# 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 of 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

- $O\left(\Gamma \cdot k(d+\log k)\right)$ 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 $\Omega(\frac{k\log n}{\varepsilon})$ 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 $(X, \mathrm{dist})$, the $(k, z)$-clustering problem asks for a set $\mathcal{S}$ of $k$ points, called *centers*, that minimizes

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

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. $\mathrm{cost}(\Omega, \mathcal{S}) \in (1 \pm \varepsilon)\mathrm{cost}(X, \mathcal{S})$. 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[1]. 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+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

[1]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 1.** *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 $(X, dist)$ 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 $cost(p, \mathcal{S}) \le \left(\frac{8z}{\varepsilon}\right)^z cost(p, \mathcal{A})$ or $cost(p, \tilde{\mathcal{S}}) \le \left(\frac{8z}{\varepsilon}\right)^z cost(p, \mathcal{A})$, it holds*

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

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 $(X, dist)$ 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*

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

*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][2].

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 A.

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

- $O\left(\Gamma \cdot k\left(d + \log k\right)\right)$ for metric spaces with doubling dimension $d$. This improves over the $O(k^3 d\varepsilon^{-2})$ from [HJLW18]. See Corollary 4.
- Since general discrete metric spaces have doubling dimension $O(\log n)$, this yields coreset of size $O\left(\Gamma \cdot k \log n\right)$. This improves on the bound from Feldman and Langberg [FL11] $O\left(\varepsilon^{-2z} k \log k \log n\right)$, and has an optimal dependency in $k$ and $n$.
- $O\left(\Gamma \cdot k\varepsilon^{-2} \cdot \log k\right)$ for Euclidean Spaces, see Corollary 7. This improves on the recent result from [HV20], who achieve $O\left(\varepsilon^{-2z} k \log^2 k\right)$. The time complexity for Euclidean spaces is $n \cdot k^{O(z^2 \log^2(1/\varepsilon)/\varepsilon^2)}$.
- $O\left(\Gamma \cdot k\left(\log^2 k + \frac{\log k}{\varepsilon^4}\right)\right)$ for a family of graphs excluding a fixed minor, see Corollary 6. This improves on [BJKW20], whose coreset has size $\widetilde{O}(k^2/\varepsilon^4)$.
- $O\left(\Gamma \cdot k\left(t + \log k\right)\right)$ in graphs with treewidth $t$, see Corollary 5. This improves upon the work of [BBH$^+$20] in two ways: their coreset is only for $k$-Median and has size $\widetilde{O}(k^2 t/\varepsilon^2)$.
- $O(k\varepsilon^{-2z} \cdot \min(d, \varepsilon^{-2}\log k))$ in $\mathbb{R}^d$ with $\ell_p$ distance, for $p \in [1, 2)$, see Corollary 8. This improves on [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$^+$20]. This means that in the case of metrics with doubling dimension $d$, our bounds are optimal up to a $\operatorname{poly}\log(1/\varepsilon)/\varepsilon$ factor. For graphs with treewidth $t$, another lower bound of $\Omega\left(\frac{kt}{\varepsilon}\right)$ from [BBH$^+$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.

**Reducing to a structured instance.** Like most coreset constructions, we initially compute a constant factor approximation $\mathcal{A}$ to the problem. We then deviate from previous importance sampling algorithms in two ways: we first augment the solution $\mathcal{A}$, by adding greedily centers as long as the cost can be decreased by a factor $(1 - \varepsilon/k)$ or the cost drops below $\varepsilon$OPT. The, the

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

| Reference | Size (Number of Points) |
|---|---|
| **Euclidean space** | |
| 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+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] | $\tilde{O}(k^2 \cdot \varepsilon^{-4})$ |
| **This paper** | $O(k \cdot \log k \cdot \varepsilon^{-2-\max(2,z)})$ |
| **General $n$-point metrics, $ddim$ denotes the doubling dimension** | |
| 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] | $O(k^3 \cdot ddim \cdot \varepsilon^{-2})$ |
| **This paper** | $O(k \cdot (ddim + \log k) \cdot \varepsilon^{-\max(2,z)})$ |
| **This paper** | $O(k \cdot \log n \cdot \varepsilon^{-\max(2,z)})$ |
| **Graph with $n$ vertices, $t$ denotes the treewidth** | |
| Baker, Braverman, Huang, Jiang, Krauthgamer, Wu (ICML'20) [BBH+20] | $\tilde{O}(k^2 \cdot t/\varepsilon^2)$ |
| **This paper** | $O(k \cdot (t + \log k) \cdot \varepsilon^{-\max(2,z)})$ |
| **Graph with $n$ vertices, excluding a fixed minor** | |
| Bravermann Jian, Krauthgamer, Wu (SODA'21) [BJKW21] | $\tilde{O}(k^2 \cdot \varepsilon^{-4})$ |
| **This paper** | $O\left(k \cdot (\log^2 k + \frac{\log k}{\varepsilon^4}) \cdot \varepsilon^{-\max(2,z)}\right)$ |

Table 1: Comparison of coreset sizes for $(k,z)$-Clustering in various metrics. Dependencies on $2^{O(z)}$ and 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+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+20], it seems that they can be generalized to any power. The main difference is in the computation of a constant factor approximation.

points are partitioned 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

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

and then applying a union bound over the validity of Eq. (1) for all solutions $S$. 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[3]. 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.

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.

[3] 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 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 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 $\max_p \left(\varepsilon^{-1} \cdot \text{dist}(p, \mathcal{A})\right)^z \cdot \frac{\text{cost}(\mathcal{A})}{\text{cost}(p,\mathcal{A})} \in O(\varepsilon^{-z})$. 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

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

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

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+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 $\mathcal{A}$, $|\text{cost}(p, \mathcal{S}) - \text{cost}(p, \tilde{\mathcal{S}})| \leq \varepsilon\left(\text{cost}(p, \mathcal{A}) + \text{cost}(p, \mathcal{S})\right)$. We call such points *interesting*.

**Metrics with doubling dimension $d$:** $\mathbb{C}$ is simply constructed taking *nets* around each input points. 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. For every point $p$, $\mathbb{C}$ contains an $\varepsilon\text{cost}(p, \mathcal{A})$-net of the points at distance at most $2\text{cost}(p, \mathcal{A})$ from $p$. If $p$ is an interesting point, there is therefore a center of $\mathbb{C}$ close to its center in $\mathcal{S}$.

**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 [BBH+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 $B$ than $c$.

$\mathbb{C}$ is therefore constructed as follows: a distance tuple to $B = \{b_1, ..., b_{|B|}\}$ is a tuple $(d_1, ..., d_{|B|})$, where $d_i$ represents 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 $\text{dist}(p, c) \leq \varepsilon^{-1}\text{dist}(p, \mathcal{A})$, rouding 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 8.2 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 separators. A construction similar in spirit to the one for treewidth is therefore possible, as presented in Section 8.3. This builds on the work of [BJKW20].

### 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 9, and the $O(k^2\varepsilon^{-2})$ construction in Appendix A. 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[4]. 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 9. 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 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

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

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$^+$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$^+$17, FS05, FGS$^+$13]. Other related work considers generalizations of $k$-median and $k$-means by either adding capacity constraints [CL19, HJV19, SSS19], or considering 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 $(X, \mathrm{dist})$, 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

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

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

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

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

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$^+$19, CS17, FSS20, MMR19, SW18] (see for instance Corollary A.2 in [MMR19]).

**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 $\varepsilon > 0$*

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

### 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 $(X, dist)$ be a metric space, $P$ a set of clients with weights $w : P \to \mathbb{R}^+$ 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*

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

*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.

We now show how to apply Theorem 1 to the input $\tilde{P}$. 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 groups that are very well structured. 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. ($\mathcal{A}$ will be chosen later to be the outcome of a greedy seeding on $P$). Let $C_1, ..., C_k$ 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* $R_I(C_i) := \cup_{j \leq z \log(\varepsilon/z)} R_{i,j}$ (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 $(\varepsilon/z)^z \Delta_{C_i}$ (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 outer rings is also partitioned into *outer groups*:

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

- We let as well $G_{min}^O = \cup_{b \leq 0} G_b^O$ and $G_{max}^O = \cup_{b \geq z \log \frac{4z}{\varepsilon}} G_b^O$.

# 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, and a set of centers $\mathcal{A}$ inducing clusters $C_1, C_2, ...$ on $G$. The output of `GroupSample` is a coreset for the group, computed as follows: a point $p \in C_i$ is sampled with probability $\frac{\delta \text{cost}(C_i,\mathcal{A})}{|C_i| \cdot \text{cost}(G,\mathcal{A})}$, and the weight of any sampled point is rescaled by a factor $\frac{|C_i| \cdot \text{cost}(G,\mathcal{A})}{\delta \text{cost}(C_i,\mathcal{A})}$.

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

**Lemma 2.** *Let $(X, dist)$ 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 $C$, all points of $G \cap C$ have the same cost in $\mathcal{A}$, up to a factor 2: $\forall p, q \in G \cap C,\ \ cost(p, \mathcal{A}) \le 2 cost(q, \mathcal{A})$.*
- *for all clusters $C$, it holds that $\frac{cost(G,\mathcal{A})}{2k} \le cost(C \cap G, \mathcal{A})$.*

*Let $\mathbb{C}$ be a $\mathcal{A}$-approximate centroid set for clients $G$.*

*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:*

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

We further require the `SensitivitySample` procedure, we which 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 $(X, dist)$ 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 clients $G$*

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

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

The final algorithm is as follows:

**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 $P' \subseteq X$ and a weight function $w : P' \mapsto \mathbb{R}_+$ such that for any set of $k$ centers $C$, $\text{cost}(P, C) = (1 \pm \varepsilon)\text{cost}(P', C)$.

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 $R_O(C) \cap G^O_{min}$, 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 $z\log(\varepsilon/z) \le j \le 2z\log(z/\varepsilon)$ and every group $G_{j,b} \notin G_{j,min}$, compute a coreset $\Omega_{j,b}$ of size

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

   using the `GroupSample` procedure.

4. **Sampling from the outer rings:** From each group $G^O_1, ..., G^O_{max}$, compute a coreset $\Omega^O_b$ of size

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

   using the `SensitivitySample` procedure.

5. **Output:**

   - A coreset consisting of $\mathcal{A} \cup \Omega_{j,b} \cup \Omega^O_i$.
   - 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.

### 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:

**Lemma 4.** *Let $(X, dist)$ 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*

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

Moreover, the partitioning ensures the following two facts:

**Fact 1.** *There exist at most $O(z\log(z/\varepsilon))$ many non-empty $R_j$ that are not in $R^{\mathcal{A}}_I$ nor in $R^{\mathcal{A}}_O$.*

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 non-empty $G_{j,b}$.*

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

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

Combining those fact, 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. 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 $b \neq max$, the cost of all clusters are equal up to a factor 2, hence for all $i$ $\frac{\text{cost}(G_{j,b},\mathcal{A})}{2k} \leq \text{cost}(C_i \cap G_{j,b}, \mathcal{A})$
- for $b = max$, it holds that $\frac{\text{cost}(G_{j,b},\mathcal{A})}{2k'} \leq \frac{\text{cost}(R_j,\mathcal{A})}{2k} \leq \text{cost}(C_i \cap G_{j,b}, \mathcal{A})$.

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 $\mathcal{S}$

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

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

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

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 $G^O_i$ implies that, with probability $1 - z^2 \log^2(z/\varepsilon) \exp\left(k \log |\mathbb{C}| - 2^{O(z \log z)} \cdot \frac{\min(\varepsilon^2,\varepsilon^z)}{\log^2 1/\varepsilon} \cdot \delta\right) - z \log(z/\varepsilon) \exp\left(k \log |\mathbb{C}| - 2^{O(z \log z)} \frac{\varepsilon^2}{\log^2 1/\varepsilon} \cdot \delta\right)$, for any solution $\mathcal{S}$,

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

where the last inequality uses that $\mathcal{A}$ is a constant-factor approximation.

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

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

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 $(X, dist)$ 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 $C$, all points of $G \cap C$ have the same cost in $\mathcal{A}$, up to a factor 2: $\forall p, q \in G \cap C$, $cost(p, \mathcal{A}) \leq 2 cost(q, \mathcal{A})$.*
- *for all clusters $C$, it holds that $\frac{cost(G,\mathcal{A})}{2k} \leq cost(C \cap G, \mathcal{A})$.*

*Let $\mathbb{C}$ be a $\mathcal{A}$-approximate centroid set for clients $G$.*

*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:*

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

The `GroupSample` merely consists of importance sampling in rounds, i.e. there are $\delta$ rounds in which one point of $G$ is sampled. The probability of sampling point $p \in C_i$ is sampled with probability $\frac{\text{cost}(C_i,\mathcal{A})}{|C_i| \cdot \text{cost}(G,\mathcal{A})}$. The weight of any sampled point is rescaled by a factor $\frac{|C_i| \cdot \text{cost}(G,\mathcal{A})}{\delta \text{cost}(C_i,\mathcal{A})}$. If there are $m$ copies of a point, it is sampled in a round with probability $\frac{m \cdot \text{cost}(C_i,\mathcal{A})}{|C_i| \cdot \text{cost}(G,\mathcal{A})}$ (which is equivalent to sampling each copy with probability $\frac{\text{cost}(C_i,\mathcal{A})}{|C_i| \cdot \text{cost}(G,\mathcal{A})}$). In what follows, each copies will be considered independently.

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

To analyse this sampling procedure, 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 $I_{tiny,\mathcal{S}}$.
- $\log \varepsilon/2 \leq \ell \leq z \log(4z/\varepsilon)$. 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*.

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 $X_1, \ldots X_\delta$ be non-negative independent random variables. Let $S = \sum_{i=1}^{\delta} X_i$. If there exists an almost-sure upper bound $M \geq X_i$, then*

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

In this paper we will simply drop the $E[X_i]^2$ 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 $\{C_1, ..., C_k\}$ be the clusters of $\mathcal{A}$ restricted to $G$. The assumptions imply the following fact:

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

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

## 5.1 Dealing with Tiny Type

**Lemma 5.** *It holds that*

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

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

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

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

## 5.2 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, $|C_i|$ is well approximated for every cluster $C_i$. This lemma will also be used later for the interesting points. We define event $\mathcal{E}$ to be: For all cluster $C_i$,

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

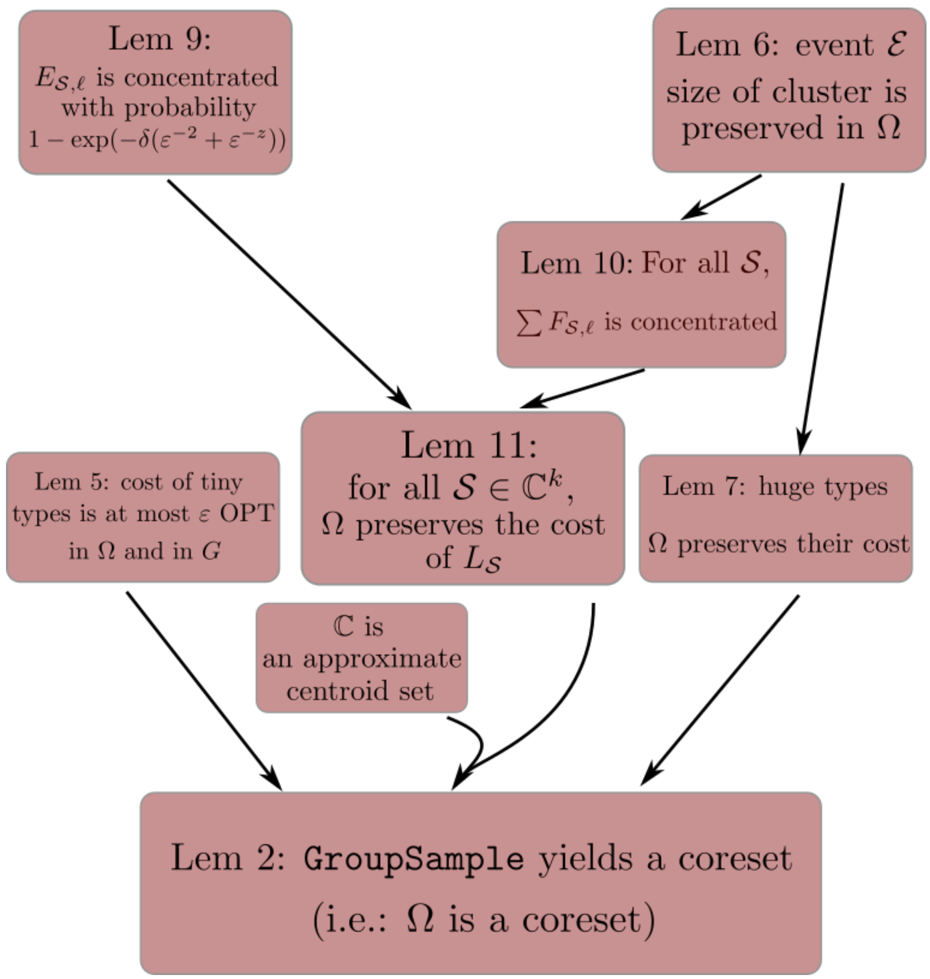


Figure 1: Arrangement of Lemmas of Section 5 to prove Lemma 2.

**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 $C_i \cap G \neq \emptyset$. The probability that a point sampled via importance sampling is from $C_i$ is then at least

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

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

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

Now, rescaling $P_i(p)$ by a factor $\frac{|C_i| \cdot \text{cost}(G, \mathcal{A})}{\text{cost}(C_i, \mathcal{A})}$ implies that approximating $\mu_i$ up to a $(1 \pm \varepsilon)$ factor also approximates $|C_i|$ 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 $G$ with a non-empty huge range, Lemma 6 implies that the cost is well approximated.

**Lemma 7.** *Condition on event* $\mathcal{E}$*. Then, for any solution* $\mathcal{S}$*, and any* $i$ *such that there exists* $\ell \geq z \log(4z/\varepsilon)$ *such that the huge range* $I_{\ell, \mathcal{S}}$ *intersects* $C_i$*, we have:*

$$\left| cost(C_i, \mathcal{S}) - \sum_{p \in \Omega \cap C_i} \frac{|C_i| \cdot cost(G, \mathcal{A})}{cost(C_i, \mathcal{A}) \cdot \delta} \cdot cost(p, \mathcal{S}) \right| \leq 7\varepsilon \cdot cost(C_i, \mathcal{S}).$$

*Proof.* Let $p \in I_{\ell, \mathcal{S}} \cap C_i$ with $I_{\ell, \mathcal{S}}$ being huge. This implies, for any $q \in C_i$: $\text{cost}(p, q) \leq (\text{dist}(p, \mathcal{A}) + \text{dist}(q, \mathcal{A}))^z \leq 3^z \cdot \text{cost}(p, \mathcal{A}) \leq 3^z \cdot 2^{(\ell - z \log(4z/\varepsilon))} \text{cost}(p, \mathcal{A}) \leq (3\varepsilon/4z)^z \cdot \text{cost}(p, \mathcal{S})$. By Lemma 1, we therefore have for any point $q \in C_i$

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

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

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

$\square$

### 5.3 Bounding Interesting $I_{\ell,\mathcal{S}}$

Now we move onto the most involved case. 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* $L_{\mathcal{S}} \subset \{C_1, ..., C_k\}$, namely the clusters that satisfy

$$\nexists p \in C_i \mid \text{cost}(p, \mathcal{S}) \geq \left(\frac{4z}{\varepsilon}\right)^z \cdot \text{cost}(p, \mathcal{A}).$$

In other words $L_{\mathcal{S}}$ contains only clusters that do not have any point in a huge $I_{\ell,\mathcal{S}}$. This restriction will be crucial to our analysis.

**Designing a Good Estimator**

As discussed in Section 1, the key technical challenge of this section is to design an estimator for $\text{cost}(\mathcal{S})$ with small variance. Before defining and describing the estimator, we require the following lemma.

**Lemma 8.** *Let $\mathcal{S}$ be an arbitrary solution and $C_i$ be a cluster of $\mathcal{A}$ where all points are at the same distance from the center, up to a factor 2. Denote by $q_{i,\mathcal{S}} = \underset{p \in C_i}{\text{argmin}}\ \text{cost}(p, \mathcal{S})$. Then for every $\ell$ and every point $p \in C_i \cap I_{\ell,\mathcal{S}}$, there exists some weight $w_p \in [0, \max(1, 2^{\ell(1-1/z)}) \cdot 2^{O(z \log z)}]$ such that*

$$\text{cost}(p, \mathcal{S}) - \text{cost}(q_{i,\mathcal{S}}, \mathcal{S}) = w_p \cdot \text{cost}(q_{i,\mathcal{S}}, \mathcal{A}).$$

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

We first note that, since $p \in I_{\ell,\mathcal{S}}$, we have by choice of $q_{i,\mathcal{S}}$: $\text{cost}(q_{i,\mathcal{S}}, \mathcal{S}) \leq \text{cost}(p, \mathcal{S}) \leq 2^{\ell+1}\text{cost}(p, \mathcal{A}) \leq 2^{\ell+2}\text{cost}(q_{i,\mathcal{S}}, \mathcal{A})$. We also have that $\text{cost}(p, q_{i,\mathcal{S}}) \leq 2^{z-1}(\text{cost}(p, \mathcal{A}) + \text{cost}(q_{i,\mathcal{S}}, \mathcal{A})) \leq 3 \cdot 2^{z-1}\text{cost}(q_{i,\mathcal{S}}, \mathcal{A})$, since points in the same cluster have up to a factor 2 the same cost.

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

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

which after rearranging implies

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

If $\max(1, 2^{\ell+1}) = 1$, then we can merely set $\alpha = 1$ and the statement is true. Otherwise 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

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

□

We build on Lemma 8 to design a cost estimator with low variance. Instead of using $\sum_{p \in I_{\ell,\mathcal{S}} \cap L_{\mathcal{S}} \cap \Omega} f(p)\text{cost}(p, \mathcal{S})$ as an estimator of the cost for interesting ranges $I_{\ell,\mathcal{S}}$, we will rather use

$$E_{\ell,\mathcal{S}} := \sum_{C_i \in L_{\mathcal{S}}} \sum_{p \in C_i \cap I_{\ell,\mathcal{S}} \cap \Omega} f(p)\text{cost}(q_{i,\mathcal{S}}, \mathcal{A}) \cdot w_{p,\mathcal{S}}, \tag{5}$$

where $q_{i,\mathcal{S}} = \underset{p \in C_i}{\text{argmin}}\, \text{cost}(p, \mathcal{S})$, and $w_{p,\mathcal{S}}$ is a weight as given by Lemma 8, so that $w_{p,\mathcal{S}} \in [0, 2^{\ell(1-1/z)} \cdot 2^{O(z \log z)}]$. $E_{\ell,\mathcal{S}}$ can be expressed differently:

$$\begin{aligned} E_{\ell,\mathcal{S}} &= \sum_{C_i \in L_{\mathcal{S}}} \sum_{p \in C_i \cap I_{\ell,\mathcal{S}} \cap \Omega} f(p)(\text{cost}(p, \mathcal{S}) - \text{cost}(q_{i,\mathcal{S}}, \mathcal{S})) \\ &= \sum_{p \in I_{\ell,\mathcal{S}} \cap L_{\mathcal{S}} \cap \Omega} f(p)\text{cost}(p, \mathcal{S}) - F_{\ell,\mathcal{S}}, \qquad (6) \\ \text{with } F_{\ell,\mathcal{S}} &:= \sum_{C_i \in L_{\mathcal{S}}} \sum_{p \in C_i \cap I_{\ell,\mathcal{S}} \cap \Omega} f(p)\text{cost}(q_{i,\mathcal{S}}, \mathcal{S}) \end{aligned}$$

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

Note that the expectation of $E_{\ell,\mathcal{S}}$ is

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

Now instead of attempting to show concentration for the generic estimator

$$\sum_{p \in I_{\ell,\mathcal{S}} \cap L_{\mathcal{S}} \cap \Omega} f(p)\text{cost}(p, \mathcal{S}),$$

we will show

1. $E_{\ell,\mathcal{S}}$ is concentrated for all $\mathcal{S}$, and
2. $F_{\ell,\mathcal{S}}$ is both small and concentrated around its expectation.

One might ask why we are not arguing on $\sum_{p\in I_{\ell,\mathcal{S}}\cap\Omega} \frac{|C_i|\cdot\text{cost}(G,\mathcal{A})}{\text{cost}(C_i,\mathcal{A})\cdot\delta}\cdot\text{cost}(p,\mathcal{S})$ directly. The reason for decoupling the two arguments is that $E_{\ell,\mathcal{S}}$ has a very small variance, for which few samples are sufficient, and while $F_{\mathcal{S}}=\sum_\ell F_{\ell,\mathcal{S}}$ does not have a small variance, a union bound for all $\mathcal{S}$ can be easily inferred via event $\mathcal{E}$ from Lemma 6.

### Concentration of the Estimator

First, we show that every estimator $E_{\ell,\mathcal{S}}$ is tightly concentrated.

**Lemma 9.** *Consider an arbitrary solution $\mathcal{S}$. Then for any estimator $E_{\ell,\mathcal{S}}$ with $\ell\le z\log 4z/\varepsilon$, it holds that:*

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

*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.* We will rely on Bernstein's inequality (Theorem 3). To do this, we need an upper bound on the variance of $E_{\ell,\mathcal{S}}$, as well as an almost sure upper bound $M$ on every sample. Any estimator $E_{\ell,\mathcal{S}}$ has weights $w_p$ in $[0,2^{\ell(1-1/z)}\cdot 2^{O(z\log z)}]$ due to Lemma 8. We write $E_{\ell,\mathcal{S}}=\sum_{i=1}^{\delta}X_i$, where $X_i=f(\Omega_i)\text{cost}(q_{\mathcal{S}},\mathcal{A})\cdot w_p$ when the $i$-th sampled point of $G$ is $\Omega_i\in I_{\ell,\mathcal{S}}\cap L_{\mathcal{S}}$ and $X_i=0$ when $\Omega_i\notin I_{\ell,\mathcal{S}}\cap L_{\mathcal{S}}$. Recall that the probability that the $i$-th sampled point is $p$ satisfies $\mathbb{P}[\Omega_i=p]=\frac{\text{cost}(C,\mathcal{A})}{|C|\cdot\text{cost}(G,\mathcal{A})}\le\frac{2\text{cost}(p,\mathcal{A})}{\text{cost}(G,\mathcal{A})}$.

We first bound $\text{E}[X_i^2]$:

$$\begin{aligned}
\text{E}[X_i^2] &= \text{E}\left[\left(\sum_{p\in\cap I_{\ell,\mathcal{S}}\cap L_{\mathcal{S}}\cap\Omega_i} f(p)\text{cost}(p,\mathcal{A})\cdot w_{p,\mathcal{S}}\right)^2\right]\\
(|\Omega_i|=1)\quad &= \sum_{p\in I_{\ell,\mathcal{S}}\cap L_{\mathcal{S}}}\text{E}\left[\left(f(p)\text{cost}(p,\mathcal{A})\cdot w_{p,\mathcal{S}}\right)^2\right]\\
&\le \sum_{p\in I_{\ell,\mathcal{S}}}\text{E}\left[\left(2\cdot\text{cost}(p,\mathcal{A})\cdot w_{p,\mathcal{S}}\cdot\frac{\text{cost}(G,\mathcal{A})}{\delta\text{cost}(p,\mathcal{A})}\right)^2\right]\\
&\qquad\text{due to Fact 5}\\
&\le \sum_{p\in I_{\ell,\mathcal{S}}}\text{E}\left[\left(2^{\ell(1-1/z)}\cdot 2^{O(z\log z)}\cdot\frac{\text{cost}(G,\mathcal{A})}{\delta}\right)^2\right]\\
&\le \sum_i\sum_{pC_i\cap I_{\ell,\mathcal{S}}} 2^{2\ell(1-1/z)}\cdot 2^{O(z\log z)}\cdot\frac{\text{cost}^2(G,\mathcal{A})}{\delta^2}\cdot\frac{\text{cost}(p,\mathcal{A})}{\text{cost}(G,\mathcal{A})}\\
&\le \sum_{p\in I_{\ell,\mathcal{S}}} 2^{2\ell(1-1/z)}\cdot 2^{O(z\log z)}\cdot\frac{\text{cost}(G,\mathcal{A})}{\delta^2}\cdot\text{cost}(p,\mathcal{A}),
\end{aligned}$$

where the second inequality follows from using Lemma 8 and the final inequality uses Fact 5.

To bound $\sum_{p\in I_{\ell,\mathcal{S}}} \mathrm{cost}(p,\mathcal{A})$, 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 $\sum_{p\in I_{\ell,\mathcal{S}}} \mathrm{cost}(p,\mathcal{A})\leq \mathrm{cost}(G,\mathcal{A})$ as an upper bound. For the latter, we use $\sum_{p\in L_{\mathcal{S}}\cap I_{\ell,\mathcal{S}}} 2^{\ell+1}\cdot \mathrm{cost}(p,\mathcal{A})\leq \mathrm{cost}(I_{\ell,\mathcal{S}},\mathcal{S})$ as an upper bound. Combining this with $\mathrm{Var}[X_i]\leq \mathrm{E}[X_i^2]$, we obtain

$$\mathrm{Var}[X_i] \quad \leq \quad \frac{\mathrm{cost}(G,\mathcal{A})}{\delta^2}\cdot 2^{O(z\log z)}\cdot \mathrm{cost}(G,\mathcal{A}) \tag{7}$$

for $z=1$, and

$$\mathrm{Var}[X_i] \quad \leq \quad \frac{\mathrm{cost}(G,\mathcal{A})}{\delta^2}\cdot 2^{O(z\log z)}2^{\ell(1-2/z)}\mathrm{cost}(I_{\ell,\mathcal{S}},\mathcal{S}) \tag{8}$$

for the other case.

The almost sure upper bound (for which no case distinction is required) can be derived similarly

$$\begin{aligned} X_i \quad \leq \quad M \quad &:= \quad 2^{\ell(1-1/z)}\cdot 2^{O(z\log z)}\cdot \frac{\mathrm{cost}(G,\mathcal{A})}{\delta} \\ &\leq \quad \frac{z}{\varepsilon}\cdot 2^{\ell(1-2/z)}\cdot 2^{O(z\log z)}\cdot \frac{\mathrm{cost}(G,\mathcal{A})}{\delta}, \end{aligned} \tag{9}$$

where the inequality holds due to $\ell\leq z\log(4z/\varepsilon)$. Applying Bernstein's inequality with Equations 7, 8, and 9, we then have

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

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 $\varepsilon^2\cdot 2^{-\ell(1-2/z)}\geq \varepsilon^{2+z-z2/z}\cdot 2^{-O(z\log z)}=\varepsilon^z\cdot 2^{-O(z\log z)}$. This yields our final desired bound of

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

□

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

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

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: we can bound the sum of $F_{\ell,\mathcal{S}}$ at once. Indeed, since we focus on *interesting clusters*, we can consider the random variable defined as follows :

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

with expectation

$$\mathbb{E}[F_{\mathcal{S}}] = \sum_{C_i \in L_{\mathcal{S}}} \sum_{p \in C_i \cap \Omega} \text{cost}(q_{i,\mathcal{S}}, \mathcal{S}) \cdot \frac{|C_i| \cdot \text{cost}(G, \mathcal{A})}{\delta \text{cost}(C_i, \mathcal{A})}$$

Showing that $F_{\mathcal{S}}$ is concentrated is now an almost direct consequence of event $\mathcal{E}$ from Lemma 7.

**Lemma 10.** *Conditioned on event $\mathcal{E}$, we have for all solutions $\mathcal{S}$*

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

*Proof.* Given a solution $\mathcal{S}$, we have

$$\mathbb{E}[F_{\mathcal{S}}] = \sum_{C_i \in L_{\mathcal{S}}} \sum_{p \in C_i} \text{cost}(q_{i,\mathcal{S}}, \mathcal{S}) \cdot \frac{|C_i| \cdot \text{cost}(G, \mathcal{A})}{\delta \text{cost}(C_i, \mathcal{A})} \Pr[p \in \Omega] = \sum_{C_i \in L_{\mathcal{S}}} |C_i| \cdot \text{cost}(q_{i,\mathcal{S}}, \mathcal{S}).$$

Due to event $\mathcal{E}$, $\sum_{C_i \cap \Omega} \frac{|C_i| \cdot \text{cost}(G,\mathcal{A})}{\delta \text{cost}(C_i,\mathcal{A})} = (1 \pm \varepsilon) \cdot |C_i|$, for every $C_i \in L_{\mathcal{S}}$. Hence $F_{\mathcal{S}} = \sum_{C_i \in L_{\mathcal{S}}} \sum_{C_i \cap \Omega} \text{cost}(q_{\mathcal{S}}, \mathcal{S}) \cdot \frac{|C_i| \cdot \text{cost}(G,\mathcal{A})}{\text{cost}(C_i,\mathcal{A})} = (1 \pm \varepsilon) \cdot \mathbb{E}[F_{\mathcal{S}}]$. Now finally observe that since $q_{i,\mathcal{S}}$ was always the point of $C_i$ whose cost in $\mathcal{S}$ is the smallest, we have $\mathbb{E}[F_{S}] \leq \text{cost}(L_{\mathcal{S}}, \mathcal{S}) \leq \text{cost}(G, \mathcal{S})$. □

## 5.4 Proving the Key Lemma

We now conclude by decomposing the term

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

into terms for which we can apply Lemmas 5, 7, 9, and 10.

First, we note that the probability of success of Lemma 9 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 $E_{\ell,\mathcal{S}}$ to $E_{\ell,\tilde{\mathcal{S}}}$, where $\tilde{\mathcal{S}}$ comes from a small set on which union-bounding is possible.

**Lemma 11.** *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 $\tilde{\mathcal{S}} \in \mathbb{C}^k$*

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

*Proof.* Taking a union-bound over the success of Lemma 9 for all possible $\tilde{\mathcal{S}} \in \mathbb{C}^k$ and all $\ell$ such that $\log(\varepsilon/2) \le \ell \le z\log(4z/\varepsilon)$, 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 $\tilde{\mathcal{S}} \in \mathbb{C}^k$ and $\ell$,

$$|E_{\ell,\tilde{\mathcal{S}}} - \mathbb{E}[E_{\ell,\tilde{\mathcal{S}}}]| \le \frac{\varepsilon}{z\log z/\varepsilon} \cdot \left(\text{cost}(G,\mathcal{A}) + \text{cost}(I_{\ell,\tilde{\mathcal{S}}},\tilde{\mathcal{S}})\right) \tag{10}$$

We now condition on that event, together with event $\mathcal{E}$. We write:

$$\begin{aligned}
&\left|\sum_{p\in L_{\tilde{\mathcal{S}}}} \text{cost}(p,\tilde{\mathcal{S}}) - \sum_{p\in L_{\tilde{\mathcal{S}}}\cap\Omega} f(p)\cdot\text{cost}(p,\tilde{\mathcal{S}})\right| \\
=\ & \left|\sum_{p\in L_{\tilde{\mathcal{S}}}} \text{cost}(p,\tilde{\mathcal{S}}) - \mathbb{E}[F_{\tilde{\mathcal{S}}}] + \mathbb{E}[F_{\tilde{\mathcal{S}}}] - F_{\tilde{\mathcal{S}}} + F_{\tilde{\mathcal{S}}} - \sum_{p\in L_{\tilde{\mathcal{S}}}\cap\Omega} f(p)\cdot\text{cost}(p,\tilde{\mathcal{S}})\right| \\
\le\ & \left|\sum_{p\in L_{\tilde{\mathcal{S}}}} \text{cost}(p,\tilde{\mathcal{S}}) - \mathbb{E}[F_{\tilde{\mathcal{S}}}] + F_{\tilde{\mathcal{S}}} - \sum_{p\in L_{\tilde{\mathcal{S}}}\cap\Omega} f(p)\cdot\text{cost}(p,\tilde{\mathcal{S}})\right| + |\mathbb{E}[F_{\tilde{\mathcal{S}}}] - F_{\tilde{\mathcal{S}}}| \\
\le\ & \sum_{\ell<\log\varepsilon/2}\left|\sum_{p\in I_{\ell,\tilde{\mathcal{S}}}\cap L_{\tilde{\mathcal{S}}}} \text{cost}(p,\tilde{\mathcal{S}}) - \mathbb{E}[F_{\ell,\tilde{\mathcal{S}}}] + F_{\ell,\tilde{\mathcal{S}}} - \sum_{p\in I_{\ell,\tilde{\mathcal{S}}}\cap L_{\tilde{\mathcal{S}}}\cap\Omega} f(p)\cdot\text{cost}(p,\tilde{\mathcal{S}})\right| && (11)\\
& + \sum_{\ell=\log\varepsilon/2}^{z\log z/4\varepsilon}\left|\sum_{p\in I_{\ell,\tilde{\mathcal{S}}}\cap L_{\tilde{\mathcal{S}}}} \text{cost}(p,\tilde{\mathcal{S}}) - \mathbb{E}[F_{\ell,\tilde{\mathcal{S}}}] + F_{\ell,\tilde{\mathcal{S}}} - \sum_{p\in I_{\ell,\tilde{\mathcal{S}}}\cap L_{\tilde{\mathcal{S}}}\cap\Omega} f(p)\cdot\text{cost}(p,\tilde{\mathcal{S}})\right| && (12)\\
& + |\mathbb{E}[F_{\tilde{\mathcal{S}}}] - F_{\tilde{\mathcal{S}}}|
\end{aligned}$$

We note that Equation 12 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 10. To bound tiny points of Equation 11, we combine Lemma 5 with the observation that $F_{\ell,\tilde{\mathcal{S}}} \le \sum_{p\in I_{\ell,\tilde{\mathcal{S}}}\cap\Omega} f(p)\text{cost}(p,\tilde{\mathcal{S}})$. This gives:

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

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

$$\left|\sum_{p\in L_{\tilde{\mathcal{S}}}} \text{cost}(p,\tilde{\mathcal{S}}) - \sum_{p\in L_{\tilde{\mathcal{S}}}\cap\Omega} f(p)\cdot\text{cost}(p,\tilde{\mathcal{S}})\right| \le 2\varepsilon\text{cost}(\mathcal{A}) + \sum_{\ell=\log\varepsilon/2}^{z\log z/4\varepsilon} \left|\mathbb{E}[E_{\ell,\tilde{\mathcal{S}}}] - E_{\ell,\tilde{\mathcal{S}}}\right| + |\mathbb{E}[F_{\tilde{\mathcal{S}}}] - F_{\tilde{\mathcal{S}}}|$$

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

where the second to last inequality used Lemma 10.

□

We can now finally turn to the proof of Lemma 2:

*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 $\text{dist}(p,\mathcal{S}) \le \frac{8z}{\varepsilon}\cdot\text{dist}(p,\mathcal{A})$ or $\text{dist}(p,\tilde{\mathcal{S}}) \le \frac{8z}{\varepsilon}\cdot\text{dist}(p,\mathcal{A})$ , we have $|\text{cost}(p,\mathcal{S}) - \text{cost}(p,\tilde{\mathcal{S}})| \le \varepsilon(\text{cost}(p,\mathcal{S}) + \text{cost}(p,\mathcal{A}))$.

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 7. All other points have distance well approximated by $\tilde{\mathcal{S}}$. Then, we can apply Lemma 11 to $L_{\tilde{\mathcal{S}}}$, since all points in $L_{\tilde{\mathcal{S}}}$ have $\text{dist}(p,\tilde{\mathcal{S}}) \le \frac{4z}{\varepsilon}\cdot\text{dist}(p,\mathcal{A})$, and so $\text{dist}(p,\mathcal{S}) \le \frac{8z}{\varepsilon}\cdot\text{dist}(p,\mathcal{A})$ 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 $\text{dist}(p,\mathcal{S}) \le \frac{8z}{\varepsilon}\cdot\text{dist}(p,\mathcal{A})$, 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.

We now make the argument formal. Let $H_{\mathcal{S}}$ be the set of all clusters that are intersecting with some huge $I_{\ell,\mathcal{S}}$ with $\ell > z\log(8z/\varepsilon)$. We decompose the cost difference as follows:

$$\begin{aligned}
\left|\text{cost}(G,\mathcal{S}) - \sum_{p\in\Omega\cap G} f(p)\cdot\text{cost}(p,\mathcal{S})\right| &\le \left|\sum_{p\in G\setminus H_{\mathcal{S}}} \text{cost}(p,\mathcal{S}) - \sum_{p\in(G\setminus H_{\mathcal{S}})\cap\Omega} f(p)\cdot\text{cost}(p,\mathcal{S})\right| && (13)\\
&\quad + \left|\sum_{p\in H_{\mathcal{S}}} \text{cost}(p,\mathcal{S}) - \sum_{p\in H_{\mathcal{S}}\cap\Omega} f(p)\cdot\text{cost}(p,\mathcal{S})\right| && (14)
\end{aligned}$$

Since we condition on event $\mathcal{E}$, we can bound term 14 by $O(\varepsilon)\cdot(\text{cost}(G,\mathcal{A}) + \text{cost}(G,\mathcal{S}))$, using Lemma 7. Now we take a closer look at term 13. For all points $p \in G\setminus H_{\mathcal{S}}$, we have $|\text{cost}(p,\mathcal{S}) - \text{cost}(p,\tilde{\mathcal{S}})| \le \varepsilon(\text{cost}(p,\mathcal{S}) + \text{cost}(p,\mathcal{A}))$. Furthermore, $L_{\tilde{\mathcal{S}}} \cap (G\setminus H_{\mathcal{S}}) = L_{\tilde{\mathcal{S}}}$. We let $H_{\tilde{\mathcal{S}}}$ be the set of clusters that are intersecting with some huge $I_{\ell,\tilde{\mathcal{S}}}$ with $\ell > z\log(4z/\varepsilon)$, and that are not in $H_{\mathcal{S}}$.

Hence,

$$\begin{aligned}
\left|\sum_{p\in G\setminus H_{\mathcal{S}}}\mathrm{cost}(p,\mathcal{S})-\sum_{p\in(G\setminus H_{\mathcal{S}})\cap\Omega}f(p)\cdot\mathrm{cost}(p,\mathcal{S})\right| &\leq \left|\sum_{p\in G\setminus H_{\mathcal{S}}}\mathrm{cost}(p,\tilde{\mathcal{S}})-\sum_{p\in(G\setminus H_{\mathcal{S}})\cap\Omega}f(p)\cdot\mathrm{cost}(p,\tilde{\mathcal{S}})\right| \\
&+ \varepsilon\left(\mathrm{cost}(\mathcal{S})+\mathrm{cost}(\mathcal{A})+\mathrm{cost}(\Omega,\mathcal{S})+\mathrm{cost}(\Omega,\mathcal{A})\right) \\
&\leq \left|\sum_{p\in L_{\tilde{\mathcal{S}}}}\mathrm{cost}(p,\tilde{\mathcal{S}})-\sum_{p\in L_{\tilde{\mathcal{S}}}\cap\Omega}f(p)\cdot\mathrm{cost}(p,\tilde{\mathcal{S}})\right| \qquad (15) \\
&+ \left|\sum_{p\in H_{\tilde{\mathcal{S}}}}\mathrm{cost}(p,\tilde{\mathcal{S}})-\sum_{p\in H_{\tilde{\mathcal{S}}}\cap\Omega}f(p)\cdot\mathrm{cost}(p,\tilde{\mathcal{S}})\right| \qquad (16) \\
&+ \varepsilon\left(\mathrm{cost}(\mathcal{S})+\mathrm{cost}(\mathcal{A})+\mathrm{cost}(\Omega,\mathcal{S})+\mathrm{cost}(\Omega,\mathcal{A})\right)
\end{aligned}$$

Eqution 15 is directly bounded using Lemma 11 by $\varepsilon(\mathrm{cost}(G,\mathcal{A})+\mathrm{cost}(L_{\tilde{\mathcal{S}}},\tilde{\mathcal{S}}))$. For Equation 16, we use again Lemma 7. Since we conditionned on event $\mathcal{E}$, for each cluster $C_i$ in $H_{\tilde{\mathcal{S}}}$ it holds that $\left|\mathrm{cost}(C_i,\tilde{\mathcal{S}})-\sum_{p\in C_i\cap\Omega}f(p)\mathrm{cost}(p,\tilde{\mathcal{S}})\right|=O(\varepsilon)\mathrm{cost}(C_i,\tilde{\mathcal{S}})$.

Since such all those points are not in $H_{\mathcal{S}}$, their cost in $\mathcal{S}$ and $\tilde{\mathcal{S}}$ satisfies $|\mathrm{cost}(p,\tilde{\mathcal{S}})-\mathrm{cost}(p,\mathcal{S})|\leq O(\varepsilon)(\mathrm{cost}(p,\mathcal{S})+\mathrm{cost}(p,\mathcal{A}))$. Hence, $\varepsilon\mathrm{cost}(L_{\tilde{\mathcal{S}}},\tilde{\mathcal{S}})=O(\varepsilon)(\mathrm{cost}(G,\mathcal{S})+\mathrm{cost}(G,\mathcal{A}))$, and summing over all clusters of $H_{\tilde{\mathcal{S}}}\cap L_{\mathcal{S}}$, we get a bound for Equation 16 of $O(\varepsilon)(\mathrm{cost}(G,\mathcal{S})+\mathrm{cost}(G,\mathcal{A}))$

Combining all the equations we now have

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

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

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

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

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

Hence, we finally conclude:

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

The probability now follows from taking a union-bound over the failure probability of Lemma 6 and Lemma 11. 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 $C_i$, the complexity of the algorithm is $O(|C_i|)$: it is both the cost of computing the scaling factor $f(p)$ for all $p\in G$, 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 $(X, dist)$ 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 clients $G$*

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

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

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 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 $4^z \cdot \text{cost}(p, \mathcal{A})$. Denote these points by $G_{close,\mathcal{S}}$. For this case, we mainly rely on Bernstein's inequality as well as properties of $R_O(\mathcal{A})$.
- 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 $G_{far,\mathcal{S}}$. For these points, we can afford to replace their distance to $S$ with the distance to the closest center $c \in \mathcal{A}$ plus the distance from $c$ to the closest center in $S$. 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 $S$.

We first analyse the points in $G_{close,\mathcal{S}}$.

*Proof of Lemma 3 for $G_{close}$.* We aim to use Bernstein's Inequality. Let $E_{close,\mathcal{S}} = \sum_{i=1}^{\delta} X_i$, where $X_i = \frac{\text{cost}(G,\mathcal{A})}{\delta \cdot \text{cost}(p,\mathcal{A})} \cdot \text{cost}(p, \mathcal{S})$ if the $i$-th sampled point $p \in G_{close,\mathcal{S}}$ and $X_i = 0$ the $i$-th sampled point $p \notin G_{close,\mathcal{S}}$. Recall that the probability of sampling a point $p$ is $\frac{\delta \cdot \text{cost}(p,\mathcal{A})}{\text{cost}(G,\mathcal{A})}$. We consider the second moment $\text{E}[X_i^2]$:

$$\begin{aligned}
E[X_i^2] &= \sum_{p\in G_{close}} \left(\frac{\text{cost}(G,\mathcal{A})}{\delta\cdot\text{cost}(p,\mathcal{A})}\cdot\text{cost}(p,\mathcal{S})\right)^2\cdot\mathbb{P}[p\in\Omega]\\
&= \text{cost}(G,\mathcal{A})\cdot\sum_{p\in G_{close}}\frac{\text{cost}(p,\mathcal{S})}{\delta^2\cdot\text{cost}(p,\mathcal{A})}\cdot\text{cost}(p,\mathcal{S})\\
&\le \text{cost}(G,\mathcal{A})\cdot\sum_{p\in G_{close}}\frac{4^z}{\delta^2}\cdot\text{cost}(p,\mathcal{S})\\
&\le \frac{4^z}{\delta^2}\cdot\text{cost}(G,\mathcal{A})\cdot\text{cost}(G,\mathcal{S})
\end{aligned}$$

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

$$X_i \le M := \max_{p\in G_{close,\mathcal{S}}}\frac{\text{cost}(G,\mathcal{A})}{\delta\cdot\text{cost}(p,\mathcal{A})}\cdot\text{cost}(p,\mathcal{S}) \le \frac{4^z}{\delta}\cdot\text{cost}(G,\mathcal{A}). \tag{17}$$

Combining both bounds with Bernstein's inequality now yields

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

Taking a union bound over all possible $\mathcal{S}\in\mathbb{C}^k$, we have with probability

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

$$\forall\mathcal{S}\in\mathbb{C}^k,\ |E_{close}-\mathbb{E}[E_{close}]| \le \frac{\varepsilon}{z\log z/\varepsilon}\cdot(\text{cost}(P,A)+\text{cost}(P,\mathcal{S}))$$

The same technique as in the final proof of Lemma 2 can be used to extend that result to any solution $\mathcal{S}$ (not restricted to $\mathbb{C}^k$). More precisely, since all points of $G_{close,\mathcal{S}}$ are of type interesting, their cost is preserved in the solution $\tilde{\mathcal{S}}$ that approximates $\mathcal{S}$. Hence, it is enough to show the coreset guarantee to $\tilde{\mathcal{S}}$. □

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

*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 $\text{E}[X_i^2]$:

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

where the final inequality follows since every cluster either has cost at least $\frac{1}{k}\text{cost}(G_{\max}^O, \mathcal{A})$, if $G = G_{\max}^O$, or the clusters in $G_b^O$ have an equal cost, up to a factor of 2.

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

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

Combining both bounds with Bernstein's inequality now yields

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

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

*Then, the there exists an algorithm* `SensitivitySample` *running in time $O(|G_j^O|)$ that constructs a set $\Omega$ of size $\delta$ such that, it holds that for all sets $\mathcal{S}$ of $k$ centers that:*

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

*Proof.* Our aim will be to show that $\mathrm{cost}(G_{far,\mathcal{S}},\mathcal{S})+\mathrm{cost}(\Omega\cap G_{far,\mathcal{S}},\mathcal{S})\leq\varepsilon\cdot\mathrm{cost}(G,\mathcal{S})$.

First, we fix a cluster $C\in\mathcal{A}$, and show that the total contribution of points of $C\cap G_{far,\mathcal{S}}$ is very cheap compared to $\mathrm{cost}(C,\mathcal{S})$, i.e. that $\mathrm{cost}(G_{far,\mathcal{S}}\cap C,\mathcal{S})\leq\frac{\varepsilon}{z\log z/\varepsilon}\cdot\mathrm{cost}(C,\mathcal{S})$. Let $c$ be the center serving $p\in G_{far,\mathcal{S}}\cap C$ in $\mathcal{A}$. Let $C_{close}$ be the points of $C$ with cost at most $\left(\frac{2z}{\varepsilon}\right)^z\cdot\frac{\mathrm{cost}(C,c)}{|C|}$. Consider an arbitrary point in $q\in C_{close}$. Due to the triangle inequality and $\mathrm{cost}(p,\mathcal{S})>4^z\cdot\mathrm{cost}(p,c)$, we have $\mathrm{dist}(c,\mathcal{S})\geq\mathrm{dist}(p,\mathcal{S})-\mathrm{dist}(p,c)\geq 4\mathrm{dist}(p,c)-\mathrm{dist}(p,c)\geq\mathrm{dist}(p,c)$. Therefore $\mathrm{cost}(c,\mathcal{S})\geq\left(\frac{4z}{\varepsilon}\right)^{2z}\cdot\frac{\mathrm{cost}(C,c)}{|C|}$. Using this and Lemma 1 we now have for any $q\in C_{close}$

$$\begin{aligned}
\mathrm{cost}(c,\mathcal{S}) &\leq (1+\varepsilon/(2z))^{z-1}\cdot\mathrm{cost}(q,\mathcal{S})+\left(\frac{2z+\varepsilon}{\varepsilon}\right)^{z-1}\cdot\mathrm{cost}(q,c)\\
&\leq (1+\varepsilon)\mathrm{cost}(q,\mathcal{S})+\left(\frac{2z+\varepsilon}{\varepsilon}\right)^{z-1}\cdot\left(\frac{4z}{\varepsilon}\right)^{z}\cdot\frac{\mathrm{cost}(C,c)}{|C|}\\
&\leq (1+\varepsilon)\mathrm{cost}(q,\mathcal{S})+\frac{\left(\frac{4z}{\varepsilon}\right)^{2z-1}\cdot\frac{\mathrm{cost}(C,c)}{|C|}}{\mathrm{cost}(p,c)}\cdot\mathrm{cost}(p,c)\\
&\leq (1+\varepsilon)\mathrm{cost}(q,\mathcal{S})+\varepsilon\cdot\mathrm{cost}(p,c) \qquad \text{since } p\in R_O(C)\\
&\leq (1+\varepsilon)\mathrm{cost}(q,\mathcal{S})+\varepsilon\cdot\mathrm{cost}(c,\mathcal{S})\\
\Rightarrow \mathrm{cost}(q,\mathcal{S}) &\geq \frac{1-\varepsilon}{1+\varepsilon}\cdot\mathrm{cost}(c,\mathcal{S}) \qquad (18)
\end{aligned}$$

We now bound the cost of $\mathrm{cost}(C,\mathcal{S})$ in terms of $\mathrm{cost}(C,c)$. We have due to Markov's inequality $|C_{close}|\geq(1-\varepsilon)\cdot|C|$ and therefore

$$\begin{aligned}
\mathrm{cost}(C,\mathcal{S}) &\geq \mathrm{cost}(C_{close},\mathcal{S})=\sum_{q\in C_{close}}\mathrm{cost}(q,\mathcal{S})\geq|C_{close}|\cdot\frac{1-\varepsilon}{1+\varepsilon}\cdot cost(c,\mathcal{S}) \qquad (19)\\
&\geq |C_{close}|\cdot\frac{1-\varepsilon}{1+\varepsilon}\cdot\left(\frac{4z}{\varepsilon}\right)^{2z}\cdot\frac{\mathrm{cost}(C,c)}{|C|}\geq\left(\frac{4z}{\varepsilon}\right)^{2z-1}\cdot\mathrm{cost}(C,c) \qquad (20)
\end{aligned}$$

which yields

$$
\begin{aligned}
\mathrm{cost}(G_{far,\mathcal{S}} \cap C, \mathcal{S}) &= \sum_{p \in G_{far,\mathcal{S}} \cap C} \mathrm{cost}(p, \mathcal{S}) \\
(\textit{Lemma } 1) \quad &\leq \sum_{p \in G_{far,\mathcal{S}} \cap C} (1+\varepsilon/2z)^{z-1} \mathrm{cost}(c, \mathcal{S}) + \left(\frac{2z+\varepsilon}{\varepsilon}\right)^{z-1} \cdot \mathrm{cost}(p, c) \\
&\leq |G_{far,\mathcal{S}} \cap C| \cdot (1+\varepsilon) \cdot \mathrm{cost}(c, \mathcal{S}) + \left(\frac{2z+\varepsilon}{\varepsilon}\right)^{z-1} \cdot \mathrm{cost}(G_{far,\mathcal{S}} \cap C, c) \\
(\text{Markov's inequality}) \quad &\leq (1+\varepsilon) \cdot \left(\frac{\varepsilon}{2z}\right)^{2z} \cdot |C| \cdot \mathrm{cost}(c, \mathcal{S}) + \left(\frac{2z+\varepsilon}{\varepsilon}\right)^{z-1} \cdot \mathrm{cost}(G_{far,\mathcal{S}} \cap C, c) \qquad (21) \\
(\text{Markov's inequality}) \quad &\leq \frac{1+\varepsilon}{1-\varepsilon} \cdot \left(\frac{\varepsilon}{2z}\right)^{2z} \cdot |C_{close}| \cdot \mathrm{cost}(c, \mathcal{S}) + \left(\frac{2z+\varepsilon}{\varepsilon}\right)^{z-1} \cdot \mathrm{cost}(G_{far,\mathcal{S}} \cap C, c) \\
(\textit{Eq. } 19) \quad &\leq \frac{(1+\varepsilon)^2}{(1-\varepsilon)^2} \cdot \left(\frac{\varepsilon}{2z}\right)^{2z} \cdot \mathrm{cost}(C, \mathcal{S}) + \left(\frac{2z+\varepsilon}{\varepsilon}\right)^{z-1} \cdot \left(\frac{\varepsilon}{4z}\right)^{2z-1} \cdot \mathrm{cost}(G_{far,\mathcal{S}} \cap C, \mathcal{S}) \\
&\leq \frac{\varepsilon}{z \log z/\varepsilon} \cdot \mathrm{cost}(C, \mathcal{S}) \qquad (22)
\end{aligned}
$$

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

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

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

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 $\mathcal{A}$

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

Therefore, we have

$$
\begin{aligned}
& \operatorname{cost}(G_{far,\mathcal{S}} \cap \Omega \cap C, \mathcal{S}) = \sum_{p \in G_{far,\mathcal{S}} \cap C} \frac{\operatorname{cost}(G,\mathcal{A})}{\delta \cdot \operatorname{cost}(p,\mathcal{A})} \cdot \operatorname{cost}(p,\mathcal{S}) \\
(Lemma\ 1) \quad \leq \quad & \sum_{p \in G_{far,\mathcal{S}} \cap \Omega \cap C} \frac{\operatorname{cost}(G,\mathcal{A})}{\delta \cdot \operatorname{cost}(p,\mathcal{A})} \cdot \left( (1+\varepsilon/2z)^{z-1} \operatorname{cost}(c,\mathcal{S}) + \left(\frac{2z+\varepsilon}{\varepsilon}\right)^{z-1} \cdot \operatorname{cost}(p,c) \right) \\
\leq \quad & (1+\varepsilon) \cdot \operatorname{cost}(c,\mathcal{S}) \cdot \sum_{p \in G_{far,\mathcal{S}} \cap \Omega \cap C} \frac{\operatorname{cost}(G,\mathcal{A})}{\delta \cdot \operatorname{cost}(p,\mathcal{A})} \\
(\mathcal{E}_{far}) \quad & + \left(\frac{2z+\varepsilon}{\varepsilon}\right)^{z-1} \cdot (1+\varepsilon) \cdot \operatorname{cost}(C \cap G, \mathcal{A}) \\
(Eq.\ 24) \quad \leq \quad & (1+\varepsilon) \cdot \operatorname{cost}(c,\mathcal{S}) \cdot \left(\frac{\varepsilon}{2z}\right)^{2z} \cdot |C| + \left(\frac{2z+\varepsilon}{\varepsilon}\right)^{z-1} \cdot (1+\varepsilon) \cdot \operatorname{cost}(C \cap G, \mathcal{A}) \\
\leq \quad & (1+\varepsilon) \cdot \left(\frac{\varepsilon}{2z}\right)^{2z} \cdot |C| \cdot \operatorname{cost}(c,\mathcal{S}) + \left(\frac{2z+\varepsilon}{\varepsilon}\right)^{z-1} \cdot \operatorname{cost}(C,c) \qquad (25) \\
\leq \quad & \frac{\varepsilon}{z \log z/\varepsilon} \cdot \operatorname{cost}(C,\mathcal{S})
\end{aligned}
$$

where the steps following Equation 25 are identical to those used to derive Equation 22 from Equation 21. Again, summing over all clusters now yields

$$
\operatorname{cost}(G_{far,\mathcal{S}}, \mathcal{S}) \leq \frac{\varepsilon}{z \log z/\varepsilon} \cdot \operatorname{cost}(P,\mathcal{S}),
$$

which yields the claim. □

The overall proof now follows by adding up the error bounds of $G_{far,\mathcal{S}}$ (Lemma 13) and $G_{close,\mathcal{S}}$ and rescaling $\varepsilon$ by a constant factor. Specifically, we have

$$
\begin{aligned}
& |\operatorname{cost}(G,\mathcal{S}) - \operatorname{cost}(\Omega,\mathcal{S})| \\
\leq \quad & |\operatorname{cost}(G_{close,\mathcal{S}},\mathcal{S}) - \operatorname{cost}(\Omega \cap G_{close,\mathcal{S}},\mathcal{S})| + |\operatorname{cost}(G_{far,\mathcal{S}},\mathcal{S}) - \operatorname{cost}(\Omega \cap G_{far,\mathcal{S}},\mathcal{S})| \\
\leq \quad & 2\frac{\varepsilon}{z \log z/\varepsilon} \cdot (\operatorname{cost}(\mathcal{S}) + \operatorname{cost}(\mathcal{A})) .
\end{aligned}
$$

# 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 $(X, dist)$ 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*

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

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 $(\varepsilon/z)^z \Delta_C$ (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 $G^O_{min}$. $P_1$ 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 $G^O_{min}$ 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 $\text{cost}(\mathcal{A}) \le c_{\mathcal{A}}\text{cost}(\text{OPT})$.

**Lemma 14.** *For any solution $\mathcal{S}$ and any cluster $C$ with center $c$ of $\mathcal{A}$,*

$$|cost(R_I(C),\mathcal{S}) - |R_I(C)| \cdot cost(c,\mathcal{S})| \le 3\varepsilon(cost(C,\mathcal{A}) + cost(R_I(C),\mathcal{S})).$$

**Lemma 15.** *For any solution $\mathcal{S}$ and any $j$,*

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

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

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

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:

$$\begin{aligned}
|\text{cost}(\mathcal{D},\mathcal{S}) - \text{cost}(P_1,\mathcal{S})| &\le \sum_{i=1}^k |\text{cost}(R_I(C_i),\mathcal{S}) - |R_I(C_i)|\text{cost}(c_i,\mathcal{S})| \\
&\quad + \sum_{j=z\log(\varepsilon/z)}^{2z\log(z/\varepsilon)} \left|\text{cost}(G_{j,min},\mathcal{S}) - \sum_{i=1}^k |C_i \cap G_{j,min}|\text{cost}(c_i,\mathcal{S})\right| \\
&\quad + \left|\text{cost}(G^O_{min},\mathcal{S}) - \sum_{i=1}^k |C_i \cap G^O_{min}|\text{cost}(c_i,\mathcal{S})\right| \\
&\le \sum_{i=1}^k 3\varepsilon(\text{cost}(C_i,\mathcal{A}) + \text{cost}(R_I(C_i),\mathcal{S})) \\
&\quad + 2\varepsilon\text{cost}(\mathcal{S}) + 2\varepsilon\text{cost}(\mathcal{A}) + \varepsilon(\text{cost}(\mathcal{S}) + \text{cost}(\mathcal{A})) \\
&\le 12\varepsilon c_{\mathcal{A}}\text{cost}(\mathcal{S}),
\end{aligned}$$

where the second inequality uses Lemmas 14 and 15. $\square$

## 7.1 The Inner Ring: Proof of Lemma 14

**Lemma 14.** *For any solution $\mathcal{S}$ and any cluster $C$ with center $c$ of $\mathcal{A}$,*

$$|cost(R_I(C), \mathcal{S}) - |R_I(C)| \cdot cost(c, \mathcal{S})| \leq 3\varepsilon(cost(C, \mathcal{A}) + cost(R_I(C), \mathcal{S})).$$

*Proof.* Let $C$ be a cluster induced by $\mathcal{A}$. For a point $p$ in the inner ring $R_I(C)$, we have by the modified triangle inequality Lemma 1

$$\begin{aligned}|\text{cost}(p, \mathcal{S}) - \text{cost}(p, c)| &\leq \varepsilon\text{cost}(c, \mathcal{S}) + (1 + z/\varepsilon)^{z-1}\text{cost}(p, c) \\ &\leq \varepsilon\text{cost}(c, \mathcal{S}) + (1 + z/\varepsilon)^{z-1} (\varepsilon/z)^z \Delta_C \\ &\leq \varepsilon(1 + \varepsilon)(\text{cost}(c, \mathcal{S}) + \Delta_C).\end{aligned}$$

Summing over all points of the inner ring yields

$$\begin{aligned}|\text{cost}(R_I(C), \mathcal{S}) - |R_I(C)| \cdot \text{cost}(c, \mathcal{S})| &\leq \sum_{p \in R_I(C)} |\text{cost}(p, \mathcal{S}) - \text{cost}(p, c)| \\ &\leq |R_I(C)|\varepsilon(1 + \varepsilon)(\text{cost}(c, \mathcal{S}) + \Delta_C) \\ &\leq \varepsilon(1 + \varepsilon)(\text{cost}(C, \mathcal{A}) + |R_I(C)|\text{cost}(c, \mathcal{S})) \\ &\leq \varepsilon(1 + \varepsilon)(\text{cost}(C, \mathcal{A}) + \text{cost}(R_I(C), \mathcal{S})) \\ &\quad + \varepsilon(1 + \varepsilon) |\text{cost}(R_I(C), \mathcal{S}) - |R_I(C)| \cdot \text{cost}(c, \mathcal{S})|\end{aligned}$$

This implies

$$|\text{cost}(R_I(C), \mathcal{S}) - |R_I(C)| \cdot \text{cost}(c, \mathcal{S})| \leq 3\varepsilon(\text{cost}(C, \mathcal{A}) + \text{cost}(R_I(C), \mathcal{S})).$$

□

## 7.2 The Cheap Groups: Proof of Lemma 15

**Lemma 15.** *For any solution $\mathcal{S}$ and any $j$,*

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

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

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

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

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

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

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

Now, either $G = G_{j,min}$ for some $j$, and $\mathrm{cost}(G, \mathcal{A}) \leq \left(\frac{\varepsilon}{2z}\right)^z \cdot \mathrm{cost}(R_j, \mathcal{A})$; or $G = G^O_{min}$, and $\mathrm{cost}(G, \mathcal{A}) \leq \left(\frac{\varepsilon}{2z}\right)^z \cdot \mathrm{cost}(R_O(\mathcal{A}), \mathcal{A}) \leq \left(\frac{\varepsilon}{2z}\right)^z \cdot \mathrm{cost}(\mathcal{A})$.

In both cases, the lemma follows. □

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

In this section, we apply the coreset framework to specifics 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 that, 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 $\mathrm{cost}(p, \mathcal{S}) \leq \left(\frac{4z}{\varepsilon}\right)^z \mathrm{cost}(p, \mathcal{A})$, it holds $|\mathrm{cost}(p, \mathcal{S}) - \mathrm{cost}(p, \tilde{\mathcal{S}})| \leq \frac{\varepsilon}{z \log(z/\varepsilon)} (\mathrm{cost}(p, \mathcal{S}) + \mathrm{cost}(p, \mathcal{A}))$.

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

**Definition 4.** *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 $\mathrm{dist}(v, x) \leq \gamma$, and for all $x, y \in X$ we have $\mathrm{dist}(x, y) > \gamma$. 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 16** (from Gupta et. al [GKL03])**.** *Let $(V, dist)$ be a metric space with doubling dimension $d$ and, diameter $D$, and let $X$ be a $\gamma$-net of $V$. Then $|X| \le 2^{d\cdot\lceil \log_2(D/\gamma)\rceil}$.*

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

**Lemma 17.** *Let $M = (X, dist)$ 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*

$$|P| \cdot (\varepsilon/c_{\mathcal{A}})^{-O(zd)}$$

A direct corollary of that lemma is the existence of a coreset in Doubling Metrics:

**Corollary 4.** *Let $M = (X, dist)$ 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*

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

*Proof.* We first compute a coreset of size $\tilde{O}(k^3 D\varepsilon^{-2})$ [HJLW18]. Then, combining Theorem 1 and Lemma 17 yields an algorithm constructing a coreset of size

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

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 17.* For each point $p \in P$, let $c$ be the center to which $p$ was assigned in $\mathcal{A}$. Let $B(p, \left(\frac{10z}{\varepsilon}\right)\text{dist}(p, c))$ be the metrics ball centered around $p$ with radius $\left(\frac{10z}{\varepsilon}\right)\cdot\text{dist}(p, c)$.

Order points $p_1, ..., p_n$ with non-decreasing value of $\text{dist}(p, \mathcal{A})$. Let $N_{p_i}$ be an $\left(\frac{\varepsilon}{2z}\right)\cdot\text{dist}(p, \mathcal{A})$-net of $B\left(p_i, \left(\frac{10z}{\varepsilon}\right)\cdot\text{dist}(p_i, \mathcal{A})\right) \ \setminus\ \cup_{j<i} B\left(p_j, \left(\frac{10z}{\varepsilon}\right)\cdot\text{dist}(p_j, \mathcal{A})\right)$, which due to Lemma 16 has size $(\varepsilon/z)^{-O(d)}$. Furthermore, let $s_f$ be a point not in any $B(p, \left(\frac{10z}{\varepsilon}\right)\text{dist}(p, \mathcal{A}))$, if such a point exist. Let $\mathcal{N} := s_f \bigcup_{p\in Y} N_p$. We claim that $\mathcal{N}$ is the desired approximate centroid set.

For a candidate solution $\mathcal{S}$, 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 $i$ be the smallest index such that $s \in B(p_i, \left(\frac{8z}{\varepsilon}\right)\text{dist}(p_i, \mathcal{A}))$. Pick $\tilde{s}$ to be the closest point to $s$ in $N_{p_i}$. If such a $i$ does not exist, pick $\tilde{s} = s_f$.

Now, let $p$ be a point such that $\text{cost}(p, \mathcal{S}) \le \left(\frac{10z}{\varepsilon}\right)^z\cdot\text{cost}(p, \mathcal{A})$. Let $s$ be the center serving $p$ in $\mathcal{S}$. Then, by construction of the $\tilde{\mathcal{S}}$, there is a center $\tilde{s}$ with $\text{dist}(s, \tilde{s}) \le \left(\frac{\varepsilon}{2z}\right)\text{dist}(p, \mathcal{A})$ and hence

$$\begin{aligned}\text{cost}(p, \tilde{\mathcal{S}}) &\le \text{cost}(p, \tilde{s}) \le (1+\varepsilon)\text{cost}(p, s) + (1+z/\varepsilon)^{z-1}\text{cost}(s, \tilde{s}) \\ &\le (1+\varepsilon)\text{cost}(p, \mathcal{S}) + (2z/\varepsilon)^{z-1}\left(\frac{\varepsilon}{2z}\right)^z\text{cost}(p, \mathcal{A}) \\ &\le (1+\varepsilon)\text{cost}(p, \mathcal{S}) + \varepsilon\text{cost}(p, \mathcal{A}). \qquad (26)\end{aligned}$$

To show the other direction, let $p$ be such that $\text{cost}(p, \tilde{\mathcal{S}}) \leq \left(\frac{10z}{\varepsilon}\right)^z \cdot \text{cost}(p, \mathcal{A})$. Let $\tilde{s}$ be the center closest to $p$ in $\tilde{\mathcal{S}}$, which is different to $s_f$ by property of $p$, and $s$ the corresponding center in $\mathcal{S}$. Let $i$ be the smallest index such that $s \in B(p_i, \left(\frac{8z}{\varepsilon}\right) \text{dist}(p_i, \mathcal{A}))$.

It must be that $\text{dist}(p, \mathcal{A}) \geq \text{dist}(p_i, \mathcal{A})$. Otherwise, $\tilde{s}$ would not be in $B(p, \left(\frac{10z}{\varepsilon}\right) \cdot \text{dist}(p, \mathcal{A}))$, by definition of $N_{p_i}$. Hence, $\text{dist}(s, \tilde{s}) \leq \frac{\varepsilon}{2z}\text{dist}(p_i, \mathcal{A}) \leq \frac{\varepsilon}{2z}\text{dist}(p, \mathcal{A})$ and so, by Lemma 1,

$$\begin{aligned}\text{cost}(p, \mathcal{S}) &\leq \text{cost}(p, s) \leq (1+\varepsilon)\text{cost}(p, \tilde{s}) + (1 + z/\varepsilon)^{z-1}\text{cost}(s, \tilde{s}) \\ &\leq (1+\varepsilon)\text{cost}(p, \tilde{\mathcal{S}}) + \varepsilon\text{cost}(p, \mathcal{A}) \qquad (27)\end{aligned}$$

We conclude using Equations 26 and 27. For a point $p$ such that $\text{cost}(p, \mathcal{S}) \leq \left(\frac{8z}{\varepsilon}\right)^z \cdot \text{cost}(p, \mathcal{A})$, Equation 26 gives $\text{cost}(p, \tilde{\mathcal{S}}) \leq (1+\varepsilon)\text{cost}(p, \mathcal{S}) + \varepsilon\text{cost}(p, \mathcal{A}) \leq \left(\frac{10z}{\varepsilon}\right)^z \cdot \text{cost}(p, \mathcal{A})$. Hence, Equation 27 gives $\text{cost}(p, \mathcal{S}) \leq (1+\varepsilon)\text{cost}(p, \tilde{\mathcal{S}}) + \varepsilon\text{cost}(p, \mathcal{A})$. Combining those two equations yields

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

The same property holds for points $p$ such that $\text{cost}(p, \tilde{\mathcal{S}}) \leq \left(\frac{8z}{\varepsilon}\right)^z \cdot \text{cost}(p, \mathcal{A})$. □

## 8.2 In 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 5.** *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 their exists a tree decomposition with maximum size bag $t + 1$.*

**Lemma 18.** *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 a set $\mathbb{C}$ of size poly$(|X|) \left(\frac{z}{\varepsilon}\right)^{O(t)}$ with the following property.*

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

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

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*

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

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

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

Lemma 18 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

$$O\left(\frac{\log^4 1/\varepsilon}{2^{O(z\log z)}\min(\varepsilon^3, \varepsilon^z)}\left(k\log |X_1| + kt\log 1/\varepsilon + \log(1/\pi)\right)\right).$$

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

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

Instead of using [BBH+20], one could apply our algorithm repeatedly as in Theorem 3.1 of [BJKW20], 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 $\log^*(x) = \log^* \log x$ for $x > 1$. The complexity of this repetition is therefore $\tilde{O}(nk)$, and the success probability $1 - 1/\pi$, as proven in [BJKW20]. □

For the proof of Lemma 18, we rely on the following structural lemma:

**Lemma 19** (Lemma 3.7 of [BBH+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. *$|\mathcal{T}| = poly(|X|)$,*
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 $A$ and $V \setminus (A \cup P_A)$[5].*

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 $P_A$.

[5]In the statement of [BBH+20], this item is slightly different. To recover our statement from theirs, take $P_A = A$ when $|A| = O(t)$.

We use this observation as follows: it would be enough if $\mathbb{C}$ contained a center at approximately the same distance of all points of $P_A$ than $s$. We make this formal in the following proof.

*Proof of Lemma 18.* Given a point $p \in V$ and a set $A \in \mathcal{T}$, we call a *distance tuple to* $A$ $\mathbf{d}_A(p) := (\text{dist}(p,x) \mid \forall x \in X \cap A) + (\text{dist}(p,x) \mid \forall x \in P_A)$. Let $q \in X$: the rounded distance tuple of $p$ with respect to $q$ is $\widetilde{\mathbf{d}_{A,q}}(p)$ defined as follows:

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

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

**Bounding the size of $\mathbb{C}$.** Fix some $A \in \mathcal{T}$. A rounded distance tuple to $A$ is made of $O(t)$ many distances. Each of them takes its value among $\frac{200z^4}{\varepsilon^4}$ possible real, 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 $\tilde{\mathcal{S}} \subseteq \mathbb{C}$. Let $A \in \mathcal{T}$ such that $s \in A$, and $q$ having the smallest $\text{cost}(q, \mathcal{A})$ value among points verifying $\text{cost}(p,s) \le \left(\frac{10z}{\varepsilon}\right)^z \text{cost}(p, \mathcal{A})$. $\tilde{s}$ 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 17, we first show that points close to $\mathcal{S}$ are close to $\tilde{\mathcal{S}}$. Showing the converse is done exactly the same way, by switching roles of $s$ and $\tilde{s}$.

Hence, let $p$ be a point served in $\mathcal{S}$ by some center $s$ such that $\text{cost}(p,s) \le \left(\frac{10z}{\varepsilon}\right)^z \text{cost}(p, \mathcal{A})$.

First, if $p \in X \cap A$, the choice of $\tilde{s}$ ensures that $\text{dist}(\tilde{s}, p) \le \text{dist}(s,p) + \frac{\varepsilon}{z}\text{dist}(x, \mathcal{A})$ and therefore $\text{cost}(p, \tilde{\mathcal{S}}) \le \text{cost}(p, \tilde{s}) \le (1+\varepsilon)\text{cost}(p,s) + (1+z/\varepsilon)^{z-1}\text{cost}(s, \tilde{s}) \le (1+\varepsilon)\text{cost}(p, \mathcal{S}) + \varepsilon\text{cost}(p, \mathcal{A})$.

When $p \notin X \cap A$, we distinguish two more subcases:

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

$$\begin{aligned}\text{dist}(p, \tilde{s}) &\le \text{dist}(p,p') + \text{dist}(p', \tilde{s}) \le \text{dist}(p,p') + \text{dist}(p', s) + \frac{\varepsilon}{z}\text{dist}(q, \mathcal{A}) \\ &\le \text{dist}(p,s) + \frac{\varepsilon}{z}\text{dist}(p, \mathcal{A}).\end{aligned}$$

  This implies that $\text{cost}(p, \tilde{\mathcal{S}}) \le (1+\varepsilon)\text{cost}(p, \mathcal{S}) + \varepsilon\text{cost}(p, \mathcal{A})$.
- 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 $\text{dist}(p,s)$. By choice of $\tilde{s}$, $\text{dist}(\tilde{s}, q) \le \text{dist}(q,s) + \frac{\varepsilon}{z}\text{dist}(q, \mathcal{A})$. Therefore, using

the properties of $q$, we get

$$\begin{aligned}
\mathrm{dist}(s,\tilde{s}) &\leq \mathrm{dist}(q,s) + \mathrm{dist}(q,\tilde{s}) \leq 2\mathrm{dist}(q,s) + \frac{\varepsilon}{z}\mathrm{dist}(p,\mathcal{A}) \\
&\leq \frac{20z}{\varepsilon} \cdot \mathrm{dist}(q,\mathcal{A}) + \frac{\varepsilon}{z}\mathrm{dist}(p,\mathcal{A}) \\
&\leq \frac{20z}{\varepsilon} \cdot \frac{\varepsilon^3}{200z^3} \cdot \mathrm{dist}(p,s) + \frac{\varepsilon}{z}\mathrm{dist}(p,\mathcal{A}) \\
&\leq \frac{20z}{\varepsilon} \cdot \frac{\varepsilon^3}{200z^3} \cdot \frac{10z}{\varepsilon}\mathrm{dist}(p,\mathcal{A}) + \frac{\varepsilon}{z}\mathrm{dist}(p,\mathcal{A}) \\
&\leq \frac{\varepsilon}{z}\mathrm{dist}(p,\mathcal{A}).
\end{aligned}$$

Hence, using Lemma 1, we conclude again that $\mathrm{cost}(p,\tilde{\mathcal{S}}) \leq \mathrm{cost}(p,\tilde{s}) \leq (1+\varepsilon)\mathrm{cost}(p,\mathcal{S}) + \varepsilon\mathrm{cost}(p,\mathcal{A})$.

Hence, in all possible cases,

$$\mathrm{cost}(p,\tilde{\mathcal{S}})(1+\varepsilon)\mathrm{cost}(p,\mathcal{S}) + \varepsilon\mathrm{cost}(p,\mathcal{A}).$$

Switching roles of $s$ and $\tilde{s}$ in the proof, one can show that for all $p$ such that $\mathrm{cost}(p,\tilde{\mathcal{S}}) \leq \left(\frac{8z}{\varepsilon}\right)^z \mathrm{cost}(p,\mathcal{A})$, then

$$\mathrm{cost}(p,\mathcal{S})(1+\varepsilon)\mathrm{cost}(p,\tilde{\mathcal{S}}) + \varepsilon\mathrm{cost}(p,\mathcal{A}).$$

As in the proof of Lemma 17, those two equations combined imply the lemma. □

### 8.3 In 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. Those 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 18.

**Lemma 20.** *Let $G = (V,E)$ be an edge-weighted graph that excludes a fixed minor, a set $X \subseteq V$ and two positive integers $k$ and $z$. Furthermore, let $\mathcal{A}$ be a solution of $(k,z)$-clustering of $X$.*

*There exists a set $\mathbb{C}$ of size $\exp(O(\log^2|X| + \log|X|/\varepsilon^4))$ with the following property.*

*For every set $\mathcal{S}$ of $k$ centers, there exists $\tilde{\mathcal{S}} \in \mathbb{C}^k$ such that for all points $p \in X$ such that cost$(p,\mathcal{S}) \leq \left(\frac{4z}{\varepsilon}\right)^z$ cost$(p,\mathcal{A})$,*

$$|cost(p,\mathcal{S}) - cost(p,\tilde{\mathcal{S}})| \leq \varepsilon^3\left(cost(p,\mathcal{S}) + cost(p,\mathcal{A})\right).$$

As for treewidth, this lemma implies the following corollary:

**Corollary 6.** *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*

$$O\left(\frac{\log^5 1/\varepsilon}{2^{O(z\log z)}\min(\varepsilon^3, \varepsilon^z)}\left(k\log^2 k\log(1/\varepsilon) + \frac{k\log k}{\varepsilon^4} + \log 1/\pi\right)\right)$$

The big picture is the same as for treewidth. Minor-free graphs have somewhat nice separators, that we can use to select centers. However, those separators do not have bounded size: they are instead made of a bounded number of shortest path, as described in the next structural lemma.

**Lemma 21** (Balanced Shortest Path Separator [AG06])**.** *Given a graph $G = (V, E)$ with positive weights on vertices and that excludes a fixed minor, there is a set of vertices $S \subseteq V$, such that*

1. *$S = P_1 \cup P_2 \cup \ldots$ where $P_i$ is a set of shortest paths in the graph formed by removing $\bigcup_{j<i} P_j$*
2. *$\sum_i |P_i| = O(1)$, where the hidden constant depends on the size of the excluded minor*
3. *the weight of every component in the graph formed by removing $S$ from $G$ is at most half the weight of $V$.*

Applying recursively that lemma on the graph $G$ with weight 1 for vertices in $X$ and 0 otherwise yields a recursive decomposition $\mathcal{R}$ (see [BJKW20] for more details about that decomposition).

The general sketch of the proof for $z = 1$ is as follows: in every leaf of $\mathcal{R}$, we would like to take a net, approximating well every center lying in the leaf. Unfortunately, small nets do not exist in minor-free graphs. To cope with that issue, we proceed slightly differently, inspired by [CPP19, BJKW20]: in short, we consider the boundary $B$ of the leaf, 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 by placing *landmarks* on it, and how to round distances from a point to a landmark.

We first picture how to discretize the boundary. Let $S$ be a cluster with center $s$ of the solution $\mathcal{S}$. Let $p \in S$ be a tiny or interesting point, and $R$ the smallest region of $\mathcal{R}$ containing both $y$ and $s$. Let $P_1, P_2, ..., P_m$ the set of shortest paths given by the application of Lemma 21 on $R$. Removing all those path separates $p$ and $s$: hence, the path $y \rightsquigarrow s$ must cross a path in some $P_i$. Let $i$ be the smallest index such that $p \rightsquigarrow s$ intersects a path $P \in P_i$. Let $x$ be the intersecting point. If there where a landmark $l$ close to $x$, we could write:

$$\begin{aligned}\text{dist}(p, s) = \text{dist}(p, x) + \text{dist}(x, s) &\approx \text{dist}(y, l) + \text{dist}(l, s)\\ &= \text{dist}(p, l) + \text{dist}(l, \tilde{s}) \geq \text{dist}(y, \tilde{s})\end{aligned}$$

where $\tilde{s} \in \mathbb{C}$ is the point realizing the same distance to the landmarks as $s$. Similarly, one can show $\text{dist}(p, \tilde{s}) \gtrapprox \text{dist}(p, s)$, so $\text{dist}(p, s) \approx \text{dist}(p, \tilde{s})$ which would be enough to conclude. We list some properties that will help us to find landmarks. First, $P$ is a shortest path in $R_j := R \setminus \cup_{j<i} P_i$. Additionally, $x$ is at distance at most $\text{dist}(y, s) \leq \text{dist}(y, \mathcal{A})/\varepsilon$ of $p$, since $p$ is of type tiny or

interesting. Hence, the landmarks could be a $\varepsilon\text{dist}(y, \mathcal{A})$-net of $P \cap B_{R_j}(y, \varepsilon^{-2}\text{dist}(y, \mathcal{A}))$: this would ensure that there is a landmark close to the $p$-to-$s$ path, allowing to write the previous line of equalities. Note as well that this net has constant size.

However, $\tilde{s}$ must be chosen consistently for every possible $p$, and so the landmarks must be the same for all $p \in S$. For this, focus first on the $p$ separated from $s$ by the path $P$, to construct a consistent set of landmarks. Among them, let $p_0$ with smallest $\text{dist}(p, \mathcal{A})$ value. Every other $p$ that intersect $P$ in $B_{R_j}(p_0, \varepsilon^{-2}\text{dist}(p_0, \mathcal{A}))$ is taken care of by landmarks defined for $p_0$. In particular, $p$ such that $\text{dist}(p, s) \leq \varepsilon^{-2}\text{dist}(p_0, \mathcal{A}) \leq \varepsilon^{-2}\text{dist}(p, \mathcal{A})$ are fine. In the case where $\text{dist}(p, s) > \varepsilon^{-2}\text{dist}(p_0, \mathcal{A})$, then we can argue that $\text{dist}(s, \tilde{s})$ is tiny compared to $\text{dist}(p, s)$: indeed, since $\tilde{s}$ has the same distances to landmarks than $s$, it must be that $d(p_0, \tilde{s}) \approx d(p_0, s)$. Hence, $\text{dist}(s, \tilde{s}) \leq \text{dist}(s, p_0) + \text{dist}(\tilde{s}, p_0) \lessapprox 2\varepsilon^{-2}\text{dist}(p_0, \mathcal{A}) \leq 2\varepsilon\text{dist}(p, s)$ and finally $|d(p, \tilde{s}) - \text{dist}(p, s)| \leq 2\varepsilon\text{dist}(p, s)$.

This argument works for all $p$ separated from $s$ by the same path. To ensure consistency over all $p$, we pick one $p_0$ for every possible such path, and choose $\tilde{s}$ to have the same distances as $s$ to all those $p_0$.

We make now the proof formal, and work for all powers $z$.

*Proof of Lemma 20.* Consider a recursive decomposition $\mathcal{R}$ of the graph, i.e. a tree where each tree node is a subset of $V$, and where a node is equal to the union of its children. In our case, $\mathcal{R}$ is constructed as follows. The root node is the whole vertex set. Then, inductively, apply Lemma 21, with weights 1 on vertices of $X$ and 0 otherwise, on a region $R$ represented by a node (simply called region in the following): the children of $R$ are

- the connected components of $R \setminus S$,
- let $S = P_1 \cup P_2 \cup ...$. Each path in $P_i$ is broken into maximal subpath each containing a point of $X$, and each subpath is added as a children of $R$.

Stop this induction when a region has weight at most 2 (i.e., there are less than 2 nodes from $X$ in it). Since the weight is divided by two at every level, the depth of that decomposition is $O(\log |X|)$.

We define sets of landmarks as follows. Consider a root-to-leaf path of the decomposition $R_1, ..., R_{\log |X|}$. Let $P_1^i, P_2^i...$ be the paths given by Lemma 21 on $R_i$. For each of those paths, choose a vertex $p_j^i \in R_i$ and let $\mathcal{L}_{i,j,p_j^i}$ be an $\frac{\epsilon}{z}\text{dist}(p_j^i, \mathcal{A})$-net of $P \cap B_{R_i^j}(p_j^i, \frac{90z^3}{\varepsilon^3}\text{dist}(p_j^i, \mathcal{A}))$, where $B_{R_i^j}(y, r)$ is the ball centered at $y$ of radius $r$ in the graph $R_i^j$, which is the graph induced by $R_i \setminus \bigcup_{j' \leq j} P_{j'}^i$.

We let $\mathbf{p}$ be any such sequence $p_1^1, p_2^1, ..., p_1^2, p_2^2, ...$, where $p_j^i \in R_i$ for all path $P_j^i$ dividing $R_i$. A set of landmarks $\mathcal{L}_{\mathbf{p}}$ is defined as:

$$\mathcal{L}_{\mathbf{p}} := \left(X \cap R_{\log |X|}\right) \bigcup_{i,j} \mathcal{L}_{i,j,p_j^i}.$$

We now describe how we round distances to landmarks, and argue that for each possible distance tuple, $\mathbb{C}$ contains a point having that distance tuple. Formally, given a point $p$ and its distance tuple $\mathbf{d}(p) = (\text{dist}(p, x) \mid \forall\ x \in X \cap R_{\log |X|}) + (\text{dist}(p, y) \mid y \in \mathcal{L}_{i,j,p_j^i}, \forall i, j)$, the *rounded distance tuple* $\tilde{\mathbf{d}}(p)$ is defined as follows :

- For $x \in X \cap A$, $\tilde{d}(p,x)$ is the multiple of $\frac{\varepsilon}{z}\text{dist}(x,\mathcal{A})$ smaller than $\frac{8z}{\varepsilon}\text{dist}(x,\mathcal{A})$ closest to $\text{dist}(p,x)$.
- For $y \in \mathcal{L}_{i,j,p^i_j}$, $\tilde{d}(p,y)$ is the multiple of $\frac{\varepsilon}{z} \cdot \text{cost}(p^i_j,\mathcal{A})$ smaller than $\frac{90z^3}{\varepsilon^3}\text{cost}(p^i_j,\mathcal{A})$ closest to $\text{dist}(p,y)$.

The set $\mathbb{C}$ is constructed as follows: for every root-to-leaf path and every sequence $\mathbf{p}$, for every rounded distance tuple $\{\tilde{\mathbf{d}}(p)\}$, add to $\mathbb{C}$ a point that realizes this rounded distance tuple.

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

**Size analysis.** $\mathbb{C}$ contains one point per rounded distance tuple. For a given set of landmarks $\mathcal{L}_\mathbf{p}$, there are $\left(\frac{90z^4}{\varepsilon^4}\right)^{|\mathcal{L}_\mathbf{p}|}$ possible rounded distances. Since the decomposition has depth $\log|X|$ and there is $O(1)$ paths for each region as prescribed by Lemma 21, there are $O(\log|X|)$ paths $P^i_j$; for each of them, there are $O(1/\varepsilon^4)$ landmarks (the size of the net). Hence, $|\mathcal{L}_\mathbf{p}| = O(\log(|X|)/\varepsilon^4)$.

The number of possible $\mathbf{p}$ is bounded by the same argument by $|X|^{O(\log|X|)}$, since there are $O(\log|X|)$ points of $X$ in $\mathbf{p}$.

Hence, the total size of $\mathbb{C}$ is at most $\exp\left(O(\log^2|X| + z^4\varepsilon^{-4}\log\varepsilon^{-1}\log|X|)\right)$.

**Error analysis.** We now show that for all solution $\mathcal{S}$, every center can be approximated by a point of $\mathbb{C}$. Let $S$ be some cluster of $\mathcal{S}$, with center $s$. We show how to find $\tilde{s} \in \mathbb{C}$ such that, for every $y \in X \cap S$ of type tiny or interesting, $|\text{cost}(y,s) - \text{cost}(y,\tilde{s})| \leq 3\varepsilon\left(\text{cost}(y,s) + \text{cost}(y,\mathcal{A})\right)$.

For this, let $R_1, ..., R_{\log|X|}$ be the path in $\mathcal{R}$ from the root to the leaf containing $s$, and $\{P^i_j\}$ be the paths given by Lemma 21. For a point $p \in R_i$, $y \notin R_{i+1}$, we say that the path $P^i_j$ separates $p$ and $s$ if the path $p \rightsquigarrow s$ is fully contained in $R^i_j$ and intersect $P^i_j$.

For every region $R_i$ and path $P^i_j$, let $p^i_j \in R_i$ be the point separated from $s$ by the path $P^i_j$ with smallest $\text{cost}(p^i_j,\mathcal{A})$ value. Let $\mathbf{p}$ be the tuple $(p^1_1, p^1_2, ..., p^2_1, p^2_2, ...)$, and $\tilde{s}$ be the point of $\mathbb{C}$ that has the same rounded distance tuple to $\mathcal{L}_\mathbf{p}$ than $s$. Let $\tilde{\mathcal{S}}$ be the solution constructed from $\mathcal{S}$ that way. We show now that $\tilde{\mathcal{S}}$ has the required properties. As in Lemma 17 and 18, we first show that points close to $\mathcal{S}$ are close to $\tilde{\mathcal{S}}$, and the converse can be shown exactly the same way, by switching roles of $s$ and $\tilde{s}$.

For this, let $p \in X$ such that $\text{cost}(p,\mathcal{S}) \leq \left(\frac{10z}{\varepsilon}\right)^z \text{cost}(p,\mathcal{A})$. We show that $\text{cost}(p,\tilde{\mathcal{S}}) \leq (1+\varepsilon)\text{cost}(p,\mathcal{S}) + \varepsilon\text{cost}(p,\mathcal{A})$. Let $s$ be the closest center of $\mathcal{S}$ to $p$.

Let $R_i$ be the smallest region containing the entire shortest-path between $p$ and $s$. If $R_i$ is a leaf, then $p \in \mathcal{L}_\mathbf{p}$ and since $\text{dist}(p,s) \leq \frac{4z}{\varepsilon}\text{dist}(p,\mathcal{A})$, the rounding of distances ensures that $\text{dist}(s,\tilde{s}) \leq \frac{\epsilon}{z}\text{dist}(p,\mathcal{A})$ and so the cost satisfies $\text{cost}(p,\tilde{s}) \leq (1+\varepsilon)\text{cost}(p,s) + \varepsilon\text{cost}(p,\mathcal{A})$. Otherwise, there exists a path $P^i_j$ separating $p$ and $s$. Note that, by choice of $P^i_j$, $\text{dist}(p,s) = \text{dist}_{R^i_j}(p,s)$. As explained in the sketch of proof, we need to distinguish two cases.

- If $\text{dist}_{R^i_j}(p,s) > \frac{8z^2}{\varepsilon^2} \cdot \text{dist}_{R^i_j}(p^i_j,s)$. Then we argue that $d(s,\tilde{s})$ is negligible. By choice of $\tilde{s}$, $s$ and $\tilde{s}$ have the same rounded distance to $p^i_j$, i.e. $\text{dist}_{R^i_j}(\tilde{s},p^i_j) \leq \text{dist}_{R^i_j}(s,p^i_j) + \frac{\varepsilon}{z} \cdot \text{dist}(p,\mathcal{A})$

and so:

$$\begin{aligned}\mathrm{dist}(s,\tilde{s}) &\le \mathrm{dist}_{R_j^i}(s,\tilde{s}) \le 2\mathrm{dist}_{R_j^i}(s,p_j^i) + \frac{\varepsilon}{z}\cdot \mathrm{dist}(p,\mathcal{A})\\ &\le \frac{\varepsilon^2}{4z^2}\mathrm{dist}(p,s) + \frac{\varepsilon}{z}\cdot\mathrm{dist}(p,\mathcal{A}) \le \frac{(1+\varepsilon)\varepsilon}{z}\cdot\mathrm{dist}(p,\mathcal{A})\end{aligned}$$

Hence, using the modified triangle inequality as above, we have :

$$\begin{aligned}\mathrm{cost}(p,\tilde{s}) &\le (1+\varepsilon(1+\varepsilon))\mathrm{cost}(p,s) + (1+{}^{z}/_{\varepsilon(1+\varepsilon)})^{z-1}\mathrm{cost}(s,\tilde{s})\\ &\le (1+\varepsilon(1+\varepsilon))\mathrm{cost}(p,s) + (1+\varepsilon)^2(\varepsilon/z)\mathrm{cost}(p,\mathcal{A})\\ &\le (1+2\varepsilon)\mathrm{cost}(p,s) + 2\varepsilon\mathrm{cost}(p,\mathcal{A}).\end{aligned}$$

- Otherwise, $\mathrm{dist}_{R_j^i}(p,s) \le \frac{8z^2}{\varepsilon^2}\cdot\mathrm{dist}_{R_j^i}(p_j^i,s)$ and we can make use of the landmarks. By definition of $j$, the path $y \rightsquigarrow s$ is entirely in $R_j^i$, and it crosses $P_j^i$ at some vertex $x$.

  First, it holds that $\mathrm{dist}_{R_j^i}(x,p_j^i) \le \mathrm{dist}_{R_j^i}(x,s) + \mathrm{dist}_{R_j^i}(s,p_j^i) \le \mathrm{dist}_{R_j^i}(p,s) + \mathrm{dist}_{R_j^i}(s,p_j^i) \le (1+\frac{8z^2}{\varepsilon^2})\mathrm{dist}_{R_j^i}(p_j^i,s) \le \frac{9z^2}{\varepsilon^2}\cdot\frac{10z}{\varepsilon}\cdot\mathrm{dist}(p_j^i,\mathcal{A})$, hence $x$ is in $P\cap B_{R_i^j}(p_j^i, \frac{90z^3}{\varepsilon^3}\mathrm{dist}(p_j^i,\mathcal{A}))$. By choice of landmarks, this implies that there is $l\in\mathcal{L}_{\mathbf{p}}$, with $\mathrm{dist}(x,l)\le\frac{\epsilon}{z}\mathrm{dist}(p_j^i,\mathcal{A})\le\varepsilon\mathrm{dist}(p,\mathcal{A})$. Furthermore,

$$\begin{aligned}\mathrm{dist}_{R_j^i}(s,l) &\le \mathrm{dist}_{R_j^i}(s,x) + \frac{\epsilon}{z}\cdot\mathrm{dist}(p_j^i,\mathcal{A}) \le \mathrm{dist}_{R_j^i}(p,s) + \frac{\epsilon}{z}\mathrm{dist}(p_j^i,\mathcal{A})\\ &\le \frac{8z^2}{\varepsilon^2}\cdot\mathrm{dist}_{R_j^i}(p_j^i,s) + \frac{\varepsilon}{z}\cdot\mathrm{dist}(p_j^i,\mathcal{A})\\ &\le (\frac{8z^2}{\varepsilon^2}\cdot\frac{10z}{\varepsilon} + \varepsilon/z)\mathrm{dist}(p_j^i,\mathcal{A}).\end{aligned}$$

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

$$\begin{aligned}\mathrm{dist}(p,\tilde{s}) &\le \mathrm{dist}_{R_j^i}(p,l) + \mathrm{dist}_{R_j^i}(l,\tilde{s})\\ &\le \mathrm{dist}_{R_j^i}(p,l) + \mathrm{dist}_{R_j^i}(l,s) + \frac{\epsilon}{z}\cdot\mathrm{dist}(p,\mathcal{A})\\ &\le \mathrm{dist}_{R_j^i}(p,x) + \mathrm{dist}_{R_j^i}(x,s) + \frac{3\epsilon}{z}\cdot\mathrm{dist}(p,\mathcal{A})\\ &= \mathrm{dist}_{R_j^i}(p,s) + \frac{3\epsilon}{z}\cdot\mathrm{dist}(p,\mathcal{A})\\ &= \mathrm{dist}(p,s) + \frac{3\epsilon}{z}\cdot\mathrm{dist}(p,\mathcal{A})\end{aligned}$$

  We can therefore conclude : $\mathrm{cost}(p,\tilde{\mathcal{S}}) \le (1+3\varepsilon)\mathrm{cost}(p,\mathcal{S}) + 3\varepsilon\mathrm{cost}(p,\mathcal{A})$.

Showing the converse, i.e., that for points $p\in X$ such that $\mathrm{cost}(p,\tilde{\mathcal{S}}) \le \left(\frac{10z}{\varepsilon}\right)^z\mathrm{cost}(p,\mathcal{A})$ it holds thatcost$(p,\mathcal{S}) \le (1+\varepsilon)\mathrm{cost}(p,\tilde{\mathcal{S}}) + \varepsilon\mathrm{cost}(p,\mathcal{A})$ is similar. As in Lemma 17, this concludes the proof.

□

# 9 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+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 6** (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 $f : \mathbb{R}^d \to \mathbb{R}^m$ 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, MMMR18, NN19], with Narayanan and Nelson [NN19] achieving an optimal target dimension of $O(\varepsilon^{-2}\log n)$, where $n$ is the number of points[6].

It was first observed by Becchetti et al. [BBC+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+19, CEM+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 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.

[6] See the paper by Larsen and Nelson for a matching lower bound [LN17]

**Corollary 7.** *There exists a coreset of size*

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

*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 coreset for $(k, z)$ clustering in an $\ell_p$ space to constructing a constructing a coreset for $(k, 2z)$ clustering in Euclidean space. Plugging in our framework into their reduction then yields the following corollary:

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

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

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

**Proposition 9.** *Suppose we have a (possibly weighted) point set $A$ in $\mathbb{R}^d$. Let $f : \mathbb{R}^d \to \mathbb{R}^m$ 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

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

and

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

Therefore,

$$(1 - \varepsilon) \cdot \operatorname{cost}(f(p), f(\mathcal{S})) \leq \operatorname{cost}(p, \mathcal{S}) \leq (1 + \varepsilon) \cdot \operatorname{cost}(f(p), f(\mathcal{S})). \tag{28}$$

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

$$\left|\sum_{p \in f(A)} w_p \cdot \operatorname{cost}(p, f(\mathcal{S})) - \sum_{q \in f(P)} w'_q \cdot \operatorname{cost}(q, f(\mathcal{S}))\right| \leq \varepsilon \cdot \sum_{p \in f(A)} w_p \cdot \operatorname{cost}(p, f(\mathcal{S})), \tag{29}$$

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 28 and 29

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

where the second inequality uses Equation 29 and the triangle inequality and the last inequality uses Equation 28. □

# 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+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+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+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.

[BJKW20] Vladimir Braverman, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for clustering in excluded-minor graphs and beyond. *CoRR*, abs/2004.07718, 2020. Accepted for publication at SODA'21.

[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+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.

[CPP19] Vincent Cohen-Addad, Marcin Pilipczuk, and Michal Pilipczuk. Efficient approximation schemes for uniform-cost clustering problems in planar graphs. In Michael A. Bender, Ola Svensson, and Grzegorz Herman, editors, *27th Annual European Symposium on Algorithms, ESA 2019, September 9-11, 2019, Munich/Garching, Germany*, volume 144 of *LIPIcs*, pages 33:1–33:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019.

[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.

[FGS+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, *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, pages 1416–1429. ACM, 2020.

[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. Nonlinear dimension reduction via outer bi-lipschitz extensions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 1088–1101, 2018.

[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+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 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 22.** *Let $(X, dist)$ 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 $\frac{\varepsilon}{z}\Delta_C$ and $\left(\frac{z}{\varepsilon}\right)^{2z}\Delta_C$, 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 $O(k \cdot 2^{O(z)}\frac{\log^3(1/\varepsilon)}{\varepsilon^2}\left(k\log k + k\log|\mathbb{C}| + \log(1/\pi)\right)$ such that, with probability $1 - 1/\pi$, for any set $\mathcal{S}$ of $k$ centers,*

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

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 $[\frac{\varepsilon}{z}\Delta_C \cdot 2^j, \frac{\varepsilon}{z}\Delta_C \cdot 2^{j+1}]$, for $j \in \{1, \dots 3z\log(z/\varepsilon)\}$.

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. $\ell \leq j \cdot \log \varepsilon$, 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. $\ell \geq j + \log(4z/\varepsilon)$, 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 23.** *It holds that, for any $R_{i,j}$ and for all solutions $S$ with at least one non-empty huge group $I_{i,j,\ell}$*

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

*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 $q \in R_{i,j}$: $\text{cost}(p,q) \leq (2 \cdot \varepsilon \cdot 2^{j+1})^z \leq 4^z \cdot \varepsilon^z \cdot 2^{(\ell - \log(4z/\varepsilon))z} \leq (\varepsilon/z)^z \cdot \text{cost}(p, \mathcal{S})$. By Lemma 1, we have therefore for any point $q \in R_{i,j}$

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

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 $I_{i,j,\ell}$. □

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 24.** *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 $I_{i,j,\ell}$:*

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

*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

$$|R_{i,j}| \cdot (\varepsilon \cdot 2^\ell)^z \leq 4^z (\text{cost}(R_{i,j}, \mathcal{S}) + \text{cost}(R_{i,j}, \mathcal{A})) \tag{30}$$

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

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{p \in I_{i,j,\ell} \cap P} \text{cost}(p, \mathcal{S}) \cdot |R_{i,j}|\right)^2\right] &= \mathbb{E}\left[\sum_{p \in I_{i,j,\ell} \cap P} \text{cost}(p, \mathcal{S})^2 \cdot |R_{i,j}|^2\right] \text{ since } |P| = 1 \\
&= \sum_{p \in I_{i,j,\ell} \cap P} \text{cost}(p, \mathcal{S})^2 \cdot |R_{i,j}| \leq |R_{i,j}| \cdot |I_{i,j,\ell}| \cdot (\varepsilon 2^\ell)^{2z} 4^z \\
&\leq \text{cost}(I_{i,j,\ell}, \mathcal{S}) \cdot (\text{cost}(R_{i,j}, \mathcal{S}) + \text{cost}(R_{i,j}, \mathcal{A})) \cdot 16^z \qquad (31)
\end{aligned}$$

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 30.

Furthermore, by the same reasoning and again using Equation 30, we have the upper bound $M$ on the (weighted) cost in $\mathcal{S}$ of every sampled point in every ring:

$$M \leq (\varepsilon \cdot 2^{\ell+1})^z \cdot |R_{i,j}| \leq (\text{cost}(R_{i,j}, \mathcal{S}) + \text{cost}(R_{i,j}, \mathcal{A})) \cdot 8^z \tag{32}$$

Applying Bernstein's inequality and Equations 31 and 32, we now have

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

where the last line uses $\mathrm{cost}(I_{i,j,\ell},\mathcal{S})\leq\mathrm{cost}(R_{i,j},\mathcal{S})$. 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 22.* As in the proof of Lemma 2, we decompose $|\mathrm{cost}(\mathcal{S})-\mathrm{cost}(\Omega,\mathcal{S})|$ 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 $\mathrm{dist}(p,\mathcal{S})\leq\frac{8z}{\varepsilon}\cdot\mathrm{dist}(p,\mathcal{A})$ or $\mathrm{dist}(p,\tilde{\mathcal{S}})\leq\frac{8z}{\varepsilon}\cdot\mathrm{dist}(p,\mathcal{A})$ , we have $|\mathrm{cost}(p,\mathcal{S})-\mathrm{cost}(p,\tilde{\mathcal{S}})|\leq\varepsilon(\mathrm{cost}(p,\mathcal{S})+\mathrm{cost}(p,\mathcal{A}))$.

Our first step is to deal with points that have $\mathrm{dist}(p,\mathcal{S})>\frac{8z}{\varepsilon}\cdot\mathrm{dist}(p,\mathcal{A})$, using Lemma 23. All other points have distance well approximated by $\tilde{\mathcal{S}}$. Then, we can apply Lemma 5 and Lemma 24 to $L_{\tilde{\mathcal{S}}}$, since all points in $L_{\tilde{\mathcal{S}}}$ have $\mathrm{dist}(p,\tilde{\mathcal{S}})\leq\frac{4z}{\varepsilon}\cdot\mathrm{dist}(p,\mathcal{A})$, and so $\mathrm{dist}(p,\mathcal{S})\leq\frac{8z}{\varepsilon}\cdot\mathrm{dist}(p,\mathcal{A})$ and were not removed by the previous step. Remaining points are those which have $\mathrm{dist}(p,\tilde{\mathcal{S}})>\frac{4z}{\varepsilon}\cdot\mathrm{dist}(p,\mathcal{A})$ and $\mathrm{dist}(p,\mathcal{S})\leq\frac{8z}{\varepsilon}\cdot\mathrm{dist}(p,\mathcal{A})$, 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 $\varepsilon^{-z}$.