgeted optimization. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, pages 737–756, 2015.

[22] Timothy M. Chan. Dynamic coresets. *Discret. Comput. Geom.*, 42(3):469–488, 2009.

[23] Ke Chen. On coresets for k-median and k-means clustering in metric and Euclidean spaces and their applications. *SIAM J. Comput.*, 39(3):923–947, 2009.

[24] Yeshwanth Cherapanamjeri and Jelani Nelson. Terminal embeddings in sublinear time. *CoRR*, abs/2110.08691, 2021.

[25] Michael B. Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 163–172, 2015.

[26] Vincent Cohen-Addad. A fast approximation scheme for low-dimensional k-means. In Artur Czumaj, editor, *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018*, pages 430–440. SIAM, 2018.

[27] Vincent Cohen-Addad, Andreas Emil Feldmann, and David Saulpic. Near-linear time approximation schemes for clustering in doubling metrics. *J. ACM*, 68(6):44:1–44:34, 2021.

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

[29] Vincent Cohen-Addad and Jason Li. On the fixed-parameter tractability of capacitated clustering. In *46th International Colloquium on Automata, Languages, and Programming, ICALP 2019, July 9-12, 2019, Patras, Greece*, pages 41:1–41:14, 2019.

[30] Vincent Cohen-Addad, Marcin Pilipczuk, and Michal Pilipczuk. Efficient approximation schemes for uniform-cost clustering problems in planar graphs. In Michael A. Bender, Ola Svensson, and Grzegorz Herman, editors, *27th Annual European Symposium on Algorithms, ESA 2019, September 9-11, 2019, Munich/Garching, Germany*, volume 144 of *LIPIcs*, pages 33:1–33:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019.

[31] Vincent Cohen-Addad and Karthik C. S. Inapproximability of clustering in lp metrics. In David Zuckerman, editor, *60th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2019, Baltimore, Maryland, USA, November 9-12, 2019*, pages 519–539. IEEE Computer Society, 2019.

[32] Vincent Cohen-Addad, Karthik C. S., and Euiwoong Lee. Johnson coverage hypothesis: Inapproximability of k-means and k-median in lp-metrics. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1493–1530.

[33] Vincent Cohen-Addad, Karthik C. S., and Euiwoong Lee. On approximability of clustering problems without candidate centers. In Dániel Marx, editor, *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021, Virtual Conference, January 10 - 13, 2021*, pages 2635–2648. SIAM, 2021.

[34] Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. Improved coresets and sublinear algorithms for power means in euclidean spaces. In Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jenn Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 7-10, 2021, Virtual Conference*, 2021.

[35] Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. A new coreset framework for clustering. In Samir Khuller and Virginia Vassilevska Williams, editors, *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*. ACM, 2021.

[36] Vincent Cohen-Addad and Chris Schwiegelshohn. On the local structure of stable clustering instances. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 49–60, 2017.

[37] Petros Drineas, Alan M. Frieze, Ravi Kannan, Santosh Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Machine Learning*, 56(1-3):9–33, 2004.

[38] Michael Elkin, Arnold Filtser, and Ofer Neiman. Terminal embeddings. *Theor. Comput. Sci.*, 697:1–36, 2017.

[39] Xiequan Fan, Ion Grama, and Quansheng Liu. Sharp large deviation results for sums of independent random variables. *Science China Mathematics*, 58(9):1939–1958, 2015.

[40] Dan Feldman. Core-sets: An updated survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 10(1), 2020.

[41] Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pages 569–578, 2011.

[42] Dan Feldman, Morteza Monemizadeh, Christian Sohler, and David P. Woodruff. Coresets and sketches for high dimensional subspace approximation problems. In Moses Charikar, editor, *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*, pages 630–649. SIAM, 2010.

[43] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca, and projective clustering. *SIAM J. Comput.*, 49(3):601–657, 2020.

[44] Zhili Feng, Praneeth Kacham, and David P. Woodruff. Strong coresets for subspace approximation and k-median in nearly linear time. *CoRR*, abs/1912.12003, 2019.

[45] Hendrik Fichtenberger, Marc Gillé, Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. BICO: BIRCH meets coresets for k-means clustering. In *Algorithms - ESA 2013 - 21st Annual European Symposium, Sophia Antipolis, France, September 2-4, 2013. Proceedings*, pages 481–492, 2013.

[46] Gereon Frahling and Christian Sohler. Coresets in dynamic geometric data streams. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing (STOC)*, pages 209–217, 2005.

[47] Fabrizio Grandoni, Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Rakesh Venkat. A refined approximation for euclidean k-means. *Inf. Process. Lett.*, 176:106251, 2022.

[48] Sudipto Guha and Samir Khuller. Greedy strikes back: Improved facility location algorithms. *J. Algorithms*, 31(1):228–248, 1999.

[49] Sariel Har-Peled and Akash Kushal. Smaller coresets for k-median and k-means clustering. *Discrete & Computational Geometry*, 37(1):3–19, 2007.

[50] Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing, Chicago, IL, USA, June 13-16, 2004*, pages 291–300, 2004.

[51] Lingxiao Huang, Shaofeng H.-C. Jiang, Jian Li, and Xuan Wu. 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*, pages 814–825, 2018.

[52] Lingxiao Huang, Shaofeng H.-C. Jiang, and Nisheeth K. Vishnoi. Coresets for clustering with fairness constraints. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 7587–7598, 2019.

[53] Lingxiao Huang, K. Sudhir, and Nisheeth K. Vishnoi. Coresets for regressions with panel data. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[54] Lingxiao Huang, K. Sudhir, and Nisheeth K. Vishnoi. Coresets for time series clustering, 2021.

[55] Lingxiao Huang and Nisheeth K. Vishnoi. Coresets for clustering in euclidean spaces: importance sampling is nearly optimal. In Konstantin Makarychev, Yury Makarychev, Madhur Tulsiani, Gautam Kamath, and Julia Chuzhoy, editors, *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, pages 1416–1429. ACM, 2020.

[56] Jonathan Huggins, Trevor Campbell, and Tamara Broderick. Coresets for scalable bayesian logistic regression. In *Advances in Neural Information Processing Systems*, pages 4080–4088, 2016.

[57] Piotr Indyk, Sepideh Mahabadi, Shayan Oveis Gharan, and Alireza Rezaei. Composable coresets for determinant maximization problems via spectral spanners. In Shuchi Chawla, editor, *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020*, pages 1675–1694. SIAM, 2020.

[58] Piotr Indyk, Sepideh Mahabadi, Mohammad Mahdian, and Vahab S. Mirrokni. Composable core-sets for diversity and coverage maximization. In Richard Hull and Martin Grohe, editors, *Proceedings of the 33rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS'14, Snowbird, UT, USA, June 22-27, 2014*, pages 100–108. ACM, 2014.

[59] Shaofeng H.-C. Jiang, Robert Krauthgamer, Jianing Lou, and Yubo Zhang. Coresets for kernel clustering. *CoRR*, abs/2110.02898, 2021.

[60] Ibrahim Jubran, Ernesto Evgeniy Sanches Shayda, Ilan Newman, and Dan Feldman. Coresets for decision trees of signals. *CoRR*, abs/2110.03195, 2021.

[61] Ibrahim Jubran, Murad Tukan, Alaa Maalouf, and Dan Feldman. Sets clustering. 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*, pages 4994–5005. PMLR, 2020.

[62] Zohar S. Karnin and Edo Liberty. Discrepancy, coresets, and sketches in machine learning. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 1975–1993. PMLR, 2019.

[63] Stavros G. Kolliopoulos and Satish Rao. A nearly linear-time approximation scheme for the euclidean k-median problem. *SIAM J. Comput.*, 37(3):757–782, June 2007.

[64] Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, pages 299–308, 2010.

[65] Michael Langberg and Leonard J. Schulman. Universal ε-approximators for integrals. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*, pages 598–607, 2010.

[66] Kasper Green Larsen and Jelani Nelson. Optimality of the Johnson-Lindenstrauss Lemma. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 633–638, 2017.

[67] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes.* Springer Science & Business Media, 2013.

[68] Euiwoong Lee, Melanie Schmidt, and John Wright. Improved and simplified inapproximability for k-means. *Inf. Process. Lett.*, 120:40–43, 2017.

[69] Shi Li and Ola Svensson. Approximating k-median via pseudo-approximation. *SIAM J. Comput.*, 45(2):530–547, 2016.

[70] Mario Lucic, Matthew Faulkner, Andreas Krause, and Dan Feldman. Training gaussian mixture models at scale via coresets. *J. Mach. Learn. Res.*, 18:160:1–160:25, 2017.

[71] Alaa Maalouf, Ibrahim Jubran, and Dan Feldman. Fast and accurate least-mean-squares solvers. In *Advances in Neural Information Processing Systems*, pages 8307–8318, 2019.

[72] Sepideh Mahabadi, Konstantin Makarychev, Yury Makarychev, and Ilya P. Razenshteyn. Nonlinear dimension reduction via outer bi-lipschitz extensions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 1088–1101, 2018.

[73] Konstantin Makarychev, Yury Makarychev, and Ilya P. Razenshteyn. Performance of johnson-lindenstrauss transform for k-means and k-medians clustering. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019*, pages 1027–1038, 2019.

[74] Konstantin Makarychev, Yury Makarychev, Maxim Sviridenko, and Justin Ward. A bi-criteria approximation algorithm for k-means. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2016, September 7-9, 2016, Paris, France*, pages 14:1–14:20, 2016.

[75] Dániel Marx and Michal Pilipczuk. Optimal parameterized algorithms for planar facility location problems using voronoi diagrams. In Nikhil Bansal and Irene Finocchi, editors, *Algorithms - ESA 2015 - 23rd Annual European Symposium, Patras, Greece, September 14-16, 2015, Proceedings*, volume 9294 of *Lecture Notes in Computer Science*, pages 865–877. Springer, 2015.

[76] Pascal Massart. Concentration inequalities and model selection. 2007.

[77] Nimrod Megiddo and Kenneth J. Supowit. On the complexity of some common geometric location problems. *SIAM J. Comput.*, 13(1):182–196, 1984.

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

[79] Alejandro Molina, Alexander Munteanu, and Kristian Kersting. Core dependency networks. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *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 (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3820–3827. AAAI Press, 2018.

[80] Alexander Munteanu and Chris Schwiegelshohn. Coresets-methods and history: A theoreticians design pattern for approximation and streaming algorithms. *Künstliche Intell.*, 32(1):37–53, 2018.

[81] Alexander Munteanu, Chris Schwiegelshohn, Christian Sohler, and David P. Woodruff. On coresets for logistic regression. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 6562–6571, 2018.

[82] Shyam Narayanan and Jelani Nelson. Optimal terminal dimensionality reduction in euclidean space. In Moses Charikar and Edith Cohen, editors, *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019*, pages 1064–1069. ACM, 2019.

[83] Jeff M. Phillips and Wai Ming Tai. Near-optimal coresets of kernel density estimates. *Discret. Comput. Geom.*, 63(4):867–887, 2020.

[84] Atri Rudra and Mary Wootters. Every list-decodable code for high noise has abundant near-optimal rate puncturings. In David B. Shmoys, editor, *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 764–773. ACM, 2014.

[85] Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. Fair coresets and streaming algorithms for fair k-means. In *Approximation and Online Algorithms - 17th International Workshop, WAOA 2019, Munich, Germany, September 12-13, 2019, Revised Selected Papers*, pages 232–251, 2019.

[86] Christian Sohler and David P. Woodruff. 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*, pages 802–813, 2018.

[87] Michel Talagrand et al. Majorizing measures: the generic chaining. *The Annals of Probability*, 24(3):1049–1103, 1996.

[88] Murad Tukan, Alaa Maalouf, and Dan Feldman. Coresets for near-convex functions. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[89] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok, editors, *Compressed Sensing*, pages 210–268. Cambridge University Press, 2012.

A Lower bound for Arbitrary Powers in Euclidean Spaces

In this section, we generalize the lower bound to arbitrary powers $z \neq 2$. The proof follows exactly the same steps as for $z = 2$, except that we make use of the following observation to handle $z \neq 2$:

Observation 1. *For any $0 < a \leq 1$ any $b > 0$ and any $x \in [0, b]$ we have $b^a(1 - x/b) \leq (b - x)^a \leq b^a(1 - xa/b)$. For any $1 \leq a$ any $b > 0$ and any $x \in [0, b]$ we have $b^a(1 - xa/b) \leq (b - x)^a \leq b^a(1 - x/b)$.*

Proof. For any $0 < a \leq 1$ any $b > 0$ and any $x \in [0, b]$, we have $(b-x)^a = b^a(1-x/b)^a = b^a \exp\left(-a\sum_{n=1}^{\infty}(x/b)^n/n\right)$. Since $a \leq 1$, this is at most $b^a \exp\left(-\sum_{n=1}^{\infty}(xa/b)^n/n\right) = b^a(1-xa/b)$. Also, since $0 \leq 1-x/b \leq 1$, it holds for any $0 < a \leq 1$ that $(1-x/b)^a \geq 1-x/b$.

For any $1 \leq a$ any $b > 0$ and any $x \in [0, b]$, we have $(b-x)^a = b^a(1-x/b)^a = b^a \exp\left(-a\sum_{n=1}^{\infty}(x/b)^n/n\right)$. Since $a \geq 1$, this is at least $b^a \exp\left(-\sum_{n=1}^{\infty}(xa/b)^n/n\right) = b^a(1-xa/b)$. Also, since $0 \leq 1-x/b \leq 1$, it holds for any $1 \leq a$ that $(1-x/b)^a \leq 1-x/b$. □

The first step of our proof is again to argue that for any coreset using few points, there is a "cheap" clustering using a single center of unit norm:

Lemma 29. *Let $r_1, \ldots, r_\ell \in \mathbb{R}^{2d}$ and let $w_1, \ldots, w_\ell \in \mathbb{R}^+$. There exists a unit vector v such that $\sum_{i=1}^{\ell} w_i \min_{\xi \in \{-1,1\}} \|r_i - \xi v\|_2^z \leq \sum_{i=1}^{\ell} w_i(\|r_i\|_2^2+1)^{z/2} - 2\min\{1, z/2\}\frac{\sum_{i=1}^{\ell} w_i(\|r_i\|_2^2+1)^{z/2-1}\|r_i\|_2}{\sqrt{\ell}}$.*

Proof. Consider the random vector $u = \sum_{i=1}^{\ell} w_i(\|r_i\|_2^2+1)^{z/2-1}\sigma_i r_i$ where the σ_i are i.i.d. uniform Rademachers. We see that

$$\begin{aligned}
\sum_{i=1}^{\ell} w_i(\|r_i\|_2^2+1)^{z/2-1}|\langle r_i, u\rangle| &= \sum_{i=1}^{\ell} w_i(\|r_i\|_2^2+1)^{z/2-1}\left|\sum_{j=1}^{\ell} w_j(\|r_j\|_2^2+1)^{z/2-1}\sigma_j\langle r_i, r_j\rangle\right| \\
&= \sum_{i=1}^{\ell} w_i(\|r_i\|_2^2+1)^{z/2-1}\left|\sum_{j=1}^{\ell} w_j(\|r_j\|_2^2+1)^{z/2-1}\sigma_i\sigma_j\langle r_i, r_j\rangle\right| \\
&\geq \sum_{i=1}^{\ell} w_i(\|r_i\|_2^2+1)^{z/2-1}\sum_{j=1}^{\ell} w_j(\|r_j\|_2^2+1)^{z/2-1}\sigma_i\sigma_j\langle r_i, r_j\rangle \\
&= \|u\|_2^2.
\end{aligned}$$

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

$$\sum_{i=1}^{\ell} w_i(\|r_i\|_2^2+1)^{z/2-1}|\langle r_i, v\rangle| \geq \|u\|_2.$$

Since $\mathbb{E}[\|u\|_2^2] = \sum_{i=1}^{\ell} w_i^2(\|r_i\|_2^2+1)^{z-2}\|r_i\|_2^2$ we conclude that there must exist a unit vector v with

$$\sum_{i=1}^{\ell} w_i(\|r_i\|_2^2+1)^{z/2-1}|\langle r_i, v\rangle| \geq \sqrt{\sum_{i=1}^{\ell} w_i^2(\|r_i\|_2^2+1)^{z-2}\|r_i\|_2^2}.$$

By Cauchy-Schwartz, we have:

$$\sum_{i=1}^{\ell} |1 \cdot w_i(\|r_i\|_2^2+1)^{z/2-1}\|r_i\|_2| \leq \sqrt{\sum_{i=1}^{\ell} w_i^2(\|r_i\|_2^2+1)^{z-2}\|r_i\|_2^2} \cdot \sqrt{\sum_{i=1}^{\ell} 1} = \sqrt{\sum_{i=1}^{\ell} w_i^2(\|r_i\|_2^2+1)^{z-2}\|r_i\|_2^2} \cdot \sqrt{\ell}$$

which finally implies

$$\sum_{i=1}^{\ell} w_i(\|r_i\|_2^2+1)^{z/2-1}|\langle r_i, v\rangle| \geq \frac{\sum_{i=1}^{\ell} w_i(\|r_i\|_2^2+1)^{z/2-1}\|r_i\|_2}{\sqrt{\ell}}.$$

For that unit vector v, consider $\sum_{i=1}^{\ell} w_i \min_{\xi\in\{-1,1\}} \|r_i - \xi v\|_2^z$:

$$\begin{aligned}\sum_{i=1}^{\ell} w_i \min_{\xi\in\{-1,1\}} \|r_i - \xi v\|_2^z &= \sum_{i=1}^{\ell} w_i(\|r_i\|_2^2 + \|v\|_2^2 - 2|\langle r_i, v\rangle|)^{z/2} \\ &= \sum_{i=1}^{\ell} w_i(\|r_i\|_2^2 + 1 - 2|\langle r_i, v\rangle|)^{z/2}.\end{aligned}$$

By Observation 1, this is at most:

$$\begin{aligned}&\leq \sum_{i=1}^{\ell} w_i(\|r_i\|_2^2+1)^{z/2}\left(1 - \frac{2\min\{1, z/2\}|\langle r_i, v\rangle|}{\|r_i\|_2^2+1}\right) \\ &\leq \sum_{i=1}^{\ell} w_i(\|r_i\|_2^2+1)^{z/2} - \sum_{i=1}^{\ell} w_i 2\min\{1, z/2\}|\langle r_i, v\rangle|(\|r_i\|_2^2+1)^{z/2-1} \\ &\leq \sum_{i=1}^{\ell} w_i(\|r_i\|_2^2+1)^{z/2} - 2\min\{1, z/2\}\frac{\sum_{i=1}^{\ell} w_i(\|r_i\|_2^2+1)^{z/2-1}\|r_i\|_2}{\sqrt{\ell}}.\end{aligned}$$

□

We now extend this to create a cheap clustering using k centers of unit norm:

Lemma 30. *Let $r_1, \ldots, r_t \in \mathbb{R}^{2d}$ and let $w_1, \ldots, w_t \in \mathbb{R}^+$. There exists a set of k unit vectors $v_1, \ldots, v_k$ such that*

$$\sum_{i=1}^{t} w_i \min_{j=1}^{k} \|r_i - v_j\|_2^z \leq \sum_{i=1}^{t} w_i(\|r_i\|_2^2+1)^{z/2} - \min\{1, z/2\}\sqrt{2k/t}\sum_{i=1}^{t} w_i(\|r_i\|_2^2+1)^{z/2-1}\|r_i\|_2.$$

and moreover, for every v_j, there is a v_i such that $v_j = -v_i$.

Proof. Partition $r_1, \ldots, r_t$ arbitrarily into $k/2$ disjoint groups $G_1, \ldots, G_{k/2}$ of at most $2t/k$ vectors each. For each group G_j, apply Lemma 29 to find a unit vector u_j with

$$\sum_{r_i\in G_j}^{\ell} w_i \min_{\xi\in\{-1,1\}} \|r_i - \xi u_j\|_2^z \leq \sum_{r_i\in G_j} w_i(\|r_i\|_2^2+1)^{z/2} - 2\min\{1, z/2\}\frac{\sum_{r_i\in G_j} w_i(\|r_i\|_2^2+1)^{z/2-1}\|r_i\|_2}{\sqrt{2t/k}}.$$

Let $v_{2j-1} = u_j$ and $v_{2j} = -u_j$. Since we always add both u_j and $-u_j$ we conclude:

$$\begin{aligned}\sum_{i=1}^{t} w_i \min_{j=1}^{k} \|r_i - v_j\|_2^z &\leq \\ \sum_{j=1}^{k/2}\sum_{r_i\in G_j} w_i \min_{\xi\in\{-1,1\}} \|r_i - \xi u_j\|_2^z &\leq \\ \sum_{i=1}^{t} w_i(\|r_i\|_2^2+1)^{z/2} - \min\{1, z/2\}\sqrt{2k/t}\sum_{i=1}^{t} w_i(\|r_i\|_2^2+1)^{z/2-1}\|r_i\|_2.\end{aligned}$$

□

We now use the orthogonality of the standard unit vectors $e_1, \ldots, e_d$ to argue that any clustering of them using unit norm centers must be expensive:

Lemma 31. *For any d, consider the point set $P = \{e_1, \ldots, e_d\}$ in $\mathbb{R}^{2d}$. For any set of k centers $c_1, \ldots, c_k \in \mathbb{R}^{2d}$, all with unit norm and satisfying that for every c_j there is an index i such that $c_j = -c_i$, it holds that $\sum_{i=1}^{d} \min_{j=1}^{k} \|e_i - c_j\|_2^z \geq 2^{z/2}d - 2^{z/2} \cdot \max\{1, z/2\} \cdot \sqrt{dk}$.*

Proof. We see that

$$\begin{aligned}\sum_{i=1}^{n} \min_{j=1}^{k} \|e_i - c_j\|_2^p &= \sum_{i=1}^{d} \min_{j=1}^{k} \left(\|e_i\|_2^2 + \|c_j\|_2^2 - 2\langle e_i, c_j\rangle\right)^{z/2} \\ &= \sum_{i=1}^{d} \left(2 - 2\max_{j=1}^{k} \langle e_i, c_j\rangle\right)^{z/2}.\end{aligned}$$

Since $c_1, \ldots, c_k$ satisfy that for every c_j there is an index h with $c_j = -c_h$, it holds that $\max_{j=1}^{k} \langle e_i, c_j\rangle \geq 0$ for every e_i. By Cauchy-Schwartz, we have $|\langle e_i, c_j\rangle| \leq 1$ hence by Observation 1, the above is at least:

$$2^{z/2}d - 2^{z/2} \cdot \max\{1, z/2\} \cdot \sum_{i=1}^{d} \max_{j=1}^{k} \langle e_i, c_j\rangle.$$

Now, for each c_j, define $\hat{c}_j$ to equal c_j, except that we set the i'th coordinate to 0 if $j \neq \operatorname{argmax}_h \langle e_i, c_h\rangle$ or $i > d$. Then:

$$\begin{aligned}2^{z/2}d - 2^{z/2} \cdot \max\{1, z/2\} \cdot \sum_{i=1}^{d} \max_{j=1}^{k} \langle e_i, c_j\rangle &= 2^{z/2}d - 2^{z/2} \cdot \max\{1, z/2\} \cdot \sum_{i=1}^{d} \sum_{j=1}^{k} \langle e_i, \hat{c}_j\rangle \\ &= 2^{z/2}d - 2^{z/2} \cdot \max\{1, z/2\} \cdot \sum_{i=1}^{d} \langle e_i, \sum_{j=1}^{k} \hat{c}_j\rangle \\ &\geq 2^{z/2}d - 2^{z/2} \cdot \max\{1, z/2\} \cdot \|\sum_{j=1}^{k} \hat{c}_j\|_1.\end{aligned}$$

By Cauchy-Schwartz, we have $\|\sum_{j=1}^{k} \hat{c}_j\|_1 \leq \|\sum_{j=1}^{k} \hat{c}_j\|_2 \cdot \sqrt{d}$. Since the $\hat{c}_j$'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 $\sum_{i=1}^{d} \min_{j=1}^{k} \|e_i - c_j\|_2^z \geq 2^{z/2}d - 2^{z/2} \cdot \max\{1, z/2\} \cdot \sqrt{dk}$. □

We also need a handle on the offset of any coreset. This is obtained by considering a clustering using a single center that is orthogonal to all points $e_1, \ldots, e_d$ and all points of a coreset:

Lemma 32. *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 ε-coreset for P, using offset Δ and with $t < d$. Then we must have $\Delta + \sum_{i=1}^{t} w_i(\|r_i\|_2^2 + 1)^{z/2} \in (1 \pm \varepsilon)2^{z/2}d$.*

Proof. Since $t + d < 2d$ there exists a unit vector v that is orthogonal to all r_i and all e_j. Consider placing all k centers at v. Then the cost of clustering P with these centers is $2^{z/2}d$. It therefore must hold that $\Delta + \sum_{i=1}^{t} w_i(\|r_i\|_2^2 + \|v\|_2^2 - 2\langle r_i, v\rangle)^{z/2} = \Delta + \sum_{i=1}^{t} w_i(\|r_i\|_2^2 + 1)^{z/2} \in (1 \pm \varepsilon)2^{z/2}d$. □

Lemma 33. *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 ε-coreset for P with $t < d$, using offset Δ. Then*

$$\sum_{i=1}^{t} w_i(\|r_i\|_2^2 + 1)^{z/2-1}\|r_i\|_2 \leq \frac{2\varepsilon 2^{z/2}d + \max\{1, z/2\}2^{z/2}\sqrt{dk}}{\sqrt{2} \cdot \min\{1, z/2\}} \cdot \sqrt{t/k}.$$

Proof. By Lemma 30, we can find k unit vectors $v_1, \ldots, v_k$ such that

$$\sum_{i=1}^{t} w_i \min_{j=1}^{k} \|r_i - v_j\|_2^z \leq \sum_{i=1}^{t} w_i(\|r_i\|_2^2 + 1)^{z/2} - \min\{1, z/2\}\sqrt{2k/t}\sum_{i=1}^{t} w_i(\|r_i\|_2^2 + 1)^{z/2-1}\|r_i\|_2.$$

Moreover, those vectors satisfy that for every v_j, there is an index i such that $v_j = -v_i$. By Lemma 31, it holds that $\sum_{p \in P} \min_{j=1}^{k} \|p - v_j\|_2^z \geq 2^{z/2}d - 2^{z/2} \cdot \max\{1, z/2\} \cdot \sqrt{dk}$. Since points $r_1, \ldots, r_t$ with respective weights $w_1, \ldots, w_t$ and offset Δ form an ε-coreset for P, it follows from Observation 1 that we must have

$$\begin{aligned}(1-\varepsilon)2^{z/2}(d - \max\{1, z/2\} \cdot \sqrt{dk}) &\leq \\ \Delta + \sum_{i=1}^{t} \min_{j=1}^{k} w_i\|r_i - v_j\|_2^z &\leq \\ \Delta + \sum_{i=1}^{t} w_i(\|r_i\|_2^2 + 1)^{z/2} - \min\{1, z/2\}\sqrt{2k/t}\sum_{i=1}^{t} w_i(\|r_i\|_2^2 + 1)^{z/2-1}\|r_i\|_2.\end{aligned}$$

By Lemma 32, this is at most:

$$(1+\varepsilon)2^{z/2}d - \min\{1, z/2\}\sqrt{2k/t}\sum_{i=1}^{t} w_i(\|r_i\|_2^2 + 1)^{z/2-1}\|r_i\|_2.$$

We have therefore shown that

$$(1-\varepsilon)2^{z/2}(d - \max\{1, z/2\} \cdot \sqrt{dk}) \leq (1+\varepsilon)2^{z/2}d - \min\{1, z/2\}\sqrt{2k/t}\sum_{i=1}^{t} w_i(\|r_i\|_2^2 + 1)^{z/2-1}\|r_i\|_2.$$

Which implies:

$$\begin{aligned}\min\{1, z/2\}\sqrt{2k/t} \cdot \sum_{i=1}^{t} w_i(\|r_i\|_2^2 + 1)^{z/2-1}\|r_i\|_2 &\leq 2\varepsilon 2^{z/2}d + \max\{1, z/2\}2^{z/2}\sqrt{dk} \Rightarrow \\ \sum_{i=1}^{t} w_i(\|r_i\|_2^2 + 1)^{z/2-1}\|r_i\|_2 &\leq \frac{2\varepsilon 2^{z/2}d + \max\{1, z/2\}2^{z/2}\sqrt{dk}}{\sqrt{2} \cdot \min\{1, z/2\}} \cdot \sqrt{t/k}.\end{aligned}$$

□

Lemma 34. *For any $0 < \varepsilon < 1/2$ and any $k > 1$, let $d = k/(\min\{1, (z/2)^2\}32^2\varepsilon^2)$ and 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 ε-coreset for P, using offset Δ. Then*

$$\sum_{h=1}^{t} w_h(\|r_h\|_2^2 + 1)^{z/2-1}\|r_h\|_2 \geq \frac{2^{z/2}d}{11\max\{1, z/2\}\min\{1, z/2\}}.$$

Proof. Consider the Hadamard basis $h_1, \ldots, h_q$ on $q = 1/(\min\{1, (z/2)^2\}32^2\varepsilon^2)$ 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 $v_1^i, \ldots, v_k^i$ obtained as follows: For each group G_j of q coordinates, copy h_i into those coordinates to obtain v_j^i. We must have that $\sum_{h=1}^d \min_{j=1}^k \|e_h - v_j^i\|_2^p = \sum_{h=1}^d \min_{j=1}^k (\|e_h\|_2^2 + \|v_j^i\|_2^2 - 2\langle e_h, v_j^i\rangle)^{z/2}$. Since $k > 1$, there is always a j such that $\langle e_h, v_j^i\rangle = 0$. Moreover, for $i = 1$, we have $\max_{j=1}^k \langle e_h, v_j^i\rangle = 1/\sqrt{q}$ (since all coordinates of h_1 are $1/\sqrt{d}$, and for $i \neq 1$, it holds for precisely half of all e_h that $\max_{j=1}^k \langle e_h, v_j^i\rangle = 1/\sqrt{q}$. Thus we have $\sum_{h=1}^d \min_{j=1}^k \|e_h - v_j^i\|_2^z \leq (d/2)2^{z/2} + (d/2)(2 - 1/\sqrt{q})^{z/2}$. By Observation 1, this is at most $(d/2)2^{z/2} + (d/2)2^{z/2}(1 - \min\{1, z/2\}/(2\sqrt{q})) = d2^{z/2} - (d\min\{1, z/2\}/(4\sqrt{q}))2^{z/2} = d2^{z/2} - 8\varepsilon d2^{z/2}$. Thus:

$$\begin{aligned}
(1+\varepsilon)(d2^{z/2} - 8\varepsilon d2^{z/2}) &\geq \Delta + \sum_{h=1}^t w_h(\|r_h\|_2^2 + 1 - 2\max_{j=1}^k \langle r_h, v_j^i\rangle)^{z/2} \\
&\geq \Delta + \sum_{h=1}^t w_h(\|r_h\|_2^2 + 1 - 2\max_{j=1}^k |\langle r_h, v_j^i\rangle|)^{z/2}
\end{aligned}$$

By Observation 1, this is at least

$$\Delta + \sum_{h=1}^t w_h(\|r_h\|_2^2 + 1)^{z/2} - 2\max\{1, z/2\} \sum_{h=1}^t w_h \max_{j=1}^k |\langle r_h, v_j^i\rangle|(\|r_h\|_2^2 + 1)^{z/2-1}$$

By Lemma 32, this is at least

$$(1-\varepsilon)2^{z/2}d - 2\max\{1, z/2\} \sum_{h=1}^t w_h \max_{j=1}^k |\langle r_h, v_j^i\rangle|(\|r_h\|_2^2 + 1)^{z/2-1}.$$

We have thus shown

$$\begin{aligned}
2\max\{1, p/2\} \sum_{h=1}^t w_h \max_{j=1}^k |\langle r_h, v_j^i\rangle|(\|r_h\|_2^2 + 1)^{p/2-1} &\geq -2\varepsilon 2^{z/2}d + (1+\varepsilon)8\varepsilon d2^{z/2} \Rightarrow \\
\sum_{h=1}^t w_h \max_{j=1}^k |\langle r_h, v_j^i\rangle|(\|r_h\|_2^2 + 1)^{z/2-1} &\geq \frac{3\varepsilon 2^{z/2}d}{\max\{1, z/2\}}.
\end{aligned}$$

Now consider any r_h with weight w_h. Collect the vectors u_h^i such that $u_h^i = v_{j^*}^i$ where $j^* = \operatorname{argmax}_j |\langle r_h, v_j^i\rangle|$. Let $\sigma_h^i = \operatorname{sign}(\langle r_h, u_h^i\rangle)$. By construction, all these q vectors are orthogonal (either disjoint support or distinct vectors from the Hadamard basis). By Cauchy-Schwartz, we then have $\langle w_h(\|r_h\|_2^2 + 1)^{z/2-1} r_h, \sum_{i=1}^q \sigma_h^i u_h^i\rangle \leq w_h(\|r_h\|_2^2 + 1)^{z/2-1}\|r_h\|_2 \|\sum_{i=1}^q \sigma_h^i u_h^i\|_2 = w_h(\|r_h\|_2^2 +$

$1)^{z/2-1}\|r_h\|_2\sqrt{q}$. We then see that

$$\begin{aligned}
\frac{3\varepsilon 2^{z/2}dq}{\max\{1,z/2\}} &\leq \sum_{i=1}^{q}\sum_{h=1}^{t} w_h \max_{j=1}^{k} |\langle r_h, v_j^i\rangle|(\|r_h\|_2^2+1)^{z/2-1} \\
&= \sum_{h=1}^{t}\sum_{i=1}^{q} w_h \max_{j=1}^{k} |\langle r_h, v_j^i\rangle|(\|r_h\|_2^2+1)^{z/2-1} \\
&= \sum_{h=1}^{t} \langle w_h(\|r_h\|_2^2+1)^{z/2-1} r_h, \sum_{i=1}^{q} \sigma_h^i u_h^i\rangle \\
&\leq \sum_{h=1}^{t} w_h(\|r_h\|_2^2+1)^{z/2-1}\|r_h\|_2\sqrt{q}.
\end{aligned}$$

We have thus shown

$$\sum_{h=1}^{t} w_h(\|r_h\|_2^2+1)^{z/2-1}\|r_h\|_2 \geq \frac{3\varepsilon 2^{z/2}d\sqrt{q}}{\max\{1,z/2\}} = \frac{3\cdot 2^{z/2}d}{32\max\{1,z/2\}\min\{1,z/2\}} \geq \frac{2^{z/2}d}{11\max\{1,z/2\}\min\{1,z/2\}}.$$

□

We finally combine it all:

Theorem 11. *For any $0 < \varepsilon < 1/2$ and any k, let $d = k/(\min\{1,(z/2)^2\}32^2\varepsilon^2)$ and 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 (ε,k,z)-coreset for P, using offset Δ. Then $t = \Omega\left(\frac{k}{\varepsilon^2\max\{1,z^4\}}\right)$.*

Proof. Combining Lemma 33 and Lemma 34, we get $\frac{2^{z/2}d}{11\max\{1,z/2\}\min\{1,z/2\}} \leq \sum_{h=1}^{t} w_h(\|r_h\|_2^2+1)^{z/2-1}\|r_h\|_2 \leq \frac{2\varepsilon 2^{z/2}d+\max\{1,z/2\}2^{z/2}\sqrt{dk}}{\sqrt{2}\cdot\min\{1,z/2\}}\cdot\sqrt{t/k}$. That is,

$$t \geq \frac{k\cdot 2\min\{1,(z/2)^2\}\cdot d^2\cdot 2^z}{(2\varepsilon 2^{z/2}d+\max\{1,z/2\}2^{z/2}\sqrt{dk})^2 11^2\max\{1,(z/2)^2\}\min\{1,(z/2)^2\}}.$$

We have $\sqrt{dk} = d\min\{1,(z/2)\}32\varepsilon$. Asymptotically, the whole bound thus becomes:

$$t = \Omega\left(\frac{k}{\varepsilon^2\max\{1,z^4\}}\right).$$

□