

# A New Coreset Framework for Clustering

Vincent Cohen-Addad\*

David Saulpic†

Chris Schwiegelshohn‡

## Abstract

Given a metric space, the  $(k, z)$ -clustering problem consists of finding  $k$  centers such that the sum of the distances raised to the power  $z$  of every point to its closest center is minimized. This encapsulates the famous  $k$ -median ( $z = 1$ ) and  $k$ -means ( $z = 2$ ) clustering problems. Designing small-space sketches of the data that approximately preserves the cost of the solutions, also known as *coresets*, has been an important research direction over the last 15 years.

In this paper, we present a new, simple coreset framework that simultaneously improves upon the best known bounds for a large variety of settings, ranging from Euclidean space, doubling metric, minor-free metric, and the general metric cases: with  $\Gamma = \min(\varepsilon^{-2} + \varepsilon^{-z}, k\varepsilon^{-2})\text{polylog}(\varepsilon^{-1})$ , this framework constructs coreset with size

- $O(\Gamma \cdot k(d + \log k))$  in doubling metrics, improving upon the recent breakthrough of [Huang, Jiang, Li, Wu, FOCS' 18], who presented a coreset with size  $O(k^3 d / \varepsilon^2)$ .
- $O(\Gamma \cdot k \cdot \min(d, \varepsilon^{-2} \log k))$  in  $d$ -dimensional Euclidean space, improving on the recent results of [Huang, Vishnoi, STOC' 20], who presented a coreset of size  $O(k \log k \cdot \varepsilon^{-2z} \cdot \min(d, \varepsilon^{-2} \log k))$ .
- $O(\Gamma \cdot k(t + \log k))$  for graphs with treewidth  $t$ , improving on [Baker, Braverman, Huang, Jiang, Krauthgamer, Wu, ICML'20], who presented a coreset of size  $O(k^2 t / \varepsilon^2)$  for  $z = 1$ .
- $O\left(\Gamma \cdot k \left(\log^2 k + \frac{\log k}{\varepsilon^4}\right)\right)$  for shortest paths metrics of graphs excluding a fixed minor. This improves on [Braverman, Jiang, Krauthgamer, Wu, SODA'21], who presented a coreset of size  $O(k^2 / \varepsilon^4)$ .
- Size  $O(\Gamma \cdot k \log n)$  in general discrete metric spaces, improving on the results of [Feldman, Lamberg, STOC'11], who presented a coreset of size  $O(k\varepsilon^{-2z} \log n \log k)$ .

A lower bound of  $\Omega\left(\frac{k \log n}{\varepsilon}\right)$  for  $k$ -Median in general metric spaces [Baker, Braverman, Huang, Jiang, Krauthgamer, Wu, ICML'20] implies that in general metrics as well as metrics with doubling dimension  $d$ , our bounds are optimal up to a  $\text{poly}(\log(1/\varepsilon))/\varepsilon$  factor. For graphs with treewidth  $t$ , the lower bound of  $\Omega\left(\frac{kt}{\varepsilon}\right)$  of [Baker, Braverman, Huang, Jiang, Krauthgamer, Wu, ICML'20] shows that our bounds are optimal up to the same factor.

## 1 Introduction

Center-based clustering problems are classic objectives for the problem of computing a “good” partition of a set of points into  $k$  parts, so that points that are “close” are in the same part. Finding a good clustering of a dataset helps extracting important information from a dataset and

---

\*Google Research, Zurich.

†Sorbonne Université, Paris

‡Aarhus University

center based clustering problems have become the cornerstones of various data analysis approaches and machine learning techniques (see formal definition in Section 3).

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. The method of choice for this is to compute a *coreset*, i.e. a weighted set of points of small size that can be used in place of the full input for algorithmic purposes. More formally, for any  $\varepsilon > 0$ , an  $\varepsilon$ -coreset (referred to simply as coreset) is a set  $Q$  of points of the metric space such that any  $\alpha$ -approximation to a clustering problem on  $Q$ , is a  $\alpha(1 + \varepsilon)$ -approximation to the clustering problem for the original point set. Hence, a small coreset is a good compression of the full input set: one can simply keep in memory a coreset and apply any given algorithm on the coreset rather than on the input to speed up performances and reduce memory consumption. Coreset constructions had been widely studied over the last 15 years.

In this paper, we specifically focus on the  $(k, z)$ -clustering problem, which encapsulates  $k$ -median ( $z = 1$ ) and  $k$ -means ( $z = 2$ ). Given two positive integers  $k$  and  $z$  and a metric space  $(X, \text{dist})$ , the  $(k, z)$ -clustering problem asks for a set  $\mathcal{S}$  of  $k$  points, called *centers*, that minimizes

$$\text{cost}(X, \mathcal{S}) := \sum_{x \in X} \min_{s \in \mathcal{S}} \text{dist}(x, s)^z$$

The method of choice for designing coreset is *importance sampling*, initiated by the seminal work of Chen [Che09]. The basic approach is to devise a non-uniform sampling distribution which picks points proportionally to their cost contribution in an arbitrary constant factor approximation. In a nutshell, the current best-known analysis shows that, for a given set  $\mathcal{S}$  of  $k$  centers, it happens with high probability that the sampled instance  $\Omega$  with appropriate weights has roughly the same cost as the original instance, i.e.  $\text{cost}(\Omega, \mathcal{S}) \in (1 \pm \varepsilon) \text{cost}(X, \mathcal{S})$ . Then, to show that the set  $\Omega$  is an  $\varepsilon$ -coreset, it is necessary to take a union-bound over these events for all possible set of  $k$  centers. Bounding the size of the union-bound is the main hurdle faced by this approach: indeed, there may be infinitely many possible set of centers.

The state-of-the-art analysis relies on VC-dimension to address this issue. Informally, the VC dimension is a complexity measure of a range space, denoting the cardinality of the largest set such that all subsets are included in the range space. The application to clustering considers weighted range spaces, where each point is weighted by its relative contribution to the cost of a given clustering<sup>1</sup>. In metric spaces where the weighted range space induced by distances to  $k$  centers has VC-dimension  $D$ , it can be shown that taking  $O_{\varepsilon, z}(k \cdot D \log k)$  samples yields a coreset [FSS20], although tighter bounds are achievable in certain cases. For instance, in  $d$  dimensional Euclidean spaces  $D$  is in  $O(kd \log k)$  [BLHK17], which would yield coresets of size  $O_{\varepsilon, z}(k^2 \cdot d \log^2 k)$ , but Huang and Vishnoi [HV20] showed the existence of a coreset with  $O(k \cdot \log^2 k \cdot \varepsilon^{-2z-2})$  points.

This analysis was proven powerful in various metric spaces, such as doubling spaces by Huang, Jiang, Li and Wu [HJLW18], graphs of bounded treewidth by Baker, Braverman, Huang, Jiang, Krauthgamer, Wu [BBH<sup>+</sup>20] or the shortest-path metric of a graph excluding a fixed minor by Braverman, Jiang, Krauthgamer and Wu [BJKW21]. However, range spaces of even heavily constrained metrics do not necessarily have small VC-dimension (e.g. bounded doubling dimension does

---

<sup>1</sup>For more on these notions, we refer to [FSS20].

not imply bounded VC-dimension or vice versa [HJLW18, LL06]), and applying previous techniques requires heavy additional machinery to adapt the VC-dimension approach to them. Moreover, the bounds provided are far from the bound obtained for Euclidean spaces: their dependency in  $k$  is at least  $\Omega(k^2)$ , leaving a significant gap to the best lower bounds of  $\Omega(k)$ . We thus ask:

**Question.** *Is it possible to design coresets whose size are near-linear in  $k$  for doubling metrics, minor-free metrics, bounded-treewidth metrics? Are the current roadblocks specific to the analysis through VC-dimension, or inherent to the problem?*

To answer positively these questions, we present a new framework to analyse importance sampling. Its analysis stems from first principles, and it can be applied in a black-box fashion to any metric space that admits an *approximate centroid set* (see Definition 1) of bounded size. We show that all previously mentioned spaces satisfy this condition, and our construction improves on the best-known coreset size. More precisely, we recover (and improve) all previous results for  $(k, z)$ -clustering such as Euclidean spaces,  $\ell_p$  spaces for  $p \in [1, 2)$ , finite  $n$ -point metrics, while also giving the first coresets with size near-linear in  $k$  and  $\varepsilon^{-z}$  for a number of other metrics such as doubling spaces, minor free metrics, and graphs with bounded treewidth.

## 1.1 Our Results

Our framework requires the existence of a particular discretization of the set of possible centers, as described in the following definition. We show in the latter sections that this is indeed the case for all the metric spaces mentioned so far.

**Definition 1.** *Let  $(X, \text{dist})$  be a metric space,  $P \subseteq X$  a set of clients and two positive integers  $k$  and  $z$ . Let  $\varepsilon > 0$  be a precision parameter. Given a set of centers  $\mathcal{A}$ , a set  $\mathbb{C}$  is an  $\mathcal{A}$ -approximate centroid set for  $(k, z)$ -clustering on  $P$  if it satisfies the following property.*

*For every set of  $k$  centers  $\mathcal{S} \in X^k$ , there exists  $\tilde{\mathcal{S}} \in \mathbb{C}^k$  such that for all points  $p \in P$  that satisfies either  $\text{cost}(p, \mathcal{S}) \leq \left(\frac{8z}{\varepsilon}\right)^z \text{cost}(p, \mathcal{A})$  or  $\text{cost}(p, \tilde{\mathcal{S}}) \leq \left(\frac{8z}{\varepsilon}\right)^z \text{cost}(p, \mathcal{A})$ , it holds*

$$|\text{cost}(p, \mathcal{S}) - \text{cost}(p, \tilde{\mathcal{S}})| \leq \frac{\varepsilon}{z \log(z/\varepsilon)} (\text{cost}(p, \mathcal{S}) + \text{cost}(p, \mathcal{A})),$$

This definition is slightly different from Matousek's one [Mat00], in that we seek to preserve distances only for interesting points, and allow an error  $\varepsilon \text{cost}(p, \mathcal{A})$ . This is crucial in some of our applications.

**Theorem 1.** *Let  $(X, \text{dist})$  be a metric space,  $P \subseteq X$  a set of clients with  $n$  distinct points and two positive integers  $k$  and  $z$ . Let  $\varepsilon > 0$  be a precision parameter. Let also  $\mathcal{A}$  be a constant-factor approximation for  $(k, z)$ -clustering on  $P$ .*

*Suppose there exists an  $\mathcal{A}$ -approximate centroid set  $\mathbb{C}$  for  $(k, z)$ -clustering on  $P$ . Then, there exists an algorithm running in time  $O(n)$  that constructs with probability at least  $1 - \pi$  a coreset of size*

$$O\left(\frac{2^{O(z \log z)} \cdot \log^4 1/\varepsilon}{\min(\varepsilon^2, \varepsilon^z)} (k \log |\mathbb{C}| + \log \log(1/\varepsilon) + \log(1/\pi))\right)$$

*with positive weights for the  $(k, z)$ -clustering problem.*

When applying this theorem to particular metric spaces, the running time is dominated by the construction of the constant-factor approximation  $\mathcal{A}$ , which can be done for instance in  $\tilde{O}(k|P|)$  given oracle access to the distances using [MP04]<sup>2</sup>.

If one wishes to trade a factor  $\varepsilon^{-z}$  for a factor  $k$ , we also present coresets of size  $O(k^2 \cdot 2^{O(z)} \frac{\log^3(1/\varepsilon)}{\varepsilon^2} (\log k + \log |\mathbb{C}| + \log(1/\pi)))$ , as explained in Appendix B.

We apply this theorem to several metric spaces, achieving the following (simplified) size bounds (we ignore  $\text{poly} \log(1/\varepsilon)$  and  $2^{O(z \log z)}$  factors): let  $\Gamma = \min(\varepsilon^{-2} + \varepsilon^{-z}, k\varepsilon^{-2})$ , see also Table 1.

- $O(\Gamma \cdot k(d + \log k))$  for metric spaces with doubling dimension  $d$ . This improves over the  $O(k^3 d \varepsilon^{-2})$  from Huang et al. [HJLW18]. See Corollary 4.
- Since general discrete metric spaces have doubling dimension  $O(\log n)$ , this yields coreset of size  $O(\Gamma \cdot k \log n)$ . This improves on the bound from Feldman and Langberg [FL11]  $O(\varepsilon^{-2z} k \log k \log n)$ .
- $O(\Gamma \cdot k \varepsilon^{-2} \cdot \log k)$  for Euclidean spaces, see Corollary 8. This improves on the recent result from Huand and Vishnoi [HV20], who achieve  $O(\varepsilon^{-2z-2} k \log^2 k)$ .
- $O\left(\Gamma \cdot k \left(\log^2 k + \frac{\log k}{\varepsilon^4}\right)\right)$  for a family of graphs excluding a fixed minor, see Corollary 7. This improves on Braverman et al. [BJKW21], whose coreset has size  $\tilde{O}(k^2/\varepsilon^4)$ .
- $O\left(\Gamma \cdot k \left(\log^2 k + \frac{\log k}{\varepsilon^3}\right)\right)$  for Planar Graphs, which is a particular family excluding a fixed minor for which we can save a  $1/\varepsilon$  factor and present a simpler, instructive proof.
- $O(\Gamma \cdot k(t + \log k))$  in graphs with treewidth  $t$ , see Corollary 5. This improves upon the work of Baker et al. [BBH<sup>+</sup>20], that construct coreset with size  $\tilde{O}(k^2 t/\varepsilon^2)$ .
- $O(k \varepsilon^{-2z} \cdot \min(d, \varepsilon^{-2} \log k))$  in  $\mathbb{R}^d$  with  $\ell_p$  distance, for  $p \in [1, 2)$ , see Corollary 9. This improves on Huang and Vishnoi [HV20], who presented a coreset of size  $O(k \log k \cdot \varepsilon^{-4z} \cdot \min(d, \varepsilon^{-2} \log k))$ .

We note the lower bound  $\Omega(\frac{k \log n}{\varepsilon})$  for  $k$ -Median in general metric spaces from [BBH<sup>+</sup>20]. This means that in the case of metrics with doubling dimension  $d$ , our bounds are optimal up to a  $\text{poly} \log(1/\varepsilon)/\varepsilon$  factor. For graphs with treewidth  $t$ , another lower bound of  $\Omega(\frac{kt}{\varepsilon})$  from [BBH<sup>+</sup>20] shows that our bounds are optimal up to the same factor.

## 1.2 Overview of Our Techniques

Our proof is arguably from first principles. We now give a quick overview of its ingredients. The approach consists in first reducing to a well structured instance, that consists of a set of centers  $\mathcal{A}$  inducing  $k$  clusters, all having roughly the same costs, and where every point is at the same distance of  $\mathcal{A}$ , up to a factor 2. Then we show it is enough to perform importance sampling on all these clusters.

---

<sup>2</sup>Although initially stated for  $z = 1$  only, this algorithm works for general  $z$  as stressed in [HV20]

Reference	Size (Number of Points)
<b>Euclidean space</b>	
Har-Peled, Mazumdar (STOC'04) [HM04]	$O(k \cdot \varepsilon^{-d} \cdot \log n)$
Har-Peled, Kushal (DCG'07) [HK07]	$O(k^3 \cdot \varepsilon^{-(d+1)})$
Chen (Sicomp'09) [Che09]	$O(k^2 \cdot d \cdot \varepsilon^{-2} \log n)$
Langberg, Schulman (SODA'10) [LS10]	$O(k^3 \cdot d^2 \cdot \varepsilon^{-2})$
Feldman, Langberg (STOC'11) [FL11]	$O(k \cdot d \cdot \log k \cdot \varepsilon^{-2z})$
Feldman, Schmidt, Sohler (Sicomp'20) [FSS20]	$O(k^3 \cdot \log k \cdot \varepsilon^{-4})$
Sohler and Woodruff (FOCS'18) [SW18]	$O(k^2 \cdot \log k \cdot \varepsilon^{-O(z)})$
Becchetti, Bury, Cohen-Addad, Grandoni, Schwiegelshohn (STOC'19) [BBC <sup>+</sup> 19]	$O(k \cdot \log^2 k \cdot \varepsilon^{-8})$
Huang, Vishnoi (STOC'20) [HV20]	$O(k \cdot \log^2 k \cdot \varepsilon^{-2-2z})$
Braverman, Jiang, Krauthgamer, Wu (SODA'21) [BJKW21]	$\tilde{O}(k^2 \cdot \varepsilon^{-4})$
<b>This paper</b>	$O(k \cdot \log k \cdot \varepsilon^{-2-\max(2,z)})$
<b>General <math>n</math>-point metrics, <math>ddim</math> denotes the doubling dimension</b>	
Chen (Sicomp'09) [Che09]	$O(k^2 \cdot \varepsilon^{-2} \cdot \log^2 n)$
Feldman, Langberg (STOC'11) [FL11]	$O(k \cdot \log k \cdot \log n \cdot \varepsilon^{-2z})$
Huang, Jiang, Li, Wu (FOCS'18) [HJLW18]	$O(k^3 \cdot ddim \cdot \varepsilon^{-2})$
<b>This paper</b>	$O(k \cdot (ddim + \log k) \cdot \varepsilon^{-\max(2,z)})$
<b>This paper</b>	$O(k \cdot \log n \cdot \varepsilon^{-\max(2,z)})$
<b>Graph with <math>n</math> vertices, <math>t</math> denotes the treewidth</b>	
Baker, Braverman, Huang, Jiang, Krauthgamer, Wu (ICML'20) [BBH <sup>+</sup> 20]	$\tilde{O}(k^2 \cdot t / \varepsilon^2)$
<b>This paper</b>	$O(k \cdot (t + \log k) \cdot \varepsilon^{-\max(2,z)})$
<b>Graph with <math>n</math> vertices, excluding a fixed minor</b>	
Bravermann Jian, Krauthgamer, Wu (SODA'21) [BJKW21]	$\tilde{O}(k^2 \cdot \varepsilon^{-4})$
<b>This paper</b>	$O\left(k \cdot (\log^2 k + \frac{\log k}{\varepsilon^4}) \cdot \varepsilon^{-\max(2,z)}\right)$

Table 1: Comparison of coresset sizes for  $(k, z)$ -Clustering in various metrics. Dependencies on  $2^{O(z)}$  and  $\text{polylog} \varepsilon^{-1}$  are omitted from all references. Additionally, we may trade a factor  $\varepsilon^{-z+2}$  for a factor  $k$  in any construction with  $z > 2$ . [HK07, HM04] only applies to  $k$ -means and  $k$ -median, [BBC<sup>+</sup>19, FSS20] only applies to  $k$ -means. [SW18] runs in exponential time, which has been addressed by Feng et al. [FKW19]. Aside from [HK07, HM04], the algorithms are randomized and succeed with constant probability. Although the results are claimed only for  $k$ -Median in [BBH<sup>+</sup>20], it seems that they can be generalized to any power. The main difference is in the computation of a constant factor approximation.

**Reducing to a structured instance.** Like most coresset constructions, we initially compute a constant factor approximation  $\mathcal{A}$  to the problem. We then deviate from previous importance sampling algorithms by partitioning points into groups such that the following conditions are satisfied, for a given group  $G$ :

- For all clusters, the cost of the intersection of the cluster with the group is at least half the

average; i.e.  $\forall C_i, \text{cost}(C_i \cap G, \mathcal{A}) \geq \frac{\text{cost}(G, \mathcal{A})}{2k}$ .

- In every cluster  $C_i$ , there exists  $r_{G,i}$  such that the points in the intersection of the cluster with the group cost  $r_{G,i}$  (up to constant factors), i.e.  $\forall p \in C_i \cap G, \text{cost}(p, \mathcal{A}) = \Theta(r_{G,i})$ .

We then compute coresets for each group and output the union. In some sense, this preprocessing step identifies canonical instances for coresets; any algorithm that produces improved coresets for instances satisfying the aforementioned regularity condition can be combined with our preprocessing steps to produce improved coreset in general.

**Importance Sampling in Groups.** The first technical challenge is to analyse the importance sampling procedure for structured instances.

The arguably simplest way to attempt to analyse importance sampling is by first showing that for any fixed solution  $\mathcal{S}$  we need a set  $\Omega$  of  $\delta$  samples to show that with good enough probability

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

and then applying a union bound over the validity of Eq. (1) for all solutions  $\mathcal{S}$ . This union bound is typically achieved via the VC-dimension.

Using this simple estimator, most analyses of importance sampling procedures require a sample size of at least  $k$  points to approximate the cost of a single given solution. To illustrate this, consider an instance where a single cluster  $C$  is isolated from all the others. Clearly, if we do not place a center close to  $C$ , the cost will be extremely large, requiring some point of  $C$  to be contained in the sample. One way to remedy this is by picking a point  $p'$  proportionate to  $\frac{\text{cost}(p', \mathcal{A})}{\text{cost}(\mathcal{A})} + \frac{1}{|C_i|}$  rather than  $\frac{\text{cost}(p', \mathcal{A})}{\text{cost}(\mathcal{A})}$ , where  $C_i$  is the cluster to which  $p'$  is assigned, see for instance [FSS20]. This analysis always leads to coreset of size quadratic in  $k$  at best<sup>3</sup>. Our analysis of importance sampling for structured instances will allow us to bypass both the quadratic dependencies on  $k$ , and the need of a bound on the VC-dimension of the range space.

Our high level idea is to use two union bounds. The first one will deal with clusters that are very expensive compared to their cost in  $\mathcal{A}$ . The second one will focus on solutions in which clusters have roughly the same cost as they do in  $\mathcal{A}$ . For the former case, we observe that if a cluster  $C_i$  is served by a center in solution  $\mathcal{S}$  that is very far away, then we can easily bound its cost in  $\mathcal{S}$  as long as our sample approximates the size of every cluster. Specifically, assume that there exists a point  $p$  in  $C_i$  with distance to  $\mathcal{S}$  at least  $\Omega(1) \cdot \varepsilon^{-1} \cdot \text{dist}(p, c_i)$ . Then, since we are working with structured instances, all points of  $C_i$  are roughly at the same distance of  $c_i$  and that this distance is negligible compared to  $\text{dist}(p, \mathcal{S})$ , all points of  $C_i$  are nearly at the same distance of  $\mathcal{S}$ . Conditioned on the event  $\mathcal{E}$  that the sample  $\Omega$  preserves the size of all clusters, the cost of  $C_i$  in solution  $\mathcal{S}$  is preserved as well. Note that this event  $\mathcal{E}$  is independent of the solution  $\mathcal{S}$  and thus we require no enumeration of solutions to preserve the cost of expensive clusters. Proving that  $\mathcal{E}$  holds is a straightforward application of concentration bounds.

---

<sup>3</sup>A linear dependency on  $k$  can be achieved using a different analysis, see [FL11, HV20] for examples. This approach does not seem to generalize to arbitrary metrics.

The second observation is that points with  $\text{dist}(p, \mathcal{S}) \leq \varepsilon/z \cdot \text{dist}(p, \mathcal{A})$  are so cheap that their cost is preserved by the sampling with an error at most  $\varepsilon \cdot \text{cost}(\mathcal{A})$ . Indeed, their cost in  $\mathcal{S}$  cannot be more than  $\varepsilon \cdot \text{cost}(\mathcal{A})$ : it is easy to show that the same bound holds for the coreset.

The intermediate cases, i.e. solutions in which  $\mathcal{S}$  serves clusters at distances further than  $\varepsilon/z \cdot \text{dist}(p, \mathcal{A})$ , but not so far as to simply use event  $\mathcal{E}$  to bound the cost, is the hardest part of the analysis. Using a geometric series, we can split the cost ranges into  $\log \frac{z}{\varepsilon^2} \in O(z \log \varepsilon^{-1})$  groups by powers of two. Due to working with a structured instance, the points within such a group have equal distances, up to a constant factors. This also implies that the cost in such a group is equal, up to a factor of  $2^{O(z)}$ . The overall variance of the cost estimator is then of the order  $\max_p (\varepsilon^{-1} \cdot \text{dist}(p, \mathcal{A}))^z \cdot \frac{\text{cost}(\mathcal{A})}{\text{cost}(p, \mathcal{A})} \in O(\varepsilon^{-z})$ . Thus, standard concentration bounds give an additive error of  $\varepsilon \cdot (\text{cost}(\mathcal{A}) + \text{cost}(\mathcal{S}))$  with at most  $O(\varepsilon^{-2-z})$  many samples for every group.

To improve this to  $O(\varepsilon^{-z})$ , we use a different estimator defined as follows. For every cluster  $C_i$ , let  $q_i$  be the point of  $C_i$  that is the closest to  $\mathcal{S}$ . We then consider

$$\sum_{p \in C_i \cap \Omega} (\text{cost}(p, \mathcal{S}) - \text{cost}(q_i, \mathcal{S})) \cdot \frac{\text{cost}(G, \mathcal{A})}{\text{cost}(p, \mathcal{A}) \cdot \delta} \quad (2)$$

$$+ \sum_{p \in C_i \cap \Omega} \text{cost}(q_i, \mathcal{S}) \cdot \frac{\text{cost}(G, \mathcal{A})}{\text{cost}(p, \mathcal{A}) \cdot \delta} \quad (3)$$

Conditioned on event  $\mathcal{E}$ , the estimator in Equation 3 is always concentrated around its expectation, as  $\text{cost}(q_i, \mathcal{S})$  is fixed for  $\mathcal{S}$ . The first estimator in Equation 2 now has a reduced variance. Specifically, at the border cases of points at distance  $\Theta(1/\varepsilon) \text{dist}(p, \mathcal{A})$  of  $\mathcal{S}$ , the Estimator 2 has variance at most  $O(1) \cdot \max(\varepsilon^{-2}, \varepsilon^{-z}) \cdot \text{cost}(\mathcal{A}) \cdot \text{cost}(\mathcal{S})$ , which ultimately allows us to show that  $O(\varepsilon^{-2} + \varepsilon^{-z})$  samples are enough to achieve an additive error of  $\varepsilon \cdot (\text{cost}(\mathcal{S}) + \text{cost}(\mathcal{A}))$ . This technique is somewhat related to (and inspired by) chaining arguments (see e.g. Talagrand [T<sup>+</sup>96] for more on chaining). The key difference is while chaining is generally applied to improve over basic union bounds, our estimator is designed to reduce the variance.

**Preserving the Cost of Points not in Well-Structured Groups** Unfortunately, it is not possible to decompose the entire point set into groups. Given an initial solution  $\mathcal{A}$  and a cluster  $C \in \mathcal{A}$ , this is possible for all the points at distance at most  $\varepsilon^{-O(z)} \cdot \frac{\text{cost}(C, c)}{|C|}$ . The remaining points are now both far from their respective center in  $\mathcal{A}$  and, due to Markov's inequality, only a small fraction of the point set. In the following, let  $P_{far}$  denote these points.

For any given subset of these far away points and a candidate solution  $\mathcal{S}$ , now use that either the points pay at most what they do in  $\mathcal{A}$ , or an increase in their cost significantly increases the overall cost. In the former case, standard sensitivity sampling preserves the cost with a very small sample size. In the latter case, a significant cost by a point  $p$  in  $P_{far}$  also implies that all points close to the center  $c$  serving  $p$  in  $\mathcal{A}$  have to significantly increase the cost.

**A Union-bound to Preserve all Solutions** As pictured in the previous paragraphs, the cost of points with either very small or very large distance to  $\mathcal{S}$  is preserved for any solution  $\mathcal{S}$  with high probability.

The guarantee we have for interesting points is weaker: their cost is preserved by the coreset with high probability for any *fixed* solution  $\mathcal{S}$ . Hence, for this to hold for any solution, we need to take a union-bound over the probability of failure for all possible solution  $\mathcal{S}$ . However, the union-bound is necessary only for these interesting points : this explains the introduction of the approximate centroid set in Definition 1. Assuming the existence of a set  $\mathbb{C}$  such as in Definition 1, one can take a union-bound over the failure of the construction for all set of  $k$  centers in  $\mathbb{C}^k$  to ensure that the cost of interesting points is preserved for all these solutions. To extend this result to *any* solution  $\mathcal{S}$ , one can take the set of  $k$  points  $\tilde{\mathcal{S}}$  in  $\mathbb{C}^k$  that approximates best  $\mathcal{S}$ , and relate the cost of interesting points in  $\mathcal{S}$  to their cost in  $\tilde{\mathcal{S}}$  with a tiny error. Since the cost of interesting points in  $\tilde{\mathcal{S}}$  is preserved in the coreset, the cost of these points in  $\mathcal{S}$  is preserved as well.

We briefly picture now how to get approximate centroid sets for specific metrics. We are looking for a set  $\mathbb{C}$  with the following property: for every solution  $\mathcal{S}$ , there exists a  $k$ -tuple  $\tilde{\mathcal{S}} \in \mathbb{C}^k$  such that for every point  $p$  with  $\text{dist}(p, \mathcal{S}) \leq \varepsilon^{-1} \text{dist}(p, \mathcal{A})$  in a given cluster  $C$  of  $\mathcal{A}$ ,  $|\text{cost}(p, \mathcal{S}) - \text{cost}(p, \tilde{\mathcal{S}})| \leq \varepsilon (\text{cost}(p, \mathcal{A}) + \text{cost}(p, \mathcal{S}))$ . We call such points *interesting*.

**Metrics with doubling dimension  $d$ :**  $\mathbb{C}$  is simply constructed taking *nets* around each input point. A  $\gamma$ -net of a metric space is a set of points that are at least at distance  $\gamma$  from each other, and such that each point of the metric is at distance at most  $\gamma$  from the net. The existence of  $\gamma$ -nets of small size is one of the key properties of doubling metrics (see Lemma 19). For every point  $p$ ,  $\mathbb{C}$  contains an  $\varepsilon \text{cost}(p, \mathcal{A})$ -net of the points at distance at most  $\frac{8\varepsilon}{\varepsilon} \cdot \text{cost}(p, \mathcal{A})$  from  $p$ . If  $p$  is an interesting point, there is therefore a center of  $\mathbb{C}$  close to its center in  $\mathcal{S}$ .

However, this only shows that centers from the solution  $\tilde{\mathcal{S}} \in \mathbb{C}^k$  are closer than those of  $\mathcal{S}$ . Showing that none gets too close is a different ballgame. We will see two ways of achieving it. The first one, that we apply for the doubling, treewidth and planar case, is based on the following observation: if a center  $s \in \mathcal{S}$  is replaced by a center  $\tilde{s}$ , that is way closer to a point  $p$  than  $s$ , then  $s$  can be discarded in the first place and be replaced by the center serving  $p$ . This is formalized in Lemma 18. The other way of ensuring that no center from  $\tilde{\mathcal{S}}$  gets too close to a point in  $p$  is based on guessing distances from points in  $\mathcal{S}$  to input points. It can be applied more broadly than Lemma 18, but yields larger centroid sets. We will use it only for minor-excluded graphs, for which Lemma 18 cannot be applied.

**Graphs with treewidth  $t$ :** The construction of  $\mathbb{C}$  is not as easy in graph metrics: we use the existence of small-size *separators*, building on ideas from Baker et al. [BBH<sup>+</sup>20]. Fix a solution  $\mathcal{S}$ , and suppose that all interesting points are in a region  $R$  of the graph, such that the boundary  $B$  of  $R$  is made of a constant number of vertices. Fix a center  $c \in \mathcal{S}$ , and suppose  $c$  is not in  $R$ . Then, to preserve the cost of interesting points, it is enough to have a center  $c'$  at the same distance to all points in the boundary  $B$  as  $c$ .

$\mathbb{C}$  is therefore constructed as follows: for a point  $p$ , its distance tuple to  $B = \{b_1, \dots, b_{|B|}\}$  is the tuple  $(d_1, \dots, d_{|B|})$ , where  $d_i = \text{dist}(p, b_i)$  is the distance to  $b_i$ . For every distance tuple to  $B$ ,  $\mathbb{C}$  contains one point having approximately that distance tuple to  $B$ .

Let  $\tilde{c}$  be the point of  $\mathbb{C}$  having approximately the same distance tuple to  $B$  as  $c$ : this ensures that  $\forall p, \text{cost}(p, c) \approx \text{cost}(p, \tilde{c})$ .

It is however necessary to limit the size of  $\mathbb{C}$ . For that, we approximate the distances to  $B$ . This



can be done for interesting points  $p$  as follows: since we have  $\text{dist}(p, c) \leq \varepsilon^{-1} \text{dist}(p, \mathcal{A})$ , rounding the distances to their closest multiple of  $\varepsilon \text{dist}(p, \mathcal{A})$  ensures that there are only  $O(1/\varepsilon^2)$  possibilities, and adds an error  $\varepsilon \text{cost}(p, \mathcal{A})$ . We show in Section 9 how to make this argument formal, and how to remove the assumption that all interesting points are in the same region.

**In minor-excluded graphs** this class of graphs, that includes planar graphs, admits as well small-size shortest-path separators. A construction similar in spirit to the one for treewidth is therefore possible, as presented in Section 11. This builds on the work of Braverman et al. [BJKW21].

However, due to the nature of the separator – which are small sets of paths, and not simply small sets of vertices – one cannot apply the idea of Lemma 18 to show that no center gets too close. Instead, we will guess the distance from input points to any point in  $\mathcal{S}$ , allowing to construct  $\tilde{\mathcal{S}}$  with the same distances. Of course, this mere idea requires way too many guesses to have a small set  $\mathbb{C}$ : we see in Section 11 how to make it work properly.

We start the section by showing two preprocessing lemmas: the first one is Lemma 18, as described above. The second one allows to apply Theorem 1 in the case the input set is weighted, so that we can assume the input has only  $\text{poly}(k, \varepsilon^{-1})$  many distinct points, by first computing a non-optimal coreset.

### 1.3 Roadmap

The paper is organized as follow: after defining the concepts used in the paper, we present formally the algorithm in Section 4. We then describe the construction of a coreset for a structured instance in Section 5, and the reduction to such an instance in Section 7. Finally, we show the existence of approximate centroid set in various metric spaces in Section 8. We furthermore explain the dimension reduction technique leading to our result for Euclidean spaces in Section 12, and the  $O(k^2 \varepsilon^{-2})$  construction in Appendix B. A deeper description of related work is made in Section 2.

## 2 Related Work

We already surveyed most of the relevant bounds for coresets for  $k$ -means and  $k$ -median. A complete overview over all of these bounds is given in Table 1, further pointers to coreset literature can be found in surveys [MS18]. For the remainder of the section, we highlight differences to previous techniques.

The early coreset results mainly considered input data embedded in constant dimensional Euclidean spaces [FS05, HK07, HM01]. These coresets relied on low-dimensional geometric decompositions inducing coresets of sizes typically of order at least  $k \cdot \varepsilon^{-d}$ . These techniques were replaced by *importance sampling* schemes, initiated by the seminal work of Chen [Che09]. The basic approach is to devise a non-uniform sampling distribution which picks points proportionately to their impact in a given constant factor approximation. A significant advantage of importance sampling over other techniques is that it generalizes to non-Euclidean metrics. While the early coreset papers [HK07, HM04] were indeed heavily reliant on the structure of Euclidean spaces, Chen gave the first coreset of size  $O(k^2 \varepsilon^{-2} \log^2 n)$  for general  $n$ -point metrics.

**Coresets via Bounded VC-Dimension** The state of the art importance sampling techniques in Euclidean spaces are based on reducing the problem of constructing a coreset to constructing an  $\varepsilon$ -net in a range space of bounded VC-dimension<sup>4</sup>. Li, Long and Srinivasan [LLS01] showed that if the VC-dimension is bounded by  $D$ , an  $\varepsilon$ -approximation of size  $O(\frac{D}{\varepsilon^2})$  exists. The remarkable aspect of these bounds is that they are independent of the number of input points. To apply the reduction, we need a bound on the VC-dimension for the range space induced by the intersection of metric balls centered around  $k$  points in a  $d$ -dimensional Euclidean space. For Euclidean  $k$ -means and  $k$ -median, an upper bound of  $D \in O(kd \log k)$  is implicit in the work of [BEHW89] and Eisenstat and Angluin [EA07]. This bound was recently shown to be tight by Csikos, Mustafa and Kupavskii [CMK19]. The dependency on  $d$  may be replaced with a dependency on  $\log k$ , as explained in more detail in Section 12. Thus  $O(k \log^2 k)$  is a natural barrier for known techniques in Euclidean spaces.

**VC-Dimension and Doubling Dimension** A further complication arises when attempting to extend sampling techniques for bounded VC-dimension in range spaces of bounded doubling dimension  $d$ . While the two notions share certain similarities and are asymptotically identical for the range space induced by the intersection of balls in Euclidean spaces, the two quantities are incomparable in general. For instance, Li and Long proved the existence of a range space with constant VC dimension and unbounded doubling dimension [LL06]. Conversely, [HJLW18] also showed that a bound on the doubling dimension does not imply a bound on the VC-dimension. Nevertheless, by carefully distorting the metric they were able to prove that a related quantity known as the shattering dimension can be bounded, yielding the first coresets for bounded doubling dimension independent of  $n$ . Even so, their bound  $\tilde{O}(k^3 d \varepsilon^{-2})$  is still far from what is currently achievable in Euclidean spaces.

Similarly, the construction from [BBH<sup>+</sup>20] for graphs with bounded treewidth uses that a graph of treewidth  $t$  has shattering dimension  $O(t)$ . They use this fact to get coreset for  $k$ -Median, of size  $\tilde{O}(k^3 t / \varepsilon^2)$ . For excluded-minor graphs, [BJKW21] proceeds similarly, but need an additional iterative procedure: they first show that in an excluded-minor graph, a subset  $X$  of the vertices has coreset of size  $O_{k,\varepsilon}(\log |X|)$ , using the shattering-dimension techniques. They show then how to iterate this construction (using that "a coreset of a coreset is a coreset") to remove dependency in  $|X|$ . This iterative procedure is of independent interest, and we use it as well for bounded treewidth and excluded-minor settings.

**Further Related Work** So far we only described works that aim at giving better coreset construction for unconstrained  $k$ -median and  $k$ -means in some metric space. Nevertheless, there is a rich literature on further related questions. As a tool for data compression, coresets feature heavily in streaming literature. Some papers consider a slightly weaker guarantee of summarizing the data set such that a  $(1 + \varepsilon)$  approximation can be maintained and extracted. Such notions are often referred to as *weak coresets* or streaming coresets, see [FL11, FMS07]. Further papers focus on maintaining coresets with little overhead in various streaming and distributed models, see [BEL13, BFLR19, BFL<sup>+</sup>17, FS05, FGS<sup>+</sup>13]. Other related work considers generalizations of  $k$ -median and  $k$ -means by either adding capacity constraints [CL19, HJV19, SSS19], or considering

---

<sup>4</sup>Strictly speaking, one has to use a generalization of VC-dimension known as the pseudo dimension. The interested reader is referred to Pollard's book [Pol12] for details.

more general objective functions [BLL18, BJKW19]. Coresets have also been studied for many other problems: we cite non-comprehensively Determinant Maximization [IMGR20], Diversity Maximization [CPP18, IMMM14] logistic regression [HCB16, MSSW18], dependency networks [MMK18], or low-rank approximation [MJF19].

### 3 Preliminaries

#### 3.1 Problem Definitions

Given an ambient metric space  $(X, \text{dist})$ , a set of points  $P \subseteq X$  called *clients*, and positive integers  $k$  and  $z$ , the goal of the  $(k, z)$ -clustering problem is to output a set  $\mathcal{S}$  of  $k$  centers (or *facilities*) chosen in  $X$  that minimizes

$$\sum_{p \in P} \min_{c \in \mathcal{S}} (\text{dist}(p, c))^z$$

**Definition 2.** An  $\varepsilon$ -coreset for the  $(k, z)$ -clustering problem in a metric space  $(X, \text{dist})$  is a weighted subset  $\Omega$  of  $X$  with weights  $w : \Omega \rightarrow \mathbb{R}_+$  such that, for any set  $\mathcal{S} \subset X$ ,  $|\mathcal{S}| = k$ ,

$$\left| \sum_{p \in X} \text{cost}(p, \mathcal{S}) - \sum_{p \in \Omega} w(p) \text{cost}(p, \mathcal{S}) \right| \leq \varepsilon \cdot \sum_{p \in X} \text{cost}(p, \mathcal{S}).$$

Given a set of point  $P$  with weights  $w : P \rightarrow \mathbb{R}^+$  on a metric space  $I = (X, \text{dist})$  and a solution  $\mathcal{S}$ , we define  $\text{cost}(P, \mathcal{S}) := \sum_{p \in P} w(p) \text{cost}(p, \mathcal{S})$  and, in the case where  $P$  contains all the points of the metric space, we define  $\text{cost}(\mathcal{S}) := \text{cost}(P, \mathcal{S})$ .

We will also make use of the following lemma, to have a weaker version of the triangle inequality for  $k$ -Means and more general distances. Proofs of this lemma (and variants thereof) can be found in [BBC<sup>+</sup>19, CS17, FSS20, MMR19, SW18]. For completeness, we provide a proof in the appendix.

**Lemma 1** (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  $\varepsilon > 0$*

$$\begin{aligned} d(a, b)^z &\leq (1 + \varepsilon)^{z-1} d(a, c)^z + \left( \frac{1 + \varepsilon}{\varepsilon} \right)^{z-1} d(b, c)^z \\ |d(a, S)^z - d(b, S)^z| &\leq \varepsilon \cdot d(a, S)^z + \left( \frac{2z + \varepsilon}{\varepsilon} \right)^{z-1} d(a, b)^z. \end{aligned}$$

#### 3.2 From Weighted to Unweighted Inputs

We start by showing a simple reduction from weighted to unweighted inputs. Essentially, we convert a point with weight  $w$  to  $w$  copies of the point.

**Corollary 2.** *Let  $\varepsilon, \pi > 0$ . Let  $(X, \text{dist})$  be a metric space,  $P$  a set of clients with weights  $w : P \rightarrow \mathbb{R}^+$  and two positive integers  $k$  and  $z$ . Let also  $\mathcal{A}$  be a constant-factor approximation for  $(k, z)$ -clustering on  $P$  with weights.*

*Suppose there exists a  $\mathcal{A}$ -approximate centroid set, denoted  $\mathbb{C}$ . Then, there exists an algorithm running in time  $O(|P|)$  that constructs with probability at least  $1 - \pi$  a positively-weighted coreset*

of size

$$O\left(\frac{2^{O(z \log z)} \cdot \log^4 1/\varepsilon}{\min(\varepsilon^3, \varepsilon^z)} (k \log |\mathbb{C}| + \log \log(1/\varepsilon) + \log(1/\pi))\right)$$

for the  $(k, z)$ -clustering problem on  $P$  with weights.

*Proof.* We start by making all weights integers: let  $w_{\min} = \min_{p \in P} w(p)$ , and  $\tilde{w}(p) = \left\lfloor 2 \frac{w(p)}{\varepsilon w_{\min}} \right\rfloor$ . This definition ensures that

$$\forall p, |w(p) - \frac{\varepsilon w_{\min}}{2} \cdot \tilde{w}(p)| \leq \frac{\varepsilon}{2} w_{\min} \leq \frac{\varepsilon}{2} w(p).$$

We denote  $\tilde{P}$  the set of points  $P$  with weight  $\tilde{w}$ . First, we note that for any solution  $\mathcal{S}$ ,

$$\left| \text{cost}(P, \mathcal{S}) - \varepsilon w_{\min} \text{cost}(\tilde{P}, \mathcal{S}) \right| \leq \frac{\varepsilon}{2} \text{cost}(P, \mathcal{S}).$$

Hence, it is enough to find an  $\varepsilon/2$ -coreset for  $\tilde{P}$ , and then scale the coreset weights of the coreset points by  $\varepsilon w_{\min}/2$ . We have that the weights in  $\tilde{P}$  are integers: a weighted point can therefore be considered as multiple copies of the same points.

By the previous equation,  $\mathcal{A}$  is a constant-factor approximation for  $\tilde{P}$  as well. The definition of a centroid set does not depend on weights, so  $\mathbb{C}$  is a  $\mathcal{A}$ -centroid set for  $\tilde{P}$  as well. Hence, we can apply Theorem 1 on  $\tilde{P}$  and scale the resulting coreset by  $\varepsilon w_{\min}/2$  to conclude the proof.  $\square$

### 3.3 Partitioning an Instance into Groups: Definitions

As sketched, the algorithm partitions the input points into structured groups. We give here the useful definitions.

Fix a metric space  $I = (X, \text{dist})$ , positive integers  $k, z$  and a set of clients  $P$ . For a solution  $\mathcal{S}$  of  $(k, z)$ -clustering on  $P$  and a center  $c \in \mathcal{S}$ ,  $c$ 's cluster consists of all points closer to  $c$  than to any other center of  $\mathcal{S}$ .

Fix as well some  $\varepsilon > 0$ , and let  $\mathcal{A}$  be any solution for  $(k, z)$ -clustering on  $P$  with  $k$  centers. Let  $C_1, \dots, C_k$  be the clusters induced by the centers of  $\mathcal{A}$ .

- the average cost of a cluster  $C_i$  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 ring*  $R_I(C_i) := \cup_{j \leq 2z \log(\varepsilon/z)} R_{i,j}$  (resp. *outer ring*  $R_O(C_i) := \cup_{j > 2z \log(z/\varepsilon)} R_{i,j}$ ) of a cluster  $C_i$  consists of the points of  $C_i$  with cost at most  $(\varepsilon/z)^{2z} \Delta_{C_i}$  (resp. at least  $(z/\varepsilon)^{2z} \Delta_{C_i}$ ). The *main ring*  $R_M(C_i)$  consists of all the other points of  $C_i$ . For a solution  $\mathcal{S}$ , we let  $R_I^{\mathcal{S}}$  and  $R_O^{\mathcal{S}}$  be the union of inner and outer rings of the clusters induced by  $\mathcal{S}$ .
- for each  $j$ ,  $R_j$  is defined to be  $\cup_{i=1}^k R_{i,j}$ .

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

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

- For any  $j$ , let  $G_{j,\min} := \cup_{b \leq 0} G_{j,b}$  be the union of the cheapest groups, and  $G_{j,\max} := \cup_{b \geq z \log \frac{4z}{\varepsilon}} G_{j,b}$  be the union of the most expensive ones. The set of interesting groups is made of  $G_{j,\min}$ ,  $G_{j,\max}$ , and  $G_{j,b}$  for all  $0 < b < z \log \frac{4z}{\varepsilon}$ .
- The set of outer rings is also partitioned into *outer groups*:

$$G_b^O = \left\{ p \mid \exists i, p \in C_i \text{ and } \left( \frac{\varepsilon}{4z} \right)^z \cdot \frac{\text{cost}(R_O^A, \mathcal{A})}{k} \cdot 2^b \leq \text{cost}(R_O(C_i), \mathcal{A}) \leq \left( \frac{\varepsilon}{4z} \right)^z \cdot 2^{b+1} \cdot \frac{\text{cost}(R_O^A, \mathcal{A})}{k} \right\}.$$

- We let as well  $G_{\min}^O = \cup_{b \leq 0} G_b^O$  and  $G_{\max}^O = \cup_{b \geq z \log \frac{4z}{\varepsilon}} G_b^O$ . The interesting outer groups are  $G_{\min}^O, G_{\max}^O$  and all  $G_b^O$  with  $0 < b < z \log \frac{4z}{\varepsilon}$ .

Intuitively, grouping points by groups is helpful, as all points in the same ring can pay the same additive error. Since there are very few groups, it turns out possible to construct a coreset for each group, and then take the union of the group's coreset. This is essentially the algorithm we propose.

We note few facts about the partitioning:

**Fact 1.** *There exist at most  $O(z \log(z/\varepsilon))$  many non-empty  $R_j$  that are not in some inner or outer ring, i.e., not in  $R_I^A$  nor in  $R_O^A$ .*

Hence, the number of different non-empty groups is bounded as well:

**Fact 2.** *There exists at most  $O(z^2 \log^2(z/\varepsilon))$  many interesting  $G_{j,b}$ .*

This is simply due to the fact that  $j$  can take only interesting values between  $2z \log(\varepsilon/z)$  and  $2z \log(z/\varepsilon)$ , and interesting  $b$  between 0 and  $z \log(4z/\varepsilon)$ .

By the definition of the outer groups, we have also that

**Fact 3.** *There exists at most  $O(z \log(z/\varepsilon))$  many interesting outer groups.*

For simplicity, we will drop mention of "interesting" : when considering any group, it will implicitly be an interesting group.

## 4 The Coreset Construction Algorithm, and Proof of Theorem 1

### 4.1 The algorithm

For an initial metric space  $(X, \text{dist})$ , set of clients  $P$  and  $\varepsilon > 0$ , our algorithm essentially consists of the following steps: given a solution  $\mathcal{A}$ , it processes the input in order to reduce the number of different groups. Then, the algorithm computes a coreset of the points inside each group using the following **GroupSample** procedure. The final coreset is made of the union of the coresets for all groups.

The **GroupSample** procedure takes as input a group of points  $G$  as defined in Section 3.3, a set of centers  $\mathcal{A}$  inducing clusters  $\tilde{C}_1, \tilde{C}_2, \dots, \tilde{C}_k$  on  $G$  and an integer  $\delta$ . Note importantly that the definition of clusters  $\tilde{C}_i$  says that they are only made of points from the group  $G$ . The output of **GroupSample** is a set of weighted points, computed as follows: a point  $p \in \tilde{C}_i$  is sampled with probability  $\frac{\delta \cdot \text{cost}(\tilde{C}_i, \mathcal{A})}{|\tilde{C}_i| \cdot \text{cost}(G, \mathcal{A})}$ , and the weight of any sampled point is rescaled by a factor  $\frac{|\tilde{C}_i| \cdot \text{cost}(G, \mathcal{A})}{\delta \cdot \text{cost}(\tilde{C}_i, \mathcal{A})}$ .<sup>5</sup>

The properties of the **GroupSample** procedure are captured by the following lemma.

**Lemma 2.** *Let  $(X, \text{dist})$  be a metric space,  $k, z$  be two positive integers and  $G$  be a group of clients and  $\mathcal{A}$  be a solution to  $(k, z)$ -clustering on  $G$  with  $k$  centers such that:*

- *for every cluster  $\tilde{C}$  induced by  $\mathcal{A}$  on  $G$ , all points of  $\tilde{C}$  have the same cost in  $\mathcal{A}$ , up to a factor 2:  $\forall p, q \in \tilde{C}, \text{cost}(p, \mathcal{A}) \leq 2\text{cost}(q, \mathcal{A})$ .*
- *for all clusters  $\tilde{C}$  induced by  $\mathcal{A}$  on  $G$ , it holds that  $\frac{\text{cost}(G, \mathcal{A})}{2k} \leq \text{cost}(\tilde{C}, \mathcal{A})$ .*

Let  $\mathbb{C}$  be a  $\mathcal{A}$ -approximate centroid set for  $(k, z)$ -clustering on  $G$ .

Then, there exists an algorithm **GroupSample**, running in time  $O(|G|)$  that constructs a set  $\Omega$  of size  $\delta$  such that, with probability  $1 - \exp\left(k \log |\mathbb{C}| - 2^{O(z \log z)} \cdot \frac{\min(\varepsilon^2, \varepsilon^z)}{\log^2 1/\varepsilon} \cdot \delta\right)$  it holds that for all set  $\mathcal{S}$  of  $k$  centers:

$$|\text{cost}(G, \mathcal{S}) - \text{cost}(\Omega, \mathcal{S})| = O(\varepsilon) (\text{cost}(G, \mathcal{S}) + \text{cost}(G, \mathcal{A})).$$

We further require the **SensitivitySample** procedure, which we will apply to some of the points not consider by the calls to **GroupSample**. From a group  $G$ , this procedure merely picks  $\delta$  points  $p$  with probability  $\frac{\text{cost}(p, \mathcal{A})}{\text{cost}(G, \mathcal{A})}$ . Each of the  $\delta$  sampled points has a weight  $\frac{\text{cost}(G, \mathcal{A})}{\delta \cdot \text{cost}(p, \mathcal{A})}$ .

The key property of **SensitivitySample** is given in the following lemma.

**Lemma 3.** *Let  $(X, \text{dist})$  be a metric space,  $k, z$  be two positive integers,  $P$  be a set of clients and  $\mathcal{A}$  be a  $c_{\mathcal{A}}$ -approximate solution solution to  $(k, z)$ -clustering on  $P$ .*

Let  $G$  be either a group  $G_b^O$  or  $G_{\max}^O$ . Suppose moreover that there is a  $\mathcal{A}$ -approximate centroid set  $\mathbb{C}$  for  $(k, z)$ -clustering on  $G$ .

Then, there exists an algorithm **SensitivitySample** running in time  $O(|G|)$  that constructs a set  $\Omega$  of size  $\delta$  such that it holds with probability  $1 - \exp\left(k \log |\mathbb{C}| - 2^{O(z \log z)} \cdot \frac{\varepsilon^2}{\log^2 1/\varepsilon} \cdot \delta\right)$  that, for all sets  $\mathcal{S}$  of  $k$  centers:

$$|\text{cost}(G, \mathcal{S}) - \text{cost}(\Omega, \mathcal{S})| = \frac{\varepsilon}{z \log z / \varepsilon} \cdot (\text{cost}(\mathcal{S}) + \text{cost}(\mathcal{A})).$$

An interesting feature of Lemma 3 is that the probability does not depend on  $\varepsilon^{-z}$ , as it does in Lemma 2.

Using the two algorithms **GroupSample** and **SensitivitySample**, we can formally present the whole algorithm:

---

<sup>5</sup>Note that this is essentially importance sampling, as each point in a cluster  $\tilde{C}_i$  have cost roughly equal to the average. We chose this different distribution for simplicity in our proofs.

**Input:** A metric space  $(X, \text{dist})$ , a set  $P \subseteq X$ ,  $k, z > 0$ , a solution  $\mathcal{A}$  to  $(k, z)$ -clustering on  $P$ , and  $\varepsilon$  such that  $0 < \varepsilon < 1/3$ .

**Output:** A coreset. Namely, a set of points  $\Omega \subseteq P \cup \mathcal{A}$  and a weight function  $w : \Omega \mapsto \mathbb{R}_+$  such that for any set of  $k$  centers  $\mathcal{S}$ ,  $\text{cost}(P, \mathcal{S}) = (1 \pm \varepsilon)\text{cost}(\Omega, \mathcal{S})$ .

1. Set the weights of all the centers of  $\mathcal{A}$  to 0.
2. **Partition the remaining instance into groups:**
  - (a) For each cluster  $C$  of  $\mathcal{A}$  with center  $c$ , remove  $R_I(C)$  and increase the weight of  $c$  by  $|R_I(C)|$ .
  - (b) For each cluster  $C$  with center  $c$  in solution  $\mathcal{A}$ , the algorithm discards also all of  $C \cap \cup_j G_{j, \min}$  and  $R_O(C) \cap G_{\min}^O$ , and increases the weight of  $c$  by the number of points discarded in cluster  $c$ .
  - (c) Let  $\mathcal{D}$  be the set of points discarded at those steps, and  $P_1$  be the weighted set of centers that have positive weights.
3. **Sampling from well structured groups:** For every  $j$  such that  $z \log(\varepsilon/z) \leq j \leq 2z \log(z/\varepsilon)$  and every group  $G_{j,b} \notin G_{j, \min}$ , compute a coreset  $\Omega_{j,b}$  of size

$$\delta = O \left( \frac{\log^2 1/\varepsilon}{2^{O(z \log z)} \min(\varepsilon^2, \varepsilon^z)} (k \log |\mathbb{C}| + \log \log(1/\varepsilon) + \log(1/\pi)) \right)$$

using the **GroupSample** procedure.

4. **Sampling from the outer rings:** From each group  $G_1^O, \dots, G_{\max}^O$ , compute a coreset  $\Omega_b^O$  of size

$$\delta = O \left( \frac{2^{O(z \log z)} \cdot \log^2(1/\varepsilon)}{\varepsilon^2} (k \log |\mathbb{C}| + \log \log(1/\varepsilon) + \log(1/\pi)) \right)$$

using the **SensitivitySample** procedure.

5. **Output:**

- A coreset consisting of  $\mathcal{A} \cup_{j,b} \Omega_{j,b} \cup_i \Omega_i^O$ .
- Weights: weights for  $\mathcal{A}$  defined throughout the algorithm, weights for  $\Omega_{j,b}$  defined by the **GroupSample** procedure, weights for  $\Omega_O$  defined by the **SensitivitySample** procedure.

**Remark 1.** *Instead of using the **GroupSample** procedure, one could use any coreset construction tailored for the well structured group. Improving on that step would improve the final coreset bound: if the size of the coreset produced for a group is  $T$ , then the total coreset has size*

$$\tilde{O} \left( T + \frac{2^{O(z \log z)}}{\varepsilon^2} \cdot k \log |\mathbb{C}| \right)$$

## 4.2 Proof of Theorem 1

As we prove in Section 7, the outcome of the partitioning step,  $\mathcal{D}$  and  $P_1$ , satisfies the following lemma, that deals with the inner ring, and the groups  $G_{j, \min}$  and  $G_{\min}^O$ :

**Lemma 4.** Let  $(X, \text{dist})$  be a metric space with a set of clients  $P$ ,  $k, z$  be two positive integers, and  $\varepsilon \in \mathbb{R}_+^*$ . For every solution  $\mathcal{S}$ , it holds that

$$|\text{cost}(\mathcal{D}, \mathcal{S}) - \text{cost}(P_1, \mathcal{S})| = O(\varepsilon) \text{cost}(\mathcal{S}),$$

where  $\mathcal{D}$  and  $P_1$  are defined in Step 2 of the algorithm.

Combining properties of the partitioning, Lemma 2, Lemma 3 and Lemma 4 allows to prove Theorem 1:

*Proof of Theorem 1.* Let  $\Omega$  be the output of the algorithm described above, and  $\delta = O\left(\frac{\log^2 1/\varepsilon}{2^{O(z \log z)} \min(\varepsilon^2, \varepsilon z)} (k \log |\mathbb{C}| + \log \log(1/\varepsilon) + \log(1/\pi))\right)$  as defined in step 3 of the algorithm. Due to Fact 2 and Fact 3,  $\Omega$  has size  $O(z^2 \log^2(z/\varepsilon) \cdot \delta + |\mathcal{A}|)$ , and non-negative weights by construction.

We now turn to analysing the quality of the coreset. Any group  $G_{j,b}$  for  $b > 0$  satisfies Lemma 2: the cost of any point  $p \in G_{j,b} \cap C_i$  satisfies  $2^j \Delta_{C_i} \leq \text{cost}(p, \mathcal{A}) \leq 2^{j+1} \Delta_{C_i}$ , and

- for  $b \in \{0, \dots, z \log \frac{4z}{\varepsilon}\}$ , the cost of all clusters induced by  $\mathcal{A}$  on  $G_{j,b}$  are equal up to a factor 2, hence for all  $i$   $\frac{\text{cost}(G_{j,b}, \mathcal{A})}{2^k} \leq \text{cost}(C_i \cap G_{j,b}, \mathcal{A})$
- for  $b = \max$ , it holds that  $\frac{\text{cost}(G_{j,\max}, \mathcal{A})}{2^k} \leq \frac{\text{cost}(R_j, \mathcal{A})}{2^k} \leq \text{cost}(C_i \cap G_{j,\max}, \mathcal{A})$ .

Hence, Lemma 2 ensures that, with probability  $1 - \exp\left(k \log |\mathbb{C}| - 2^{O(z \log z)} \cdot \frac{\min(\varepsilon^2, \varepsilon z)}{\log^2 1/\varepsilon} \cdot \delta\right)$ , the coreset  $\Omega_{j,b}$  constructed for  $G_{j,b}$  satisfies for any solution  $\mathcal{S}$

$$|\text{cost}(G_{j,b}, \mathcal{S}) - \text{cost}(\Omega_{j,b}, \mathcal{S})| = O(\varepsilon) (\text{cost}(G_{j,b}, \mathcal{S}) + \text{cost}(G_{j,b}, \mathcal{A})).$$

Similarly, Lemma 3 ensures that, with probability  $1 - \exp\left(\log |\mathbb{C}| - 2^{O(z \log z)} \cdot \frac{\varepsilon^2}{\log^2 1/\varepsilon} \cdot \delta\right)$ , the coreset  $\Omega_b^O$  constructed for  $G_b^O$  satisfies for any solution  $\mathcal{S}$

$$|\text{cost}(G_b^O, \mathcal{S}) - \text{cost}(\Omega_b^O, \mathcal{S})| = \frac{\varepsilon}{z \log(z/\varepsilon)} (\text{cost}(\mathcal{S}) + \text{cost}(\mathcal{A})).$$

Taking a union-bound over the failure probability of Lemma 3 and of Lemma 2 applied to all groups  $G_{j,b}$  with  $z \log(\varepsilon/z) \leq j \leq 2z \log(z/\varepsilon)$  and all  $G_i^O$  implies that, with probability

$$1 - z^2 \log^2(z/\varepsilon) \exp\left(k \log |\mathbb{C}| - 2^{O(z \log z)} \cdot \frac{\min(\varepsilon^2, \varepsilon z)}{\log^2 1/\varepsilon} \cdot \delta\right) \\ - z \log(z/\varepsilon) \exp\left(\log |\mathbb{C}| - 2^{O(z \log z)} \frac{\varepsilon^2}{\log^2 1/\varepsilon} \cdot \delta\right)$$

for any solution  $\mathcal{S}$ ,

$$\begin{aligned} & |\text{cost}(\mathcal{S}) - \text{cost}(\Omega, \mathcal{S})| \\ & \leq |\text{cost}(\mathcal{D}, \mathcal{S}) - \text{cost}(P_1, \mathcal{S})| + \sum_{j,b} |\text{cost}(G_{j,b}, \mathcal{S}) - \text{cost}(G_{j,b} \cap \Omega, \mathcal{S})| \\ & \quad + \sum_i |\text{cost}(G_b^O, \mathcal{S}) - \text{cost}(G_b^O \cap \Omega, \mathcal{S})| \\ & \leq O(\varepsilon) \text{cost}(\mathcal{S}) + O(\varepsilon) \text{cost}(\mathcal{A}) \leq O(\varepsilon) \text{cost}(\mathcal{S}) \end{aligned}$$



where the penultimate inequality uses Lemma 4, and the last one that  $\mathcal{A}$  is a constant-factor approximation.

For  $\delta = \frac{\log^2 1/\varepsilon}{2^{O(z \log z)} \min(\varepsilon^2, \varepsilon^z)}$  ( $k \log |\mathbb{C}| + \log \log(1/\varepsilon) + \log(1/\pi)$ ), this probability can be simplified to

$$1 - \exp \left( 2(\log z + \log \log(z/\varepsilon)) + k \log |\mathbb{C}| - 2^{O(z \log z)} \cdot \frac{\min(\varepsilon^2, \varepsilon^z)}{\log^2 1/\varepsilon} \cdot \delta \right) = 1 - \pi.$$

The complexity of this algorithm is:

- $O(n)$  to compute the groups: given all distances from a client to its center, computing the average cost of all clusters costs  $O(n)$ , hence partitioning into  $R_j$  cost  $O(n)$  as well, and then decomposing  $R_j$  into groups is also done in  $O(n)$  time;
- plus the cost to compute the coreset in the groups, which is  $\sum_{j,b} O(|G_{j,b}|) + \sum_i O(|G_b^O|) = O(n)$

Hence, the total complexity is  $O(n)$ .  $\square$

## 5 Sampling inside Groups: Proof of Lemma 2

The goal of this section is to prove Lemma 2:

**Lemma 2.** *Let  $(X, \text{dist})$  be a metric space,  $k, z$  be two positive integers and  $G$  be a group of clients and  $\mathcal{A}$  be a solution to  $(k, z)$ -clustering on  $G$  with  $k$  centers such that:*

- *for every cluster  $\tilde{C}$  induced by  $\mathcal{A}$  on  $G$ , all points of  $\tilde{C}$  have the same cost in  $\mathcal{A}$ , up to a factor 2:  $\forall p, q \in \tilde{C}, \text{cost}(p, \mathcal{A}) \leq 2\text{cost}(q, \mathcal{A})$ .*
- *for all clusters  $\tilde{C}$  induced by  $\mathcal{A}$  on  $G$ , it holds that  $\frac{\text{cost}(G, \mathcal{A})}{2k} \leq \text{cost}(\tilde{C}, \mathcal{A})$ .*

*Let  $\mathbb{C}$  be a  $\mathcal{A}$ -approximate centroid set for  $(k, z)$ -clustering on  $G$ .*

*Then, there exists an algorithm **GroupSample**, running in time  $O(|G|)$  that constructs a set  $\Omega$  of size  $\delta$  such that, with probability  $1 - \exp \left( k \log |\mathbb{C}| - 2^{O(z \log z)} \cdot \frac{\min(\varepsilon^2, \varepsilon^z)}{\log^2 1/\varepsilon} \cdot \delta \right)$  it holds that for all set  $\mathcal{S}$  of  $k$  centers:*

$$|\text{cost}(G, \mathcal{S}) - \text{cost}(\Omega, \mathcal{S})| = O(\varepsilon) (\text{cost}(G, \mathcal{S}) + \text{cost}(G, \mathcal{A})).$$

### 5.1 Description of the GroupSample Algorithm

The **GroupSample** merely consists of importance sampling in rounds, i.e. there are  $\delta$  rounds in which one point of  $G$  is sampled. Let  $\tilde{C}_1, \tilde{C}_2, \dots$  be the clusters induced by  $\mathcal{A}$  on  $G$ : the probability of sampling point  $p \in \tilde{C}_i$  is  $\frac{\text{cost}(\tilde{C}_i, \mathcal{A})}{|\tilde{C}_i| \cdot \text{cost}(G, \mathcal{A})}$  – recall that all clusters  $\tilde{C}_i$  contain only points from the group  $G$ . The weight of any sampled point is rescaled by a factor  $\frac{|\tilde{C}_i| \cdot \text{cost}(G, \mathcal{A})}{\delta \text{cost}(\tilde{C}_i, \mathcal{A})}$ . If there are  $m$  copies of a point, it is sampled in a round with probability  $\frac{m \cdot \text{cost}(\tilde{C}_i, \mathcal{A})}{|\tilde{C}_i| \cdot \text{cost}(G, \mathcal{A})}$  (which is equivalent to sampling each copy with probability  $\frac{\text{cost}(\tilde{C}_i, \mathcal{A})}{|\tilde{C}_i| \cdot \text{cost}(G, \mathcal{A})}$ ). In what follows, each copies will be considered independently.

**Definition 3.** We denote  $f(p) := \frac{|\tilde{C}_i| \cdot \text{cost}(G, \mathcal{A})}{\delta \text{cost}(\tilde{C}_i, \mathcal{A})}$  the scaling factor of the weight of a point  $p \in \tilde{C}_i$ .

## 5.2 Organization of the Proof

To analyze the sampling procedure of **GroupSample**, we consider different cost ranges  $I_{\ell, \mathcal{S}}$  induced by a solution  $\mathcal{S}$  as follows. A point  $p$  of  $G$  is in  $I_{\ell, \mathcal{S}}$  if  $2^\ell \cdot \text{cost}(p, \mathcal{A}) \leq \text{cost}(p, \mathcal{S}) \leq 2^{\ell+1} \cdot \text{cost}(p, \mathcal{A})$ . We distinguish between the following cases.

- $\ell \leq \log \varepsilon / 2$ . We call all  $I_{\ell, \mathcal{S}}$  in this range *tiny*. The union of all tiny  $I_{\ell, \mathcal{S}}$  is denoted by  $I_{\text{tiny}, \mathcal{S}}$ .
- $\log \varepsilon / 2 \leq \ell \leq z \log(8z/\varepsilon)$ . We call all  $I_{\ell, \mathcal{S}}$  in this range *interesting*.
- $\ell \geq z \log(4z/\varepsilon)$ . We call all  $I_{\ell, \mathcal{S}}$  in this range *huge*.

Note that interesting and huge ranges intersect. This is to give us some slack in the proof: for a solution  $\mathcal{S}$ , we will deal with huge ranges before relating  $\mathcal{S}$  to its representative  $\tilde{\mathcal{S}}$  from  $\mathbb{C}^k$ . Due to the approximation, some non-huge range for  $\mathcal{S}$  can become huge for  $\tilde{\mathcal{S}}$ : however, due to our definition, they stay in the interesting ranges.

A simple observation leads to the next fact.

**Fact 4.** Given a solution  $\mathcal{S}$ , there are at most  $O(z \log z / \varepsilon)$  interesting  $I_{\ell, \mathcal{S}}$ .

Bounding the difference in cost of  $G \cap I_{\ell, \mathcal{S}}$  requires different arguments depending on the type of  $I_{\ell, \mathcal{S}}$ . The two easy cases are tiny and huge, so we will first proceed to prove those. Proving the interesting case is arguably both the main challenge and our main technical contribution.

For the proof, we will rely on Bernstein's concentration inequality:

**Theorem 3** (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

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

In this paper we will simply drop the  $\mathbb{E}[X_i]^2$  terms from the denominator, as the second moment will dominate in all important cases.

In what follows, we fix  $k, z, G$  and  $\mathcal{A}$ , as in the assumptions of Lemma 2. Let  $\tilde{C}_1, \dots, \tilde{C}_k$  be the clusters induced by  $\mathcal{A}$  on  $G$ . The assumptions imply the following fact:

**Fact 5.** For any  $p \in \tilde{C}_i$ ,  $\frac{\text{cost}(\tilde{C}_i, \mathcal{A})}{2|\tilde{C}_i|} \leq \text{cost}(p, \mathcal{A}) \leq \frac{2\text{cost}(\tilde{C}_i, \mathcal{A})}{|\tilde{C}_i|}$ .

We will start with the tiny type, as it is mostly divorced from the others.

## 5.3 Dealing with Tiny Type

**Lemma 5.** It holds that, for any solution  $\mathcal{S}$ ,

$$\max \left( \sum_{p \in I_{\text{tiny}, \mathcal{S}}} \text{cost}(p, \mathcal{S}), \sum_{p \in I_{\text{tiny}, \mathcal{S}} \cap \Omega} f(p) \text{cost}(p, \mathcal{S}) \right) \leq \varepsilon \cdot \text{cost}(G, \mathcal{A}).$$

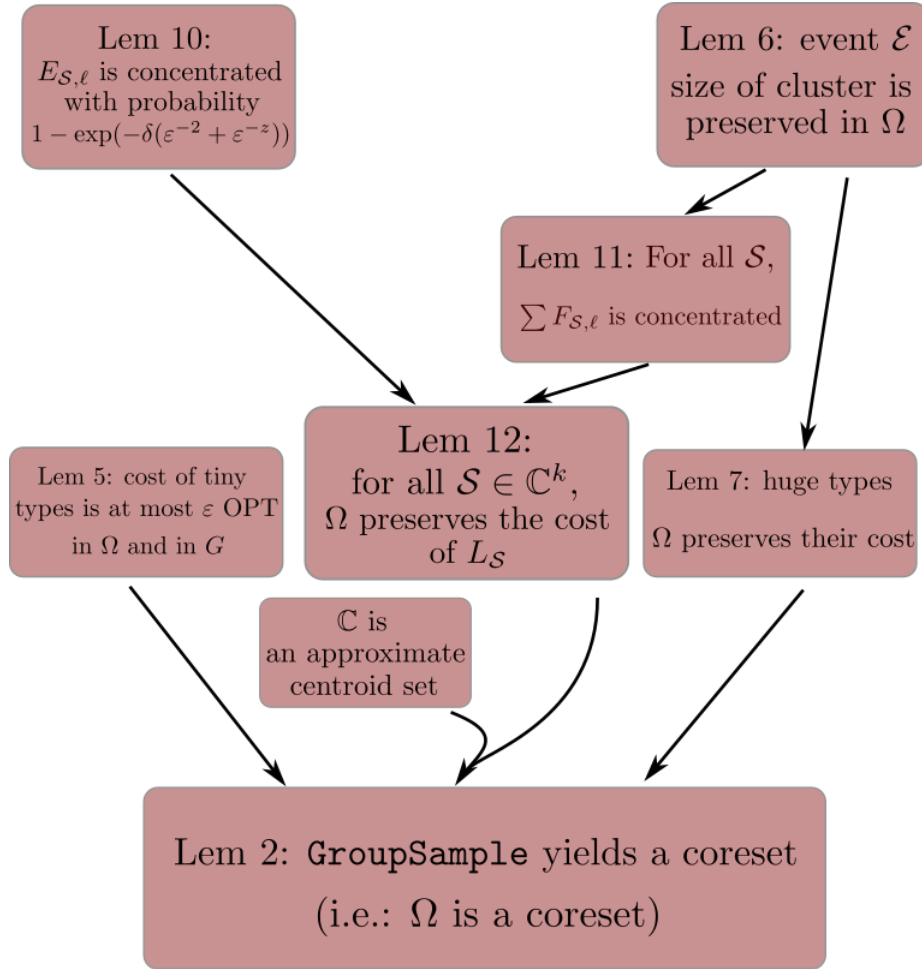


Figure 1: Arrangement of Lemmas of Section 5 to prove Lemma 2.

*Proof.* By definition of  $I_{\text{tiny}, \mathcal{S}}$ ,  $\sum_{p \in I_{\text{tiny}, \mathcal{S}}} \text{cost}(p, \mathcal{S}) \leq \sum_{p \in I_{\text{tiny}, \mathcal{S}}} \frac{\varepsilon}{2} \cdot \text{cost}(p, \mathcal{A}) \leq \frac{\varepsilon}{2} \cdot \text{cost}(G, \mathcal{A})$ . Similarly, we have for the other term

$$\begin{aligned}
\sum_{p \in I_{\text{tiny}, \mathcal{S}} \cap \Omega} f(p) \cdot \text{cost}(p, \mathcal{S}) &\leq \sum_{p \in I_{\text{tiny}, \mathcal{S}} \cap \Omega} f(p) \frac{\varepsilon}{2} \cdot \text{cost}(p, \mathcal{A}) \\
&\leq \frac{\varepsilon}{2} \sum_{i=1}^k \sum_{p \in \tilde{C}_i \cap I_{\text{tiny}, \mathcal{S}} \cap \Omega} \frac{|\tilde{C}_i| \cdot \text{cost}(G, \mathcal{A})}{\delta \text{cost}(\tilde{C}_i, \mathcal{A})} \cdot \frac{2 \cdot \text{cost}(\tilde{C}_i, \mathcal{A})}{|\tilde{C}_i|} \\
&\leq \varepsilon \cdot \frac{|I_{\text{tiny}, \mathcal{S}} \cap \Omega|}{\delta} \text{cost}(G, \mathcal{A}) \\
&\leq \varepsilon \cdot \text{cost}(G, \mathcal{A}).
\end{aligned}$$

where the last inequality uses that  $\Omega$  contains  $\delta$  points.  $\square$

#### 5.4 Preserving the Weight of Clusters, and the Huge Type

We now consider the huge ranges. For this, we first show that, given we sampled enough points,  $|\tilde{C}_i|$  is well approximated for every cluster  $\tilde{C}_i$ . This lemma will also be used later for the interesting points. We define event  $\mathcal{E}$  to be: For all cluster  $\tilde{C}_i$  induced by  $\mathcal{A}$  on  $G$ ,

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

**Lemma 6.** *We have that with probability at least  $1 - k \cdot z^2 \log^2(z/\varepsilon) \exp\left(-O(1)\frac{\varepsilon^2}{k}\delta\right)$ , event  $\mathcal{E}$  happens.*

*Proof.* Consider any cluster  $\tilde{C}_i$  induced by  $\mathcal{A}$  on  $G$ . The expected number of points sampled from  $\tilde{C}_i$  is then at least

$$\mu_i := \sum_{p \in \tilde{C}_i} \frac{\delta \text{cost}(\tilde{C}_i, \mathcal{A})}{|\tilde{C}_i| \cdot \text{cost}(G, \mathcal{A})} = \frac{\delta \text{cost}(\tilde{C}_i, \mathcal{A})}{\text{cost}(G, \mathcal{A})} \geq \frac{\delta}{2k},$$

where the inequality holds by assumption on  $G$ . Define the indicator variable of point  $p$  from the sample being drawn from  $\tilde{C}_i$  as  $P_i(p)$ . Using Chernoff bounds, we therefore have

$$\mathbb{P} \left[ \left| \sum_{p \in G \cap \Omega} P_i(p) - \mu_i \right| \geq \varepsilon \cdot \mu_i \right] \leq \exp\left(-\frac{\varepsilon^2 \cdot \mu_i}{3}\right) \leq \exp\left(-\frac{\varepsilon^2 \delta}{6k}\right). \quad (4)$$

Now, rescaling  $P_i(p)$  by a factor  $\frac{|\tilde{C}_i| \cdot \text{cost}(G, \mathcal{A})}{\delta \text{cost}(\tilde{C}_i, \mathcal{A})}$  implies that approximating  $\mu_i$  up to a  $(1 \pm \varepsilon)$  factor also approximates  $|\tilde{C}_i|$  up to a  $(1 \pm \varepsilon)$  factor.

The final result follows by applying a union bound for all clusters in all groups.  $\square$

We now show that for any cluster  $\tilde{C}_i$  with a non-empty huge range, Lemma 6 implies that the cost is well approximated – without the need of going through the approximate solution  $\tilde{\mathcal{S}}$ .

**Lemma 7.** *Condition on event  $\mathcal{E}$ . Then, for any solution  $\mathcal{S}$ , and any  $i$  such that there exists  $\ell \geq z \log(4z/\varepsilon)$  and a point  $p \in \tilde{C}_i$  with  $\text{cost}(p, \mathcal{S}) \geq 2^\ell \text{cost}(p, \mathcal{A})$ , we have:*

$$\left| \text{cost}(\tilde{C}_i, \mathcal{S}) - \sum_{p \in \Omega \cap \tilde{C}_i} f(p) \cdot \text{cost}(p, \mathcal{S}) \right| \leq 7\varepsilon \cdot \text{cost}(\tilde{C}_i, \mathcal{S}).$$

*Proof.* Let  $p \in \tilde{C}_i$  as given in the statement. Using the structure of clusters in a group, this implies for any  $q \in \tilde{C}_i$ :  $\text{cost}(p, q) \leq (\text{dist}(p, \mathcal{A}) + \text{dist}(q, \mathcal{A}))^z \leq 3^z \cdot \text{cost}(p, \mathcal{A}) \leq 3^z \cdot 2^{(\ell - z \log(4z/\varepsilon))} \text{cost}(p, \mathcal{A}) \leq (3\varepsilon/4z)^z \cdot \text{cost}(p, \mathcal{S})$ . By Lemma 1, we therefore have for any point  $q \in \tilde{C}_i$

$$\begin{aligned} \text{cost}(p, \mathcal{S}) &\leq (1 + \varepsilon/2z)^{z-1} \text{cost}(q, \mathcal{S}) + (1 + 2z/\varepsilon)^{z-1} \text{cost}(p, q) \\ &\leq (1 + \varepsilon) \text{cost}(q, \mathcal{S}) + \varepsilon \cdot \text{cost}(p, \mathcal{S}) \\ \Rightarrow \text{cost}(q, \mathcal{S}) &\geq \frac{1 - \varepsilon}{1 + \varepsilon} \text{cost}(p, \mathcal{S}) \geq (1 - 2\varepsilon) \text{cost}(p, \mathcal{S}) \end{aligned}$$

By a similar calculation, we can also derive an upper bound of  $\text{cost}(q, \mathcal{S}) \leq \text{cost}(p, \mathcal{S}) \cdot (1 + 2\varepsilon)$ . Hence, we have

$$\begin{aligned} \sum_{q \in \Omega \cap \tilde{C}_i} \frac{|\tilde{C}_i| \cdot \text{cost}(G, \mathcal{A})}{\text{cost}(\tilde{C}_i, \mathcal{A}) \cdot \delta} \cdot \text{cost}(q, \mathcal{S}) &= (1 \pm 2\varepsilon) \cdot \text{cost}(p, \mathcal{S}) \cdot \sum_{q \in \Omega \cap \tilde{C}_i} \frac{|\tilde{C}_i| \cdot \text{cost}(G, \mathcal{A})}{\text{cost}(\tilde{C}_i, \mathcal{A}) \cdot \delta} \\ &= (1 \pm 2\varepsilon) \cdot \text{cost}(p, \mathcal{S}) \cdot (1 \pm \varepsilon) \cdot |\tilde{C}_i| \quad (\text{Event } \mathcal{E}) \\ &= (1 \pm 2\varepsilon) \cdot (1 \pm \varepsilon) \cdot (1 \pm 2\varepsilon) \cdot \text{cost}(\tilde{C}_i, \mathcal{S}) \\ &= (1 \pm 7\varepsilon) \cdot \text{cost}(\tilde{C}_i, \mathcal{S}). \end{aligned}$$

□

## 5.5 Bounding Interesting $I_{\ell, \mathcal{S}}$ : a Simple but Suboptimal Analysis.

Now we move onto the most involved case, presenting first a suboptimal analysis of **GroupSample** for the interesting types. As explained in the introduction, our main goal is to design a good estimator and apply Bernstein's inequality to it.

Since the clusters intersecting a huge  $I_{\ell, \mathcal{S}}$  are dealt with by Lemma 7, we only need to focus on the *interesting clusters*, namely clusters  $\tilde{C}$  that satisfy

$$\nexists p \in \tilde{C} \mid \text{cost}(p, \mathcal{S}) \geq \left( \frac{8z}{\varepsilon} \right)^z \cdot \text{cost}(p, \mathcal{A}). \quad (5)$$

In other words, a clustering is interesting only if it does not have any point in a huge  $I_{\ell, \mathcal{S}}$ . This restriction will be crucial to our analysis. Let  $L_{\mathcal{S}}$  be a set of interesting clusters (possibly not all of them).<sup>6</sup> For simplicity, we will assimilate  $L_{\mathcal{S}}$  and the points contained in the clusters of  $L_{\mathcal{S}}$ .

We present here a first attempt to show that the cost of interesting points is preserved. Although suboptimal, it serves as a good warm-up for our improved bound.

---

<sup>6</sup>We define  $L_{\mathcal{S}}$  to contain only huge clusters but not all of them in order to relate the cost of solutions from the approximate centroid set  $\mathbb{C}$  to the cost of any solution, as it will become clear in Section 5.7.

In this first attempt, we will use the simple estimator  $E(L_S) := \sum_{p \in L_S \cap \Omega} f(p) \text{cost}(p, \mathcal{S})$  as an estimator of the cost for points in  $L_S$ . Note that by choice of the weights  $f(p)$ , this estimator is unbiased:  $\mathbb{E}[E(L_S)] = \sum_{p \in L_S} \text{cost}(p, \mathcal{S})$ , precisely the quantity we seek to estimate.

To show concentration, we rely on Bernstein's inequality from Theorem 3. Hence, the key part of our proof is to bound the variance of the estimator.

**Lemma 8.** *Let  $G$  be a group of points, and  $\mathcal{A}$  be a solution. Let  $\mathbb{C}$  be an  $\mathcal{A}$ -approximate centroid set, as in Definition 1. It holds with probability*

$$1 - \exp\left(k \log |\mathbb{C}| - \frac{\varepsilon^{2+z}}{2^{O(z \log z)} \log^2 1/\varepsilon} \cdot \delta\right)$$

that, for all solution  $\tilde{\mathcal{S}} \in \mathbb{C}^k$  and any set of interesting clusters  $L_{\tilde{\mathcal{S}}}$  induced by  $\mathcal{A}$  on  $G$ :

$$|E(L_S) - \mathbb{E}[E(L_S)]| \leq \frac{\varepsilon}{z \log z / \varepsilon} \cdot \text{cost}(G, \mathcal{A})$$

*Proof.* First, we fix some solution  $\mathcal{S}$  and some set of interesting clusters  $L_S$ , verifying Eq. (5). We express  $E(L_S)$  as a sum of i.i.d variables :  $E(L_S) = \sum_{j=1}^{\delta} X_j$ , where  $X_j = f(\Omega_j) \text{cost}(\Omega_j, \mathcal{S})$  when the  $j$ -th sampled point is  $\Omega_j \in L_S$ ,  $X_j = 0$  otherwise. Recall that, due to Fact 5, the probability that the  $j$ -th sampled point is  $p$  from some cluster  $\tilde{C}$  satisfies  $\mathbb{P}[\Omega_j = p] = \frac{\text{cost}(\tilde{C}, \mathcal{A})}{|\tilde{C}| \cdot \text{cost}(G, \mathcal{A})} \leq \frac{2 \text{cost}(p, \mathcal{A})}{\text{cost}(G, \mathcal{A})}$ . From the same fact,  $f(p) \leq \frac{2 \text{cost}(G, \mathcal{A})}{\delta \text{cost}(p, \mathcal{A})}$ .

We will rely on Bernstein's inequality (Theorem 3). To do this, we need an upper bound on the variance of  $E(L_S)$ , as well as an almost sure upper bound  $M$  on every sample. We first bound  $\mathbb{E}[X_i^2]$ :

$$\begin{aligned} \mathbb{E}[X_i^2] &= \mathbb{E}\left[(f(\Omega_i) \text{cost}(\Omega_i, \mathcal{S}))^2\right] \\ &= \sum_{p \in L_S} (f(p) \text{cost}(p, \mathcal{S}))^2 \Pr[\Omega_i = p] \\ &\leq \sum_{p \in L_S} \text{cost}(p, \mathcal{S}) \cdot \left(\frac{4z}{\varepsilon}\right)^z \cdot \text{cost}(p, \mathcal{A}) \cdot \left(\frac{2 \text{cost}(G, \mathcal{A})}{\delta \text{cost}(p, \mathcal{A})}\right)^2 \frac{2 \text{cost}(p, \mathcal{A})}{\text{cost}(G, \mathcal{A})} \\ &\leq \left(\frac{4z}{\varepsilon}\right)^z \cdot \frac{\text{cost}(G, \mathcal{A})}{\delta^2} \sum_{p \in L_S} \text{cost}(p, \mathcal{S}) \\ &\leq \left(\frac{4z}{\varepsilon}\right)^z \cdot \frac{\text{cost}(G, \mathcal{A}) \text{cost}(G, \mathcal{S})}{\delta^2} \\ &\leq \left(\frac{4z}{\varepsilon}\right)^z \cdot \frac{(\text{cost}(G, \mathcal{A}) + \text{cost}(G, \mathcal{S}))^2}{\delta^2} \end{aligned}$$

Where, in the third line, we upper bounded only one of the  $\text{cost}(p, \mathcal{S})$  by  $(4z/\varepsilon)^z \text{cost}(p, \mathcal{A})$ . Hence, it holds that  $\sum_{i=1}^{\delta} \mathbb{E}[X_i^2] \leq \left(\frac{4z}{\varepsilon}\right)^z \cdot \frac{(\text{cost}(G, \mathcal{A}) + \text{cost}(G, \mathcal{S}))^2}{\delta}$ .

To apply Bernstein's inequality, we also need an upper-bound on the value of  $X_i$ : using  $\text{cost}(p, \mathcal{S}) \leq \left(\frac{4z}{\varepsilon}\right)^z \text{cost}(p, \mathcal{A})$  and  $f(p) \leq \frac{2 \text{cost}(G, \mathcal{A})}{\delta \text{cost}(p, \mathcal{A})}$  we get

$$X_i \leq M := 2^{O(z \log z)} \cdot \varepsilon^{-z} \frac{(\text{cost}(G, \mathcal{A}) + \text{cost}(G, \mathcal{S}))}{\delta}$$

Applying Bernstein's inequality with those bounds on the variance and the value of the  $X_i$ , we then have:

$$\begin{aligned}
& \mathbb{P} \left[ |E(L_{\mathcal{S}}) - \mathbb{E}[E(L_{\mathcal{S}})]| > \frac{\varepsilon}{z \log z / \varepsilon} \cdot (\text{cost}(G, \mathcal{A}) + \text{cost}(G, \mathcal{S})) \right] \\
& \leq \exp \left( - \frac{\frac{\varepsilon^2}{z^2 \log^2 z / \varepsilon} \cdot (\text{cost}(G, \mathcal{A}) + \text{cost}(G, \mathcal{S}))^2}{2 \sum_{i=1}^{\delta} \text{Var}[X_i] + \frac{1}{3} M \cdot \frac{\varepsilon}{z \log z / \varepsilon} \cdot (\text{cost}(G, \mathcal{A}) + \text{cost}(G, \mathcal{S}))} \right) \\
& \leq \exp \left( - \frac{\varepsilon^{2+z}}{2^{O(z \log z)} \log^2 1/\varepsilon} \cdot \delta \right)
\end{aligned}$$

Hence, for a fixed solution  $\mathcal{S}$  and a fixed set of interesting clusters  $L_{\mathcal{S}}$ , it holds with probability  $1 - \exp \left( - \frac{\varepsilon^{2+z}}{2^{O(z \log z)} \log^2 1/\varepsilon} \cdot \delta \right)$  that  $|E(L_{\mathcal{S}}) - \mathbb{E}[E(L_{\mathcal{S}})]| > \frac{\varepsilon}{z \log z / \varepsilon} \cdot (\text{cost}(G, \mathcal{A}) + \text{cost}(G, \mathcal{S}))$ .

Doing a union-bound over the  $\mathbb{C}^k$  many solutions  $\mathcal{S}$  and the  $2^k$  many sets of interesting clusters concludes the lemma: it holds with probability  $1 - \exp \left( k \log \mathbb{C} - \frac{\varepsilon^{2+z}}{2^{O(z \log z)} \log^2 1/\varepsilon} \cdot \delta \right)$  that, for any solution  $\mathcal{S} \in \mathbb{C}^k$  and any set of interesting clusters  $L_{\mathcal{S}}$ ,  $|E(L_{\mathcal{S}}) - \mathbb{E}[E(L_{\mathcal{S}})]| > \frac{\varepsilon}{z \log z / \varepsilon} \cdot (\text{cost}(G, \mathcal{A}) + \text{cost}(G, \mathcal{S}))$ .  $\square$

In order to apply Lemma 8, note that the quantity  $|E(L_{\mathcal{S}}) - \mathbb{E}[E(L_{\mathcal{S}})]|$  is equal to  $|\text{cost}(L_{\mathcal{S}} \cap \Omega, \mathcal{S}) - \text{cost}(L_{\mathcal{S}}, \mathcal{S})|$ , namely the difference between the cost in the full input and the cost in the coreset of points in  $L_{\mathcal{S}}$ .

This lemma is enough to conclude that the outcome of **GroupSample** is a coreset, once combined with Lemmas 5 and 7. To see the end of the proof, one can jump directly to the proof of Lemma 2 (in Section 5.7) and use Lemma 8 instead of Lemma 12. This would give a coreset of size  $\tilde{O}(k\varepsilon^{-2-z})$ , instead of  $\tilde{O}(k\varepsilon^{-\max(2,z)})$ .

## 5.6 Bounding Interesting $I_{\ell, \mathcal{S}}$ : Improved Analysis

The shortcoming of the previous estimator is its huge variance, with dependency in  $\varepsilon^{-z}$ . We present an alternate estimator with small variance, allowing in turn to increase the success probability of the algorithm.

As for the previous estimator, we only need to focus on some interesting clusters  $L_{\mathcal{S}}$ , namely clusters that do not have any point in a huge  $I_{\ell, \mathcal{S}}$  and satisfy Eq. (5), important enough to be recalled here: all clusters in  $L_{\mathcal{S}}$  verify

$$\nexists p \in \tilde{C} \mid \text{cost}(p, \mathcal{S}) \geq \left( \frac{8z}{\varepsilon} \right)^z \cdot \text{cost}(p, \mathcal{A}). \quad (6)$$

### 5.6.1 Designing a Good Estimator: Reducing the Variance

Our first observation is that we can estimate the cost of points in  $I_{\ell, \mathcal{S}} \cap L_{\mathcal{S}}$ , for each  $\ell$  independently, instead of estimating directly the cost of  $L_{\mathcal{S}}$  as in previous section. For them, we will use the following estimator:

**Definition 4.** Let  $G$  be a group of points, and  $\tilde{C}_i$  be the clusters induced by a solution  $\mathcal{A}$  on  $G$ . For a given set of interesting clusters  $L_S$ , we let

$$E_{\ell,S}(L_S) := \sum_{\tilde{C}_i \in L_S} \sum_{p \in \tilde{C}_i \cap I_{\ell,S} \cap \Omega} f(p)(\text{cost}(p, \mathcal{S}) - \text{cost}(q_{i,S}, \mathcal{S})), \quad (7)$$

where  $q_{i,S} = \underset{p \in \tilde{C}_i}{\text{argmin}} \text{cost}(p, \mathcal{S})$ .

$E_{\ell,S}(L_S)$  can be expressed differently:

$$\begin{aligned} E_{\ell,S}(L_S) &= \sum_{\tilde{C}_i \in L_S} \sum_{p \in \tilde{C}_i \cap I_{\ell,S} \cap \Omega} f(p)(\text{cost}(p, \mathcal{S}) - \text{cost}(q_{i,S}, \mathcal{S})) \\ &= \sum_{p \in I_{\ell,S} \cap L_S \cap \Omega} f(p)\text{cost}(p, \mathcal{S}) - F_{\ell,S}(L_S), \\ \text{with } F_{\ell,S}(L_S) &:= \sum_{\tilde{C}_i \in L_S} \sum_{p \in \tilde{C}_i \cap I_{\ell,S} \cap \Omega} f(p)\text{cost}(q_{i,S}, \mathcal{S}) \end{aligned} \quad (8)$$

$F_{\ell,S}(L_S)$  is a random variable whose value depends on the randomly sampled points  $\Omega$  (we will discuss  $F_{\ell,S}(L_S)$  in more detail later).

Note that the expectation of  $E_{\ell,S}(L_S)$  is

$$\begin{aligned} \mathbb{E}[E_{\ell,S}(L_S)] &= \sum_{p \in I_{\ell,S} \cap L_S} \frac{\delta \text{cost}(\tilde{C}_i, \mathcal{G})}{|\tilde{C}_i| \text{cost}(G, \mathcal{G})} \cdot f(p)\text{cost}(p, \mathcal{S}) - \mathbb{E}[F_{\ell,S}(L_S)] \\ &= \sum_{p \in I_{\ell,S} \cap L_S} \frac{\delta \text{cost}(\tilde{C}_i, \mathcal{G})}{|\tilde{C}_i| \text{cost}(G, \mathcal{G})} \cdot \frac{|\tilde{C}_i| \text{cost}(G, \mathcal{G})}{\delta \text{cost}(\tilde{C}_i, \mathcal{G})} \cdot \text{cost}(p, \mathcal{S}) - \mathbb{E}[F_{\ell,S}(L_S)] \\ &= \text{cost}(I_{\ell,S} \cap L_S, \mathcal{S}) - \mathbb{E}[F_{\ell,S}(L_S)], \end{aligned}$$

Now instead of attempting to show directly concentration of all  $\text{cost}(I_{\ell,S} \cap L_S \cap \Omega, \mathcal{S})$ , we will instead show that:

1.  $E_{\ell,S}(L_S)$  is concentrated for all  $\mathcal{S}$ , and
2.  $\sum_{\ell} F_{\ell,S}(L_S)$  is concentrated around its expectation.

The reason for decoupling the two arguments is that  $E_{\ell,S}(L_S)$  has a very small variance, for which few samples are sufficient: each term of the sum has magnitude  $\text{cost}(p, \mathcal{S}) - \text{cost}(q_{i,S}, \mathcal{S})$  instead of simply  $\text{cost}(p, \mathcal{S})$ . This difference is crucial to our analysis. Furthermore, event  $\mathcal{E}$  from Lemma 6 easily leads to a concentration bound on  $F_S(L_S) = \sum_{\ell} F_{\ell,S}(L_S)$ .

To establish the gain in variance obtained by subtracting  $\text{cost}(q_{i,S}, \mathcal{S})$ , we have the following lemma.



**Lemma 9.** Let  $G$  be a group of points, and  $\mathcal{S}$  be an arbitrary solution and  $\tilde{C}_i$  be a cluster induced  $\mathcal{A}$  on  $G$  where all points have same cost, up to a factor 2. Denote by  $q_{i,\mathcal{S}} = \underset{p \in \tilde{C}_i}{\operatorname{argmin}} \operatorname{cost}(p, \mathcal{S})$ .

Then for every interesting range with  $\ell \geq \log \varepsilon / 2$  and every point  $p \in \tilde{C}_i \cap I_{\ell, \mathcal{S}}$ ,

$$w_p := \frac{\operatorname{cost}(p, \mathcal{S}) - \operatorname{cost}(q_{i,\mathcal{S}}, \mathcal{S})}{\operatorname{cost}(q_{i,\mathcal{S}}, \mathcal{A})} \in \left[0, 2^{\ell(1-1/z)} \cdot 2^{O(z \log z)}\right]$$

*Proof.* Let  $w_p = \frac{\operatorname{cost}(p, \mathcal{S}) - \operatorname{cost}(q_{i,\mathcal{S}}, \mathcal{S})}{\operatorname{cost}(q_{i,\mathcal{S}}, \mathcal{A})}$ . By choice of  $q_{i,\mathcal{S}}$ ,  $w_p \geq 0$ , so we consider the upper bound.

We first show useful inequalities, relating the different solutions. Since  $p \in I_{\ell, \mathcal{S}}$ , we have:

$$\begin{aligned} \operatorname{cost}(q_{i,\mathcal{S}}, \mathcal{S}) &\leq \operatorname{cost}(p, \mathcal{S}) \leq 2^{\ell+1} \operatorname{cost}(p, \mathcal{A}) \\ &\leq 2^{\ell+2} \operatorname{cost}(q_{i,\mathcal{S}}, \mathcal{A}), \end{aligned}$$

where the last inequality holds since  $p$  and  $q_{i,\mathcal{S}}$  are in the same cluster and have up to a factor 2 the same cost. We also have that  $\operatorname{cost}(p, q_{i,\mathcal{S}}) \leq 2^{z-1}(\operatorname{cost}(p, \mathcal{A}) + \operatorname{cost}(q_{i,\mathcal{S}}, \mathcal{A})) \leq 3 \cdot 2^{z-1} \operatorname{cost}(q_{i,\mathcal{S}}, \mathcal{A})$ .

Now, using Lemma 1, for any  $\alpha \leq 1$ ,

$$\operatorname{cost}(p, \mathcal{S}) \leq (1 + \alpha/z)^{z-1} \operatorname{cost}(q_{i,\mathcal{S}}, \mathcal{S}) + \left(1 + \frac{z}{\alpha}\right)^{z-1} \operatorname{cost}(p, q_{i,\mathcal{S}})$$

which after rearranging implies

$$\begin{aligned} \operatorname{cost}(p, \mathcal{S}) - \operatorname{cost}(q_{i,\mathcal{S}}, \mathcal{S}) &\leq 2\alpha \cdot \operatorname{cost}(q_{i,\mathcal{S}}, \mathcal{S}) + \left(\frac{2z}{\alpha}\right)^{z-1} \operatorname{cost}(p, q_{i,\mathcal{S}}) \\ &\leq \alpha \cdot 2^{\ell+3} \cdot \operatorname{cost}(q_{i,\mathcal{S}}, \mathcal{A}) + \left(\frac{2z}{\alpha}\right)^{z-1} \cdot 3 \cdot 2^{z-1} \operatorname{cost}(q_{i,\mathcal{S}}, \mathcal{A}) \\ &\leq 2^{z+1} \cdot \left(\alpha \cdot 2^{\ell+3} + \left(\frac{2z}{\alpha}\right)^{z-1}\right) \cdot \operatorname{cost}(q_{i,\mathcal{S}}, \mathcal{A}). \end{aligned}$$

We optimize the final term with respect to  $\alpha$ , which leads to  $\alpha = 2^{-\frac{\ell}{z}}$  (ignoring constants that depend on  $z$ ) and hence an upper bound of

$$\operatorname{cost}(p, \mathcal{S}) - \operatorname{cost}(q_{i,\mathcal{S}}, \mathcal{S}) \leq 2^{O(z \log z)} 2^{\ell(1-1/z)} \cdot \operatorname{cost}(q_{i,\mathcal{S}}, \mathcal{A}).$$

□

### 5.6.2 Concentration of the Estimator $E_{\ell, \mathcal{S}}(L_{\mathcal{S}})$

First, we show that every estimator  $E_{\ell, \mathcal{S}}(L_{\mathcal{S}})$  is tightly concentrated. This follows the lines of the proof of Lemma 8, incorporating carefully the result of Lemma 9.

**Lemma 10.** Let  $G$  be a group of points, and  $\mathcal{A}$  be a solution. Consider an arbitrary solution  $\mathcal{S}$ . Then for any set of interesting clusters  $L_{\mathcal{S}}$  induced by  $\mathcal{A}$  on  $G$ , and any estimator  $E_{\ell, \mathcal{S}}(L_{\mathcal{S}})$  with  $\ell \leq z \log 4z/\varepsilon$ , it holds that:

$$|E_{\ell, \mathcal{S}}(L_{\mathcal{S}}) - \mathbb{E}[E_{\ell, \mathcal{S}}(L_{\mathcal{S}})]| \leq \frac{\varepsilon}{z \log z / \varepsilon} \cdot (\operatorname{cost}(G, \mathcal{A}) + \operatorname{cost}(I_{\ell, \mathcal{S}}, \mathcal{S})),$$

with probability at least

$$1 - \exp\left(-2^{O(z \log z)} \cdot \frac{\min(\varepsilon^2, \varepsilon^z)}{\log^2 1/\varepsilon} \cdot \delta\right).$$

*Proof.* In order to simplify the notations, we drop mention of  $L_S$  and define  $E_{\ell, S} = E_{\ell, S}(L_S)$ .

Lemma 9 allows to write slightly differently  $E_{\ell, S}$ :

$$E_{\ell, S} = \sum_{\tilde{C}_i \in L_S} \sum_{p \in \tilde{C}_i \cap I_{\ell, S} \cap \Omega} f(p) \cdot w_p \text{cost}(q_i, S),$$

with all the weights  $w_p$  are in  $[0, 2^{\ell(1-1/z)} \cdot 2^{O(z \log z)}]$ .

We can also write  $E_{\ell, S}$  as a sum of independent random variables:  $E_{\ell, S} = \sum_{j=1}^{\delta} X_j$ , where  $X_j = f(\Omega_j) \cdot w_{\Omega_j} \text{cost}(q_i, S, \mathcal{A})$  when the  $j$ -th sampled point of  $G$  is  $\Omega_j \in \tilde{C}_i \cap I_{\ell, S} \cap L_S$  and  $X_j = 0$  when  $\Omega_j \notin I_{\ell, S} \cap L_S$ . Recall that, due to Fact 5, the probability that the  $j$ -th sampled point is  $p$ , where  $p \in \tilde{C}_i$  satisfies  $\mathbb{P}[\Omega_j = p] = \frac{\text{cost}(\tilde{C}_i, \mathcal{A})}{|\tilde{C}_i| \cdot \text{cost}(G, \mathcal{A})} \leq \frac{2 \text{cost}(p, \mathcal{A})}{\text{cost}(G, \mathcal{A})}$ . From the same fact,  $f(p) \leq \frac{2 \text{cost}(G, \mathcal{A})}{\delta \text{cost}(p, \mathcal{A})}$ .

We will rely on Bernstein's inequality (Theorem 3). To do this, we need an upper bound on the variance of  $E_{\ell, S}$ , as well as an almost sure upper bound  $M$  on every sample. We first bound  $\mathbb{E}[X_j^2]$ : in the second line, we use that  $\Omega_j$  consists of a single point to move the square inside the sum.

$$\begin{aligned} \mathbb{E}[X_j^2] &= \mathbb{E}\left[\left(f(\Omega_j) \text{cost}(\Omega_j, \mathcal{A}) \cdot w_{\Omega_j, S}\right)^2\right] \\ &= \sum_{p \in I_{\ell, S} \cap L_S} (f(p) \text{cost}(p, \mathcal{A}) \cdot w_{p, S})^2 \cdot \Pr[\Omega_j = p] \\ &\leq \sum_{p \in I_{\ell, S}} \left(\frac{2 \text{cost}(G, \mathcal{A})}{\delta \text{cost}(p, \mathcal{A})} \cdot \text{cost}(p, \mathcal{A}) \cdot w_{p, S}\right)^2 \cdot \Pr[\Omega_j = p] \\ &\leq \sum_{p \in I_{\ell, S}} 2^{2\ell(1-1/z)} \cdot 2^{O(z \log z)} \cdot \frac{\text{cost}^2(G, \mathcal{A})}{\delta^2} \cdot \frac{\text{cost}(p, \mathcal{A})}{\text{cost}(G, \mathcal{A})} \\ &\leq \sum_{p \in I_{\ell, S}} 2^{2\ell(1-1/z)} \cdot 2^{O(z \log z)} \cdot \frac{\text{cost}(G, \mathcal{A})}{\delta^2} \cdot \text{cost}(p, \mathcal{A}), \end{aligned}$$

where the fourth line follows from using Lemma 9 to replace the value of  $w_{p, S}$ .

To bound  $\sum_{p \in I_{\ell, S}} \text{cost}(p, \mathcal{A})$ , we need to deal with the cases  $z = 1$  (i.e.  $k$ -median) and  $z \geq 2$  ( $k$ -means and higher powers) separately. For the former, we have  $2^{2\ell(1-1/1)} = 1$ , so we can use  $\sum_{p \in I_{\ell, S}} \text{cost}(p, \mathcal{A}) \leq \text{cost}(G, \mathcal{A})$  as an upper bound. For the latter, we use  $\sum_{p \in I_{\ell, S}} 2^\ell \cdot \text{cost}(p, \mathcal{A}) \leq \text{cost}(I_{\ell, S}, S)$  as an upper bound. Combining this with  $\text{Var}[X_i] \leq \mathbb{E}[X_i^2]$ , we obtain for  $z = 1$ :

$$\text{Var}[X_i] \leq \frac{\text{cost}(G, \mathcal{A})}{\delta^2} \cdot 2^{O(z \log z)} \cdot \text{cost}(G, \mathcal{A}), \quad (9)$$

and for  $z > 1$ :

$$\text{Var}[X_i] \leq \frac{\text{cost}(G, \mathcal{A})}{\delta^2} \cdot 2^{O(z \log z)} 2^{\ell(1-2/z)} \text{cost}(I_{\ell, S}, S). \quad (10)$$

The almost sure upper bound (for which no case distinction is required) can be derived similarly , using  $X_i \leq \sup \frac{2\text{cost}(G, \mathcal{A})}{\delta \text{cost}(p, \mathcal{A})} \cdot \text{cost}(p, \mathcal{A}) \cdot w_{p, \mathcal{S}}$ :

$$\begin{aligned} X_i &\leq M := 2^{\ell(1-1/z)} \cdot 2^{O(z \log z)} \cdot \frac{\text{cost}(G, \mathcal{A})}{\delta} \\ &\leq \frac{z}{\varepsilon} \cdot 2^{\ell(1-2/z)} \cdot 2^{O(z \log z)} \cdot \frac{\text{cost}(G, \mathcal{A})}{\delta}, \end{aligned} \quad (11)$$

where the inequality holds due to  $\ell \leq z \log(4z/\varepsilon)$ . Applying Bernstein's inequality with Equations 9, 10, and 11, we then have

$$\begin{aligned} &\mathbb{P} \left[ |E_{\ell, \mathcal{S}} - \mathbb{E}[E_{\ell, \mathcal{S}}]| \leq \frac{\varepsilon}{z \log z / \varepsilon} \cdot (\text{cost}(G, \mathcal{A}) + \text{cost}(I_{\ell, \mathcal{S}}, \mathcal{S})) \right] \\ &\leq \exp \left( - \frac{\frac{\varepsilon^2}{z^2 \log^2 z / \varepsilon} \cdot (\text{cost}(G, \mathcal{A}) + \text{cost}(I_{\ell, \mathcal{S}}, \mathcal{S}))^2}{2 \sum_{i=1}^{\delta} \text{Var}[X_i] + \frac{1}{3} M \cdot \frac{\varepsilon}{z \log z / \varepsilon} \cdot (\text{cost}(G, \mathcal{A}) + \text{cost}(I_{\ell, \mathcal{S}}, \mathcal{S}))} \right) \\ &\leq \exp \left( - \frac{\frac{\varepsilon^2}{z^2 \log^2 z / \varepsilon} \cdot \delta}{2^{O(z \log z)} \cdot \begin{cases} 1 & \text{if } z = 1 \\ 2^{\ell(1-2/z)} & \text{if } z \geq 2 \end{cases}} \right) \end{aligned}$$

For  $z = 1$  this becomes  $\exp \left( - \frac{\varepsilon^2 \cdot \delta}{2^{O(z \log z)} \log^2 1 / \varepsilon} \right)$ . For  $z = 2$ , we have  $2^{\ell(1-2/z)} = 1$ , so the same bound as for  $z = 1$ . For  $z > 2$ , we use  $\ell \leq z \log 4z/\varepsilon$ , which implies  $\varepsilon^2 \cdot 2^{-\ell(1-2/z)} \geq \varepsilon^{2+z-2/z} \cdot 2^{-O(z \log z)} = \varepsilon^z \cdot 2^{-O(z \log z)}$ . This yields our final desired bound of

$$\exp \left( - \frac{\min(\varepsilon^2, \varepsilon^z)}{2^{O(z \log z)} \log^2 1 / \varepsilon} \cdot \delta \right).$$

□

### 5.6.3 Concentration of $F_{\ell, \mathcal{S}}(L_{\mathcal{S}})$

We now turn our attention to bounding the random variable  $F_{\ell, \mathcal{S}}(L_{\mathcal{S}})$ . It turns out that bounding

$$F_{\ell, \mathcal{S}}(L_{\mathcal{S}}) = \sum_{\tilde{C}_i \in L_{\mathcal{S}}} \sum_{p \in \tilde{C}_i \cap \Omega \cap I_{\ell, \mathcal{S}}} \text{cost}(q_{i, \mathcal{S}}, \mathcal{S}) \cdot \frac{|\tilde{C}_i| \cdot \text{cost}(G, \mathcal{A})}{\delta \text{cost}(\tilde{C}_i, \mathcal{A})}$$

is rather hard, and in fact no easier than bounding  $\text{cost}(I_{\ell, \mathcal{S}} \cap \Omega, \mathcal{S})$ . Fortunately, this is not necessary, as it turns out that we can merely bound the sum of  $F_{\ell, \mathcal{S}}(L_{\mathcal{S}})$ . We consider the random variable defined as follows:

$$F_{\mathcal{S}}(L_{\mathcal{S}}) = \sum_{\ell \leq z \log(4z/\varepsilon)} F_{\ell, \mathcal{S}}(L_{\mathcal{S}})$$

with expectation

$$\mathbb{E}[F_S(L_S)] = \sum_{\tilde{C}_i \in L_S} \sum_{p \in \tilde{C}_i \cap \Omega} \text{cost}(q_{i,S}, \mathcal{S}) \cdot \frac{|\tilde{C}_i| \cdot \text{cost}(G, \mathcal{A})}{\delta \text{cost}(\tilde{C}_i, \mathcal{A})}$$

Showing that  $F_S(L_S)$  is concentrated is now an almost direct consequence of event  $\mathcal{E}$  from Lemma 6, which says that  $\sum_{p \in \tilde{C}_i \cap \Omega} \frac{|\tilde{C}_i| \cdot \text{cost}(G, \mathcal{A})}{\delta \text{cost}(\tilde{C}_i, \mathcal{A})} = (1 \pm \varepsilon) |\tilde{C}_i|$ .

**Lemma 11.** *Let  $G$  be a group of points, and  $\mathcal{A}$  be a solution. Conditioned on event  $\mathcal{E}$ , we have for all solutions  $\mathcal{S}$  and all sets of interesting clusters  $L_S$  induced by  $\mathcal{A}$  on  $G$ :*

$$|F_S(L_S) - \mathbb{E}[F_S(L_S)]| \leq \varepsilon \cdot \text{cost}(G, \mathcal{S}).$$

*Proof.* Given a solution  $\mathcal{S}$  and any set of interesting clusters  $L_S$  induced by  $\mathcal{A}$  on  $G$ , we have

$$\mathbb{E}[F_S(L_S)] = \sum_{\tilde{C}_i \in L_S} \sum_{p \in \tilde{C}_i} \text{cost}(q_{i,S}, \mathcal{S}) \cdot \frac{|\tilde{C}_i| \cdot \text{cost}(G, \mathcal{A})}{\delta \text{cost}(\tilde{C}_i, \mathcal{A})} \Pr[p \in \Omega] = \sum_{\tilde{C}_i \in L_S} |\tilde{C}_i| \cdot \text{cost}(q_{i,S}, \mathcal{S}).$$

Event  $\mathcal{E}$  ensures that the mass of each cluster is preserved in the coreset, i.e., that  $\sum_{p \in \tilde{C}_i \cap \Omega} \frac{|\tilde{C}_i| \cdot \text{cost}(G, \mathcal{A})}{\delta \text{cost}(\tilde{C}_i, \mathcal{A})} = (1 \pm \varepsilon) \cdot |\tilde{C}_i|$ , for every cluster  $\tilde{C}_i \in L_S$ . Hence

$$F_S(L_S) = \sum_{\tilde{C}_i \in L_S} \sum_{p \in \tilde{C}_i \cap \Omega} \text{cost}(q_{i,S}, \mathcal{S}) \cdot \frac{|\tilde{C}_i| \cdot \text{cost}(G, \mathcal{A})}{\delta \text{cost}(\tilde{C}_i, \mathcal{A})} = (1 \pm \varepsilon) \cdot \mathbb{E}[F_S(L_S)].$$

Now finally observe that since  $q_{i,S}$  was always the point of  $\tilde{C}_i$  whose cost in  $\mathcal{S}$  is the smallest, we have  $\mathbb{E}[F_S(L_S)] \leq \text{cost}(L_S, \mathcal{S}) \leq \text{cost}(G, \mathcal{S})$ .  $\square$

## 5.7 Combining Them All

We can now show that the sample  $\Omega$  indeed verifies Lemma 2. To do that, we naturally follow the structure of previous lemmas, and decompose

$$\left| \text{cost}(G, \mathcal{S}) - \sum_{p \in \Omega} f(p) \cdot \text{cost}(p, \mathcal{S}) \right|$$

into terms for which we can apply Lemmas 5, 7, 10, and 11.

First, we note that the probability of success of Lemma 10 is too small to take a union-bound over its success for all  $\mathcal{S}$ . To cope with that issue, we use the approximate centroid set, in order to relate  $E_{\ell, \mathcal{S}}(L_S)$  to  $E_{\ell, \tilde{\mathcal{S}}}(L_S)$ , where  $\tilde{\mathcal{S}}$  comes from a small set on which union-bounding is possible.

**Lemma 12.** *Let  $G$  be a group of points, and  $\mathcal{A}$  be a solution. Let  $\mathbb{C}$  be an  $\mathcal{A}$ -approximate centroid set, as in Definition 1. It holds with probability*

$$1 - \exp \left( k \log |\mathbb{C}| - 2^{O(z \log z)} \cdot \frac{\min(\varepsilon^2, \varepsilon^z)}{\log^2 1/\varepsilon} \cdot \delta \right)$$

that, for all solution  $\tilde{\mathcal{S}} \in \mathbb{C}^k$  and any set of interesting clusters  $L_{\tilde{\mathcal{S}}}$  induced by  $\mathcal{A}$  on  $G$ :

$$\left| \text{cost}(L_{\tilde{\mathcal{S}}}, \tilde{\mathcal{S}}) - \text{cost}(\Omega \cap L_{\tilde{\mathcal{S}}}, \tilde{\mathcal{S}}) \right| \leq \varepsilon \left( \text{cost}(G, \mathcal{A}) + \text{cost}(L_{\tilde{\mathcal{S}}}, \tilde{\mathcal{S}}) \right).$$

*Proof.* Taking a union-bound over the success of Lemma 10 for all possible  $\tilde{\mathcal{S}} \in \mathbb{C}^k$ , all choice of interesting clusters  $L_{\tilde{\mathcal{S}}}$  and all  $\ell$  such that  $\log(\varepsilon/2) \leq \ell \leq z \log(4z/\varepsilon)$ , it holds with probability  $1 - \exp(k \log |\mathbb{C}|) \exp\left(-2^{O(z \log z)} \cdot \frac{\min(\varepsilon^2, \varepsilon^z)}{\log^2 1/\varepsilon} \cdot \delta\right)$  that, for every  $\tilde{\mathcal{S}} \in \mathbb{C}^k$ ,  $L_{\tilde{\mathcal{S}}}$  and  $\ell$ ,

$$|E_{\ell, \tilde{\mathcal{S}}}(L_{\tilde{\mathcal{S}}}) - \mathbb{E}[E_{\ell, \tilde{\mathcal{S}}}(L_{\tilde{\mathcal{S}}})]| \leq \frac{\varepsilon}{z \log z / \varepsilon} \cdot \left( \text{cost}(G, \mathcal{A}) + \text{cost}(I_{\ell, \tilde{\mathcal{S}}}, \tilde{\mathcal{S}}) \right) \quad (12)$$

For simplicity, we drop again the mention of  $L_{\tilde{\mathcal{S}}}$  and write  $E_{\ell, \tilde{\mathcal{S}}} = E_{\ell, \tilde{\mathcal{S}}}(L_{\tilde{\mathcal{S}}})$ ,  $F_{\tilde{\mathcal{S}}} = F_{\tilde{\mathcal{S}}}(L_{\tilde{\mathcal{S}}})$ . We now condition on that event, together with event  $\mathcal{E}$ . We write:

$$\begin{aligned} & \left| \sum_{p \in L_{\tilde{\mathcal{S}}}} \text{cost}(p, \tilde{\mathcal{S}}) - \sum_{p \in L_{\tilde{\mathcal{S}}} \cap \Omega} f(p) \cdot \text{cost}(p, \tilde{\mathcal{S}}) \right| \\ &= \left| \sum_{p \in L_{\tilde{\mathcal{S}}}} \text{cost}(p, \tilde{\mathcal{S}}) - \mathbb{E}[F_{\tilde{\mathcal{S}}}] + \mathbb{E}[F_{\tilde{\mathcal{S}}}] - F_{\tilde{\mathcal{S}}} + F_{\tilde{\mathcal{S}}} - \sum_{p \in L_{\tilde{\mathcal{S}}} \cap \Omega} f(p) \cdot \text{cost}(p, \tilde{\mathcal{S}}) \right| \\ &\leq \left| \sum_{p \in L_{\tilde{\mathcal{S}}}} \text{cost}(p, \tilde{\mathcal{S}}) - \mathbb{E}[F_{\tilde{\mathcal{S}}}] + F_{\tilde{\mathcal{S}}} - \sum_{p \in L_{\tilde{\mathcal{S}}} \cap \Omega} f(p) \cdot \text{cost}(p, \tilde{\mathcal{S}}) \right| + |\mathbb{E}[F_{\tilde{\mathcal{S}}}] - F_{\tilde{\mathcal{S}}}| \\ &\leq \sum_{\ell < \log \varepsilon / 2} \left| \sum_{p \in I_{\ell, \tilde{\mathcal{S}}} \cap L_{\tilde{\mathcal{S}}}} \text{cost}(p, \tilde{\mathcal{S}}) - \mathbb{E}[F_{\ell, \tilde{\mathcal{S}}}] + F_{\ell, \tilde{\mathcal{S}}} - \sum_{p \in I_{\ell, \tilde{\mathcal{S}}} \cap L_{\tilde{\mathcal{S}}} \cap \Omega} f(p) \cdot \text{cost}(p, \tilde{\mathcal{S}}) \right| \quad (13) \end{aligned}$$

$$\begin{aligned} &+ \sum_{\ell = \log \varepsilon / 2}^{z \log z / 4\varepsilon} \left| \sum_{p \in I_{\ell, \tilde{\mathcal{S}}} \cap L_{\tilde{\mathcal{S}}}} \text{cost}(p, \tilde{\mathcal{S}}) - \mathbb{E}[F_{\ell, \tilde{\mathcal{S}}}] + F_{\ell, \tilde{\mathcal{S}}} - \sum_{p \in I_{\ell, \tilde{\mathcal{S}}} \cap L_{\tilde{\mathcal{S}}} \cap \Omega} f(p) \cdot \text{cost}(p, \tilde{\mathcal{S}}) \right| \quad (14) \\ &+ |\mathbb{E}[F_{\tilde{\mathcal{S}}}] - F_{\tilde{\mathcal{S}}}| \end{aligned}$$

We note that Equation 14 is  $\sum_{\ell = \log \varepsilon / 2}^{z \log z / 4\varepsilon} |E_{\ell, \tilde{\mathcal{S}}} - \mathbb{E}[E_{\ell, \tilde{\mathcal{S}}}]|$  and can be directly bounded using Equation 12. To bound tiny points of Equation 13, we combine Lemma 5 with the observation that

$F_{\ell, \tilde{\mathcal{S}}} \leq \sum_{p \in I_{\ell, \tilde{\mathcal{S}}} \cap \Omega} f(p) \text{cost}(p, \tilde{\mathcal{S}})$ . This gives:

$$\begin{aligned}
& \sum_{\ell < \log \varepsilon / 2} \left| \sum_{p \in I_{\ell, \tilde{\mathcal{S}}} \cap L_{\tilde{\mathcal{S}}}} \text{cost}(p, \tilde{\mathcal{S}}) - \mathbb{E}[F_{\ell, \tilde{\mathcal{S}}}] + F_{\ell, \tilde{\mathcal{S}}} - \sum_{p \in I_{\ell, \tilde{\mathcal{S}}} \cap \Omega} f(p) \cdot \text{cost}(p, \tilde{\mathcal{S}}) \right| \\
& \leq \sum_{\ell < \log \varepsilon / 2} \left( \sum_{p \in I_{\ell, \tilde{\mathcal{S}}}} \text{cost}(p, \tilde{\mathcal{S}}) + \mathbb{E}[F_{\ell, \tilde{\mathcal{S}}}] + F_{\ell, \tilde{\mathcal{S}}} + \sum_{p \in I_{\ell, \tilde{\mathcal{S}}} \cap \Omega} f(p) \cdot \text{cost}(p, \tilde{\mathcal{S}}) \right) \\
& \leq 2 \sum_{\ell < \log \varepsilon / 2} \left( \sum_{p \in I_{\ell, \tilde{\mathcal{S}}}} \text{cost}(p, \tilde{\mathcal{S}}) + \sum_{p \in I_{\ell, \tilde{\mathcal{S}}} \cap \Omega} f(p) \cdot \text{cost}(p, \tilde{\mathcal{S}}) \right) \\
& \leq 4\varepsilon \text{cost}(G, \mathcal{A}),
\end{aligned}$$

where the last equation uses Lemma 5. Plugging this result into the previous inequality, we have:

$$\begin{aligned}
& \left| \sum_{p \in L_{\tilde{\mathcal{S}}}} \text{cost}(p, \tilde{\mathcal{S}}) - \sum_{p \in L_{\tilde{\mathcal{S}}} \cap \Omega} f(p) \cdot \text{cost}(p, \tilde{\mathcal{S}}) \right| \\
& \leq 4\varepsilon \text{cost}(G, \mathcal{A}) + \sum_{\ell = \log \varepsilon / 2}^{z \log z / 4\varepsilon} \left| \mathbb{E}[E_{\ell, \tilde{\mathcal{S}}}] - E_{\ell, \tilde{\mathcal{S}}} \right| + |\mathbb{E}[F_{\tilde{\mathcal{S}}}] - F_{\tilde{\mathcal{S}}}| \\
& \leq 4\varepsilon \text{cost}(G, \mathcal{A}) + \sum_{\ell = \log \varepsilon / 2}^{z \log z / 4\varepsilon} \frac{\varepsilon}{z \log z / \varepsilon} \cdot \left( \text{cost}(G, \mathcal{A}) + \text{cost}(I_{\ell, \tilde{\mathcal{S}}}, \tilde{\mathcal{S}}) \right) + |\mathbb{E}[F_{\tilde{\mathcal{S}}}] - F_{\tilde{\mathcal{S}}}| \\
& \leq 4\varepsilon \text{cost}(G, \mathcal{A}) + (z \log(z/4\varepsilon) - \log \varepsilon / 2) \cdot \frac{\varepsilon}{z \log z / \varepsilon} \cdot \left( \text{cost}(G, \mathcal{A}) + \text{cost}(L_{\tilde{\mathcal{S}}}, \tilde{\mathcal{S}}) \right) \\
& \quad + \varepsilon \cdot \text{cost}(G, \tilde{\mathcal{S}}) \\
& \leq O(\varepsilon) \cdot (\text{cost}(G, \mathcal{A}) + \text{cost}(L_{\tilde{\mathcal{S}}}, \tilde{\mathcal{S}})),
\end{aligned}$$

where the second to last inequality used Lemma 11. □

**From the approximate centroid set to any solution.** We can now finally turn to the proof of Lemma 2: it combines the result we show previously for the huge type, and the use of approximate centroid set with the Lemma 12 for the interesting and tiny types.

*Proof of Lemma 2.* Let  $X, k, z, G$  and  $\mathcal{A}$  as in the lemma statement. We condition on event  $\mathcal{E}$  happening. Let  $\mathcal{S}$  be a set of  $k$  points, and  $\tilde{\mathcal{S}} \in \mathbb{C}^k$  that approximates best  $\mathcal{S}$ , as given by the definition of  $\mathbb{C}$  (see Definition 1). This ensures that for all points  $p$  with  $\text{dist}(p, \mathcal{S}) \leq \frac{8z}{\varepsilon} \cdot \text{dist}(p, \mathcal{A})$  or  $\text{dist}(p, \tilde{\mathcal{S}}) \leq \frac{8z}{\varepsilon} \cdot \text{dist}(p, \mathcal{A})$ , we have  $|\text{cost}(p, \mathcal{S}) - \text{cost}(p, \tilde{\mathcal{S}})| \leq \frac{\varepsilon}{z \log(z/\varepsilon)} (\text{cost}(p, \mathcal{S}) + \text{cost}(p, \mathcal{A}))$ .

Our first step is to deal with points that have  $\text{dist}(p, \mathcal{S}) > \frac{4z}{\varepsilon} \cdot \text{dist}(p, \mathcal{A})$ , using Lemma 7. None of the remaining points is huge with respect to  $\tilde{\mathcal{S}}$ : hence, they all are in interesting clusters with

respect to  $\tilde{\mathcal{S}}$ . Let  $L_{\tilde{\mathcal{S}}}$  be this set of cluster: it can be handled with Lemma 12. The remaining of the proof formalizes the argument.

Let  $H_{\mathcal{S}}$  be the set of all clusters that are intersecting with some  $I_{\ell, \mathcal{S}}$  with  $\ell > z \log(4z/\varepsilon)$ . We also denote  $H_{\mathcal{S}}$  the points contained in those clusters. We decompose the cost difference as follows:

$$\left| \text{cost}(G, \mathcal{S}) - \sum_{p \in \Omega \cap G} f(p) \cdot \text{cost}(p, \mathcal{S}) \right| \leq \left| \sum_{p \in G \setminus H_{\mathcal{S}}} \text{cost}(p, \mathcal{S}) - \sum_{p \in (G \setminus H_{\mathcal{S}}) \cap \Omega} f(p) \cdot \text{cost}(p, \mathcal{S}) \right| \quad (15)$$

$$+ \left| \sum_{p \in H_{\mathcal{S}}} \text{cost}(p, \mathcal{S}) - \sum_{p \in H_{\mathcal{S}} \cap \Omega} f(p) \cdot \text{cost}(p, \mathcal{S}) \right| \quad (16)$$

Since we condition on event  $\mathcal{E}$ , the term 16 is  $O(\varepsilon) \cdot (\text{cost}(G, \mathcal{A}) + \text{cost}(G, \mathcal{S}))$ , using Lemma 7. Now we take a closer look at term 15. By definition of  $\tilde{\mathcal{S}}$ , it holds for all points  $p \in G \setminus H_{\mathcal{S}}$  that  $|\text{cost}(p, \mathcal{S}) - \text{cost}(p, \tilde{\mathcal{S}})| \leq \varepsilon(\text{cost}(p, \mathcal{S}) + \text{cost}(p, \mathcal{A}))$ . Therefore:

$$\begin{aligned} & \left| \sum_{p \in G \setminus H_{\mathcal{S}}} \text{cost}(p, \mathcal{S}) - \sum_{p \in (G \setminus H_{\mathcal{S}}) \cap \Omega} f(p) \cdot \text{cost}(p, \mathcal{S}) \right| \\ & \leq \left| \sum_{p \in G \setminus H_{\mathcal{S}}} \text{cost}(p, \tilde{\mathcal{S}}) - \sum_{p \in (G \setminus H_{\mathcal{S}}) \cap \Omega} f(p) \cdot \text{cost}(p, \tilde{\mathcal{S}}) \right| \\ & + \varepsilon (\text{cost}(G, \mathcal{S}) + \text{cost}(G, \mathcal{A}) + \text{cost}(\Omega, \mathcal{S}) + \text{cost}(\Omega, \mathcal{A})). \end{aligned}$$

This allows us to focus on bounding the cost difference to solution  $\tilde{\mathcal{S}}$  instead of  $\mathcal{S}$ .

For the remaining points in  $G \setminus H_{\mathcal{S}}$ , we aim at using Lemma 12: for that, we show that  $L_{\tilde{\mathcal{S}}} := G \setminus H_{\mathcal{S}}$  contains only interesting clusters with respect to  $\tilde{\mathcal{S}}$ . Indeed, for any  $p \in L_{\tilde{\mathcal{S}}}$ , we have  $|\text{cost}(p, \mathcal{S}) - \text{cost}(p, \tilde{\mathcal{S}})| \leq \frac{\varepsilon}{z \log z / \varepsilon} (\text{cost}(p, \mathcal{S}) + \text{cost}(p, \mathcal{A}))$  by definition of  $\tilde{\mathcal{S}}$ . Hence,

$$\begin{aligned} \text{cost}(p, \tilde{\mathcal{S}}) & \leq \text{cost}(p, \mathcal{S}) + \frac{\varepsilon}{z \log z / \varepsilon} (\text{cost}(p, \mathcal{S}) + \text{cost}(p, \mathcal{A})) \\ & \leq \left( (1 + \varepsilon) \left( \frac{4\varepsilon}{z} \right)^z + \varepsilon \right) \text{cost}(p, \mathcal{A}) \\ & \leq \left( \frac{8\varepsilon}{z} \right)^z \text{cost}(p, \mathcal{A}), \end{aligned}$$

and  $p$  is indeed not huge with respect to  $\tilde{\mathcal{S}}$ . Therefore, we can apply Lemma 12 to get:

$$\begin{aligned} \left| \sum_{p \in G \setminus H_{\mathcal{S}}} \text{cost}(p, \tilde{\mathcal{S}}) - \sum_{p \in (G \setminus H_{\mathcal{S}}) \cap \Omega} f(p) \cdot \text{cost}(p, \tilde{\mathcal{S}}) \right| & = \left| \sum_{p \in L_{\tilde{\mathcal{S}}}} \text{cost}(p, \tilde{\mathcal{S}}) - \sum_{p \in L_{\tilde{\mathcal{S}}} \cap \Omega} f(p) \cdot \text{cost}(p, \tilde{\mathcal{S}}) \right| \\ & \leq \varepsilon (\text{cost}(G, \mathcal{A}) + \text{cost}(L_{\tilde{\mathcal{S}}}, \tilde{\mathcal{S}})) \\ & = O(\varepsilon) (\text{cost}(L_{\tilde{\mathcal{S}}}, \mathcal{S}) + \text{cost}(G, \mathcal{A})) \end{aligned}$$

Combining all the equations yields

$$|\text{cost}(G, \mathcal{S}) - \text{cost}(\Omega, \mathcal{S})| \leq O(\varepsilon) \cdot (\text{cost}(G, \mathcal{A}) + \text{cost}(G, \mathcal{S}) + \text{cost}(\Omega, \mathcal{A}) + \text{cost}(\Omega, \mathcal{S})).$$

To conclude the proof, it only remains to remove the term  $\text{cost}(\Omega, \mathcal{A}) + \text{cost}(\Omega, \mathcal{S})$  from the right-hand-side. Applying this inequality for  $\mathcal{S} = \mathcal{A}$  and using  $\text{cost}(\Omega, \mathcal{A}) \leq \text{cost}(G, \mathcal{A}) + |\text{cost}(G, \mathcal{A}) - \text{cost}(\Omega, \mathcal{A})|$  yields first

$$\text{cost}(\Omega, \mathcal{A}) = O(1) \cdot \text{cost}(G, \mathcal{A}).$$

Similarly, we can use  $\text{cost}(\Omega, \mathcal{S}) \leq \text{cost}(G, \mathcal{S}) + |\text{cost}(G, \mathcal{S}) - \text{cost}(\Omega, \mathcal{S})|$  to get

$$\text{cost}(\Omega, \mathcal{S}) = O(1) \cdot (\text{cost}(G, \mathcal{S}) + \text{cost}(G, \mathcal{A})).$$

Hence, we finally conclude:

$$|\text{cost}(G, \mathcal{S}) - \text{cost}(\Omega, \mathcal{S})| \leq O(\varepsilon) \cdot (\text{cost}(G, \mathcal{A}) + \text{cost}(G, \mathcal{S})).$$

The probability now follows from taking a union-bound over the failure probability of Lemma 6 and Lemma 12. Specifically

$$1 - \exp\left(k \log |\mathbb{C}| - 2^{O(z \log z)} \cdot \frac{\min(\varepsilon^2, \varepsilon^z)}{\log^2 1/\varepsilon} \cdot \delta\right) - k \cdot z^2 \log^2(z/\varepsilon) \exp\left(-O(1) \frac{\varepsilon^2}{k} \delta\right)$$

In a given cluster  $\tilde{C}_i$  induced by  $\mathcal{A}$  on  $G$ , the complexity of the algorithm is  $O(|\tilde{C}_i|)$ : it is both the cost of computing the scaling factor  $f(p)$  for all  $p \in \tilde{C}_i$ , and the cost of sampling  $\delta$  points using reservoir sampling [Vit85]. Hence, the cost of this algorithm for all clusters is  $O(|G|)$ .  $\square$

## 6 Sampling from Outer Rings

In this section we prove Lemma 3:

**Lemma 3.** *Let  $(X, \text{dist})$  be a metric space,  $k, z$  be two positive integers,  $P$  be a set of clients and  $\mathcal{A}$  be a  $c_{\mathcal{A}}$ -approximate solution solution to  $(k, z)$ -clustering on  $P$ .*

*Let  $G$  be either a group  $G_b^O$  or  $G_{\max}^O$ . Suppose moreover that there is a  $\mathcal{A}$ -approximate centroid set  $\mathbb{C}$  for  $(k, z)$ -clustering on  $G$ .*

*Then, there exists an algorithm **SensitivitySample** running in time  $O(|G|)$  that constructs a set  $\Omega$  of size  $\delta$  such that it holds with probability  $1 - \exp\left(k \log |\mathbb{C}| - 2^{O(z \log z)} \cdot \frac{\varepsilon^2}{\log^2 1/\varepsilon} \cdot \delta\right)$  that, for all sets  $\mathcal{S}$  of  $k$  centers:*

$$|\text{cost}(G, \mathcal{S}) - \text{cost}(\Omega, \mathcal{S})| = \frac{\varepsilon}{z \log z / \varepsilon} \cdot (\text{cost}(\mathcal{S}) + \text{cost}(\mathcal{A})).$$



Recall that the **SensitivitySample** procedure merely picks  $\delta$  points  $p$  with probability  $\frac{\text{cost}(p, \mathcal{A})}{\text{cost}(G, \mathcal{A})}$ . Each of the  $\delta$  sampled points has a weight  $\frac{\text{cost}(G, \mathcal{A})}{\delta \cdot \text{cost}(p, \mathcal{A})}$ . The procedure runs in time  $O(|G|)$ .

The main steps of the proof are as follows.

- First, we consider the cost of the points in  $G$  such that  $\text{cost}(p, \mathcal{S})$  is at most  $4^z \cdot \text{cost}(p, \mathcal{A})$ . For this case, we can (almost) directly apply Bernstein's inequality as in the previous section.
- Second, we consider the cost of the points in  $G$  such that  $\text{cost}(p, \mathcal{S}) > 4^z \cdot \text{cost}(p, \mathcal{A})$ . Denote this set by  $G_{\text{far}, \mathcal{S}}$ . For these points, we can afford to replace their cost in  $\mathcal{S}$  with the distance to the closest center  $c \in \mathcal{A}$  plus the distance from  $c$  to the closest center in  $\mathcal{S}$ . The latter part can be charged to the remaining points of the cluster from the original dataset (i.e., not restricted to group  $G$ ) which are in much larger number and already paying a similar value in  $\mathcal{S}$ .

We first analyse the points not in  $G_{\text{far}, \mathcal{S}}$ . For that, we will go through the approximate centroid set  $\mathbb{C}$  to afford a union-bound: we show the following lemma.

**Lemma 13.** *Let  $\tilde{\mathcal{S}} \in \mathbb{C}^k$ , and define  $G_{\text{close}, \tilde{\mathcal{S}}}$  to be the set of points of  $G$  such that  $\text{cost}(p, \tilde{\mathcal{S}}) \leq 5^z \cdot \text{cost}(p, \mathcal{A})$ . It holds with probability*

$$1 - \exp\left(-2^{-O(z)} \left(\frac{\varepsilon}{\log 1/\varepsilon}\right)^2 \delta\right)$$

that

$$|\text{cost}(G_{\text{close}, \tilde{\mathcal{S}}}, \tilde{\mathcal{S}}) - \text{cost}(\Omega \cap G_{\text{close}, \tilde{\mathcal{S}}}, \tilde{\mathcal{S}})| \leq \frac{\varepsilon}{z \log z / \varepsilon} \left( \text{cost}(G, \mathcal{A}) + \text{cost}(G_{\text{close}, \tilde{\mathcal{S}}}, \tilde{\mathcal{S}}) \right)$$

*Proof.* We aim to use Bernstein's Inequality. Let  $E_{\text{close}, \tilde{\mathcal{S}}} = \sum_{i=1}^{\delta} X_i$ , where  $X_i = \frac{\text{cost}(G, \mathcal{A})}{\delta \cdot \text{cost}(p, \mathcal{A})} \cdot \text{cost}(p, \tilde{\mathcal{S}})$  if the  $i$ -th sampled point is  $p \in G_{\text{close}, \tilde{\mathcal{S}}}$  and  $X_i = 0$  if the  $i$ -th sampled point is  $p \notin G_{\text{close}, \tilde{\mathcal{S}}}$ . Recall that the probability that  $p$  is the  $i$ -th sampled point is  $\frac{\text{cost}(p, \mathcal{A})}{\text{cost}(G, \mathcal{A})}$ . We consider the second moment  $\mathbb{E}[X_i^2]$ :

$$\begin{aligned} \mathbb{E}[X_i^2] &= \sum_{p \in G_{\text{close}, \tilde{\mathcal{S}}}} \left( \frac{\text{cost}(G, \mathcal{A})}{\delta \cdot \text{cost}(p, \mathcal{A})} \cdot \text{cost}(p, \tilde{\mathcal{S}}) \right)^2 \cdot \mathbb{P}[p \in \Omega] \\ &= \text{cost}(G, \mathcal{A}) \cdot \sum_{p \in G_{\text{close}, \tilde{\mathcal{S}}}} \frac{\text{cost}(p, \tilde{\mathcal{S}})}{\delta^2 \cdot \text{cost}(p, \mathcal{A})} \cdot \text{cost}(p, \tilde{\mathcal{S}}) \\ &\leq \text{cost}(G, \mathcal{A}) \cdot \sum_{p \in G_{\text{close}, \tilde{\mathcal{S}}}} \frac{5^z}{\delta^2} \cdot \text{cost}(p, \tilde{\mathcal{S}}) \\ &\leq \frac{5^z}{\delta^2} \cdot \text{cost}(G, \mathcal{A}) \cdot \text{cost}(G_{\text{close}, \tilde{\mathcal{S}}}, \tilde{\mathcal{S}}) \end{aligned}$$

Furthermore, we have the following upper bound for the maximum value any of the  $X_i$ :

$$X_i \leq M := \max_{p \in G_{\text{close}, \tilde{\mathcal{S}}}} \frac{\text{cost}(G, \mathcal{A})}{\delta \cdot \text{cost}(p, \mathcal{A})} \cdot \text{cost}(p, \tilde{\mathcal{S}}) \leq \frac{5^z}{\delta} \cdot \text{cost}(G, \mathcal{A}). \quad (17)$$

Combining both bounds with Bernstein's inequality now yields

$$\begin{aligned}
& \mathbb{P}[|E_{close, \tilde{\mathcal{S}}} - \mathbb{E}[E_{close, \tilde{\mathcal{S}}}]| \leq \frac{\varepsilon}{z \log z / \varepsilon} \cdot (\text{cost}(G, \mathcal{A}) + \text{cost}(G_{close, \tilde{\mathcal{S}}}, \tilde{\mathcal{S}}))] \\
& \leq \exp \left( - \frac{\left( \frac{\varepsilon}{z \log z / \varepsilon} \right)^2 \cdot (\text{cost}(G, \mathcal{A}) + \text{cost}(G_{close, \tilde{\mathcal{S}}}, \tilde{\mathcal{S}}))^2}{2 \sum_{i=1}^{\delta} \text{Var}[X_i] + \frac{1}{3} M \cdot \varepsilon \cdot (\text{cost}(G, \mathcal{A}) + \text{cost}(G_{close, \tilde{\mathcal{S}}}, \tilde{\mathcal{S}}))} \right) \\
& \leq \exp \left( - \frac{\left( \frac{\varepsilon}{z \log z / \varepsilon} \right)^2 \cdot \delta \cdot (\text{cost}(G, \mathcal{A}) + \text{cost}(G_{close, \tilde{\mathcal{S}}}, \tilde{\mathcal{S}}))^2}{24^z \cdot \text{cost}(G, \mathcal{A}) \cdot \text{cost}(G_{close, \tilde{\mathcal{S}}}, \tilde{\mathcal{S}}) + 4^z \cdot \text{cost}(G, \mathcal{A}) \cdot \varepsilon \cdot (\text{cost}(G, \mathcal{A}) + \text{cost}(G_{close, \tilde{\mathcal{S}}}, \tilde{\mathcal{S}}))} \right) \\
& \leq \exp \left( - 2^{-O(z)} \cdot \left( \frac{\varepsilon}{z \log z / \varepsilon} \right)^2 \cdot \delta \right)
\end{aligned}$$

Noting that  $\text{cost}(\Omega \cap G_{close, \tilde{\mathcal{S}}}, \tilde{\mathcal{S}}) = \mathbb{E}[E_{close, \tilde{\mathcal{S}}}]$ , concludes: we have with probability

$$1 - \exp \left( - 2^{-O(z)} \cdot \left( \frac{\varepsilon}{\log 1/\varepsilon} \right)^2 \cdot \delta \right) \text{ that:}$$

$$|\text{cost}(G_{close, \tilde{\mathcal{S}}}, \tilde{\mathcal{S}}) - \text{cost}(\Omega \cap G_{close, \tilde{\mathcal{S}}}, \tilde{\mathcal{S}})| \leq \frac{\varepsilon}{z \log z / \varepsilon} \cdot (\text{cost}(G, \mathcal{A}) + \text{cost}(G_{close, \tilde{\mathcal{S}}}, \tilde{\mathcal{S}}))$$

□

Now we turn our attention to  $G_{far, \mathcal{S}}$ . For this, we analyse the following event  $\mathcal{E}_{far}$ , similar to  $\mathcal{E}$ : For all cluster  $C$  of solution  $\mathcal{A}$  such that  $C \cap G \neq \emptyset$

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

**Lemma 14.** *Event  $\mathcal{E}_{far}$  happens with probability at least*

$$1 - k \exp \left( \frac{\varepsilon^2}{6k} \cdot \delta \right).$$

*Proof.* We aim to use Bernstein's Inequality. Let  $E_C = \sum_{i=1}^{\delta} X_i$ , where  $X_i = \frac{\text{cost}(G, \mathcal{A})}{\delta \cdot \text{cost}(p, \mathcal{A})} \cdot \text{cost}(p, \mathcal{A})$  if the  $i$ -th sampled point  $p \in C$  and  $X_i = 0$  the  $i$ -th sampled point  $p \notin C$ . Recall that the probability that the  $i$ -th sampled point is  $p$  is  $\frac{\text{cost}(p, \mathcal{A})}{\text{cost}(G, \mathcal{A})}$ . We consider the second moment  $\mathbb{E}[X_i^2]$ :

$$\begin{aligned}
\mathbb{E}[X_i^2] &= \sum_{p \in C \cap G} \left( \frac{\text{cost}(G, \mathcal{A})}{\delta \cdot \text{cost}(p, \mathcal{A})} \cdot \text{cost}(p, \mathcal{A}) \right)^2 \cdot \mathbb{P}[p \text{ is the } i\text{-th sampled point}] \\
&= \frac{\text{cost}(G, \mathcal{A})}{\delta^2} \cdot \sum_{p \in C \cap G} \text{cost}(p, \mathcal{A}) \\
&= \frac{\text{cost}(G, \mathcal{A})}{\delta^2} \text{cost}(C \cap G, \mathcal{A}) \\
&\leq \frac{2k}{\delta^2} \cdot \text{cost}^2(C \cap G, \mathcal{A})
\end{aligned}$$

where the final inequality follows since every cluster has cost at least half the average. Indeed, either the group considered is  $G_{\max}^O$ , and then any cluster verifies  $\text{cost}(C \cap G) \geq \frac{1}{k} \text{cost}(R_O^A, \mathcal{A}) \geq \frac{1}{k} \text{cost}(G_{\max}^O, \mathcal{A})$ , or all the clusters in  $G_b^O$  have an equal cost, up to a factor of 2 – hence none cost less than half of the average.

Furthermore, we have by the same argument the following upper bound for the maximum value any of the  $X_i$ :

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

Combining both bounds with Bernstein's inequality now yields

$$\begin{aligned} & \mathbb{P}[|\text{cost}(C \cap G \cap \Omega, \mathcal{A}) - \text{cost}(C \cap G, \mathcal{A})| \leq \varepsilon \cdot \text{cost}(C \cap G, \mathcal{A})] \\ & \leq \exp\left(-\frac{\varepsilon^2 \cdot \text{cost}^2(C \cap G, \mathcal{A})}{2 \sum_{i=1}^{\delta} \text{Var}[X_i] + \frac{1}{3} M \cdot \varepsilon \cdot \text{cost}(C \cap G, \mathcal{A})}\right) \leq \exp\left(-\frac{\varepsilon^2}{6k} \cdot \delta\right) \end{aligned}$$

Reformulating, we now have

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

□

**Lemma 15.** *Let  $(X, \text{dist})$  be a metric space,  $k, z$  be two positive integers. Suppose  $G$  is either a group  $G_b^O$  or  $G_{\max}^O$ . Let  $G_{\text{far}, \mathcal{S}} \subset G$  be the set of all clients such that  $\text{cost}(p, \mathcal{S}) > 4^z \cdot \text{cost}(p, \mathcal{A})$ . Condition on event  $\mathcal{E}_{\text{far}}$ .*

*Then, the set  $\Omega$  of size  $\delta$  constructed by **SensitivitySample** verifies the following. It holds for all sets  $\mathcal{S}$  of  $k$  centers that:*

$$\text{cost}(G_{\text{far}, \mathcal{S}}, \mathcal{S}) + \text{cost}(\Omega \cap G_{\text{far}, \mathcal{S}}, \mathcal{S}) \leq \frac{2\varepsilon}{z \log z / \varepsilon} \cdot \text{cost}(\mathcal{S}).$$

*Proof.* Our aim will be to show that  $\max(\text{cost}(G_{\text{far}, \mathcal{S}}, \mathcal{S}), \text{cost}(\Omega \cap G_{\text{far}, \mathcal{S}}, \mathcal{S})) \leq \frac{\varepsilon}{z \log z / \varepsilon} \cdot \text{cost}(\mathcal{S})$ . It is key here that we compare to the cost of the full input in  $\mathcal{S}$ , and not simply the cost of the group  $G$ .

First, we fix a cluster  $C \in \mathcal{A}$ , and show that the total contribution of points of  $C \cap G_{\text{far}, \mathcal{S}}$  is very cheap compared to  $\text{cost}(C, \mathcal{S})$ , i.e. that  $\text{cost}(G_{\text{far}, \mathcal{S}} \cap C, \mathcal{S}) \leq \frac{\varepsilon}{z \log z / \varepsilon} \cdot \text{cost}(C, \mathcal{S})$ .

For this, fix a point  $p \in G_{\text{far}, \mathcal{S}} \cap C$ , and let  $c$  be the center of cluster  $C$ .

Let  $C_{\text{close}}$  be the points of  $C$  with cost at most  $\left(\frac{z}{\varepsilon}\right)^z \cdot \frac{\text{cost}(C, \mathcal{A})}{|C|}$ . Due to Markov's inequality, most of  $C$ 's points are in  $C_{\text{close}}$ :  $|C_{\text{close}}| \geq (1 - \varepsilon/z) \cdot |C|$ .

Using that the point  $p$  is both in the outer ring of  $C$  and in  $G_{\text{far}, \mathcal{S}}$ , we can lower bound the distance from  $c$  to  $\mathcal{S}$  as follows. Triangle inequality and  $\text{cost}(p, \mathcal{S}) > 4^z \cdot \text{cost}(p, c)$ , yield  $\text{dist}(c, \mathcal{S}) \geq \text{dist}(p, \mathcal{S}) - \text{dist}(p, c) \geq 4\text{dist}(p, c) - \text{dist}(p, c) \geq 3\text{dist}(p, c)$ . Since  $p$  is from an outer group, it verifies

$\text{cost}(p, c) \geq \left(\frac{z}{\varepsilon}\right)^{2z} \cdot \frac{\text{cost}(C, c)}{|C|}$ . Combining those two observations yields:  $\text{cost}(c, \mathcal{S}) \geq 3^z \text{cost}(p, \mathcal{A}) \geq 3^z \cdot \left(\frac{z}{\varepsilon}\right)^{2z} \cdot \frac{\text{cost}(C, c)}{|C|}$ .

Using this and Lemma 1, we now have for any  $q \in C_{\text{close}}$ :

$$\begin{aligned}
\text{cost}(c, \mathcal{S}) &\leq (1 + \varepsilon/(2z))^{z-1} \cdot \text{cost}(q, \mathcal{S}) + \left(\frac{2z + \varepsilon}{\varepsilon}\right)^{z-1} \cdot \text{cost}(q, c) \\
&\leq (1 + \varepsilon) \text{cost}(q, \mathcal{S}) + \left(\frac{2z + \varepsilon}{\varepsilon}\right)^{z-1} \cdot \left(\frac{z}{\varepsilon}\right)^z \cdot \frac{\text{cost}(C, c)}{|C|} \\
&\leq (1 + \varepsilon) \text{cost}(q, \mathcal{S}) + 3^{z-1} \cdot \left(\frac{z}{\varepsilon}\right)^{2z-1} \cdot \frac{\text{cost}(C, c)}{|C|} \\
&\leq (1 + \varepsilon) \text{cost}(q, \mathcal{S}) + \frac{\varepsilon}{3^z} \cdot \text{cost}(c, \mathcal{S}) \\
\Rightarrow \text{cost}(q, \mathcal{S}) &\geq \frac{1 - \varepsilon}{1 + \varepsilon} \cdot \text{cost}(c, \mathcal{S}) \\
\Rightarrow \text{cost}(C, \mathcal{S}) &\geq \text{cost}(C_{\text{close}}, \mathcal{S}) \geq |C_{\text{close}}| \cdot \frac{1 - \varepsilon}{1 + \varepsilon} \cdot \text{cost}(c, \mathcal{S}). \tag{18}
\end{aligned}$$

Using additionally that  $|C_{\text{close}}| \geq (1 - \frac{\varepsilon}{z}) \cdot |C|$  and  $\text{cost}(c, \mathcal{S}) \geq 3^z \cdot \left(\frac{z}{\varepsilon}\right)^{2z} \cdot \frac{\text{cost}(C, c)}{|C|}$ , we get:

$$\text{cost}(C, \mathcal{S}) \geq |C_{\text{close}}| \cdot \frac{1 - \varepsilon}{1 + \varepsilon} \cdot 3^z \cdot \left(\frac{z}{\varepsilon}\right)^{2z} \cdot \frac{\text{cost}(C, \mathcal{A})}{|C|} \geq 3^z \cdot \left(\frac{z}{\varepsilon}\right)^{2z-1} \cdot \text{cost}(C, \mathcal{A}). \tag{19}$$

We are now equipped to show the first part of the lemma, namely  $\text{cost}(G_{\text{far}, \mathcal{S}}, \mathcal{S}) \leq \frac{\varepsilon}{z \log z / \varepsilon} \cdot \text{cost}(\mathcal{S})$ .

Since  $G \cap C$  contains only points from the outer ring of  $C$ , with distance at least  $(z/\varepsilon)^2$  times the average, Markov's inequality implies that  $|G \cap C| \leq \left(\frac{\varepsilon}{z}\right)^2 \cdot |C|$ . Hence,  $|G_{\text{far}, \mathcal{S}} \cap C| \leq \frac{1}{1 - \varepsilon/z} \cdot \left(\frac{\varepsilon}{z}\right)^2 \cdot |C_{\text{close}}|$ . This yields

$$\begin{aligned}
& \text{cost}(G_{far,S} \cap C, \mathcal{S}) = \sum_{p \in G_{far,S} \cap C} \text{cost}(p, \mathcal{S}) \\
(Lem.1) \quad & \leq \sum_{p \in G_{far,S} \cap C} (1 + \varepsilon/2z)^{z-1} \text{cost}(c, \mathcal{S}) + \left(\frac{2z + \varepsilon}{\varepsilon}\right)^{z-1} \cdot \text{cost}(p, c) \\
& \leq |G_{far,S} \cap C| \cdot (1 + \varepsilon) \cdot \text{cost}(c, \mathcal{S}) \\
& \quad + \left(\frac{2z + \varepsilon}{\varepsilon}\right)^{z-1} \cdot \text{cost}(G_{far,S} \cap C, \mathcal{A}) \tag{20} \\
& \leq \frac{1 + \varepsilon}{1 - \varepsilon/z} \cdot \left(\frac{\varepsilon}{z}\right)^2 \cdot |C_{close}| \text{cost}(c, \mathcal{S}) + \left(\frac{2z + \varepsilon}{\varepsilon}\right)^{z-1} \text{cost}(G_{far,S} \cap C, \mathcal{A}) \\
(Eq. 18) \quad & \leq \frac{(1 + \varepsilon)^2}{(1 - \varepsilon)^2} \cdot \left(\frac{\varepsilon}{z}\right)^2 \cdot \text{cost}(C, \mathcal{S}) + \left(\frac{2z + \varepsilon}{\varepsilon}\right)^{z-1} \cdot \text{cost}(G_{far,S} \cap C, \mathcal{A}) \\
(Eq. 19) \quad & \leq \frac{(1 + \varepsilon)^2}{(1 - \varepsilon)^2} \cdot \left(\frac{\varepsilon}{z}\right)^2 \cdot \text{cost}(C, \mathcal{S}) \\
& \quad + \left(\frac{2z + \varepsilon}{\varepsilon}\right)^{z-1} \cdot \frac{1}{3^z} \cdot \left(\frac{\varepsilon}{z}\right)^{2z-1} \cdot \text{cost}(G_{far,S} \cap C, \mathcal{S}) \tag{21} \\
& \leq \frac{\varepsilon}{z \log z / \varepsilon} \cdot \text{cost}(C, \mathcal{S}) \tag{22}
\end{aligned}$$

Summing this up over all clusters  $C$ , we therefore have

$$\text{cost}(G_{far,S}, \mathcal{S}) \leq \frac{\varepsilon}{z \log z / \varepsilon} \cdot \text{cost}(\mathcal{S}) \tag{23}$$

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

In particular, event  $\mathcal{E}_{far}$  implies that with probability  $1 - k' \cdot \exp\left(-O(1) \cdot \frac{\varepsilon^2}{k'} \cdot \delta\right)$  for all clusters  $C$  induced by  $\mathcal{A}$

$$\begin{aligned}
\sum_{p \in C \cap G \cap \Omega} \frac{\text{cost}(G, \mathcal{A})}{\delta \cdot \text{cost}(p, \mathcal{A})} \cdot \left(\frac{2z}{\varepsilon}\right)^{2z} \cdot \frac{\text{cost}(C, \mathcal{A})}{|C|} & \leq \sum_{p \in C \cap G \cap \Omega} \frac{\text{cost}(G, \mathcal{A})}{\delta \cdot \text{cost}(p, \mathcal{A})} \text{cost}(p, \mathcal{A}) \\
& \leq (1 + \varepsilon) \cdot \text{cost}(C \cap G, \mathcal{A}) \\
\Rightarrow \sum_{p \in C \cap G \cap \Omega} \frac{\text{cost}(G, \mathcal{A})}{\delta \cdot \text{cost}(p, \mathcal{A})} & \leq (1 + \varepsilon) \cdot \left(\frac{\varepsilon}{2z}\right)^{2z} \cdot |C| \frac{\text{cost}(C \cap G, \mathcal{A})}{\text{cost}(C, \mathcal{A})} \tag{24} \\
& \leq (1 + \varepsilon) \cdot \left(\frac{\varepsilon}{2z}\right)^{2z} \cdot |C| \tag{25}
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
& \text{cost}(G_{far, \mathcal{S}} \cap \Omega \cap C, \mathcal{S}) \\
&= \sum_{p \in G_{far, \mathcal{S}} \cap C} \frac{\text{cost}(G, \mathcal{A})}{\delta \cdot \text{cost}(p, \mathcal{A})} \cdot \text{cost}(p, \mathcal{S}) \\
(Lem. 1) \quad &\leq \sum_{p \in G_{far, \mathcal{S}} \cap \Omega \cap C} \frac{\text{cost}(G, \mathcal{A})}{\delta \cdot \text{cost}(p, \mathcal{A})} \cdot \left( \left(1 + \frac{\varepsilon}{2z}\right)^{z-1} \text{cost}(c, \mathcal{S}) \right. \\
&\quad \left. + \left(\frac{2z + \varepsilon}{\varepsilon}\right)^{z-1} \cdot \text{cost}(p, c) \right) \\
&\leq (1 + \varepsilon) \cdot \text{cost}(c, \mathcal{S}) \cdot \sum_{p \in G_{far, \mathcal{S}} \cap \Omega \cap C} \frac{\text{cost}(G, \mathcal{A})}{\delta \cdot \text{cost}(p, \mathcal{A})} \\
(\mathcal{E}_{far}) \quad &+ \left(\frac{2z + \varepsilon}{\varepsilon}\right)^{z-1} \cdot (1 + \varepsilon) \cdot \text{cost}(C \cap G, \mathcal{A}) \\
(Eq. 25) \quad &\leq (1 + \varepsilon)^2 \text{cost}(c, \mathcal{S}) \cdot \left(\frac{\varepsilon}{2z}\right)^{2z} \cdot |C| + \left(\frac{2z + \varepsilon}{\varepsilon}\right)^{z-1} \cdot (1 + \varepsilon) \cdot \text{cost}(C \cap G, \mathcal{A}) \\
&\leq (1 + \varepsilon)^2 \cdot \left(\frac{\varepsilon}{2z}\right)^{2z} \cdot |C| \cdot \text{cost}(c, \mathcal{S}) + \left(\frac{2z + \varepsilon}{\varepsilon}\right)^{z-1} \cdot \text{cost}(C, \mathcal{A}) \tag{26} \\
&\leq \frac{\varepsilon}{z \log z / \varepsilon} \cdot \text{cost}(C, \mathcal{S})
\end{aligned}$$

where the steps following Equation 26 are identical to those used to derive Equation 22 from Equation 20. Again, summing over all clusters now yields

$$\text{cost}(G_{far, \mathcal{S}} \cap \Omega, \mathcal{S}) \leq \frac{\varepsilon}{z \log z / \varepsilon} \cdot \text{cost}(\mathcal{S}),$$

which yields the claim.  $\square$

**Combining far and close to show Lemma 3** The overall proof follows from those lemmas.

*Proof of Lemma 3.* First, we condition on event  $\mathcal{E}_{far}$ , and on the success of Lemma 13 for all solution in  $\mathbb{C}^k$ . This happens with probability

$$1 - k \exp\left(\frac{\varepsilon^2}{k} \cdot \delta\right) - \exp\left(k \log |\mathbb{C}| - 2^{-O(z)} \left(\frac{\varepsilon}{\log 1/\varepsilon}\right)^2 \delta\right).$$

Let  $\mathcal{S}$  be a solution, and  $\tilde{\mathcal{S}}$  its corresponding solution in  $\mathbb{C}^k$ . We break the cost of  $\mathcal{S}$  into two parts: points with  $\text{cost}(p, \tilde{\mathcal{S}}) \leq 5^z \cdot \text{cost}(p, \mathcal{A})$ , on which we can apply Lemma 13, on the others, on which we will apply Lemma 15.

From Lemma 13, we directly get

$$|\text{cost}(G_{close, \tilde{\mathcal{S}}} - \text{cost}(\Omega \cap G_{close, \tilde{\mathcal{S}}}, \tilde{\mathcal{S}})| \leq \frac{\varepsilon}{z \log z / \varepsilon} \cdot \left(\text{cost}(G, \mathcal{A}) + \text{cost}(G, \tilde{\mathcal{S}})\right).$$

Since any point in  $G_{close, \tilde{\mathcal{S}}}$  verifies  $|\text{cost}(p, \mathcal{S}) - \text{cost}(p, \tilde{\mathcal{S}})| \leq \frac{\varepsilon}{z \log z / \varepsilon} (\text{cost}(p, \mathcal{S}) + \text{cost}(p, \mathcal{A}))$  we can relate this to  $\text{cost}(G_{close, \tilde{\mathcal{S}}}, \mathcal{S})$  as follows. First, this implies  $\text{cost}(G_{close, \tilde{\mathcal{S}}}, \tilde{\mathcal{S}}) \leq (1 + \varepsilon) \text{cost}(G_{close, \tilde{\mathcal{S}}}, \mathcal{S}) + \varepsilon \text{cost}(G, \mathcal{A})$ . Hence:

$$\begin{aligned}
& |\text{cost}(G_{close, \tilde{\mathcal{S}}}, \mathcal{S}) - \text{cost}(\Omega \cap G_{close, \tilde{\mathcal{S}}}, \mathcal{S})| \\
& \leq |\text{cost}(G_{close, \tilde{\mathcal{S}}}, \tilde{\mathcal{S}}) - \text{cost}(\Omega \cap G_{close, \tilde{\mathcal{S}}}, \tilde{\mathcal{S}})| \\
& \quad + \frac{\varepsilon}{z \log z / \varepsilon} (\text{cost}(G_{close, \tilde{\mathcal{S}}}, \mathcal{S}) + \text{cost}(G_{close, \tilde{\mathcal{S}}}, \mathcal{A})) \\
& \quad + \frac{\varepsilon}{z \log z / \varepsilon} (\text{cost}(G_{close, \tilde{\mathcal{S}}} \cap \Omega, \mathcal{S}) + \text{cost}(G_{close, \tilde{\mathcal{S}}} \cap \Omega, \mathcal{A})) \\
& \leq \frac{O(\varepsilon)}{z \log z / \varepsilon} (\text{cost}(G, \mathcal{A}) + \text{cost}(G, \mathcal{S})) \\
& \quad + \frac{\varepsilon}{z \log z / \varepsilon} (\text{cost}(G_{close, \tilde{\mathcal{S}}} \cap \Omega, \mathcal{S}) + \text{cost}(G_{close, \tilde{\mathcal{S}}} \cap \Omega, \mathcal{A}))
\end{aligned}$$

We now deal with the other far points. For this, note that  $G \setminus G_{close, \tilde{\mathcal{S}}} \subseteq G_{far, \mathcal{S}}$ . Indeed, any point  $p \in G \setminus G_{far, \mathcal{S}}$  has its cost preserved by  $\tilde{\mathcal{S}}$ , and therefore verifies

$$\begin{aligned}
\text{cost}(p, \tilde{\mathcal{S}}) & \leq (1 + \frac{\varepsilon}{z \log z / \varepsilon}) \text{cost}(p, \mathcal{S}) + \frac{\varepsilon}{z \log z / \varepsilon} \text{cost}(p, \mathcal{A}) \\
& \leq (1 + \varepsilon) \cdot 4^z \text{cost}(p, \mathcal{A}) + \varepsilon \text{cost}(p, \mathcal{A}) \leq 5^z \text{cost}(p, \mathcal{A}).
\end{aligned}$$

Consequently,  $G \setminus G_{far, \mathcal{S}} \subseteq G_{close, \tilde{\mathcal{S}}}$ , which implies  $G \setminus G_{close, \tilde{\mathcal{S}}} \subseteq G_{far, \mathcal{S}}$ . Hence, we can use Lemma 15:

$$\begin{aligned}
& |\text{cost}(G \setminus G_{close, \tilde{\mathcal{S}}}, \mathcal{S}) - \text{cost}(\Omega \cap (G \setminus G_{close, \tilde{\mathcal{S}}}), \mathcal{S})| \\
& \leq \text{cost}(G \setminus G_{close, \tilde{\mathcal{S}}}, \mathcal{S}) + \text{cost}(\Omega \cap (G \setminus G_{close, \tilde{\mathcal{S}}}), \mathcal{S}) \\
& \leq \text{cost}(G_{far, \mathcal{S}}, \mathcal{S}) + \text{cost}(\Omega \cap G_{far, \mathcal{S}}, \mathcal{S}) \\
& \leq \frac{\varepsilon}{z \log z / \varepsilon} \cdot \text{cost}(\mathcal{S}).
\end{aligned}$$

Hence, adding the two inequalities gives that

$$\begin{aligned}
& |\text{cost}(G, \mathcal{S}) - \text{cost}(\Omega \cap G, \mathcal{S})| \\
& \leq \frac{O(\varepsilon)}{z \log z / \varepsilon} \cdot \text{cost}(\mathcal{S}) + \frac{\varepsilon}{z \log z / \varepsilon} (\text{cost}(G \cap \Omega, \mathcal{S}) + \text{cost}(G \cap \Omega, \mathcal{A})).
\end{aligned}$$

To remove the terms depending on  $\Omega$  from the right hand side, one can proceed as in the end of Lemma 2, applying grossly the previous inequality to get  $\text{cost}(G \cap \Omega, \mathcal{S}) = O(1) \text{cost}(\mathcal{S})$  and  $\text{cost}(G \cap \Omega, \mathcal{A}) = O(1) \text{cost}(\mathcal{A})$ . This concludes the theorem:

$$|\text{cost}(G, \mathcal{S}) - \text{cost}(\Omega \cap G, \mathcal{S})| \leq \frac{O(\varepsilon)}{z \log z / \varepsilon} \cdot (\text{cost}(G, \mathcal{S}) + \text{cost}(G, \mathcal{A})).$$

□

This concludes the coreset construction for the outer groups.

## 7 Partitioning into Well Structured Groups

In this section, we show that the outcome of the partitioning step satisfies Lemma 4, that we restate for convenience.

**Lemma 4.** *Let  $(X, \text{dist})$  be a metric space with a set of clients  $P$ ,  $k, z$  be two positive integers, and  $\varepsilon \in \mathbb{R}_+^*$ . For every solution  $\mathcal{S}$ , it holds that*

$$|\text{cost}(\mathcal{D}, \mathcal{S}) - \text{cost}(P_1, \mathcal{S})| = O(\varepsilon) \text{cost}(\mathcal{S}),$$

where  $\mathcal{D}$  and  $P_1$  are defined in Step 2 of the algorithm.

Recall that the *inner ring*  $R_I(C)$  (resp. *outer ring*  $R_O(C)$ ) of a cluster  $C$  consists of the points of  $C$  with cost at most  $(\varepsilon/z)^{2z} \Delta_C$  (resp. at least  $(z/\varepsilon)^{2z} \Delta_C$ ). The *main ring*  $R_M(C)$  consist of all the other points of  $C$ .

Recall also that  $\mathcal{D}$  contains all points that are either in some inner ring, in some group  $G_{j,\min}$  or in  $G_{\min}^O$ .  $P_1$  contains center of  $\mathcal{A}$  weighted by the number of points from  $\mathcal{D}$  in their clusters.

To prove Lemma 4, we treat separately the inner ring and the groups  $G_{j,\min}$  and  $G_{\min}^O$  in the next two lemmas. Their proof are deferred to next sections. For all those lemmas, we fix a metric space  $I$  a set of clients  $P$ , two positive integers  $k$  and  $z$ , and  $\varepsilon \in \mathbb{R}_+^*$ . We also fix  $\mathcal{A}$ , a solution to  $(k, z)$ -clustering on  $P$  with cost  $\text{cost}(\mathcal{A}) \leq c_{\mathcal{A}} \text{cost}(\text{OPT})$ .

**Lemma 16.** *For any solution  $\mathcal{S}$  and any cluster  $C$  with center  $c$  of  $\mathcal{A}$ ,*

$$|\text{cost}(R_I(C), \mathcal{S}) - |R_I(C)| \cdot \text{cost}(c, \mathcal{S})| \leq \varepsilon (\text{cost}(C, \mathcal{A}) + \text{cost}(R_I(C), \mathcal{S})).$$

**Lemma 17.** *For any solution  $\mathcal{S}$  and any  $j$ ,*

$$\left| \text{cost}(G_{j,\min}, \mathcal{S}) - \sum_{i=1}^k |C_i \cap G_{j,\min}| \cdot \text{cost}(c_i, \mathcal{S}) \right| \leq \varepsilon \cdot \text{cost}(R_j, \mathcal{S}) + \varepsilon \cdot \text{cost}(R_j, \mathcal{A}).$$

Moreover, for any solution  $\mathcal{S}$ ,

$$\left| \text{cost}(G_{\min}^O, \mathcal{S}) - \sum_{i=1}^k |C_i \cap G_{\min}^O| \cdot \text{cost}(c_i, \mathcal{S}) \right| \leq \varepsilon \cdot \text{cost}(\mathcal{S}) + \varepsilon \cdot \text{cost}(\mathcal{A}).$$

The proof of Lemma 4 combines those lemmas.

*Proof of Lemma 4.* We decompose  $|\text{cost}(\mathcal{D}, \mathcal{S}) - \text{cost}(P_1, \mathcal{S})|$  into terms corresponding to the previous lemmas:



$$\begin{aligned}
|\text{cost}(\mathcal{D}, \mathcal{S}) - \text{cost}(P_1, \mathcal{S})| &\leq \sum_{i=1}^k |\text{cost}(R_I(C_i), \mathcal{S}) - |R_I(C_i)| \text{cost}(c_i, \mathcal{S})| \\
&\quad + \sum_{j=2z \log(z/\varepsilon)}^{2z \log(z/\varepsilon)} \left| \text{cost}(G_{j, \min}, \mathcal{S}) - \sum_{i=1}^k |C_i \cap G_{j, \min}| \text{cost}(c_i, \mathcal{S}) \right| \\
&\quad + \left| \text{cost}(G_{\min}^O, \mathcal{S}) - \sum_{i=1}^k |C_i \cap G_{\min}^O| \text{cost}(c_i, \mathcal{S}) \right| \\
&\leq \sum_{i=1}^k \varepsilon (\text{cost}(C_i, \mathcal{A}) + \text{cost}(R_I(C_i), \mathcal{S})) \\
&\quad + 2\varepsilon \text{cost}(\mathcal{S}) + 2\varepsilon \text{cost}(\mathcal{A}) + \varepsilon (\text{cost}(\mathcal{S}) + \text{cost}(\mathcal{A})) \\
&\leq 8\varepsilon c_{\mathcal{A}} \text{cost}(\mathcal{S}),
\end{aligned}$$

where the second inequality uses Lemmas 16 and 17. □

## 7.1 The Inner Ring: Proof of Lemma 16

**Lemma 16.** *For any solution  $\mathcal{S}$  and any cluster  $C$  with center  $c$  of  $\mathcal{A}$ ,*

$$|\text{cost}(R_I(C), \mathcal{S}) - |R_I(C)| \cdot \text{cost}(c, \mathcal{S})| \leq \varepsilon (\text{cost}(C, \mathcal{A}) + \text{cost}(R_I(C), \mathcal{S})).$$

*Proof.* Let  $C$  be a cluster induced by  $\mathcal{A}$ , and  $p$  be a point in the inner ring  $R_I(C)$ . We start by bounding  $|\text{cost}(p, \mathcal{S}) - \text{cost}(c, \mathcal{S})|$ . Let  $\mathcal{S}(p)$  (resp.  $\mathcal{S}(c)$ ) be the closest point from  $\mathcal{S}$  to  $p$  (resp.  $c$ ).

Using Lemma 1, we get

$$|\text{cost}(p, \mathcal{S}) - \text{cost}(c, \mathcal{S})| \leq \varepsilon \cdot \text{cost}(p, \mathcal{S}) + (1 + 2z/\varepsilon)^{z-1} \cdot \text{cost}(c, p).$$

Since  $p$  is from the inner ring of its cluster,  $\text{cost}(c, p) \leq (\frac{\varepsilon}{z})^{2z} \Delta_C$ , hence  $(1 + 2z/\varepsilon)^{z-1} \text{cost}(c, p) \leq (2 + \varepsilon)^{z-1} \cdot (\varepsilon/z)^{z+1} \cdot \Delta_C \leq \varepsilon \Delta_C$ , for small enough  $\varepsilon$ .

Summing this over all points of the inner ring yields

$$\begin{aligned}
|\text{cost}(R_I(C), \mathcal{S}) - |R_I(C)| \cdot \text{cost}(c, \mathcal{S})| &\leq \sum_{p \in R_I(C)} |\text{cost}(p, \mathcal{S}) - \text{cost}(c, \mathcal{S})| \\
&\leq \sum_{p \in R_I(C)} \varepsilon \text{cost}(p, \mathcal{S}) + \varepsilon \Delta_C \\
&\leq \varepsilon \text{cost}(R_I(C), \mathcal{S}) + \varepsilon |R_I(C)| \Delta_C \\
&\leq \varepsilon \text{cost}(R_I(C), \mathcal{S}) + \varepsilon \text{cost}(C, \mathcal{A})
\end{aligned}$$

This implies

$$|\text{cost}(R_I(C), \mathcal{S}) - |R_I(C)| \cdot \text{cost}(c, \mathcal{S})| \leq \varepsilon (\text{cost}(C, \mathcal{A}) + \text{cost}(R_I(C), \mathcal{S})).$$

□

## 7.2 The Cheap Groups: Proof of Lemma 17

**Lemma 17.** *For any solution  $\mathcal{S}$  and any  $j$ ,*

$$\left| \text{cost}(G_{j,\min}, \mathcal{S}) - \sum_{i=1}^k |C_i \cap G_{j,\min}| \cdot \text{cost}(c_i, \mathcal{S}) \right| \leq \varepsilon \cdot \text{cost}(R_j, \mathcal{S}) + \varepsilon \cdot \text{cost}(R_j, \mathcal{A}).$$

Moreover, for any solution  $\mathcal{S}$ ,

$$\left| \text{cost}(G_{\min}^O, \mathcal{S}) - \sum_{i=1}^k |C_i \cap G_{\min}^O| \cdot \text{cost}(c_i, \mathcal{S}) \right| \leq \varepsilon \cdot \text{cost}(\mathcal{S}) + \varepsilon \cdot \text{cost}(\mathcal{A}).$$

*Proof.* Using Lemma 1, for a point  $p$  in cluster  $C_i$

$$|\text{cost}(c_i, \mathcal{S}) - \text{cost}(p, \mathcal{S})| \leq \varepsilon \text{cost}(p, \mathcal{S}) + \left(1 + \frac{2z}{\varepsilon}\right)^{z-1} \text{cost}(p, c_i).$$

Let  $G$  be a group, either  $G_{j,\min}$  or  $G_{\min}^O$ . Summing for all cluster  $C_i$  and all  $p \in G \cap C_i$ , we now get

$$\begin{aligned} & \left| \sum_{i=1}^k |C_i \cap G| \cdot \text{cost}(c_i, \mathcal{S}) - \text{cost}(G, \mathcal{S}) \right| \\ & \leq \varepsilon \cdot \text{cost}(G, \mathcal{S}) + \sum_{i=1}^k \sum_{p \in G \cap C_i} \left(1 + \frac{2z}{\varepsilon}\right)^{z-1} \text{cost}(p, \mathcal{A}) \\ & \leq \varepsilon \cdot \text{cost}(G, \mathcal{S}) + \sum_{i=1}^k \left(\frac{3z}{\varepsilon}\right)^{z-1} \text{cost}(C_i \cap G, \mathcal{A}) \\ & \leq \varepsilon \cdot \text{cost}(G, \mathcal{S}) + \left(\frac{3z}{\varepsilon}\right)^{z-1} \text{cost}(G, \mathcal{A}) \end{aligned}$$

Now, either  $G = G_{j,\min}$  for some  $j$ , and  $\text{cost}(G, \mathcal{A}) \leq \left(\frac{\varepsilon}{4z}\right)^z \cdot \text{cost}(R_j, \mathcal{A})$ ; or  $G = G_{\min}^O$ , and  $\text{cost}(G, \mathcal{A}) \leq \left(\frac{\varepsilon}{4z}\right)^z \cdot \text{cost}(R_O(\mathcal{A}), \mathcal{A}) \leq \left(\frac{\varepsilon}{4z}\right)^z \cdot \text{cost}(\mathcal{A})$ .

In both cases, the lemma follows.  $\square$

## 8 Application of the Framework: New Coreset Bounds for Various Metric Spaces

In this section, we apply the coreset framework to specific metric spaces. For each of them, we show the existence of a small approximate centroid set, and apply Theorem 1 to prove the existence of small coresets.

We recall the definition of a centroid set (Definition 1): given an instance of  $(k, z)$ -clustering and a set of centers  $\mathcal{A}$ , an  $\mathcal{A}$ -approximate centroid set  $\mathbb{C}$  is a set that satisfies the following: for every

solution  $\mathcal{S}$ , there exists  $\tilde{\mathcal{S}} \in \mathbb{C}^k$  such that for all points  $p$  that verifies  $\text{cost}(p, \mathcal{S}) \leq \left(\frac{8z}{\varepsilon}\right)^z \text{cost}(p, \mathcal{A})$  or  $\text{cost}(p, \tilde{\mathcal{S}}) \leq \left(\frac{8z}{\varepsilon}\right)^z \text{cost}(p, \mathcal{A})$ , it holds  $|\text{cost}(p, \mathcal{S}) - \text{cost}(p, \tilde{\mathcal{S}})| \leq \frac{\varepsilon}{z \log(z/\varepsilon)} (\text{cost}(p, \mathcal{S}) + \text{cost}(p, \mathcal{A}))$ .

Theorem 1 states that in case there is an  $\mathcal{A}$ -approximate centroid set  $\mathbb{C}$ , then there is a linear-time algorithm that constructs with probability  $1 - \pi$  a coreset of size

$$O\left(\frac{2^{O(z \log z)} \cdot \log^4 1/\varepsilon}{\min(\varepsilon^2, \varepsilon^z)} (k \log |\mathbb{C}| + \log \log(1/\varepsilon) + \log(1/\pi))\right)$$

## 8.1 Structural Property on Solutions

We also show a structural property on solutions, that we will use in order to show the existence of small approximate centroid sets. Essentially, when replacing a center  $s$  by a center in  $\mathbb{C}$  we will make an error  $\varepsilon \text{cost}(q, \mathcal{A})$  for some  $q$  that we can choose: it is necessary to ensure this error is tiny compared to any  $\text{cost}(p, s) + \text{cost}(p, \mathcal{A})$ .

Given a point  $q$  and a center  $s$ , we say that a point  $p$  is *problematic* with respect to  $q$  and  $s$  when  $\text{dist}(p, \mathcal{A}) + \text{dist}(p, s) \leq \frac{\varepsilon^2}{8z^2} (\text{dist}(q, \mathcal{A}) + \text{dist}(q, s))$ . In that case, we cannot bound the error  $\text{dist}(q, \mathcal{A}) + \text{dist}(q, s)$  by some quantity depending on  $\text{cost}(p, s) + \text{cost}(p, \mathcal{A})$ . However, we show the following:

**Lemma 18.** *Let  $\mathcal{S}$  be a solution, such that any input point  $p$  verifies  $\text{dist}(p, \mathcal{S}) \leq \frac{8z}{\varepsilon} \cdot \text{dist}(p, \mathcal{A})$ . There exists a solution  $\mathcal{S}' \subseteq \mathcal{S}$  such that*

- *for all  $p$ , it holds that  $|\text{cost}(p, \mathcal{S}) - \text{cost}(p, \mathcal{S}')| \leq \frac{\varepsilon}{z \log z/\varepsilon} (\text{cost}(p, \mathcal{S}) + \text{cost}(p, \mathcal{A}))$ , and*
- *for any center  $s \in \mathcal{S}'$ , let  $q = \text{argmin}_{p: \text{dist}(p, s) \leq \frac{10z}{\varepsilon} \text{dist}(p, \mathcal{A})} \text{dist}(p, \mathcal{A}) + \text{dist}(p, s)$ . There is no problematic point with respect to  $q$  and  $s$ .*

*Proof.* First, we show that in case there is a problematic point  $p$  with respect to some  $s$  and  $q$ , then we can serve the whole cluster of  $s$  by  $\mathcal{S}(p)$ , the point that serves  $p$  in  $\mathcal{S}$ . We work in this proof with particular solutions, where points are not necessarily assigned to their closest center. This simplifies the proof, but needs particular care at some moments. In particular, we will ensure that  $\text{dist}(p, \mathcal{S}(p)) \leq \frac{10z}{\varepsilon} \cdot \text{dist}(p, \mathcal{A})$  is always verified. We will then remove inductively centers with problematic points to construct  $\mathcal{S}'$ .

### Removing a center that has a problematic point.

Let  $s \in \mathcal{S}$ , and  $q = \text{argmin}_{p: \text{dist}(p, s) \leq \frac{10z}{\varepsilon} \text{dist}(p, \mathcal{A})} \text{dist}(p, \mathcal{A}) + \text{dist}(p, s)$  as in the statement. Let  $p$  be a problematic point with respect to  $s$  and  $q$ , and  $\mathcal{S}(p)$  its the center serving  $p$  in  $\mathcal{S}$ . First, note that since  $p$  is problematic, it must be that  $\text{dist}(p, \mathcal{S}(p)) \leq \text{dist}(s, p)$ : otherwise,  $p$  would verify  $\text{dist}(p, s) \leq \frac{10z}{\varepsilon} \text{dist}(p, \mathcal{A})$ , and the minimality of  $q$  would ensure that  $p$  is not problematic. Thus, it holds that:

$$\begin{aligned} \text{dist}(s, \mathcal{S}(p)) &\leq \text{dist}(s, p) + \text{dist}(p, \mathcal{S}(p)) \leq 2\text{dist}(s, p) \\ &\leq 2(\text{dist}(s, p) + \text{dist}(p, \mathcal{A})) \leq \frac{\varepsilon^2}{4z^2} (\text{dist}(q, \mathcal{A}) + \text{dist}(q, s)). \end{aligned}$$

Now, let  $p'$  be served by  $s$ . Using the triangle inequality, we immediately get

$$\text{dist}(p', \mathcal{S}(p)) \leq \text{dist}(p', s) + \text{dist}(s, \mathcal{S}(p)) \leq \text{dist}(p', s) + \frac{\varepsilon^2}{4z^2}(\text{dist}(q, \mathcal{A}) + \text{dist}(q, s)).$$

Additionally, it holds that

$$\begin{aligned} \min_{q': \text{dist}(q', \mathcal{S}(p)) \leq \frac{10z}{\varepsilon} \text{dist}(q', \mathcal{A})} \text{dist}(q', \mathcal{A}) + \text{dist}(q', \mathcal{S}(p)) &\leq \text{dist}(p, \mathcal{S}(p)) + \text{dist}(p, \mathcal{A}) \\ &\leq \text{dist}(s, p) + \text{dist}(p, \mathcal{A}) \\ &\leq \frac{\varepsilon^2}{8z^2}(\text{dist}(q, \mathcal{A}) + \text{dist}(q, s)) \end{aligned} \quad (27)$$

Hence, if  $\mathcal{S}(p)$  is removed as well, the error for points served by  $s$  will be an  $\frac{\varepsilon^2}{8z^2}$ -fraction of the initial error. This implies that the total error will not accumulate, as we will now see.

**Constructing  $\mathcal{S}'$ .** To construct  $\mathcal{S}'$ , we proceed iteratively: start with  $\mathcal{S}' = \mathcal{S}$ , and as long as there exists a center  $s$  that have a problematic point  $p$  with respect to it, remove  $s$  and reassign the whole cluster of  $s$  to  $\mathcal{S}'(p)$ , the closest point to  $p$  in the current solution. This process must end, as there is no problematic point when there is a single center.

For a point  $p$ , let  $s_1, \dots, s_j$  be the successive cluster it is reassigned to, with corresponding  $q_1, \dots, q_j$ . Using Eq. (27), it holds that  $\text{dist}(q_{i+1}, \mathcal{A}) + \text{dist}(q_{i+1}, s_{i+1}) \leq \frac{\varepsilon^2}{8z^2}(\text{dist}(q_i, \mathcal{A}) + \text{dist}(q_i, s_i))$ . Hence, the distance increase for  $p$  is geometric: using that  $\text{dist}(q_1, \mathcal{A}) + \text{dist}(q_1, s_1) \leq \text{dist}(p, \mathcal{A}) + \text{dist}(p, \mathcal{S})$  (which holds by minimality of  $q_1$ ), we get that at any given step  $i$  it holds that

$$\begin{aligned} \text{dist}(p, s_i) &\leq \text{dist}(p, \mathcal{S}) + (\text{dist}(p, \mathcal{A}) + \text{dist}(p, \mathcal{S})) \sum_{j=1}^i \left( \frac{\varepsilon^2}{8z^2} \right)^j \\ &\leq \text{dist}(p, \mathcal{S}) + \frac{\varepsilon^2}{4z^2}(\text{dist}(p, \mathcal{A}) + \text{dist}(p, \mathcal{S})) \\ &\leq \frac{8z}{\varepsilon} \cdot \text{dist}(p, \mathcal{A}) + \frac{\varepsilon^2}{4z^2} \cdot \left(1 + \frac{8z}{\varepsilon}\right) \text{dist}(p, \mathcal{A}) \\ &\leq \frac{10z}{\varepsilon} \cdot \text{dist}(p, \mathcal{A}), \end{aligned}$$

as promised in order to remove centers.

Last, we show that the first bullet of the lemma holds. We consider now the standard assignment:  $p$  is assigned to its closest center of  $\mathcal{S}'$ , instead of  $s_j$ . First, since we only removed centers, it holds that  $\text{cost}(p, \mathcal{S}) \leq \text{cost}(p, \mathcal{S}')$ . Second, using Lemma 1, we have for any  $\varepsilon' = \frac{\varepsilon}{z \log z / \varepsilon}$ :

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

Hence, we conclude that

$$|\text{cost}(p, \mathcal{S}') - \text{cost}(p, \mathcal{S})| \leq \frac{\varepsilon}{z \log z / \varepsilon} (\text{cost}(p, \mathcal{S}) + \text{cost}(p, \mathcal{A})),$$

which concludes the lemma.  $\square$

## 8.2 In Metrics with Bounded Doubling Dimension

We start by defining the Doubling Dimension of a metric space, and stating a key lemma.

Consider a metric space  $(X, \text{dist})$ . For a point  $p \in X$  and an integer  $r \geq 0$ , we let  $\beta(p, r) = \{x \in X \mid \text{dist}(p, x) \leq r\}$  be the *ball* around  $p$  with radius  $r$ .

**Definition 5.** *The doubling dimension of a metric is the smallest integer  $d$  such that any ball of radius  $2r$  can be covered by  $2^d$  balls of radius  $r$ .*

Notably, the Euclidean space  $\mathbb{R}^d$  has doubling dimension  $\theta(d)$ .

A  $\gamma$ -net of  $V$  is a set of points  $X \subseteq V$  such that for all  $v \in V$  there is an  $x \in X$  such that  $\text{dist}(v, x) \leq \gamma$ , and for all  $x, y \in X$  we have  $\text{dist}(x, y) > \gamma$ . A net is therefore a set of points not too close to each other, such that every point of the metric is close to a net point. The following lemma bounds the cardinality of a net in doubling metrics.

**Lemma 19** (from Gupta et. al [GKL03]). *Let  $(V, \text{dist})$  be a metric space with doubling dimension  $d$  and, diameter  $D$ , and let  $X$  be a  $\gamma$ -net of  $V$ . Then  $|X| \leq 2^{d \cdot \lceil \log_2(D/\gamma) \rceil}$ .*

The goal of this section is to prove the following lemma. Combined with Theorem 1, it ensures the existence of small coresets in graphs with small doubling dimension.

**Lemma 20.** *Let  $M = (X, \text{dist})$  be a metric space with doubling dimension  $d$ , let  $P \subset X$ , let  $k$  and  $z$  be positive integers and let  $\varepsilon > 0$ . Further, let  $\mathcal{A}$  be a  $c_{\mathcal{A}}$ -approximate solution with at most  $k$  centers. There exists an  $\mathcal{A}$ -approximate centroid set for  $P$  of size*

$$|P| \cdot \left(\frac{z}{\varepsilon}\right)^{O(d)}$$

A direct corollary of that lemma is the existence of a coreset in Doubling Metrics, as it is enough to show the mere existence of a small centroid set for applying Corollary 2.

**Corollary 4.** *Let  $M = (X, \text{dist})$  be a metric space with doubling dimension  $d$ , and two positive integers  $k$  and  $z$ .*

*There exists an algorithm with running time  $\tilde{O}(nk)$  that constructs an  $\varepsilon$ -coreset for  $(k, z)$ -clustering on  $P \subseteq X$  with size*

$$O\left(\frac{\log^5 1/\varepsilon}{2^{O(z \log z)} \min(\varepsilon^2, \varepsilon^z)} (kd + \log 1/\pi)\right)$$

*Proof.* We first compute a coreset of size  $\tilde{O}(k^3 d \varepsilon^{-2})$  [HJLW18]. Then, combining Theorem 1 and Lemma 20 yields an algorithm constructing a coreset of size

$$O\left(\frac{\log^4 1/\varepsilon}{2^{O(z \log z)} \min(\varepsilon^2, \varepsilon^z)} (kd \log 1/\varepsilon + k \log kd/\varepsilon + \log 1/\pi)\right).$$

If  $\log k > d$  then  $O(\log kd) = O(\log k)$ . If  $d > \log k$  then  $O(kd + k \log kd) = O(kd)$ , hence the claimed bound follows.  $\square$

*Proof of Lemma 20.* For each point  $p \in P$ , let  $c$  be the center to which  $p$  was assigned in  $\mathcal{A}$ . Let  $B(p, (\frac{8z}{\varepsilon}) \text{dist}(p, c))$  be the metric ball centered around  $p$  with radius  $(\frac{8z}{\varepsilon}) \cdot \text{dist}(p, c)$ , and let  $N_p$  be an  $(\frac{\varepsilon}{4z}) \cdot \text{dist}(p, \mathcal{A})$ -net of that ball.

Due to Lemma 19,  $N_p$  has size  $(\varepsilon/z)^{-O(d)}$ . Additionally, let  $s_f$  be a point not in any  $B(p, (\frac{10z}{\varepsilon}) \text{dist}(p, \mathcal{A}))$ , if such a point exist.

Let  $\mathcal{N} := s_f \cup_{p \in Y} N_p$ . We claim that  $\mathcal{N}$  is the desired approximate centroid set.

For a candidate solution  $\mathcal{S}$ , apply first Lemma 18, so that we can assume that for any center  $s \in \mathcal{S}$ , and  $q = \text{argmin}_{p: \text{dist}(p, s) \leq \frac{10z}{\varepsilon} \text{dist}(p, \mathcal{A})} \text{dist}(p, \mathcal{A}) + \text{dist}(p, s)$ , there is no problematic point with respect to  $q$  and  $s$ .

let  $\tilde{\mathcal{S}}$  be the solution obtained by replacing every center  $s \in \mathcal{S}$  by  $\tilde{s} \in \mathbb{C}$  as follows: let  $q = \text{argmin}_{p: \text{dist}(p, s) \leq \frac{10z}{\varepsilon} \text{dist}(p, \mathcal{A})} \text{dist}(p, \mathcal{A}) + \text{dist}(p, s)$ . Pick  $\tilde{s}$  to be the closest point to  $s$  in  $N_q$ . If such a  $q$  does not exist, pick  $\tilde{s} = s_f$ .

Now, let  $p$  be a point such that  $\text{cost}(p, \mathcal{S}) \leq (\frac{8z}{\varepsilon})^z \cdot \text{cost}(p, \mathcal{A})$ , let  $s$  be any center in  $\mathcal{S}$  and  $q$  defined as previously. Then, by construction of the  $\tilde{\mathcal{S}}$ , there is a center  $\tilde{s}$  with  $\text{dist}(s, \tilde{s}) \leq (\frac{\varepsilon}{4z}) \text{dist}(q, \mathcal{A})$  and therefore, using that  $p$  is no problematic:

$$\begin{aligned} \text{cost}(p, \tilde{\mathcal{S}}) &\leq \text{cost}(p, \tilde{s}) \leq (1 + \varepsilon) \text{cost}(p, s) + (1 + z/\varepsilon)^{z-1} \text{cost}(s, \tilde{s}) \\ &\leq (1 + \varepsilon) \text{cost}(p, \mathcal{S}) + (2z/\varepsilon)^{z-1} \left(\frac{\varepsilon}{2z}\right)^z \text{cost}(q, \mathcal{A}) \\ &\leq (1 + \varepsilon) \text{cost}(p, \mathcal{S}) + \varepsilon \text{cost}(q, \mathcal{A}) \end{aligned} \tag{28}$$

$$\leq (1 + \varepsilon) \text{cost}(p, \mathcal{S}) + \varepsilon \text{cost}(p, \mathcal{A}). \tag{29}$$

To show the other direction, for any point in  $\tilde{\mathcal{S}}$  there is a center  $s$  with  $\text{dist}(s, \tilde{s}) \leq (\frac{\varepsilon}{4z}) \text{dist}(q, \mathcal{A})$ . Hence the previous equations apply as well, and we can conclude: for a point  $p$  such that  $\text{cost}(p, \mathcal{S}) \leq (\frac{8z}{\varepsilon})^z \cdot \text{cost}(p, \mathcal{A})$ ,

$$|\text{cost}(p, \mathcal{S}) - \text{cost}(p, \tilde{\mathcal{S}})| \leq \varepsilon (\text{cost}(p, \mathcal{S}) + \text{cost}(p, \mathcal{A})).$$

Rescaling  $\varepsilon$  concludes the lemma: there is an  $\mathcal{A}$ -approximate centroid set with size  $|P| \left(\frac{z^2 \log z/\varepsilon}{\varepsilon}\right)^{O(d)} = |P| \left(\frac{z}{\varepsilon}\right)^{O(d)}$ .  $\square$

## 9 Graphs with Bounded Treewidth

In this section, we show that for graphs with treewidth  $t$ , there exists a small approximate centroid set. Hence, the main framework provides an algorithm computing a small coreset. We first define the treewidth of a graph:

**Definition 6.** A tree decomposition of a graph  $G = (V, E)$  is a tree  $\mathcal{T}$  where each node  $b$  (call a bag) is a subset of  $V$  and the following conditions hold:

- The union of bags is  $V$ ,
- $\forall v \in V$ , the nodes containing  $v$  in  $\mathcal{T}$  form a connected subtree of  $\mathcal{T}$ , and
- for all edge  $(u, v) \in E$ , there is one bag containing  $u$  and  $v$ .

The treewidth of a graph  $G$  is the smallest integer  $t$  such that there exists a tree decomposition with maximum size bag  $t + 1$ .

**Lemma 21.** Let  $G = (V, E)$  be a graph with treewidth  $t$ ,  $X \subseteq V$  and  $k, z > 0$ . Furthermore, let  $\mathcal{A}$  be solution to  $(k, z)$ -clustering for  $X$ . Then, there exists an  $\mathcal{A}$ -approximate centroid set for  $(k, z)$ -clustering on  $V$  of size  $\text{poly}(|X|) \left( \frac{z^2 \log z / \varepsilon}{\varepsilon} \right)^{O(t)}$ .

Applying this lemma with  $X$  yields the direct corollary:

**Corollary 5.** Let  $G = (V, E)$  be a graph with treewidth  $t$ ,  $X \subseteq V$ ,  $k$  and  $z > 0$ .

There exists an algorithm running time  $\tilde{O}(nk)$  that constructs an  $\varepsilon$ -coreset for  $(k, z)$ -clustering on  $X$ , with size

$$O \left( \frac{\log^5 1/\varepsilon}{2^{O(z \log z)} \min(\varepsilon^2, \varepsilon^z)} (k \log k + kt + \log(1/\pi)) \right).$$

*Proof.* Let  $X \subseteq V$ . We start by computing a  $(k, \varepsilon)$ -coreset  $X_1$  of size  $O(\text{poly}(k, 1/\varepsilon, t))$ , using the algorithm from [BBH<sup>+</sup>20]

We now apply our framework to  $X_1$ . Computing an approximation on  $X_1$  takes time  $\tilde{O}(|X_1|k)$ , using the algorithm from Mettu and Plaxton [MP04].

Lemma 21 ensure the existence of an approximate centroid set for  $X_1$  with size  $\text{poly}(|X_1|) \left( \frac{z}{\varepsilon} \right)^{O(t)}$ . Hence, Corollary 2 and the framework developed in the previous sections gives an algorithm that computes an  $\varepsilon$ -coreset of  $X$  with size

$$O \left( \frac{\log^4 1/\varepsilon}{2^{O(z \log z)} \min(\varepsilon^2, \varepsilon^z)} (k \log |X_1| + kt \log 1/\varepsilon + \log(1/\pi)) \right).$$

Using that  $|X_1| = O(\text{poly}(k, \varepsilon, t))$  yields a coreset of size

$$O \left( \frac{\log^5 1/\varepsilon}{2^{O(z \log z)} \min(\varepsilon^2, \varepsilon^z)} (k \log k + kt + \log(1/\pi)) \right).$$

Instead of using [BBH<sup>+</sup>20], one could apply our algorithm repeatedly as in Theorem 3.1 of [BJKW21], to reduce iteratively the number of distinct point consider and to eventually get the same coreset size. The number of repetition needed to achieve that size bound is  $O(\log^* n)$ , where  $\log^*(x)$  is the number of times  $\log$  is applied to  $x$  before the result is at most 1; formally  $\log^*(x) = 0$  for  $x \leq 1$ , and  $\log^*(x) = 1 + \log^* \log x$  for  $x > 1$ . The complexity of this repetition is therefore  $\tilde{O}(nk)$ , and the success probability  $1 - \pi$ , as proven in [BJKW21].  $\square$

For the proof of Lemma 21, we rely on the following structural lemma:<sup>7</sup>.

**Lemma 22** (Lemma 3.7 of [BBH<sup>+</sup>20]). *Given a graph  $G = (V, E)$  of treewidth  $t$ , and  $X \subseteq V$ , there exists a collection  $\mathcal{T}$  of subsets of  $V$  such that:*

1.  $\cup_{A \in \mathcal{T}} A = V$ ,
2.  $|\mathcal{T}| = \text{poly}(|X|)$ ,
3. *For each  $A \in \mathcal{T}$ ,  $|A \cap X| = O(t)$ , and there exists  $P_A \subseteq V$  with  $|P_A| = O(t)$  such that there is no edge between  $A \setminus P_A$  and  $V \setminus (A \cup P_A)$ .*

Our construction relies on the following simple observation. Let  $s$  be a possible center, and  $p$  be a vertex such that  $\text{cost}(p, s) \leq \left(\frac{4z}{\varepsilon}\right)^z \text{cost}(p, \mathcal{A})$ . Let  $A \in \mathcal{T}$  such that  $p \in A$ . Then, either  $s \in A$ , or the path connecting  $p$  to  $s$  has to go through  $P_A$ .

We use this observation as follows: it would be enough to replace a center  $s$  from solution  $\mathcal{S}$  by one that has approximately the same distance of all points of  $P_A$ . The main question is : how should we round the distances to  $P_A$ ? The goal is to classify the potential centers into few classes, such that taking one representative per class gives an approximate centroid set. The previous observation indicates that classifying the centers according to their distances to points of  $P_A$  is enough. However, there are too many different classes: instead, we round those distances.

Ideally, this rounding would ensure that for any point  $p$  and any center  $s$ , all centers in  $s$ 's class have same distance to  $p$ , up to an additive error  $\varepsilon(\text{cost}(p, s) + \text{cost}(p, \mathcal{A}))$ . This would mean rounding the distance from  $s$  to any point in  $P_A$  by that amount – for instance, rounding to the closest multiple of  $\varepsilon(\text{cost}(p, s) + \text{cost}(p, \mathcal{A}))$ . Nonetheless, this way of rounding depends on each point  $p$ : a rounding according to  $p$  may not be suited for another point  $q$ . To cope with that, we will quite naturally round distances according to the point  $p$  that minimizes  $\text{cost}(p, s) + \text{cost}(p, \mathcal{A})$ . Additionally, to ensure that the number of classes stays bounded, it is not enough to round to the closest multiple of  $\varepsilon(\text{cost}(p, s) + \text{cost}(p, \mathcal{A}))$ : we also show that distances bigger than  $\frac{1}{\varepsilon}(\text{cost}(p, s) + \text{cost}(p, \mathcal{A}))$  can be trimmed down to  $\frac{1}{\varepsilon}(\text{cost}(p, s) + \text{cost}(p, \mathcal{A}))$ . That way, for each point of  $P_A$  there are only  $1/\varepsilon^2$  many possible rounded distances.

Hence, a class is defined by a certain point  $p$ , a part  $A$  and by  $|P_A| = t$  many rounded distances: in total, that makes  $\text{poly}(|X|)\varepsilon^{-O(t)}$  many classes. The approximate centroid set contains one representative of each class: this would prove Lemma 21. We now make the argument formal, in particular to show that the error incurred by the trimming is affordable.

*Proof of Lemma 21.* Given a point  $s \in V$  and a set  $A \in \mathcal{T}$ , we call a *distance tuple to  $A$*   $\mathbf{d}_A(s) := (\text{dist}(s, x) \mid \forall x \in X \cap A) + (\text{dist}(s, x) \mid \forall x \in P_A)$ . Let  $q \in X$ : the rounded distance tuple of  $s$  with respect to  $q$  is  $\widetilde{\mathbf{d}}_{A,q}(s)$  defined as follows:

1. For  $x \in X \cap A$  or  $x = q$ ,  $\widetilde{d}(s, x)$  is the multiple of  $\frac{\varepsilon}{z} \cdot \text{dist}(x, \mathcal{A})$  smaller than  $\frac{10z}{\varepsilon} \text{dist}(x, \mathcal{A})$  closest to  $\text{dist}(s, x)$ .
2. For  $y \in P_A$ ,  $\widetilde{d}(s, y)$  is the multiple of  $\frac{\varepsilon^3}{8z^3} \cdot \text{dist}(q, \mathcal{A})$  smaller than  $\frac{200z^3}{\varepsilon^3} \text{dist}(q, \mathcal{A})$  closest to  $\text{dist}(s, y)$ .

---

<sup>7</sup>In the statement of [BBH<sup>+</sup>20], the third item is slightly different. To recover our statement from theirs, take  $P_A = A$  when  $|A| = O(t)$ .



Now, for every  $A \in \mathcal{T}$ ,  $q \in X$  and every rounded distance tuple  $T$  to  $A$  with respect to  $q$  such that  $\exists s : T = \widetilde{\mathbf{d}}_A(s)$ ,  $\mathbb{C}$  contains one point  $s \in A$  having that rounded distance tuple.

**Bounding the size of  $\mathbb{C}$ .** Fix some  $A \in \mathcal{T}$ , and  $q \in X$ . A rounded distance tuple to  $A$  is made of  $O(t)$  many distances. Each of them takes its value among  $\text{poly}(z/\varepsilon)$  possible numbers, due to the rounding. Hence, there are at most  $\left(\frac{z}{\varepsilon}\right)^{O(t)}$  possible rounded distance tuple to  $A$ , and so at most that many points in  $\mathbb{C}$ . Since there are  $\text{poly}(|X|)$  different choices for  $A$  and  $q$ , the total size of  $\mathbb{C}$  is  $\text{poly}(|X|) \left(\frac{z}{\varepsilon}\right)^{O(t)}$ .

**Bounding the error.** We now bound the error induced by approximating a solution  $\mathcal{S}$  by a solution  $\tilde{\mathcal{S}} \subseteq \mathbb{C}$ .

First, by applying Lemma 18, we can assume that for any center  $s \in \mathcal{S}$ , and  $q = \text{argmin}_{p: \text{dist}(p,s) \leq \frac{10z}{\varepsilon} \text{dist}(p,\mathcal{A})} \text{dist}(p,\mathcal{A}) + \text{dist}(p,s)$ , there is no problematic point with respect to  $q$  and  $s$ .

Let  $A \in \mathcal{T}$  such that  $s \in A$ , and  $q = \text{argmin}_{p: \text{dist}(p,s) \leq \frac{10z}{\varepsilon} \text{dist}(p,\mathcal{A})} \text{dist}(p,\mathcal{A}) + \text{dist}(p,s)$ .  $\tilde{s}$  is chosen to have the same rounded distance tuple to  $A$  with respect to  $q$  as  $s$ .  $\tilde{\mathcal{S}}$  is the solution made of all such  $\tilde{s}$ , for  $s \in \mathcal{S}$ .

As in the proof of Lemma 20, we first show that points close to  $s$  have cost preserved in  $\tilde{s}$ . We will later show that points with large distance to  $s$  have also large distance to  $\tilde{s}$ , to ensure that their distance to  $\tilde{\mathcal{S}}$  does not decrease.

Let  $p \in X$  be an input point. By Lemma 18,  $p$  is not problematic with respect to  $s$  and  $q$ . Note that  $s$  is not necessarily the closest center to  $p$ . We aim at showing that  $|\text{cost}(p, s) - \text{cost}(p, \tilde{s})| \leq \varepsilon(\text{cost}(p, s) + \text{cost}(p, \mathcal{A}))$ .

First, when  $p \notin X \cap A$ , we distinguish two subcases:

- either  $\text{dist}(p, s) \leq \frac{200z^3}{\varepsilon^3} \text{dist}(q, \mathcal{A})$ : in that case, let  $x \in p_A$  that is on the shortest path between  $p$  and  $s$ . We have  $\text{dist}(s, x) \leq \frac{200z^3}{\varepsilon^3} \text{dist}(q, \mathcal{A})$ , and so  $s$  and  $\tilde{s}$  have the same rounded distance to  $x$ . Hence,

$$\begin{aligned} \text{dist}(p, \tilde{s}) &\leq \text{dist}(p, x) + \text{dist}(x, \tilde{s}) \leq \text{dist}(p, x) + \text{dist}(x, s) + \frac{\varepsilon^3}{8z^3} \text{dist}(q, \mathcal{A}) \\ &\leq \text{dist}(p, s) + \frac{\varepsilon^3}{8z^3} \cdot \left(\frac{8z^2}{\varepsilon^2}\right) (\text{dist}(p, \mathcal{A}) + \text{dist}(p, s)) \\ &\leq \left(1 + \frac{\varepsilon}{z}\right) \text{dist}(p, s) + \frac{\varepsilon}{z} \text{dist}(p, \mathcal{A}), \end{aligned}$$

The first line implies that  $\text{dist}(p, \tilde{s}) \leq \frac{200z^3}{\varepsilon^3} \text{dist}(q, \mathcal{A})$  as well: we can therefore repeat the argument, choosing  $x$  to be on the shortest path between  $p$  and  $\tilde{s}$  instead, to show that  $\text{dist}(p, s) \leq \left(1 + \frac{\varepsilon}{z}\right) \text{dist}(p, \tilde{s}) + \frac{\varepsilon}{z} \text{dist}(p, \mathcal{A})$ . This implies, using Lemma 1, that  $|\text{cost}(p, \tilde{s}) - \text{cost}(p, s)| \leq \frac{\varepsilon}{z} \cdot (\text{cost}(p, s) + \text{cost}(p, \mathcal{A}))$ .

- Otherwise,  $\text{dist}(p, s) > \frac{200z^3}{\varepsilon^3} \text{dist}(q, \mathcal{A})$ . In that case, we can argue that  $\text{dist}(s, \tilde{s})$  is negligible compared to  $\text{dist}(p, s)$ . Recall that  $\text{dist}(q, s) \leq \frac{10z}{\varepsilon} \text{dist}(q, \mathcal{A})$ .

The rounding ensures that the distance to  $q$  is preserved:  $\text{dist}(q, \tilde{s}) \leq \text{dist}(q, s) + \frac{\varepsilon}{z} \text{dist}(q, \mathcal{A})$ . Hence, we get that

$$\begin{aligned} \text{dist}(s, \tilde{s}) &\leq 2\text{dist}(q, s) + \frac{\varepsilon}{z} \text{dist}(q, \mathcal{A}) \\ &\leq \left( \frac{100z^2}{\varepsilon^2} + \frac{\varepsilon}{z} \right) \cdot \text{dist}(q, \mathcal{A}) \\ &\leq \frac{200z^2}{\varepsilon^2} \cdot \frac{\varepsilon^3}{200z^3} \cdot \text{dist}(p, s) \leq \frac{\varepsilon}{z} \cdot \text{dist}(p, s). \end{aligned}$$

Finally, using Lemma 1, we conclude again that  $|\text{cost}(p, \tilde{s}) - \text{cost}(p, s)| \leq \varepsilon \text{cost}(p, s) + \varepsilon \text{cost}(p, \mathcal{A})$ .

Now, in the other case where  $p \in X \cap A$ , if  $\text{dist}(p, s) \leq \frac{10z}{\varepsilon} \text{dist}(p, \mathcal{A})$ , then the choice of  $\tilde{s}$  ensures that  $|\text{dist}(\tilde{s}, p) - \text{dist}(s, p)| \leq \frac{\varepsilon}{z} \text{dist}(x, \mathcal{A})$  and therefore  $|\text{cost}(p, s) - \text{cost}(p, \tilde{s})| \leq \varepsilon \text{cost}(p, s) + (1 + z/\varepsilon)^{z-1} \text{cost}(s, \tilde{s}) \leq \varepsilon \text{cost}(p, \mathcal{S}) + \varepsilon \text{cost}(p, \mathcal{A})$ . In the last case when  $\text{dist}(p, s) > \frac{10z}{\varepsilon} \text{dist}(p, \mathcal{A})$ , then the rounding enforces  $\text{dist}(p, \tilde{s}) = \frac{10z}{\varepsilon} \text{dist}(p, \mathcal{A})$ .

Hence, in all possible cases, it holds that either  $\text{dist}(p, s)$  and  $\text{dist}(p, \tilde{s})$  are bigger than  $\frac{10z}{\varepsilon} \text{dist}(p, \mathcal{A})$ , or:

$$|\text{cost}(p, \tilde{s}) - \text{cost}(p, s)| \leq \varepsilon \text{cost}(p, s) + \varepsilon \text{cost}(p, \mathcal{A}). \quad (30)$$

To extend that result to the full solutions  $\mathcal{S}$  and  $\tilde{\mathcal{S}}$  instead of a particular center, we note that since  $p$  is interesting,  $\text{dist}(p, \mathcal{S}) \leq \frac{8z}{\varepsilon} \text{dist}(p, \mathcal{A})$ . Hence, we can apply Eq. (30) with  $s$  being the closest point to  $p$  in  $\mathcal{S}$ :  $\text{cost}(p, \tilde{\mathcal{S}}) \leq (1 + \varepsilon) \text{cost}(p, \mathcal{S}) + \varepsilon \text{cost}(p, \mathcal{A})$ .

In particular, this implies that  $\text{dist}(p, \tilde{s}) \leq \frac{10z}{\varepsilon} \text{dist}(p, \mathcal{A})$ . Choose now  $\tilde{s}$  to be the closest point to  $p$  in  $\tilde{\mathcal{S}}$  and  $s$  its corresponding center in  $\mathcal{S}$ . Using Eq. (30) therefore gives:

$$\begin{aligned} \text{cost}(p, s) &\leq \text{cost}(p, \tilde{\mathcal{S}}) + \varepsilon(\text{cost}(p, \mathcal{A}) + \text{cost}(p, s)) \\ \implies \text{cost}(p, s) &\leq \frac{1}{1 - \varepsilon} \text{cost}(p, \tilde{\mathcal{S}}) + \frac{\varepsilon}{1 - \varepsilon} \text{cost}(p, \mathcal{A}) \\ &\leq (1 + 2\varepsilon) \text{cost}(p, \tilde{\mathcal{S}}) + 2\varepsilon \text{cost}(p, \mathcal{A}) \\ \implies \text{cost}(p, \mathcal{S}) &\leq (1 + 2\varepsilon) \text{cost}(p, \tilde{\mathcal{S}}) + 3\varepsilon \text{cost}(p, \mathcal{A}). \end{aligned}$$

Hence, combining those two inequalities yields

$$|\text{cost}(p, \mathcal{S}) - \text{cost}(p, \tilde{\mathcal{S}})| \leq \varepsilon \text{cost}(p, \mathcal{S}) + 2\varepsilon \text{cost}(p, \tilde{\mathcal{S}}) + 4\varepsilon \text{cost}(p, \mathcal{A}).$$

To remove the dependency in  $\text{cost}(p, \tilde{\mathcal{S}})$  from the right hand side, one can upper bound it with  $\text{cost}(p, \mathcal{S}) + |\text{cost}(p, \mathcal{S}) - \text{cost}(p, \tilde{\mathcal{S}})|$ , which yields the following:

$$\begin{aligned} |\text{cost}(p, \mathcal{S}) - \text{cost}(p, \tilde{\mathcal{S}})| &\leq \varepsilon \text{cost}(p, \mathcal{S}) + 2\varepsilon \left( \text{cost}(p, \mathcal{S}) + |\text{cost}(p, \mathcal{S}) - \text{cost}(p, \tilde{\mathcal{S}})| \right) \\ &\quad + 4\varepsilon \text{cost}(p, \mathcal{A}) \\ \iff |\text{cost}(p, \mathcal{S}) - \text{cost}(p, \tilde{\mathcal{S}})| &\leq \frac{1}{1 - 2\varepsilon} (3\varepsilon \text{cost}(p, \mathcal{S}) + 4\varepsilon \text{cost}(p, \mathcal{A})) \\ &\leq 9\varepsilon \text{cost}(p, \mathcal{S}) + 12\varepsilon \text{cost}(p, \mathcal{A}) \end{aligned}$$

Finally, rescaling  $\varepsilon$  concludes. □

## 10 Planar Graphs

The goal of this section is to prove the existence of small centroid sets for planar graph, analogously to the treewidth case. This is the following lemma:

**Lemma 23.** *Let  $G = (V, E)$  be an edge-weighted planar graph, a set  $X \subseteq V$  and two positive integers  $k$  and  $z$ . Furthermore, let  $\mathcal{A}$  be a solution of  $(k, z)$ -clustering of  $X$ .*

*Then, there exists an  $\mathcal{A}$ -approximate centroid set for  $(k, z)$ -clustering on  $V$  of size  $\text{poly}(|X|) \cdot \exp(O(z^3 \varepsilon^{-3} \log z / \varepsilon))$ .*

As for treewidth, this lemma implies the following corollary:

**Corollary 6.** *Let  $G = (V, E)$  be an edge-weighted planar graph, a set  $X \subseteq V$ , and two positive integers  $k$  and  $z$ .*

*There exists an algorithm with running time  $\tilde{O}(nk)$  that constructs an  $\varepsilon$ -coreset for  $(k, z)$ -clustering on  $X$  with size*

$$O\left(\frac{\log^5 1/\varepsilon}{2^{O(z \log z)} \min(\varepsilon^2, \varepsilon z)} \left(k \log^2 k + \frac{k \log k}{\varepsilon^3} + \log 1/\pi\right)\right)$$

The big picture is the same as for treewidth. As in the treewidth case, planar graph can be broken into  $\text{poly}(X)$  pieces, each containing at most 2 vertices of  $X$ . The main difference is in the nature of the separators: while treewidth admit small vertex separators, the region in the planar decomposition are bounded by a few number of shortest path instead. This makes the previous argument void: we cannot round distances to all vertices in the boundary of a region. We show how to bypass this, using the fact that separators are shortest paths: it is enough to round distances to a well-chosen subset of the paths, as we will argue in the proof.

Formally, the decomposition is as follows:

**Lemma 24** (Lemma 4.5 of [BJKW21], see also [EKM14]). *For every edge-weighted planar graph  $G = (V, E)$  and subset  $X \subseteq V$ , there exists a collection of subsets of  $V$   $\Pi := \{V_i\}$  with  $|\Pi| = \text{poly}(|X|)$  and  $\cup V_i = V$  such that, for every  $V_i \in \Pi$ :*

- $|V_i \cap X| = O(1)$ , and
- *there exists a collection of shortest paths  $\mathcal{P}_i$  with  $|\mathcal{P}_i| = O(1)$  such removing the vertices of all paths of  $\mathcal{P}_i$  disconnects  $V_i$  from  $V \setminus V_i$ .*

As for treewidth, we proceed as follows: given the decomposition of Lemma 24, for any center  $s \in V_i$ , we identify a point  $q$  and round distances from  $s$  to  $\mathcal{P}_i$  according to  $\text{dist}(q, \mathcal{A})$ .  $\mathbb{C}$  contains one point  $\tilde{s}$  with the same rounded distances as  $s$ , and we will argue that  $\tilde{s}$  can replace  $s$ . As mentioned, we cannot round distances to the whole shortest-paths  $\mathcal{P}_i$ . Instead, we show that it is enough to round distances from  $s$  to points on the boundary of  $V_i$  that are close to  $q$ : since the boundary consists of shortest path, it is possible to discretize that set.

*Proof of Lemma 23.* Let  $\Pi = \{V_i\}$  be the decomposition given by Lemma 24. For any  $V_i$  and any  $q \in X$ , we define a set of *landmarks*  $\mathcal{L}_{i,q}$  as follows: for any  $P \in \mathcal{P}_i$ , let  $\mathcal{L}_{i,q,P}$  be a  $\frac{\varepsilon}{z} \cdot \text{dist}(q, \mathcal{A})$ -net of  $P \cap B\left(q, \frac{90z^2}{\varepsilon^2} \cdot \text{dist}(q, \mathcal{A})\right)$ . Note that since  $P$  is a shortest path, the total length of  $P \cap B\left(q, \frac{90z^2}{\varepsilon^2} \cdot \text{dist}(q, \mathcal{A})\right)$  is at most  $\frac{180z^2}{\varepsilon^2} \cdot \text{dist}(q, \mathcal{A})$ , and so the net has size at most  $\frac{180z^3}{\varepsilon^3}$ . We define  $\mathcal{L}_{i,q} = (V_i \cap X) \cup_{P \in \mathcal{P}_i} \mathcal{L}_{i,q,P}$ .

**Rounding the distances to  $\mathcal{L}_{i,q}$**  We now describe how we round distances to landmarks, and define  $\mathbb{C}$  such that for each possible distance tuple,  $\mathbb{C}$  contains a point having that distance tuple. Formally, given a point  $s \in V_i$  and a point  $q \in X$ , the distance tuple  $\mathbf{d}_q(s)$  of  $s$  is defined as  $\mathbf{d}_q(s) = (\text{dist}(s, x) \mid \forall x \in X \cap V_i) + (\text{dist}(s, y) \mid \forall y \in \mathcal{L}_{i,q}, \forall i)$ . The *rounded distance tuple*  $\tilde{\mathbf{d}}_q(s)$  of  $s$  is defined as follows :

- For  $x \in X \cap V_i$  or  $x = q$ ,  $\tilde{d}(s, x)$  is the multiple of  $\frac{\varepsilon}{z} \text{dist}(x, \mathcal{A})$  smaller than  $\frac{10z}{\varepsilon} \text{dist}(x, \mathcal{A})$  closest to  $\text{dist}(s, x)$ .
- For  $y \in \mathcal{L}_{i,q}$ ,  $\tilde{d}(s, y)$  is the multiple of  $\frac{\varepsilon}{z} \cdot \text{dist}(q, \mathcal{A})$  smaller than  $\frac{90z^2}{\varepsilon^2} \text{dist}(q, \mathcal{A})$  closest to  $\text{dist}(s, y)$ .

The set  $\mathbb{C}$  is constructed as follows: for every  $V_i$  and every  $q$ , for every rounded distance tuple  $\{\tilde{\mathbf{d}}_q(p)\}$ , add to  $\mathbb{C}$  a point that realizes this rounded distance tuple (if such a point exists).

It remains to show both that  $\mathbb{C}$  has size  $\text{poly}(|X|) \exp(O(z^3 \varepsilon^{-3} \log z / \varepsilon))$ , and that  $\mathbb{C}$  contains good approximation of each center of any given solution.

**Size analysis.** For any given  $V_i$  and  $q$ , there are  $\left(\frac{90z^3}{\varepsilon^3}\right)^{|\mathcal{L}_{i,q}|}$  possible rounded distances. As explained previously,  $|\mathcal{L}_{i,q}| = O(z^3 / \varepsilon^3)$ .

There are  $|V|$  choices of  $q$ , and Lemma 24 ensures that there are  $\text{poly}(|X|)$  choices for  $V_i$ .

Hence, the total size of  $\mathbb{C}$  is at most  $\text{poly}(|X|) \cdot \exp(O(z^3 \varepsilon^{-3} \log z / \varepsilon))$ .

**Error analysis.** We now show that for all solution  $\mathcal{S}$ , every center can be approximated by a point of  $\mathbb{C}$ . First, by applying Lemma 18, we can assume that for any center  $s \in \mathcal{S}$ , and  $q = \text{argmin}_{p: \text{dist}(p,s) \leq \frac{10z}{\varepsilon} \text{dist}(p,\mathcal{A})} \text{dist}(p, \mathcal{A}) + \text{dist}(p, s)$ , there is no problematic point with respect to  $q$  and  $s$ .

Let  $S$  be some cluster of  $\mathcal{S}$ , with center  $s$ . As in Lemma 20 and 21, we aim at showing how to find  $\tilde{s} \in \mathbb{C}$  such that, for every  $p \in X \cap S$  with  $\text{dist}(p, \mathcal{S}) \leq \frac{10z}{\varepsilon} \cdot \text{dist}(p, \mathcal{A})$ , we have  $|\text{cost}(p, s) - \text{cost}(p, \tilde{s})| \leq 3\varepsilon (\text{cost}(p, s) + \text{cost}(p, \mathcal{A}))$ .

For this, let  $V_i$  be a part of  $\Pi$  containing  $s$ , and  $\mathcal{P}_i$  be the paths given by Lemma 24. We let  $q := \text{argmin}_{p \in X: \text{dist}(p,s) \leq \frac{10z}{\varepsilon} \text{dist}(p,\mathcal{A})} \text{dist}(p, s) + \text{dist}(p, \mathcal{A})$ . We define  $\tilde{s}$  to be the point of  $\mathbb{C}$  that has the same rounded distance tuple to  $\mathcal{L}_{i,q}$  as  $s$ . Let  $\tilde{\mathcal{S}}$  be the solution constructed from  $\mathcal{S}$  that way. We show now that  $\tilde{\mathcal{S}}$  has the required properties.

First, if  $p \notin V_i$ , then we show how to use that  $s$  and  $\tilde{s}$  have the same rounded distances to  $\mathcal{L}_{i,q}$ .

- If  $\text{dist}(p, s) > \frac{21z^2}{\varepsilon^2} \cdot \text{dist}(q, \mathcal{A})$ , we argue that  $d(s, \tilde{s})$  is negligible. The argument is exactly alike the one from Lemma 21, we repeat it for completeness.

The rounding ensures that the distance to  $q$  is preserved:  $\text{dist}(q, \tilde{s}) \leq \text{dist}(q, s) + \frac{\varepsilon}{z} \text{dist}(q, \mathcal{A})$ , and therefore:

$$\begin{aligned} \text{dist}(s, \tilde{s}) &\leq 2\text{dist}(q, s) + \frac{\varepsilon}{z} \cdot \text{dist}(q, \mathcal{A}) \\ &\leq \left( \frac{20z}{\varepsilon} + \frac{\varepsilon}{z} \right) \cdot \text{dist}(q, \mathcal{A}) \\ &\leq \frac{21z}{\varepsilon} \cdot \frac{\varepsilon^2}{21z^2} \cdot \text{dist}(p, s) \leq \frac{\varepsilon}{z} \cdot \text{dist}(p, s). \end{aligned}$$

Hence, using the modified triangle inequality Lemma 1, we can conclude:  $|\text{cost}(p, \tilde{s}) - \text{cost}(p, s)| \leq \varepsilon \text{cost}(p, s) + \varepsilon \text{cost}(p, \mathcal{A})$ .

- Otherwise,  $\text{dist}(p, s) \leq \frac{21z^2}{\varepsilon^2} \cdot \text{dist}(q, \mathcal{A})$  and we can make use of the landmarks. Since  $p \notin V_i$  the shortest-path  $p \rightsquigarrow s$  and crosses  $\mathcal{P}_i$  at some vertex  $x$ .

First, it holds that  $\text{dist}(x, q) \leq \text{dist}(x, s) + \text{dist}(s, q) \leq \text{dist}(p, s) + \text{dist}(s, q) \leq (\frac{10z}{\varepsilon} + \frac{8z^2}{\varepsilon^2}) \text{dist}(q, \mathcal{A})$ , hence  $x$  is in  $P \cap B(q, \frac{90z^2}{\varepsilon^2} \text{dist}(q, \mathcal{A}))$ . By choice of landmarks, this implies that there is  $\ell \in \mathcal{L}_{i,q}$ , with  $\text{dist}(x, \ell) \leq \frac{\varepsilon}{z} \text{dist}(q, \mathcal{A})$ . To show that  $s$  and  $\tilde{s}$  have the same distance to  $\ell$ , it is necessary to show that  $s$  is not too far away from  $\ell$ :

$$\begin{aligned} \text{dist}(s, \ell) &\leq \text{dist}(s, x) + \frac{\varepsilon}{z} \cdot \text{dist}(q, \mathcal{A}) \leq \text{dist}(p, s) + \frac{\varepsilon}{z} \text{dist}(q, \mathcal{A}) \\ &\leq \frac{21z^2}{\varepsilon^2} \cdot \text{dist}(q, \mathcal{A}) + \frac{\varepsilon}{z} \cdot \text{dist}(q, \mathcal{A}) \end{aligned}$$

Hence,  $s$  is close enough to  $\ell$  to ensure that  $\tilde{s}$  has the same rounded distance to  $\ell$  as  $s$ , and we get:

$$\begin{aligned} \text{dist}(p, \tilde{s}) &\leq \text{dist}(p, \ell) + \text{dist}(\ell, \tilde{s}) \\ &\leq \text{dist}(p, \ell) + \text{dist}(\ell, s) + \frac{\varepsilon}{z} \cdot \text{dist}(q, \mathcal{A}) \\ &\leq \text{dist}(p, x) + \text{dist}(x, s) + 2\text{dist}(x, \ell) + \frac{\varepsilon}{z} \cdot \text{dist}(q, \mathcal{A}) \\ &= \text{dist}(p, s) + \frac{3\varepsilon}{z} \cdot \text{dist}(q, \mathcal{A}) \end{aligned}$$

First, this ensures that  $\text{dist}(p, \tilde{s}) \leq \frac{8z^2}{\varepsilon^2} \cdot \text{dist}(q, \mathcal{A})$ , and so we can repeat the argument switching roles of  $s$  and  $\tilde{s}$ , to get  $|\text{dist}(p, \tilde{s}) - \text{dist}(p, s)| \leq \frac{3\varepsilon}{z} \cdot \text{dist}(q, \mathcal{A})$ . Using that  $p$  is not problematic with respect to  $q$  and  $s$ , we can conclude that

$$|\text{dist}(p, \tilde{s}) - \text{dist}(p, s)| \leq \frac{3\varepsilon}{z} \cdot (\text{dist}(p, \mathcal{A}) + \text{dist}(p, s)).$$

In turn, using Lemma 1, we conclude:

$$|\text{cost}(p, \tilde{s}) - \text{cost}(p, s)| \leq \varepsilon \cdot (\text{cost}(p, \mathcal{A}) + \text{cost}(p, s)).$$

Finally, in the case where  $p \in V_i$ , then we get either  $|\text{dist}(p, \tilde{s}) - \text{dist}(p, s)| \leq \frac{\varepsilon}{z} \text{dist}(p, \mathcal{A})$  and we are done, or both  $\text{dist}(p, \tilde{s})$  and  $\text{dist}(p, s)$  are bigger than  $\frac{8z}{\varepsilon} \text{dist}(p, \mathcal{A})$ .

We can now conclude, exactly as in the treewidth case: in all possible cases, it holds that either  $\text{dist}(p, s)$  and  $\text{dist}(p, \tilde{s})$  are bigger than  $\frac{10z}{\varepsilon} \text{dist}(p, \mathcal{A})$ , or:

$$|\text{cost}(p, \tilde{s}) - \text{cost}(p, s)| \leq \varepsilon \text{cost}(p, s) + \varepsilon \text{cost}(p, \mathcal{A}). \quad (31)$$

To extend that result to the full solutions  $\mathcal{S}$  and  $\tilde{\mathcal{S}}$  instead of a particular center, we note that since  $p$  is interesting,  $\text{dist}(p, \mathcal{S}) \leq \frac{8z}{\varepsilon} \text{dist}(p, \mathcal{A})$ . Hence, we can apply Eq. (31) with  $s$  being the closest point to  $p$  in  $\mathcal{S}$ :  $\text{cost}(p, \tilde{\mathcal{S}}) \leq (1 + \varepsilon) \text{cost}(p, \mathcal{S}) + \varepsilon \text{cost}(p, \mathcal{A})$ .

In particular, this implies that  $\text{dist}(p, \tilde{s}) \leq \frac{10z}{\varepsilon} \text{dist}(p, \mathcal{A})$ . Chose now  $\tilde{s}$  to be the closest point to  $p$  in  $\tilde{\mathcal{S}}$  and  $s$  its corresponding center in  $\mathcal{S}$ . Using Eq. (31) therefore gives:

$$\begin{aligned} \text{cost}(p, \mathcal{S}) &\leq (1 + \varepsilon) \text{cost}(p, \tilde{\mathcal{S}}) + \varepsilon \text{cost}(p, \mathcal{A}) \\ &\leq (1 + \varepsilon)^2 \text{cost}(p, \mathcal{S}) + \varepsilon(2 + \varepsilon) \text{cost}(p, \mathcal{A}) \\ &\leq (1 + 3\varepsilon) \text{cost}(p, \mathcal{S}) + 3\varepsilon \text{cost}(p, \mathcal{A}). \end{aligned}$$

Rescaling  $\varepsilon$  and combining the two inequality concludes.  $\square$

## 11 Minor-Excluded Graphs

A graph  $H$  is a *minor* of a graph  $G$  if it can be obtained from  $G$  by deleting edges and vertices and contracting edges.

We are interested here in families of graph excluding a fixed minor  $H$ , i.e. none of the graph in the family contains  $H$  as a minor. The graphs are weighted: we assume that for each edge, its value is equal to shortest-path distance between its two endpoints.

The goal of this section is to prove the following lemma, analogous to Lemma 23.

**Lemma 25.** *Let  $G = (V, E)$  be an edge-weighted graph that excludes a minor of fixed size, a set  $X \subseteq V$  and two positive integers  $k$  and  $z$ . Furthermore, let  $\mathcal{A}$  be a solution of  $(k, z)$ -clustering of  $X$ .*

*Then, there exists an  $\mathcal{A}$ -approximate centroid set for  $(k, z)$ -clustering on  $V$  of size  $\exp(O(\log^2 |X| + \log |X|/\varepsilon^4))$ .*

As for the bounded treewidth and planar cases, this lemma implies the following corollary:

**Corollary 7.** *Let  $G = (V, E)$  be an edge-weighted graph that excludes a fixed minor, and two positive integers  $k$  and  $z$ .*

*There exists an algorithm with running time  $\tilde{O}(nk)$  that constructs an  $\varepsilon$ -coreset for  $(k, z)$ -clustering on  $V$  with size*

$$O\left(\frac{\log^5 1/\varepsilon}{2^{O(z \log z)} \min(\varepsilon^2, \varepsilon^z)} \left(k \log^2 k \log(1/\varepsilon) + \frac{k \log k}{\varepsilon^4} + \log 1/\pi\right)\right)$$

The big picture is the same as for planar graphs. Minor-free graphs have somewhat nice separators, that we can use to select centers. However, those separators are not shortest paths in the original graph, as described in the next structural lemma.

**Lemma 26** (Lemma 4.12 in [BJKW21], from Theorem 1 in [AG06]). *For every edge-weighted graph  $G = (V, E)$  excluding some fixed minor, and subset  $X \subseteq V$ , there exists a collection of subsets of  $V$   $\Upsilon := \{\Pi_i\}$  with  $|\Upsilon| = \text{poly}(|X|)$  and  $\cup \Pi_i = V$  such that, for every  $\Pi_i \in \Upsilon$ :*

- $|\Pi_i \cap X| = O(1)$ , and
- *there exists a groups of paths  $\{\mathcal{P}_j^i\}$  with  $|\cup \mathcal{P}_j^i| = O(\log |X|)$  such that removing the vertices of all paths of  $\mathcal{P}_i$  disconnects  $\Pi_i$  from  $V \setminus \Pi_i$ , and such that paths in  $\mathcal{P}_j^i$  are shortest-paths in the graph  $G_j^i := G \setminus (\cup_{j' < j} \mathcal{P}_{j'}^i)$ .*

The general sketch of the proof is as follows: we consider the boundary  $B$  of a region  $\Pi_i$ , and enumerate all possible tuple of distances from a point inside the leaf to the boundary. For each tuple, we include in  $\mathbb{C}$  a point realizing it. Of course, this would lead to a set  $\mathbb{C}$  way too big: the boundary of each leaf consists of too many points, and there are too many distances possible. For that, we show how to discretize the boundary, and how to round distances from a point to the boundary.

Discretizing the boundary is not as easy as in the planar case, as the separating paths are not shortest paths in the original graph  $G$ . A separating path  $P \in \mathcal{P}_j$ , however, is a shortest path in the graph  $G_j^i := G \setminus (\cup_{j' < j} \mathcal{P}_{j'}^i)$ .

As in the planar case, we therefore start from the point  $q$  closest to  $s$  in the graph  $G_j^i$ . Note here that we cannot infer much on the distances in the original graph  $G$ : for this reason, we are not able to apply Lemma 18, and we need to present a whole different argument.

We will assume that we know  $D = \text{dist}_j(q, s)$ , where  $\text{dist}_j$  is the distance in the graph  $G_j^i$ . In that case, we can simply take an  $\varepsilon D$ -net of  $P \cap B_j(q, D)$ , where  $B_j(q, D)$  is the ball centered at  $q$  and of radius  $D$  in  $G_j^i$ . This net has size  $O(1/\varepsilon^2)$ , as  $P$  is a shortest path in  $G_j^i$ . Then, if  $\tilde{s}$  has same distances to this net as  $s$ , we are able to show as in the previous cases that for any point separated from  $s$  by  $P$ ,  $\text{dist}(p, \tilde{s}) \lesssim \text{dist}(p, s)$ ; and for any point separated from  $\tilde{s}$  by  $P$ ,  $\text{dist}(p, s) \lesssim \text{dist}(p, \tilde{s})$ .

To estimate  $\text{dist}_i(q, s)$ , we proceed as follows: either  $\text{dist}_i(q, s) \approx \text{dist}_i(q, q_2)$  for some  $q_2 \in X$ , or not. In the first case, we can pick such a  $q_2$ . In the second case, we will need to ensure that when  $p$  is such that  $\text{dist}_i(p, q) \gg \text{dist}_i(q, s)$ , then  $\tilde{s}$  stays close to  $q$ . When  $p$  is such that  $\text{dist}_i(q, p) \ll \text{dist}_i(q, s)$ , then  $p$  and  $q$  are essentially located at the same spot, and we ensure that  $\tilde{s}$  stays far from  $q$ .

### 11.1 Construction of the centroid set.

From Lemma 26, we have a decomposition into regions  $\Upsilon = \{\Pi_i\}$ . In this argument, we fix a region  $\Pi_j \in \Upsilon$ .  $\Pi_j$  is bounded by  $O(\log |X|)$  paths  $P_1, \dots, P_m$  and  $P_i$  is a shortest path in some graph  $G_i$ , subgraph of  $G$ : if  $P_i \in \mathcal{P}_\ell^j$ , then  $G_i := G_\ell^j$ . We change the indexing for simplicity, and let  $\Pi = \Pi_j$ . We let  $\text{dist}_i$  be the distances in the graph  $G_i$ .

We consider two ways of rounding the distances. The first starts from a point  $q_1 \in X$ , and is useful when there is  $q_2 \in X$  such that  $\varepsilon \text{dist}_i(q_1, s) \leq \text{dist}_i(q_1, q_2) \leq \frac{1}{\varepsilon} \text{dist}_i(q_1, s)$ .

Along each paths, we designate portals as follows. Consider a path  $P_i$ . For any pair of vertices  $q_1, q_2 \in X$ , let  $D = \text{dist}_i(q_1, q_2) + \text{dist}(q_2, \mathcal{A})$  and let  $N_{i,q_1,q_2}$  be an  $\varepsilon^2 D$ -net of  $P_i \cap B_i(q_1, \frac{D}{\varepsilon^2})$ , where  $B_i(q, \frac{D}{\varepsilon^2})$  is the ball centered at  $q$  and of radius  $\frac{D}{\varepsilon^2}$  in  $G_i$ .

For each possible  $q_1, q_2$  and any point  $s \in \Pi$ , we consider the following distance tuple:  $(\text{dist}_i(s, n), \forall n \in N_{i,q_1,q_2}) \cup (\text{dist}_i(s, q_1)) \cup (\text{dist}_i(x, s), \forall x \in \Pi \cap X)$ . We define the rounded tuple  $\tilde{d}^1(q_1, q_2) := \left( \tilde{d}^1(s, n), \forall n \in N_{i,q_1,q_2} \right) \cup \left( \tilde{d}^1(s, q_1) \right) \cup \left( \tilde{d}^1(x, s), \forall x \in \Pi \cap X \right)$ , where

- $\tilde{d}^1(s, n)$  is the multiple of  $\varepsilon^2 D$  closest to  $\min\left(\frac{3D}{\varepsilon^2}, \text{dist}_i(s, n)\right)$ .
- $\tilde{d}^1(s, q_1)$  is the multiple of  $\varepsilon D$  closest to  $\text{dist}_i(s, q_1)$  and smaller than  $\frac{3D}{\varepsilon}$ .
- for any  $x \in \Pi \cap X$ ,  $\tilde{d}^1(x, s)$  is the closest multiple of  $\varepsilon \text{dist}(x, \mathcal{A})$  to  $\text{dist}_i(x, s)$  smaller than  $\frac{1}{\varepsilon} \cdot \text{dist}(x, \mathcal{A})$ .

We also consider another rounding, which will be helpful when for all points,  $\text{dist}(p, \mathcal{A}) + \text{dist}_i(q, q_1) \notin [\varepsilon \text{dist}_i(q_1, s), \frac{1}{\varepsilon} \text{dist}_i(q_1, s)]$ .

For any  $q_1, q_3$ , and  $q_4$  in  $X$ ,  $\tilde{d}^2(q_1, q_3, q_4) = \top$  when  $\frac{1}{\varepsilon} \cdot (\text{dist}_i(q_1, q_4) + \text{dist}(q_4, \mathcal{A})) < \text{dist}_i(q_1, s) < \varepsilon \cdot (\text{dist}_i(q_1, q_3) + \text{dist}(q_3, \mathcal{A}))$ , and  $\tilde{d}^2(q_1, q_3, q_4) = \perp$  otherwise.  $q_3$  or  $q_4$  may be unspecified. In that case, the corresponding part of the inequality is dropped.<sup>8</sup>

To construct  $\mathbb{C}$ , we proceed as follows: for any region  $\Pi \in \Upsilon$  given by Lemma 26, and for any path  $P_i$  in the boundary of  $\Pi$ , select a rounding  $\tilde{d}_i^1(q_1^i, q_2^i)$  or  $\tilde{d}_i^2(q_1^i, q_3^i, q_4^i)$ . If there is any, pick one point  $s$  achieving all those rounding distances, and add  $s$  to  $\mathbb{C}$ .

We will show Lemma 25 using this centroid set. For that, we break the proof into two parts: first, the size of  $\mathbb{C}$  is the desired one; then,  $\mathbb{C}$  is indeed an approximate centroid set.

## 11.2 $\mathbb{C}$ has Small Size

**Lemma 27.**  $\mathbb{C}$  constructed as previously has size  $\exp\left(O(\log^2 |X| + \log |X|/\varepsilon^4)\right)$ .

*Proof.* Fix a region  $\Pi$ , a path  $P_i$  on  $\Pi$ 's boundary, and points  $q_1, q_2$ . There are  $O(1/\varepsilon^4)$  points in the net  $N_{i,q_1,q_2}$ , and  $O(1)$  in  $\Pi \cap X$ . For each of those points, there are at most  $3/\varepsilon^4$  many choices of distances.

For a fixed region  $\Pi$ , path  $P_i$  on  $\Pi$ 's boundary, and points  $q_1, q_2, q_3$ , there only 2 possible different  $\tilde{d}^2(q_1, q_2, q_3)$ .

Now, there are  $\text{poly}(|X|)$  many regions  $\Pi$ , and for each of them  $O(\log |X|)$  many paths  $P_i$ . For each path, there are at most  $|X|^3$  choices of  $q_j$  for it, so in total  $|X|^{O(\log |X|)}$  possible choices. Each choice gives rise to  $O(\log |X|) \cdot O(1/\varepsilon^4)$  many net points, each having at most  $3/\varepsilon^4$  many choices of distances.

So, in total, there are

$$|X|^{O(\log |X|)} \cdot (1/\varepsilon)^{O(\log |X|/\varepsilon^4)}$$

---

<sup>8</sup>When  $q_3$  is unspecified,  $\tilde{d}^2(q_1, q_3, q_4) = \top$  when  $\frac{1}{\varepsilon} \cdot (\text{dist}_i(q_1, q_4) + \text{dist}(q_4, \mathcal{A})) < \text{dist}_i(q_1, s)$ , and  $\tilde{d}^2(q_1, q_3, q_4) = \perp$  otherwise. When  $q_4$  is unspecified,  $\tilde{d}^2(q_1, q_3, q_4) = \top$  when  $\text{dist}_i(q_1, s) < \varepsilon \cdot (\text{dist}_i(q_1, q_3) + \text{dist}(q_3, \mathcal{A}))$ , and  $\tilde{d}^2(q_1, q_3, q_4) = \perp$  otherwise.



many choices of rounded distances tuples. That upper bounds the size of  $\mathbb{C}$ , as there is at most one point per rounded distance tuple.  $\square$

### 11.3 $\mathbb{C}$ is an Approximate Centroid Set

**Construction of solution  $\tilde{\mathcal{S}}$ .** Now, for a point  $s \in \mathcal{S}$ , we construct  $\tilde{s}$  as follows, using the rounded distance tuples. Let  $\Pi$  be a region of  $\Upsilon$  that contains  $s$ . For each path  $P_i$  in the boundary of  $\Pi$ , we define the tuple  $\tilde{d}_i$  as follows. Let  $q_1^i$  be the point minimizing  $\text{dist}_i(p, s)$ . Now, we distinguish two cases:

- either there is some  $q_2^i$  such that  $\varepsilon \text{dist}_i(q_1^i, s) \leq \text{dist}_i(q_1^i, q_2^i) + \text{dist}(q_2^i, \mathcal{A}) \leq \frac{1}{\varepsilon} \text{dist}_i(q_1^i, s)$ . Then  $\tilde{d}_i$  is the tuple  $\tilde{d}^1(q_1^i, q_2^i)$ .
- or there exists points  $q$  with  $\text{dist}_i(q_1^i, q) + \text{dist}(q, \mathcal{A}) > \frac{1}{\varepsilon} \text{dist}_i(q_1^i, s)$ : let  $q_3^i$  be such a point, with smallest  $\text{dist}_i(q_1^i, q_3^i) + \text{dist}(q_3^i, \mathcal{A})$  value. If there are no such points,  $q_3^i$  is unspecified.

If there are points  $q$  with  $\text{dist}_i(q_1^i, q) + \text{dist}(q, \mathcal{A}) < \varepsilon \text{dist}_i(q_1^i, s)$ , then let  $q_4^i$  be the point with largest  $\text{dist}_i(q_1^i, q_4^i) + \text{dist}(q_4^i, \mathcal{A})$  value. Otherwise,  $q_4^i$  is unspecified. Note that since we are not in the first case, either  $q_3^i$  or  $q_4^i$  is specified.

Then  $\tilde{d}_i$  is the tuple  $\tilde{d}^2(q_1^i, q_3^i, q_4^i)$ .

$\tilde{s}$  is chosen to be in  $\mathbb{C} \cap \Pi$  and to have the same rounded distance tuples as  $s$ , for *all* the rounded tuples  $\tilde{d}_i$ .  $\tilde{\mathcal{S}}$  is the union of all those  $\tilde{s}$  for  $s \in \mathcal{S}$ .

**Lemma 28.** *Let  $\mathcal{S}$  be a solution, and  $s \in \mathcal{S}$ . Let  $\tilde{s}$  defined as previously. For any point  $p \in X$ , either  $|\text{cost}(p, s) - \text{cost}(p, \tilde{s})| \leq \varepsilon(\text{cost}(p, s) + \text{cost}(p, \mathcal{A}))$  or both  $\text{dist}(p, s)$  and  $\text{dist}(p, \tilde{s})$  are bigger than  $\frac{10z \cdot \text{dist}(p, \mathcal{A})}{\varepsilon}$ .*

*Proof.* Fix  $s \in \mathcal{S} \cap \Pi$ , and let  $\tilde{s}$  be its corresponding point in  $\tilde{\mathcal{S}}$ . Let  $s_1 \in \{s, \tilde{s}\}$ , and  $s_2$  the other choice: we will show that  $\text{dist}(p, s_1) \leq (1 + \varepsilon)\text{dist}(p, s_2) + \varepsilon \text{dist}(p, \mathcal{A})$ . This implies that the costs verify the same inequality, which will allow us to conclude, switching the roles of  $s_1$  and  $s_2$ .

First, in the case where  $p \in \Pi \cap X$ , then the rounding directly ensures that either  $\text{dist}(p, s) > 1/\varepsilon \cdot \text{dist}(x, \mathcal{A})$ , in which case it holds as well than  $\text{dist}(p, \tilde{s}) > 1/\varepsilon \cdot \text{dist}(x, \mathcal{A})$ , or  $|\text{dist}(p, s) - \text{dist}(p, \tilde{s})| \leq \varepsilon \text{dist}(x, \mathcal{A})$ .

Otherwise,  $p$  is separated from  $s_1$  by some path among  $\{P_1, \dots, P_m\}$ . Let  $i$  be the smallest integer such that  $P_i$  intersects the shortest path between  $p$  and  $s_1$ . Our argument depends on the type of tuple  $\tilde{d}_i$  chosen for  $s$ . Since  $s$  and  $\tilde{s}$  have the same rounded distance tuples  $\tilde{d}_1, \tilde{d}_2, \dots$ , they have in particular the same rounded distance  $\tilde{d}_i$ . Let  $q_1^i$  be the point with smallest  $\text{dist}_i(q, s)$  value (importantly, the  $q_1^i, q_2^i, q_3^i$  and  $q_4^i$  appearing in the proof are defined with respect to  $s$ , not to  $s_1$ ).

**If we can estimate  $\text{dist}_i(q_1^i, s)$ .** In the first case, there is a  $q_2^i$  such that  $\varepsilon \text{dist}_i(q_1^i, s) \leq \text{dist}_i(q_1^i, q_2^i) + \text{dist}(q_2^i, \mathcal{A}) \leq \frac{1}{\varepsilon} \text{dist}_i(q_1^i, s)$ . We let  $D := \text{dist}_i(q_1^i, q_2^i) + \text{dist}(q_2^i, \mathcal{A})$  our (rough) estimate on the distance  $\text{dist}_i(q_1^i, s)$ .

Then, our argument goes as follows. Let  $x$  be a point in the intersection of  $P_i$  and the shortest path  $s_1 \rightsquigarrow p$ . We have the following properties: by choice of  $i$ ,  $\text{dist}_i(p, s_1) = \text{dist}(p, s_1)$  and

$\text{dist}_i(x, s_1) = \text{dist}(x, s_1)$ . By choice of  $x$ ,  $\text{dist}(p, x) + \text{dist}(x, s_1) = \text{dist}(p, s_1)$ . Last, by choice of  $q_1^i$ ,  $\text{dist}_i(q_1^i, s) \leq \text{dist}_i(p, s)$ , and  $D \leq \frac{\text{dist}_i(p, s)}{\varepsilon}$ .

- First, if  $\text{dist}_i(x, q_1^i) \leq \frac{D}{\varepsilon^2}$ . Then there is a point  $n$  from  $N_{i, q_1^i, q_2^i}$  with  $\text{dist}_i(n, x) \leq \varepsilon^2 D$ . Furthermore,  $\text{dist}_i(s_1, n) = \text{dist}_i(s_2, n) \pm \varepsilon^2 D$ , as  $\text{dist}_i(s, n) \leq \text{dist}_i(s, x) + \text{dist}_i(x, q_1^i) + \text{dist}_i(q_1^i, s) \leq \frac{3D}{\varepsilon^2}$  and so  $n$  has same rounded distances to  $s_1$  and  $s_2$ . Hence, we get:

$$\begin{aligned} \text{dist}_i(p, s_2) &\leq \text{dist}_i(p, x) + \text{dist}_i(x, n) + \text{dist}_i(n, s_2) \\ &\leq \text{dist}_i(p, x) + \text{dist}_i(x, n) + \text{dist}_i(n, s_1) + \varepsilon^2 D \\ &\leq \text{dist}_i(p, x) + \text{dist}_i(x, s_1) + 2\text{dist}_i(x, n) + \varepsilon^2 D \\ &\leq \text{dist}_i(p, s_1) + 3\varepsilon^2 D \\ &\leq \text{dist}(p, s_1) + 3\varepsilon \text{dist}_i(p, s). \end{aligned}$$

Now, two cases: either  $s = s_1$  and  $\text{dist}_i(p, s) = \text{dist}(p, s)$ , and then we get  $\text{dist}(p, s_2) \leq (1 + 3\varepsilon)\text{dist}(p, s_1)$ . Or  $s = s_2$ , and we have  $(1 - 3\varepsilon)\text{dist}(p, s) \leq \text{dist}(p, \tilde{s})$  which implies  $\text{dist}(p, s_2) \leq (1 + 6\varepsilon)\text{dist}(p, s_1)$ .

- Otherwise,  $\text{dist}_i(x, q_1^i) > \frac{D}{\varepsilon^2}$ : we first show that  $\text{dist}(s, \tilde{s}) \leq 3\varepsilon(\text{dist}_i(p, s) + \text{dist}(p, \mathcal{A}))$ , which will allow to conclude. It holds that  $\text{dist}_i(q_1^i, s) = \text{dist}_i(q_1^i, \tilde{s}) \pm \varepsilon D$ , as by definition of  $D$ ,  $\text{dist}_i(q_1^i, s) \leq D/\varepsilon$ . Hence,

$$\begin{aligned} \text{dist}_i(s, \tilde{s}) &\leq \text{dist}_i(s, q_1^i) + \text{dist}_i(\tilde{s}, q_1^i) \leq 2\text{dist}_i(s, q_1^i) + \varepsilon D \\ &\leq \frac{2 + \varepsilon^2}{\varepsilon} D \leq 3\varepsilon \text{dist}_i(x, q_1^i) \\ &\leq 3\varepsilon(\text{dist}_i(x, s_1) + \text{dist}_i(s_1, s_2) + \text{dist}_i(s, q_1^i)) \\ \Rightarrow \text{dist}(s, \tilde{s}) &\leq 9\varepsilon(\text{dist}(p, s_1) + \text{dist}_i(p, s)). \end{aligned}$$

Hence,

$$\begin{aligned} \text{dist}_i(p, s_2) &\leq \text{dist}_i(p, s_1) + \text{dist}_i(s, \tilde{s}) \\ &\leq \text{dist}(p, s_1) + 9\varepsilon(\text{dist}(p, s_1) + \text{dist}_i(p, s)) \end{aligned}$$

Similarly as in the previous case, either  $s_1 = s$  and the right hand side is  $(1 + 18\varepsilon)\text{dist}(p, s_1)$ , or  $s_2 = s$  and we infer  $\text{dist}(p, s_2) \leq (1 + 27\varepsilon)\text{dist}(p, s_1)$ .

**When we can only overestimate or underestimate  $\text{dist}_i(q_1^i, s)$**  In the second case,  $q_3^i$  is such that  $\text{dist}(q_3^i, \mathcal{A}) + \text{dist}_i(q_1^i, q_3^i) > \frac{1}{\varepsilon}\text{dist}_i(q_1^i, s)$ , and has minimal  $\text{dist}_i(q_1^i, q_3^i) + \text{dist}(q_3^i, \mathcal{A})$  value among those. Similarly,  $q_4^i$  is the point with largest  $\text{dist}_i(q_1^i, q_4^i) + \text{dist}(q_4^i, \mathcal{A})$  value among those verifying  $\text{dist}_i(q_1^i, q) + \text{dist}(q, \mathcal{A}) < \varepsilon \text{dist}_i(q_1^i, s)$ .

By choice of  $q_3^i$  and  $q_4^i$ , it must be that

$$\frac{1}{\varepsilon} \cdot (\text{dist}_i(q_1^i, q_4^i) + \text{dist}(q_4^i, \mathcal{A})) < \text{dist}_i(q_1^i, s) < \varepsilon \cdot (\text{dist}_i(q_1^i, q_3^i) + \text{dist}(q_3^i, \mathcal{A})).$$

Hence,  $\tilde{d}^2(q_1^i, q_3^i, q_4^i) = \top$ , and  $\tilde{s}$  is chosen such that

$$\frac{1}{\varepsilon} \cdot (\text{dist}_i(q_1^i, q_4^i) + \text{dist}(q_4^i, \mathcal{A})) < \text{dist}_i(q_1^i, \tilde{s}) < \varepsilon \cdot (\text{dist}_i(q_1^i, q_3^i) + \text{dist}(q_3^i, \mathcal{A})).$$

Since we are not in the first case where we can estimate  $\text{dist}_i(q_1^i, s)$ ,  $p$  verifies either  $\text{dist}(p, \mathcal{A}) + \text{dist}_i(p, q_1^i) < \varepsilon \text{dist}_i(q_1^i, s)$  or  $\text{dist}(p, \mathcal{A}) + \text{dist}_i(p, q_1^i) > \frac{1}{\varepsilon} \text{dist}_i(q_1^i, s)$ .

First, if  $\text{dist}_i(p, q_1^i) + \text{dist}(p, \mathcal{A}) > \frac{1}{\varepsilon} \text{dist}_i(q_1^i, s)$ . Then we have, by choice of  $q_3^i$ :

$$\begin{aligned} \text{dist}_i(s, \tilde{s}) &\leq \text{dist}_i(s, q_1^i) + \text{dist}_i(\tilde{s}, q_1^i) \\ &\leq 2\varepsilon(\text{dist}_i(q_1^i, q_3^i) + \text{dist}(q_3^i, \mathcal{A})) \\ &\leq 2\varepsilon(\text{dist}_i(q_1^i, p) + \text{dist}(p, \mathcal{A})) \\ &\leq 2\varepsilon(\text{dist}_i(p, s) + \text{dist}_i(q_1^i, s) + \text{dist}(p, \mathcal{A})) \\ &\leq 4\varepsilon(\text{dist}_i(p, s) + \text{dist}(p, \mathcal{A})) \end{aligned}$$

and therefore, we can conclude just as before (distinguishing whether  $s = s_1$  or  $s = s_2$ ) that

$$\text{dist}(p, s_2) \leq (1 + 12\varepsilon)\text{dist}(p, s_1) + 12\varepsilon\text{dist}(p, \mathcal{A}).$$

Lastly, in the case where  $\text{dist}_i(p, q_1^i) + \text{dist}(p, \mathcal{A}) < \varepsilon \text{dist}_i(q_1^i, s)$ , we use that  $\text{dist}_i(q_1^i, s_1) > \frac{1}{\varepsilon} \cdot (\text{dist}_i(q_1^i, q_4^i) + \text{dist}(q_4^i, \mathcal{A}))$  (as both  $s$  and  $\tilde{s}$  verifies this) to get:

$$\begin{aligned} \text{dist}(p, s_1) &= \text{dist}_i(p, s_1) \geq \text{dist}_i(q_1^i, s_1) - \text{dist}_i(p, q_1^i) \\ &\geq \frac{1}{\varepsilon} \cdot (\text{dist}_i(q_1^i, q_4^i) + \text{dist}(q_4^i, \mathcal{A})) - \text{dist}_i(p, q_1^i) \\ &\geq \frac{1}{\varepsilon} \cdot (\text{dist}_i(q_1^i, p) + \text{dist}(p, \mathcal{A})) - \text{dist}_i(p, q_1^i) \\ &\geq \frac{\text{dist}(p, \mathcal{A})}{\varepsilon}. \end{aligned}$$

**Conclusion.** Rescaling  $\varepsilon$  by  $1/27z$ , the previous inequalities gives us that either  $\text{dist}(p, s_1) \geq \frac{27z \cdot \text{dist}(p, \mathcal{A})}{\varepsilon}$ , or  $\text{dist}(p, s_2) \leq (1 + \varepsilon/z)\text{dist}(p, s) + \varepsilon/z \cdot \text{dist}(p, \mathcal{A})$ . The second inequality combined with Lemma 1 implies that  $\text{cost}(p, s_2) \leq (1 + \varepsilon)\text{cost}(p, s_1) + \varepsilon \cdot \text{cost}(p, \mathcal{A})$ . Therefore, using this result with  $s_1 = s, s_2 = \tilde{s}$  and then  $s_1 = \tilde{s}, s_2 = s$  shows that:

- either  $\text{dist}(p, s) \geq \frac{27z \cdot \text{dist}(p, \mathcal{A})}{\varepsilon}$ , or  $\text{cost}(p, \tilde{s}) \leq (1 + \varepsilon)\text{cost}(p, s) + \varepsilon \cdot \text{cost}(p, \mathcal{A})$
- either  $\text{dist}(p, \tilde{s}) \geq \frac{27z \cdot \text{dist}(p, \mathcal{A})}{\varepsilon}$ , or  $\text{cost}(p, s) \leq (1 + \varepsilon)\text{cost}(p, \tilde{s}) + \varepsilon \cdot \text{cost}(p, \mathcal{A})$ .

Therefore, if  $p$  is such that  $\text{dist}(p, s) \leq \frac{5z \cdot \text{dist}(p, \mathcal{A})}{\varepsilon}$ , then  $\text{dist}(p, \tilde{s}) \leq \frac{10z \cdot \text{dist}(p, \mathcal{A})}{\varepsilon}$ , and reciprocally when  $\text{dist}(p, \tilde{s}) \leq \frac{5z \cdot \text{dist}(p, \mathcal{A})}{\varepsilon}$ , then  $\text{dist}(p, s) \leq \frac{10z \cdot \text{dist}(p, \mathcal{A})}{\varepsilon}$ .

Thus we conclude: either both  $\text{dist}(p, \tilde{s})$  and  $\text{dist}(p, s)$  are bigger than  $\frac{10z \cdot \text{dist}(p, \mathcal{A})}{\varepsilon}$ , and we are done. Or both are smaller than  $\frac{20z \cdot \text{dist}(p, \mathcal{A})}{\varepsilon}$ , and then using the previous inequalities we get:

$$|\text{cost}(p, s) - \text{cost}(p, \tilde{s})| \leq 2\varepsilon(\text{cost}(p, s) + \text{cost}(p, \tilde{s}) + \text{cost}(p, \mathcal{A}))$$

Which, using  $\text{cost}(p, \tilde{s}) \leq \text{cost}(p, s) + |\text{cost}(p, s) - \text{cost}(p, \tilde{s})|$ , yields

$$|\text{cost}(p, s) - \text{cost}(p, \tilde{s})| \leq 7\varepsilon(\text{cost}(p, s) + \text{cost}(p, \mathcal{A})).$$

□

Lemma 28 gives exactly the same guarantee as Eq. (30): hence, as in the proof for treewidth, we can conclude from that inequality that for any solution  $\mathcal{S}$  and any interesting point  $p$ ,  $|\text{cost}(p, \mathcal{S}) - \text{cost}(p, \tilde{\mathcal{S}})| \leq \varepsilon(\text{cost}(p, \mathcal{S}) + \text{cost}(p, \mathcal{A}))$ .

Combining the guarantees from Lemma 28 and Lemma 27 concludes the proof of Lemma 25.

## 12 A Note on Euclidean Spaces

Lastly, we briefly want to survey the state of the art results for eliminating the dependency on the dimension in Euclidean spaces.

In a nutshell, the frameworks by both Feldman and Langberg [FL11] and us only yield coresets of size  $O(kd\text{poly}(\log k, \varepsilon^{-1}))$ . To eliminate the dependency on the dimension, we typically have to use some form of dimension reduction.

In a landmark paper, [FSS20] showed that one can replace the dependency on  $d$  with a dependency on  $k/\varepsilon^2$  for the  $k$ -means problem, see also [CEM<sup>+</sup>15] for further improvements on this idea. Subsequently, Sohler and Woodruff [SW18] gave a construction for arbitrary  $k$ -clustering objectives which lead to the first existence proof of dimension independent coresets for these problems. Unfortunately, there were a few caveats; most notably a running time exponential in both  $k$ . Huang and Vishnoi [HV20] showed that the mere existence of the Sohler-Woodruff construction was enough to compute coresets of size  $\text{poly}(k/\varepsilon)$ . Recently, the Sohler-Woodruff result was made constructive in the work of Feng, Kacham and Woodruff [FKW19].

Having obtained a  $\text{poly}(k/\varepsilon)$ -sized coreset, one can now use a terminal embedding to replace the dependency on  $d$  by a dependency  $\varepsilon^{-2} \log k/\varepsilon$ . Terminal embeddings are defined as follows:

**Definition 7** (Terminal Embeddings). *Let  $\varepsilon \in (0, 1)$  and let  $A \subset \mathbb{R}^d$  be arbitrary with  $|A|$  having size  $n > 1$ . Define the Euclidean norm of a  $d$ -dimensional vector  $\|x\| = \sqrt{\sum_{i=1}^d x_i^2}$ . Then a mapping  $f : \mathbb{R}^d \rightarrow \mathbb{R}^m$  is a terminal embedding if*

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

Terminal embeddings were studied by [EFN17, MMR18, NN19], with Narayanan and Nelson [NN19] achieving an optimal target dimension of  $O(\varepsilon^{-2} \log n)$ , where  $n$  is the number of points<sup>9</sup>.

It was first observed by Becchetti et al. [BBC<sup>+</sup>19] how terminal embeddings can be combined with the Feldman-Langberg [FL11] (or indeed our) framework. Specifically, given the existence of a  $\text{poly}(k/\varepsilon)$ -sized coreset, applying a terminal embedding with  $n$  being the number of distinct points in the coreset now allows us to further reduce the dimension. At the time, the only problem with such a coreset bound was  $k$ -means. The generalization to arbitrary  $k$ -clustering objectives is now immediate following the results by Huang and Vishnoi [HV20] and Feng et al. [FKW19].

It should be noted that more conventional Johnson-Lindenstrauss type embeddings proposed in [BBC<sup>+</sup>19, CEM<sup>+</sup>15, MMR19] do not (obviously) imply the same guarantee as terminal embeddings. We appended a short proof showing that terminal embeddings are sufficient at the end of this

---

<sup>9</sup>See the paper by Larsen and Nelson for a matching lower bound [LN17]

section. For a more in-depth discussion as to why normal Johnson-Lindenstrauss transforms may not be sufficient, we refer to Huang and Vishnoi [HV20].

Combining our  $O(k(d + \log k) \cdot \varepsilon^{-\max(2, z)})$  bound for general Euclidean spaces with either the Huang and Vishnoi [HJV19] or the Feng et al. [FKW19] constructions and terminal embeddings now immediately imply the following corollary.

**Corollary 8.** *There exists a coresset of size*

$$O\left(k \log k \cdot \left(\varepsilon^{-2-\max(2, z)}\right) \cdot 2^{O(z \log z)} \cdot \text{polylog}(\varepsilon^{-1})\right)$$

for  $(k, z)$ -clustering in Euclidean spaces.

Huang and Vishnoi further considered clustering in  $\ell_p$  metrics for  $p \in [1, 2)$ , i.e. non-Euclidean spaces. For this they reduced constructing a coresset for  $(k, z)$  clustering in an  $\ell_p$  space to constructing a coresset for  $(k, 2z)$  clustering in Euclidean space. Plugging in our framework into their reduction then yields the following corollary:

**Corollary 9.** *There exists a coresset of size*

$$O\left(k \log k \cdot (\varepsilon^{-2-2z}) \cdot 2^{O(z \log z)} \cdot \text{polylog}(\varepsilon^{-1})\right)$$

for  $(k, z)$ -clustering in any  $\ell_p$  space for  $p \in [1, 2)$ .

**Proposition 10.** *Suppose we have a (possibly weighted) point set  $A$  in  $\mathbb{R}^d$ . Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}^m$  with  $m \in O(\varepsilon^{-2} \cdot z^2 \log n)$  be a terminal embedding for  $A$  and let  $f(A)$  be the projected point set. Then if  $f(P) \subset f(A)$  is an  $\varepsilon$ -coreset for  $f(A)$ ,  $P \subset A$  is an  $O(\varepsilon)$ -coreset for  $A$ . Conversely, if  $P \subset A$  is an  $\varepsilon$ -coreset for  $A$ , then  $f(P) \subset f(A)$  is an  $O(\varepsilon)$ -coreset for  $f(A)$ .*

*Proof.* We prove the result for the first direction, the other direction is analogous. Consider an arbitrary solution  $S$  in  $\mathbb{R}^d$ . We first notice that for any point  $p \in A$ , we have

$$(1 - \varepsilon/2z)^z \cdot \text{cost}(f(p), f(S)) \leq (1 - \varepsilon) \cdot \text{cost}(f(p), f(S))$$

and

$$(1 + \varepsilon/2z)^z \cdot \text{cost}(f(p), f(S)) \geq (1 + \varepsilon) \cdot \text{cost}(f(p), f(S))$$

Therefore,

$$(1 - \varepsilon) \cdot \text{cost}(f(p), f(S)) \leq \text{cost}(p, S) \leq (1 + \varepsilon) \cdot \text{cost}(f(p), f(S)). \quad (32)$$

Now suppose  $f(P)$  is a coresset for  $f(A)$ , which means for any set of  $k$  points  $f(S) \subset \mathbb{R}^m$

$$\left| \sum_{p \in f(A)} w_p \cdot \text{cost}(p, f(S)) - \sum_{q \in f(P)} w'_q \cdot \text{cost}(q, f(S)) \right| \leq \varepsilon \cdot \sum_{p \in f(A)} w_p \cdot \text{cost}(p, f(S)), \quad (33)$$

where  $w$  and  $w'$  are the weights assigned to points in  $f(A)$  and  $f(P)$ , respectively. Let us now consider a solution  $S$  in the original  $d$ -dimensional space. Since  $P$  is a subset of  $A$ , we have by

combining Equations 32 and 33

$$\begin{aligned}
& \left| \sum_{p \in A} w_p \cdot \text{cost}(p, \mathcal{S}) - \sum_{q \in P} w'_q \cdot \text{cost}(p, \mathcal{S}) \right| \\
& \leq \varepsilon \cdot \sum_{p \in A} w_p \cdot \text{cost}(f(p), f(\mathcal{S})) + \varepsilon \cdot \sum_{q \in P} w'_q \cdot \text{cost}(f(q), f(\mathcal{S})) \\
& \quad + \left| \sum_{p \in A} w_p \cdot \text{cost}(f(p), f(\mathcal{S})) - \sum_{q \in P} w'_q \cdot \text{cost}(f(q), f(\mathcal{S})) \right| \\
& \leq 2\varepsilon \cdot \sum_{p \in A} w_p \cdot \text{cost}(f(p), f(\mathcal{S})) + \varepsilon \cdot \sum_{q \in P} w'_q \cdot \text{cost}(f(q), f(\mathcal{S})) \\
& \leq (3 + \varepsilon)\varepsilon \cdot \sum_{p \in A} w_p \cdot \text{cost}(f(p), f(\mathcal{S})) \\
& \leq (3 + 3\varepsilon)\varepsilon \cdot \sum_{p \in A} w_p \cdot \text{cost}(p, \mathcal{S}),
\end{aligned}$$

where the second inequality uses Equation 33 and the triangle inequality and the last inequality uses Equation 32.  $\square$

## References

- [AG06] Ittai Abraham and Cyril Gavoille. Object location using path separators. In Eric Ruppert and Dahlia Malkhi, editors, *Proceedings of the Twenty-Fifth Annual ACM Symposium on Principles of Distributed Computing, PODC 2006, Denver, CO, USA, July 23-26, 2006*, pages 188–197. ACM, 2006.
- [BBC<sup>+</sup>19] 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.
- [BBH<sup>+</sup>20] Daniel Baker, Vladimir Braverman, Lingxiao Huang, Shaofeng H. C. Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for clustering in graphs of bounded treewidth, 2020.
- [BEHW89] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989.
- [BEL13] Maria-Florina Balcan, Steven Ehrlich, and Yingyu Liang. Distributed  $k$ -means and  $k$ -median clustering on general communication topologies. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 1995–2003, 2013.
- [BFL<sup>+</sup>17] 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.
- [BFLR19] 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.

- [BJKW19] 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.
- [BJKW21] 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.
- [BLHK17] Olivier Bachem, Mario Lucic, S. Hamed Hassani, and Andreas Krause. Uniform deviation bounds for k-means clustering. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 283–291. PMLR, 2017.
- [BLL18] Olivier Bachem, Mario Lucic, and Silvio Lattanzi. One-shot coresets: The case of k-clustering. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pages 784–792, 2018.
- [CEM<sup>+</sup>15] 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.
- [Che09] 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.
- [CL19] Vincent Cohen-Addad and Jason Li. On the fixed-parameter tractability of capacitated clustering. In Christel Baier, Ioannis Chatzigiannakis, Paola Flocchini, and Stefano Leonardi, editors, *46th International Colloquium on Automata, Languages, and Programming, ICALP 2019, July 9-12, 2019, Patras, Greece*, volume 132 of *LIPIcs*, pages 41:1–41:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019.
- [CMK19] Mónika Csikós, Nabil H. Mustafa, and Andrey Kupavskii. Tight lower bounds on the vc-dimension of geometric set systems. *J. Mach. Learn. Res.*, 20:81:1–81:8, 2019.
- [CPP18] Matteo Ceccarello, Andrea Pietracaprina, and Geppino Pucci. Fast coreset-based diversity maximization under matroid constraints. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 81–89, 2018.
- [CS17] 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.
- [EA07] David Eisenstat and Dana Angluin. The VC dimension of k-fold union. *Inf. Process. Lett.*, 101(5):181–184, 2007.
- [EFN17] Michael Elkin, Arnold Filtser, and Ofer Neiman. Terminal embeddings. *Theor. Comput. Sci.*, 697:1–36, 2017.
- [EKM14] David Eisenstat, Philip N. Klein, and Claire Mathieu. Approximating k-center in planar graphs. In Chandra Chekuri, editor, *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014*, pages 617–627. SIAM, 2014.

- [FGS<sup>+</sup>13] 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.
- [FKW19] 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.
- [FL11] 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.
- [FMS07] Dan Feldman, Morteza Monemizadeh, and Christian Sohler. A PTAS for k-means clustering based on weak coresets. In *Proceedings of the 23rd ACM Symposium on Computational Geometry, Gyeongju, South Korea, June 6-8, 2007*, pages 11–18, 2007.
- [FS05] 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.
- [FSS20] 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.
- [GKL03] Anupam Gupta, Robert Krauthgamer, and James R. Lee. Bounded geometries, fractals, and low-distortion embeddings. In *44th Symposium on Foundations of Computer Science (FOCS 2003), 11-14 October 2003, Cambridge, MA, USA, Proceedings*, pages 534–543, 2003.
- [HCB16] Jonathan Huggins, Trevor Campbell, and Tamara Broderick. Coresets for scalable bayesian logistic regression. In *Advances in Neural Information Processing Systems*, pages 4080–4088, 2016.
- [HJLW18] 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.
- [HJV19] 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.
- [HK07] Sariel Har-Peled and Akash Kushal. Smaller coresets for k-median and k-means clustering. *Discrete & Computational Geometry*, 37(1):3–19, 2007.
- [HM01] Pierre Hansen and Nenad Mladenovic. J-means: a new local search heuristic for minimum sum of squares clustering. *Pattern Recognition*, 34(2):405–413, 2001.
- [HM04] 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.
- [HV20] 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, *Proceedings 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.



- [IMGR20] Piotr Indyk, Sepideh Mahabadi, Shayan Oveis Gharan, and Alireza Rezaei. Composable core-sets 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.
- [IMMM14] 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.
- [LL06] Yi Li and Philip M. Long. Learnability and the doubling dimension. In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 889–896, 2006.
- [LLS01] Yi Li, Philip M. Long, and Aravind Srinivasan. Improved bounds on the sample complexity of learning. *J. Comput. Syst. Sci.*, 62(3):516–527, 2001.
- [LN17] 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.
- [LS10] Michael Langberg and Leonard J. Schulman. Universal  $\varepsilon$ -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.
- [Mat00] Jirí Matousek. On approximate geometric  $k$ -clustering. *Discrete & Computational Geometry*, 24(1):61–84, 2000.
- [MJF19] 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.
- [MMK18] 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.
- [MMMR18] Sepideh Mahabadi, Konstantin Makarychev, Yury Makarychev, and Ilya P. Razenshteyn. Non-linear 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.
- [MMR19] 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.
- [MP04] Ramgopal R. Mettu and C. Greg Plaxton. Optimal time bounds for approximate clustering. *Mach. Learn.*, 56(1-3):35–60, 2004.
- [MS18] 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.

- [MSSW18] 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.
- [NN19] 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.
- [Pol12] David Pollard. *Convergence of stochastic processes*. Springer Science & Business Media, 2012.
- [SSS19] 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.
- [SW18] 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.
- [T<sup>+</sup>96] Michel Talagrand et al. Majorizing measures: the generic chaining. *The Annals of Probability*, 24(3):1049–1103, 1996.
- [Vit85] Jeffrey Scott Vitter. Random sampling with a reservoir. *ACM Trans. Math. Softw.*, 11(1):37–57, 1985.

## A Missing Proof

**Lemma 1** (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  $\varepsilon > 0$*

$$d(a, b)^z \leq (1 + \varepsilon)^{z-1} d(a, c)^z + \left( \frac{1 + \varepsilon}{\varepsilon} \right)^{z-1} d(b, c)^z$$

$$|d(a, S)^z - d(b, S)^z| \leq \varepsilon \cdot d(a, S)^z + \left( \frac{2z + \varepsilon}{\varepsilon} \right)^{z-1} d(a, b)^z.$$

*Proof.* The proof of the first inequality is appears in [MMR19], Corollary A.2.

For the second part, let  $S(a), S(b)$  be the closest point to  $a$  and  $b$  from  $S$ , and assume that  $d(b, S) \leq d(a, S)$ . Then:

$$\begin{aligned} d(a, S)^z &\leq d(a, S(b))^z \\ &\leq \left( 1 + \frac{\varepsilon}{2z} \right)^{z-1} \cdot d(b, S(b))^z + \left( 1 + \frac{2z}{\varepsilon} \right)^{z-1} \cdot d(a, b)^z \\ &\leq (1 + \varepsilon) \cdot d(b, S(b))^z + \left( 1 + \frac{2z}{\varepsilon} \right)^{z-1} \cdot d(a, b)^z \\ &\leq d(b, S)^z + \varepsilon \cdot d(a, S(a))^z + \left( 1 + \frac{2z}{\varepsilon} \right)^{z-1} \cdot d(a, b)^z, \end{aligned}$$

and so

$$|d(a, S)^z - d(b, S)^z| = d(a, S)^z - d(b, S)^z \leq \varepsilon \cdot d(a, S)^z + \left(\frac{2z + \varepsilon}{\varepsilon}\right)^{z-1} d(a, b)^z.$$

In the other case, when  $d(a, S) \leq d(b, S)$ :

$$\begin{aligned} d(b, S)^z &\leq d(b, S(a))^z \\ &\leq \left(1 + \frac{\varepsilon}{2z}\right)^{z-1} \cdot d(a, S(a))^z + \left(1 + \frac{2z}{\varepsilon}\right)^{z-1} \cdot d(a, b)^z \\ &\leq (1 + \varepsilon) \cdot d(a, S)^z + \left(1 + \frac{2z}{\varepsilon}\right)^{z-1} \cdot d(a, b)^z, \end{aligned}$$

and so

$$|d(a, S)^z - d(b, S)^z| = d(b, S)^z - d(a, S)^z \leq \varepsilon \cdot d(a, S)^z + \left(\frac{2z + \varepsilon}{\varepsilon}\right)^{z-1} d(a, b)^z.$$

□

## B A Coreset of Size $k^2 \varepsilon^{-2}$

In this section, we show how to trade a factor  $\varepsilon^{-z}$  for a factor  $k$  in the coreset size.

**Lemma 29.** *Let  $(X, \text{dist})$  be a metric space,  $P$  be a set of points,  $k, z$  two positive integers and  $\mathcal{A}$  a set of  $O(k)$  centers such that each for each cluster with center  $c$  induced by  $\mathcal{A}$ , all points of the cluster are at distance between  $(\frac{\varepsilon}{z})^2 \Delta_C$  and  $(\frac{z}{\varepsilon})^2 \Delta_C$ , for some  $\Delta_C$ .*

*Suppose there exists an  $\mathcal{A}$ -approximate centroid set  $\mathbb{C}$  for  $P$ .*

*Then, there exists an algorithm running in time  $O(|P|)$  that constructs a set  $\Omega$  of size  $O(k \cdot 2^{O(z)} \frac{\log^3(1/\varepsilon)}{\varepsilon^2} (k \log k + k \log |\mathbb{C}| + \log(1/\pi)))$  such that, with probability  $1 - 1/\pi$ , for any set  $\mathcal{S}$  of  $k$  centers,*

$$|\text{cost}(\mathcal{S}) - \text{cost}(\Omega, \mathcal{S})| = O(\varepsilon) \text{cost}(\mathcal{S}).$$

Suppose we initially computed a set of  $k'$  centers  $\mathcal{A}$ . Our aim is to define a sampling distribution that approximates the cost of any solution  $\mathcal{S}$  with high probability. While the basic idea is related to importance sampling (i.e. sampling proportionate to  $\text{cost}(p, \mathcal{A})$ ), we add a few modifications that are crucial.

Compared to the framework described in the main body, we change slightly the definition of ring. For every cluster  $\mathcal{C}_i$  of  $\mathcal{A}$ , we partition the points of  $\mathcal{C}_i$  into rings  $R_{i,j}$  from between distances  $[(\frac{\varepsilon}{z})^2 \Delta_C \cdot 2^j, (\frac{\varepsilon}{z})^2 \Delta_C \cdot 2^{j+1}]$ , for  $j \in \{1, \dots, 4z \log(z/\varepsilon)\}$ .

The algorithm is as follows: from every  $R_{i,j}$ , sample  $\delta$  points uniformly at random (if  $|R_{i,j}| \leq \delta$ , simply add the whole  $R_{i,j}$ ).

The analysis of this algorithm follows the same line as the main one. Rings are divided into tiny, interesting and huge types; tiny and huge are dealt with as in Lemmas 5 and 7, and interesting points slightly differently.

From the definition of  $R_{i,j}$ , we immediately get the following observation.

**Fact 6.** *For every cluster we have at most  $O(z \cdot \log z / \varepsilon)$  non-empty rings in total.*

Given a solution  $\mathcal{S}$ , we consider the groups  $I_{i,j,\ell} \subset \mathcal{C}_i$  consisting of the points of  $R_{i,j}$  served in  $\mathcal{S}$  by a center at distance  $[\varepsilon \cdot 2^\ell, \varepsilon \cdot 2^{\ell+1}]$ . As before, we let  $\text{cost}(I_{i,j,\ell}, \mathcal{S}) = \sum_{p \in I_{i,j,\ell}} \text{cost}(p, \mathcal{S})$  and  $\text{cost}(I_{j,\ell}, \mathcal{S}) = \sum_{i=1}^{k'} \text{cost}(I_{i,j,\ell}, \mathcal{S})$ .

Our analysis will distinguish between three cases:

1.  $\ell \leq j + \log \varepsilon$ , in which case we say that  $I_{i,j,\ell}$  is *tiny*.
2.  $j \cdot \log \varepsilon \leq \ell \leq j + \log(4z/\varepsilon)$ , in which case we say  $I_{i,j,\ell}$  is *interesting*.
3.  $\ell \geq j + \log(16z/\varepsilon)$ , in which case we say  $I_{i,j,\ell}$  is *huge*.

We first consider the huge case. For this, we show that the weight of every ring is preserved with high probability, which implies that the huge groups are well approximated.

**Lemma 30.** *It holds that, for any  $R_{i,j}$  and for all solutions  $\mathcal{S}$  with at least one non-empty huge group  $I_{i,j,\ell}$*

$$\left| \text{cost}(R_{i,j}, \mathcal{S}) - \sum_{p \in \Omega \cap R_{i,j}} \frac{|R_{i,j}|}{\delta} \cdot \text{cost}(p, \mathcal{S}) \right| \leq 3\varepsilon \cdot \text{cost}(R_{i,j}, \mathcal{S}).$$

*Proof.* Fix a ring  $R_{i,j}$  and let  $I_{i,j,\ell}$  be a huge group. First, the weight of  $R_{i,j}$  is preserved in  $\Omega$ : since  $\delta$  points are sampled from  $R_{i,j}$ , it holds that

$$\sum_{p \in \Omega \cap R_{i,j}} \frac{|R_{i,j}|}{\delta} = |R_{i,j}|$$

Now, let  $\mathcal{S}$  be a solution, and  $p \in I_{i,j,\ell}$  with  $I_{i,j,\ell}$  being huge. This implies, for any  $q \in R_{i,j}$ :  $\text{cost}(p, q) \leq (2 \cdot \varepsilon \cdot 2^{j+1})^z \leq 4^z \cdot \varepsilon^z \cdot 2^{(\ell - \log(16z/\varepsilon))z} \leq \left(\frac{\varepsilon}{4z}\right)^z \cdot \text{cost}(p, \mathcal{S})$ . By Lemma 1, we have therefore for any point  $q \in R_{i,j}$

$$\begin{aligned} \text{cost}(p, \mathcal{S}) &\leq (1 + \varepsilon/2z)^{z-1} \text{cost}(q, \mathcal{S}) + (1 + 2z/\varepsilon)^{z-1} \text{cost}(p, q) \\ &\leq (1 + \varepsilon) \text{cost}(q, \mathcal{S}) + \varepsilon \cdot \text{cost}(p, \mathcal{S}) \\ \Rightarrow \text{cost}(q, \mathcal{S}) &\geq \frac{1 - \varepsilon}{1 + \varepsilon} \text{cost}(p, \mathcal{S}) \geq (1 - 2\varepsilon) \text{cost}(p, \mathcal{S}) \end{aligned}$$

Moreover, by a similar calculation, we can also derive an upper bound of  $\text{cost}(q, \mathcal{S}) \leq \text{cost}(p, \mathcal{S}) \cdot (1 + 2\varepsilon)$ . Hence, combined with  $\sum_{p \in \Omega \cap R_{i,j}} \frac{|R_{i,j}|}{\delta} = |R_{i,j}|$ , this is sufficient to approximate  $\text{cost}(R_{i,j}, \mathcal{S})$ .

Therefore, the cost of  $R_{i,j}$  is well approximated for any solution  $\mathcal{S}$  such that there is a non-empty huge group  $I_{i,j,\ell}$ .  $\square$

Next, we consider the interesting cases. The main observation here is that there are only  $O(\log 1/\varepsilon)$  many rings per cluster, hence a coarser estimation using Bernstein's inequality is actually sufficient to bound the cost.

**Lemma 31.** *Consider an  $R_{i,j}$  and any solution  $\mathcal{S}$  such that all huge  $I_{i,j,\ell}$  are empty. It holds with probability at least  $1 - \log(z/\varepsilon) \exp(-\frac{\varepsilon^2}{2 \cdot 16^z \log^2 z/\varepsilon} \cdot \delta)$  that, for all interesting  $I_{i,j,\ell}$ :*

$$\left| \text{cost}(I_{i,j,\ell}, \mathcal{S}) - \sum_{p \in I_{i,j,\ell} \cap \Omega} \text{cost}(p, \mathcal{S}) \cdot \frac{|R_{i,j}|}{\delta} \right| \leq \frac{\varepsilon}{\log(z/\varepsilon)} \cdot (\text{cost}(R_{i,j}, \mathcal{A}) + \text{cost}(R_{i,j}, \mathcal{S})).$$

*Proof.* We start by bounding  $|R_{i,j}| \cdot (\varepsilon \cdot 2^\ell)^z$  in terms of  $\text{cost}(R_{i,j}, \mathcal{S}) + \text{cost}(R_{i,j}, \mathcal{A})$ .

If  $I_{i,j,\ell}$  for some  $\ell \geq j+3$  is non-empty, then  $\varepsilon \cdot 2^\ell - \varepsilon \cdot 2^{j+2} \leq d(q, \mathcal{S})$ , for any point  $q$ . Hence,  $|R_{i,j}| \cdot (\varepsilon \cdot 2^\ell)^z \leq \text{cost}(R_{i,j}, \mathcal{S}) \cdot 2^z$ . If  $\ell \leq j+2$ , then  $|R_{i,j}| \cdot (\varepsilon \cdot 2^\ell)^z \leq |R_{i,j}| \cdot (\varepsilon \cdot 2^{j+2})^z \leq \text{cost}(R_{i,j}, \mathcal{A}) \cdot 4^z$ . Putting both bounds together, we have

$$|R_{i,j}| \cdot (\varepsilon \cdot 2^\ell)^z \leq 4^z (\text{cost}(R_{i,j}, \mathcal{S}) + \text{cost}(R_{i,j}, \mathcal{A})) \quad (34)$$

Since we aim to apply Bernstein's inequality, we now require a bound on the second moment of our cost estimator. We have for a single randomly chosen point  $P$ :

$$\mathbb{E} \left[ \sum_{p \in I_{i,j,\ell} \cap P} \text{cost}(p, \mathcal{S}) \cdot |R_{i,j}| \right] = \text{cost}(I_{i,j,\ell}, \mathcal{S})$$

and

$$\begin{aligned} \mathbb{E} \left[ \left( \sum_{p \in I_{i,j,\ell} \cap P} \text{cost}(p, \mathcal{S}) \cdot |R_{i,j}| \right)^2 \right] &= \mathbb{E} \left[ \sum_{p \in I_{i,j,\ell} \cap P} \text{cost}(p, \mathcal{S})^2 \cdot |R_{i,j}|^2 \right] \text{ since } |P| = 1 \\ &= \sum_{p \in I_{i,j,\ell} \cap P} \text{cost}(p, \mathcal{S})^2 \cdot |R_{i,j}| \leq |R_{i,j}| \cdot |I_{i,j,\ell}| \cdot (\varepsilon \cdot 2^\ell)^{2z} 4^z \\ &\leq \text{cost}(I_{i,j,\ell}, \mathcal{S}) \cdot (\text{cost}(R_{i,j}, \mathcal{S}) + \text{cost}(R_{i,j}, \mathcal{A})) \cdot 16^z \quad (35) \end{aligned}$$

where the final equation follows from by lower bounding the cost in  $\mathcal{S}$  of any point in  $I_{i,j,\ell}$  with  $(\varepsilon \cdot 2^\ell)^z$  and using Equation 34.

Furthermore, by the same reasoning and again using Equation 34, we have the upper bound  $M$  on the (weighted) cost in  $\mathcal{S}$  of every sampled point in every ring:

$$M \leq (\varepsilon \cdot 2^{\ell+1})^z \cdot |R_{i,j}| \leq (\text{cost}(R_{i,j}, \mathcal{S}) + \text{cost}(R_{i,j}, \mathcal{A})) \cdot 8^z \quad (36)$$

Applying Bernstein's inequality and Equations 35 and 36, we now have

$$\begin{aligned}
& \mathbb{P} \left[ \left| \delta \cdot \text{cost}(I_{i,j,\ell}, \mathcal{S}) - \sum_{p \in I_{i,j,\ell} \cap \Omega} \text{cost}(p, \mathcal{S}) \cdot |R_{i,j}| \right| > \frac{\varepsilon \cdot \delta}{r} \cdot (\text{cost}(R_{i,j}, \mathcal{S}) + \text{cost}(R_{i,j}, \mathcal{A})) \right] \\
& \leq \exp \left( - \frac{\frac{\varepsilon^2 \cdot \delta}{r^2} \cdot (\text{cost}(R_{i,j}, \mathcal{S}) + \text{cost}(R_{i,j}, \mathcal{A}))}{\text{cost}(I_{i,j,\ell}, \mathcal{S}) \cdot 16^z + \frac{4\varepsilon}{3r} \cdot (\text{cost}(R_{i,j}, \mathcal{S}) + \text{cost}(R_{i,j,\ell})) \cdot 8^z} \right) \leq \exp \left( - \frac{\varepsilon^2 \cdot \delta}{2r^2 16^z} \right),
\end{aligned}$$

where the last line uses  $\text{cost}(I_{i,j,\ell}, \mathcal{S}) \leq \text{cost}(R_{i,j}, \mathcal{S})$ . Applying a union bound over all  $r$  interesting sets  $I_{i,j,\ell}$ , we obtain the above guarantee for all  $I_{i,j,\ell}$  simultaneously with probability

$$1 - r \cdot \exp \left( - \frac{\varepsilon^2 \cdot \delta}{2r^2 16^z} \right).$$

□

Finally, we conclude:

*Proof of Lemma 29.* As in the proof of Lemma 2, we decompose  $|\text{cost}(\mathcal{S}) - \text{cost}(\Omega, \mathcal{S})|$  into terms corresponding to points of tiny, interesting or huge groups. We only sketch the proof here, the details are the same as for Lemma 2. We condition on event  $\mathcal{E}$  happening. Let  $\mathcal{S}$  be a set of  $k$  points, and  $\tilde{\mathcal{S}} \in \mathbb{C}^k$  that approximates best  $\mathcal{S}$ , as given by the definition of  $\mathbb{C}$  (see Definition 1). This ensures that for all points  $p$  with  $\text{dist}(p, \mathcal{S}) \leq \frac{8z}{\varepsilon} \cdot \text{dist}(p, \mathcal{A})$  or  $\text{dist}(p, \tilde{\mathcal{S}}) \leq \frac{8z}{\varepsilon} \cdot \text{dist}(p, \mathcal{A})$ , we have  $|\text{cost}(p, \mathcal{S}) - \text{cost}(p, \tilde{\mathcal{S}})| \leq \varepsilon(\text{cost}(p, \mathcal{S}) + \text{cost}(p, \mathcal{A}))$ .

Our first step is to deal with points that have  $\text{dist}(p, \mathcal{S}) > \frac{8z}{\varepsilon} \cdot \text{dist}(p, \mathcal{A})$ , using Lemma 30. All other points have distance well approximated by  $\tilde{\mathcal{S}}$ . Then, we can apply Lemma 5 and Lemma 31 to  $L_{\tilde{\mathcal{S}}}$ , since all points in  $L_{\tilde{\mathcal{S}}}$  have  $\text{dist}(p, \tilde{\mathcal{S}}) \leq \frac{4z}{\varepsilon} \cdot \text{dist}(p, \mathcal{A})$ , and so  $\text{dist}(p, \mathcal{S}) \leq \frac{8z}{\varepsilon} \cdot \text{dist}(p, \mathcal{A})$  and were not removed by the previous step. Remaining points are those which have  $\text{dist}(p, \tilde{\mathcal{S}}) > \frac{4z}{\varepsilon} \cdot \text{dist}(p, \mathcal{A})$  and  $\text{dist}(p, \mathcal{S}) \leq \frac{8z}{\varepsilon} \cdot \text{dist}(p, \mathcal{A})$ , i.e., their distance is preserved in  $\tilde{\mathcal{S}}$  and they are huge with respect to  $\tilde{\mathcal{S}}$ . We apply Lemma 7 to them as well. □

Combining this lemma and Lemma 4 gives an analogous to Theorem 1. Now, using this lemma instead of Theorem 1 in all proofs of section Section 8 gives bound with a factor  $k$  instead of a  $\varepsilon^{-z}$ .