Towards Optimal Coreset Construction for \((k, z)\)-CLUSTERING: 
Breaking the Quadratic Dependency on \(k\)

Lingxiao Huang\(^*\) \hspace{1cm} Jian Li\(^†\) \hspace{1cm} Xuan Wu\(^‡\)

Huawei TCS Lab → Nanjing University \hspace{1cm} Tsinghua University \hspace{1cm} Huawei TCS Lab

November 28, 2022

Abstract

Constructing small-sized coresets for various clustering problems has attracted significant attention recently. We provide efficient coreset construction algorithms for \((k, z)\)-CLUSTERING with improved coreset sizes in several metric spaces. In particular, we provide an \(\tilde{O}(k^{(2z+2)/(z+2)}\epsilon^{-2})\)-sized coreset for \((k, z)\)-CLUSTERING for all \(z \geq 1\) in Euclidean space, improving upon the best known \(\tilde{O}(k^2\epsilon^{-2})\) size upper bound [Cohen-Addad, Larsen, Saulpic, Schwiegelshohn. STOC’22], breaking the quadratic dependency on \(k\) for the first time (when \(k \leq \epsilon^{-1}\)). For example, our coreset size for Euclidean \(k\)-MEDIAN is \(\tilde{O}(k^{4/3}\epsilon^{-2})\), improving the best known result \(\tilde{O}(\min\{k^2\epsilon^{-2}, k\epsilon^{-3}\})\) by a factor \(k^{2/3}\) when \(k \leq \epsilon^{-1}\); for Euclidean \(k\)-MEANS, our coreset size is \(\tilde{O}(k^{3/2}\epsilon^{-2})\), improving the best known result \(\tilde{O}(\min\{k^2\epsilon^{-2}, k\epsilon^{-4}\})\) by a factor \(k^{1/2}\) when \(k \leq \epsilon^{-2}\). We also obtain optimal or improved coreset sizes for general metric space, metric space with bounded doubling dimension, and shortest path metric when the underlying graph has bounded treewidth, for all \(z \geq 1\). Our algorithm largely follows the framework developed by Cohen-Addad et al. with some minor but useful changes. Our technical contribution mainly lies in the analysis. An important improvement in our analysis is a new notion of \(\alpha\)-covering of distance vectors with a novel error metric, which allows us to provide a tighter variance bound. Cohen-Addad et al. explicitly mentioned that the variance bound they can obtain is tight in their analysis framework, which is the main obstacle to improving their \(O(k^2\epsilon^{-2})\) bound. Another useful technical ingredient is terminal embedding with additive errors, for bounding the covering number in the Euclidean case.

\(^*\)Email: huanglingxiao1990@126.com  
\(^†\)Email: lapordge@gmail.com  
\(^‡\)Email: wu3412790@gmail.com
1 Introduction

We study the problem of constructing small-sized coresets for the classic \((k, z)\)-Clustering problem, where \(z \geq 1\) is a given constant. The \((k, z)\)-Clustering problem is defined as follows.

\((k, z)\)-Clustering. In the \((k, z)\)-Clustering problem, the input consists of a metric space \((X, d)\), where \(X\) is a ground set (continuous or discrete) and \(d : X \times X \to \mathbb{R}_{\geq 0}\) is a distance function, and a dataset \(P \subseteq X\) of \(n\) points. The goal is to find a set \(C \subseteq X\) of \(k\) points, called center set, that minimizes the objective function

\[
\text{cost}_z(P, C) := \sum_{x \in P} d^z(x, C),
\]

where \(d(x, C) := \min\{d(x, c) : c \in C\}\) is the distance from \(x\) to center set \(C\) and \(d^z\) denotes the distance raised to power \(z \geq 1\).

This formulation captures several classical clustering problems, including the well studied \(k\)-Median as \(z = 1\) and \(k\)-Means as \(z = 2\). We also study several metrics \((X, d)\), including Euclidean metric \(X = \mathbb{R}^d\), metric with bounded doubling dimension and shortest path metric defined over a finite graph. The \((k, z)\)-Clustering problem has numerous applications in a variety of domains, including data analysis, approximation algorithms, unsupervised learning and computational geometry [31, 37, 1, 11].

Coresets. Motivated by the ever increasing volume of data, a powerful data-reduction technique, called coresets, has been developed for harnessing large datasets for various problems [21, 19, 20]. Roughly speaking, for an optimization problem, a coreset is a small-sized subset of (weighted) data points that can be used to compute an approximation of the optimization objective for every possible solution. In the context of \((k, z)\)-Clustering, we use \(X^k\) to denote the collection of all \(k\)-center sets in \(X\). A coreset for \((k, z)\)-Clustering is formally defined as follows.

**Definition 1.1 (Coreset [29, 19])** Given a metric space \((X, d)\) together with a dataset \(P \subseteq X\) of \(n\) points, an \(\varepsilon\)-coreset for the \((k, z)\)-Clustering problem is a weighted subset \(S \subseteq P\) with weight \(w : S \to \mathbb{R}_{\geq 0}\), such that

\[
\forall C \in X^k, \quad \sum_{x \in S} w(x) \cdot d^z(x, C) \in (1 \pm \varepsilon) \cdot \text{cost}_z(P, C). \tag{2}
\]

From the above definition, one can see that if we run any existing approximation or exact algorithm on the coreset instead of the full dataset, the resulting solution provides almost the same performance guarantee in terms of the clustering objective, but the running time can be much smaller. Hence, constructing small-sized coresets for \((k, z)\)-Clustering has been an important research topic and studied extensively in the literature, for various metric spaces including Euclidean metrics [19, 20, 27, 17, 14], doubling metrics [22, 17, 14], shortest path metrics in graphs [2, 7, 17] and general metrics [19, 17, 14].

Despite the substantial effort, there is still a gap between the current best known upper and lower bounds for most metric spaces studied in the literature. Now we briefly mention some of existing results. Please refer to Table 1 for the best known upper and lower bounds, and [14, Table 1] for many earlier results. For example, Cohen-Addad et al. [14] recently obtained a lower bound of \(\Omega(k \varepsilon^{-2})\) for Euclidean \((k, z)\)-Clustering, and the best upper bound they can achieve is
Table 1: Comparison of the state-of-the-art coreset sizes and our results for \((k, z)\)-Clustering. We assume \(z \geq 1\) is constant and ignore \(2^{O(z)}\) or \(z^{O(z)}\) factors in the coreset size. “ddim” denotes the doubling dimension. Our results for doubling metrics and general discrete metrics are nearly optimal.

\[
\tilde{O}(\min\{k\varepsilon^{-3}, k^2\varepsilon^{-2}\}) \quad \text{for Euclidean } k\text{-MEDIAN.} \quad \text{Given that } \varepsilon^{-2} \text{ is likely to be the right dependency on } \varepsilon, \text{ a natural question is whether the quadratic dependency on } k \text{ is necessary (when } k \leq \varepsilon^{-1}). \quad \text{Other metric include doubling metrics, shortest path metrics in graphs, and general metrics for } z \geq 3 \quad \text{[17, 14]}, \text{known upper bounds all have such quadratic dependency on } k. \text{In this paper, we focus on the regime } k \leq \varepsilon^{-1} \text{ and study following natural question:}
\]

**Problem 1** Is it possible to improve the quadratic dependency on \(k\) in the upper bounds of the coreset sizes for Euclidean metrics, doubling metrics, shortest path metrics in graphs, and general metrics? Furthermore, can we reduce the \(k^2\) factor in the known upper bounds to near-linear in \(k\), thus achieving nearly optimal coreset sizes that match the lower bounds?

### 1.1 Our Contributions

In this paper, we address Problem 1. We adopt the existing importance sampling-based coreset framework (Algorithm 1), proposed by [17]. Our main contribution is a unified and tighter analysis that leads to improved coreset size upper bounds for several metrics (also summarized in Table 1). Throughout this paper, we assume there exists an oracle that answers \(d(p, q)\) in \(O(1)\) time for any \(p, q \in X\).

**Theorem 1.2 (Euclidean metrics; see also Theorem 4.1)** Given a finite set \(P \subseteq \mathbb{R}^d\) of \(n\) points, there exists a randomized algorithm that constructs an \(\varepsilon\)-coreset of \(P\) of size \(2^{O(z)} \cdot \tilde{O}(k^{k^2/2} \varepsilon^{-2})\) for \((k, z)\)-Clustering. Given an \(O(1)\)-approximate solution (the center set) for \(P\), the algorithm runs in \(O(nk)\) time.

Consequently, our coreset size for Euclidean \(k\text{-MEDIAN} is \(\tilde{O}(k^{k^2/2} \varepsilon^{-2})\), which improves the prior result \(\tilde{O}(\min\{k^2\varepsilon^{-2}, k^2\varepsilon^{-3}\})\) [14] by a factor \(k^{2/3}\) when \(k \leq \varepsilon^{-1}\). Our coreset size for Euclidean \(k\text{-MEANS}\) is \(\tilde{O}(k^{k^2/2} \varepsilon^{-2})\), which improves the prior result \(\tilde{O}(\min\{k^2\varepsilon^{-2}, k^2\varepsilon^{-4}\})\) [14] by a factor \(k^{1/2}\) when \(k \leq \varepsilon^{-2}\). For general \(z \geq 1\), we are the first result that improves the \(k^2\) upper bound in the coreset size. Further closing the gap between our coreset size and the lower bound \(\Omega(k\varepsilon^{-2})\) [14] is interesting.

**Theorem 1.3 (Doubling metrics; see also Theorem 5.1)** Suppose the doubling dimension of the metric space \((X, d)\) is \(\text{ddim}\). Given a dataset \(P \subseteq X\) of \(n\) points, there exists a randomized algorithm that constructs an \(\varepsilon\)-coreset of \(P\) of size \(2^{O(z)} \cdot \tilde{O}(k \cdot \text{ddim} \cdot \varepsilon^{-2})\) for \((k, z)\)-Clustering. Provided an \(O(1)\)-approximation \(A^* \in X^k\) of \(P\) for \((k, z)\)-Clustering, the algorithm runs in \(O(nk)\) time.
Our coreset size in doubling metrics matches the lower bound \( \Omega(k \cdot \text{ddim} \cdot \varepsilon^{-2}) \) \([14]\) up to logarithmic factor, and hence is nearly optimal. Compared to \([17]\), our coreset size saves a \( \min\{k, 1 + \varepsilon^{-z+2}\} \) factor. Specifically, when \( z > 2 \), the improvement is non-trivial.

For a discrete metric space \((\mathcal{X}, d)\), the doubling dimension is at most \( O(\log |\mathcal{X}|) \). By the above theorem, we directly have the following corollary for an arbitrary discrete metric space.

**Corollary 1.4 (General discrete metrics; see also Corollary 5.3)** Let \((\mathcal{X}, d)\) be a discrete metric space. Given a dataset \( P \subseteq \mathcal{X} \) of \( n \) points, there exists a randomized algorithm that constructs an \( \varepsilon \)-coreset of \( P \) of size \( 2^{O(z)} \cdot \tilde{O}(k \cdot \log |\mathcal{X}| \cdot \varepsilon^{-2}) \) and runs in \( O(nk) \) time, provided an \( O(1) \)-approximate solution for \((k, z)\)-CLUSTERING.

Similar to the doubling metrics, our coreset size in general discrete metrics matches the lower bound \( \Omega(k \cdot \log |\mathcal{X}| \cdot \varepsilon^{-2}) \) \([14]\) up to logarithmic factor, and hence is nearly optimal. Compared to \([17]\), our coreset size also saves a \( \min\{k, 1 + \varepsilon^{-z+2}\} \) factor, which is nontrivial for any \( z > 2 \).

**Theorem 1.5 (Shortest path metric on graphs with bounded treewidth; see also Theorem 6.1)** Let \( G = (\mathcal{X}, E) \) be an edge-weighted graph with treewidth at most \( tw \geq 1 \), \((\mathcal{X}, d)\) be the shortest path metric on \( G \). Given a dataset \( P \subseteq \mathcal{X} \) of \( n \) points, there exists a randomized algorithm that constructs an \( \varepsilon \)-coreset of \( P \) of size \( 2^{O(z)} \cdot \tilde{O}(ktw \varepsilon^{-2}) \) for \((k, z)\)-CLUSTERING. If we have already computed all-pairs shortest distances, and are given an \( O(1) \)-approximation of \( P \), our algorithm runs in \( O(nk) \) time.

For shortest path metrics in graphs with treewidth \( tw \geq 1 \), our coreset size again saves a \( \min\{k, 1 + \varepsilon^{-z+2}\} \) factor compared to \([17]\). The best known lower bound is \( \Omega(ktw \varepsilon^{-1}) \) \([2]\). It is an interesting open question to narrow the gap further.

### 1.2 Technical Overview

For convenience, we provide an overview of our techniques for \( k \)-MEDIAN (\( z = 1 \)). We use cost as a shorthand notation of cost\(_1\) in this section. Extension to general \( z \geq 1 \) is standard via the relaxed triangle inequality (Lemma D.3).

**The importance sampling-based framework in [17].** Importance sampling is an important technique for constructing coresets, initially developed in \([29, 19]\). Our results are based on a more recent importance sampling framework (Algorithm 1), developed in \([17]\), which we briefly review now. We first compute an \( O(1) \)-approximate solution \( A^* \in \mathcal{X}^k \) for \( k \)-MEDIAN, and partition the point set \( P \) into \( P_1, \ldots, P_k \) according to their distances to \( A^* \). Then adopting the idea of \([8]\), we decompose each \( P_i \) into a set of rings (Definition 2.1), and then group rings with similar costs to \( A^* \) together (Definition 2.2). Now we have a collection \( \mathcal{G} \) of groups with \(|\mathcal{G}| = \tilde{O}(1) \) (Lemma 2.4), where each group \( G \in \mathcal{G} \) consists of at most \( k \) rings, each from a different partition \( P_i \), with similar costs. The key step is to use importance sampling to get a small weighted subset \( S_G \) from each group \( G \). Let \( |S_G| = \Gamma_G \) be the size of \( S_G \). \( \Gamma_G \) should be large enough such that \( \sum_{p \in S_G} w(p) \cdot d(p, C) \) is a good approximation of \( \text{cost}(G, C) \) for all center sets \( C \). We use \( \text{cost}(G, C + A^*) \) as a shorthand notation for \( \text{cost}(G, C) + \text{cost}(G, A^*) \). More precisely, the following uniform convergence over all possible center sets should hold:

\[
\sup_{C \in \mathcal{X}^k} \mathbb{E}_{S_G} \left[ \frac{1}{\text{cost}(G, C + A^*)} \cdot \sum_{p \in S_G} w(p) \cdot d(p, C) - \text{cost}(G, C) \right] \leq \varepsilon.
\] (3)
Analysis via chaining: relating $\Gamma_G$ with variance and covering number. To show the uniform convergence bound (3), a standard idea is to discretize the space $\mathcal{X}^k$ and apply union bound over the finite discretization. Instead of discretizing $\mathcal{X}^k$ directly, we construct a certain $\varepsilon$-covering $V \subset \mathbb{R}^{|G|}$ for tuples $u^C = (d(p,C))_{p \in G}$ over all center sets $C \in \mathcal{X}^k$. Then, one can show the convergence guarantee for each $v \in V$ using standard concentration inequality, and then apply the union bound over all tuples in $V$. For discretization, [14] roughly showed that it suffices to consider those tuples $u^C$ such that for any $p \in G$, $u^C_p = d(p,C) \leq \varepsilon^{-1} \cdot d(p,A^*)$, which constrains the range of $V$ and leads to a finite covering number $|V|$. For a specific tuple $v = u^C \in V$ ($C \in \mathcal{X}^k$), we can apply the Chernoff bound to show that $\text{Var}_C \cdot \varepsilon^{-2}$ samples are enough to ensure that

$$\frac{|\sum_{p \in S_G} w(p) \cdot u_p - \|v\|_1|}{\text{cost}(G, C + A^*)} = \frac{|\sum_{p \in S_G} w(p) \cdot d(p, C) - \text{cost}(G, C)|}{\text{cost}(G, C + A^*)} \leq \varepsilon$$

holds with high probability, where $\text{Var}_C$ is the variance of the estimation error (i.e., the LHS) and is upper bounded by $O(1)$. Then by the union bound, we need $\Gamma_G = \log |V| \cdot \varepsilon^{-2}$ samples to ensure that Inequality (3) holds. Unfortunately, $\log |V|$ is usually much larger than $k$. For instance, $\log |V| = k \cdot \min \{\varepsilon^{-2}, d\}$ in Euclidean metrics, which leads to an upper bound $\Gamma_G = k\varepsilon^{-2} \cdot \min \{\varepsilon^{-2}, d\}$.

To further improve on this idea, one can observe that the uniform upper bound for $\text{Var}_C \leq O(1)$ may not be tight for most of $C$s. This observation motivates [14] to introduce a chaining argument in Euclidean metrics $\mathbb{R}^d$. Roughly speaking, they construct $2^{-h}$-coverings $V_h \subset \mathbb{R}^{|G|}$ at different scales $h \geq 0$. This allows them to write every tuple $u^C$ as a telescoping sum of its covering vectors $u^{C,h} \in V_h$ at different scales, namely

$$u^C = \sum_{h=0}^{\infty} (u^{C,h+1} - u^{C,h}),$$

where $u^{C,h} \in V_h$ is the closest tuple of $u^C$. By this point of view, [14] separately upper bounds

$$\mathbb{E}_{S_G} \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}(G, C + A^*)} \sum_{p \in S_G} w(p) \cdot (u^{C,h+1}_p - u^{C,h}_p) - (\|u^{C,h+1}_p - u^{C,h}_p\|_1) \right]$$

for each $h \geq 0$, and use their sum as an upper bound of the total estimation error. Compared to $\|u^C\|_1$, the scale of $\|\|u^{C,h+1} - u^{C,h}\|_1|$ roughly reduces by a multiplicative factor $2^h$, which leads to a variance $\text{Var}_{C,h} \approx 2^{-2h} \cdot \min \{\varepsilon^{-1}, k\}$.

Since there are $|V_{h+1}| \times |V_h| \leq \exp(O(k \cdot 2^h))$ many different possible pairs $(u^{C,h+1}, u^{C,h})$, [14] showed that the required sample number $\Gamma_G$ (ignoring polylog factor) is at most

$$\varepsilon^{-2} \cdot \text{Var}_{C,h} \cdot \log(|V_{h+1}| \times |V_h|) \approx k\varepsilon^{-2} \cdot \min \{\varepsilon^{-1}, k\}.$$ 

Overall, we can see that $\Gamma_G$ is mainly determined by the design of coverings, which decides both the covering number and the variance upper bound. The unexpected term $\min \{\varepsilon^{-1}, k\}$ is due to the upper bound for $\text{Var}_{C,h}$. As admitted in [14], “bounding the variance in this setting is highly nontrivial and requires a number of new ideas. The lower variance we could show for estimating $\|u^C\|_1$ is only of the order $\min(\varepsilon, k), \ldots$, and this bound on the variance is tight.” To improve the coreset size, [14] mentioned that “further ideas will be necessary”, specifically, on the construction of coverings.

1The additional term $\min \{\varepsilon^{-1}, k\}$ in the variance bound appears because [14] considered different scales $2^{-h}$ instead of only $\varepsilon$-coverings, which renders the variance reduction argument for $\varepsilon$-coverings in [17] inapplicable.
**Tighten the variance: a relative covering error.** In this paper, we first extend the chaining argument in [14] from Euclidean metrics to general metric space $\mathcal{X}$, and obtain a unified theorem that relates the size of coreset to the number of samples $\Gamma_G$ for a group (see Theorem 3.5). Our main improvement is a new covering which is motivated by the following observation: for each $i \in [k]$ and $p \in P_i \cap G$, and any center set $C \in \mathcal{X}^k$,

$$d(p, C) = d(p, C) - d(a_i^*, C) + d(a_i^*, C),$$

and the fact that $\sum_{q \in P_i \cap S_G} w(q) \cdot d(a_i^*, C) \approx \sum_{q \in P_i \cap G} d(a_i^*, C)$ always holds (Lemma 3.11). Thus, it suffices to upper bound the estimation error induced by $y^C = (d(p, C) - d(a_i^*, C))_{i,p}$ (Lemma 3.12). More precisely, we need to following error bound:

$$\mathbb{E}_{S_G} \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}(G,C + A^*)} \left| \sum_{p \in S_G} w(p) \cdot y^C_p - \|y^C\|_1 \right| \right] \leq \varepsilon.$$ 

Subsequently, we apply the chaining argument for $y^C$ and construct coverings $V_h$s for tuples $y^C$ instead of $u^C$. The benefit of introducing $y^C$ is that $|d(p, C) - d(a_i^*, C)| \leq d(p, a_i^*)$ by triangle inequality, which can be much smaller than $d(p, C)$. Moreover, considering the space of $y^C$ enables us to design a tighter relative covering error (see Definition 3.2). Roughly, we say $v \in \mathbb{R}^{|G|}$ is an $\alpha$-covering w.r.t. $C$ if for any $p \in G$,

$$|d(p, C) - d(a_i^*, C) - v_p| \leq \alpha \cdot \text{err}(p, C),$$

where $\text{err}(p, C)$ is the relative covering error of $p$ to $C$ defined as follows

$$\text{err}(p, C) := (\sqrt{d(p, C) \cdot d(p, A^*)} + d(p, A^*) \cdot \sqrt{\frac{\text{cost}(G, C + A^*)}{\text{cost}(G, A^*)}}).$$

Compared to the covering error defined in [14] (which is $d(p, C) + d(p, A^*)$), the new $\text{err}(p, C)$ is smaller and leads to a smaller variance upper bound (Lemma 3.14) for the chaining argument. Specifically, our variance $\text{Var}_{C,h}$ is exactly proportional to $2^{2h}$ and avoid the additional term $\min \{ \varepsilon^{-1}, k \}$ that appears in the bound in [14], and our variance bound is almost tight for all center sets $C$ (Remark 3.15). Similar ideas for reducing the variance have been also applied in existing coreset literature, e.g., [19, 17].

The remaining issue is to bound the covering number $|V_h|$s with respect to the new relative covering error $\text{err}(p, C)$.

**Bound the covering number in Euclidean metric.** We first briefly recall the approach in [14] for upper bounding $|V_h|$ induced by their covering error $\text{err}(p, C) = d(p, C) + d(p, A^*)$. Their key idea is to utilize a $2^{-h}$-terminal embedding [33, 35] on $G$ with target dimension being $m = \tilde{O}(2^{2h})$, and then construct an $\ell_\infty$-net in the embedded space. As a consequence, they can construct a covering $V_h$ of size $\exp(\tilde{O}(km)) = \exp(\tilde{O}(k \cdot 2^{2h}))$. However, terminal embedding introduces a multiplicative error $2^{-h} \cdot d(p, C)$, which is larger than our relative covering error $2^{-h} \cdot \text{err}(p, C)$. Hence, directly following their idea to ensure an error at most $2^{-h} \cdot \text{err}(p, C)$, the target dimension of terminal embedding needs to be as large as $\tilde{O}(2^{2h} \cdot \min \{ k, \varepsilon^{-1} \})$, which results in the same coreset size as in [14]. Thus, further improvement requires additional ideas.
For ease of exposition, we consider the following special case here: fix a sub-collection \( B \subset [k] \) such that \( P_i \cap G \neq \emptyset \) for all \( i \in [k] \), i.e., there is a ring of \( P_i \) contained in group \( G \), and let \( G_B = \bigcup_{i \in B} P_i \cap G \) be the collection of these rings. We only show how to construct \( 2^{-h} \)-coverings \( V \subset \mathbb{R}^{[S_G \cap G_B]} \) of \( G_B \) for those center sets \( C \in \mathcal{X}^k \) with \( d(p, C) \approx \ell \cdot d(p, A^*) \) for all \( p \in S_G \cap G_B \) \( (1 \leq \ell \leq \epsilon^{-1} \) is some fixed number). The complete construction of \( V_h \) for all \( C \in \mathcal{X}^k \) is via a hierarchical decomposition of rings in \( G \) w.r.t. \( C \) (Definition 4.9 and Lemma 4.11). We first note that for every \( p \in G_B \), \( \text{err}(p, C) \geq d(p, C)/\sqrt{\ell} \). Hence, using the same approach as in [14], we can perform a \( 2^{-h}/\sqrt{\ell} \)-terminal embedding on \( G_B \) with target dimension \( m = \tilde{O}(2^{2h} \cdot \ell) \). The resulting error is as desired, say \( 2^{-h} \cdot d(p, C)/\sqrt{\ell} \leq 2^{-h} \cdot \text{err}(p, C) \).

On the other hand, we introduce a novel dimension reduction notion, called additive terminal embedding (see Definition 4.5), which embeds a ring of radius \( r > 0 \) to low dimensional spaces, the required distance distortion should be an additive error \( 2^{-h}r \) (proportional to the radius), instead of a multiplicative error in ordinary terminal embedding. Using the approach developed in [33, 35], we can prove that the target dimension of additive terminal embedding can be bounded by \( \tilde{O}(2^{2h}) \) (Theorem 4.6). Therefore, for each ring in \( G_B \), we first perform a \( \sqrt{|B|}k^{-1}\ell 2^{-h} \)-additive terminal embedding with target dimension \( \tilde{O}(|B|^{-1}k\ell^{-2}2^{2h}) \). By a direct product of these additive terminal embeddings for all rings in \( B \), we obtain a “combined target dimension” \( m = \tilde{O}(k\ell^{-2}2^{2h}) \). It remains to check the induced error, say

\[
\sqrt{|B|}k^{-1}\ell 2^{-h} \cdot d(p, A^*) \leq O(2^{-h}) \cdot \text{err}(p, C),
\]

which holds since \( d(p, C) \approx \ell \cdot d(p, A^*) \).

Taking the minimum of the above two target dimensions, we conclude that the target dimension is at most \( \tilde{O}(k^{1/3}2^{2h}) \), which leads to a covering number \( \exp(\tilde{O}(km)) = \exp(\tilde{O}(k^{1/3}2^{2h})) \). Consequently, the required sample number \( \Gamma_C \) is at most \( \tilde{O}(k^{4/3}\epsilon^{-2}) \) (see Theorem 1.2). The notion of additive terminal embedding might be of independent interest.

**Bounding the covering number in discrete metrics.** Upper bounding the covering number in discrete metrics is much easier than that in Euclidean metrics, since we can directly construct coverings for \( y^C \) without dimension reduction. Our smaller relative covering error only reduces the base of the covering number instead of the exponent. Concretely, the covering number \( |V_h| \) in doubling metrics with doubling dimension \( \text{ddim} \) is upper bounded by \( (\frac{1}{2^n})^{O(k^{\text{ddim}})} \) (Lemma 5.2) instead of \( (\frac{1}{2^n})^{O(k^{\text{ddim}})} \) in [17], which only increases \( \log |V_h| \) by a logarithmic factor. Combining with our improvement in the variance, we can obtain optimal coreset sizes (up to polylog factors) in doubling metrics (Theorem 1.3) and general discrete metrics (Corollary 1.4). Furthermore, for shortest path metrics with treewidth \( \text{tw} \), the covering number is upper bounded by \( (\frac{1}{2^n})^{O(k^{\text{tw}})} \) (Lemma 6.2) instead of \( (\frac{1}{2^n})^{O(k^{\text{tw}})} \) in [17], which results in an improvement for the coreset size (Theorem 1.5) together with our smaller variance bound.

### 1.3 Other Related Work

**Coresets for variants of clustering.** Coresets for several variants of clustering have also been studied. Cohen-Addad and Li [16] first constructed a coreset of size \( \tilde{O}(k^2 \log^2 n \epsilon^{-3}) \) for capacitated \( k \)-Median in \( \mathbb{R}^d \), and the coreset size was improved to \( \tilde{O}(k^3 \epsilon^{-6}) \) by [5]. A generalization of capacitated clustering, called fair clustering, has also been shown to admit coresets of small size [36, 24, 5]. Another important variant of clustering is robust clustering, in which we can exclude at
most $m$ points as outliers from the clustering objective. Very recently, Huang et al. [23] provided a coreset construction for robust $(k, z)$-CLUSTERING in Euclidean spaces of size $m + \text{poly}(k, \epsilon^{-1})$. Other variants of clustering that admit small-sized coresets include fuzzy clustering [3], ordered weighted clustering [6], and time-series clustering [26].

**Coresets for other problems.** Coresets have also been applied to a wide range of optimization and machine learning problems, including regression [18, 30, 4, 12, 28, 10], low rank approximation [13], projective clustering [19, 38], and mixture model [32, 25].

**Concurrent work.** Very recently, a concurrent work [15] claimed a coreset size bound $\tilde{O}(k^{1.5}\epsilon^{-2})$ for both Euclidean $k$-MEANS and $k$-MEDIAN in their NeurIPS version. For $k$-MEANS, their bound is the same as ours. For $k$-MEDIAN, their bound is worse than our $\tilde{O}(k^{4/3}\epsilon^{-2})$ bound. Their analysis also tries to tighten the variance bound. But our details are quite different from theirs (e.g., we use different relative errors and the overall idea of bounding the covering numbers is very different). Our other results, i.e., the improved result for general $z > 0$, the optimal coreset size bounds for doubling metric and discrete metric, and the improved coreset size for the shortest path metric with bounded treewidth, are new to the best of our knowledge.

1.4 Roadmap

In Section 2, we review the importance sampling framework (Algorithm 1) for $(k, z)$-CLUSTERING proposed by [17]. In Section 3, we introduce a unified analysis of the the performance guarantee for Algorithm 1 (Theorem 3.5). In Sections 4 to 6, we apply Theorem 3.5 to several metrics including Euclidean metrics (Theorem 4.1), doubling metrics (Theorem 5.1) and general metrics (Corollary 5.3), and shortest path metrics in graphs with bounded treewidth (Theorem 6.1), and achieve tighter coreset sizes.

2 The Coreset Construction Algorithm for $(k, z)$-Clustering

In this section, we present a unified coreset construction algorithm (Algorithm 1) for $(k, z)$-CLUSTERING via importance sampling for all metric spaces considered in this paper. Our algorithm large follows the one proposed in [17, 14], with some minor variations. For ease of presentation, we first show how to construct coresets for $k$-MEDIAN ($z = 1$) and use cost to denote cost$_1$. The general $z > 1$ case will be discussed in Appendix A. We first introduce some useful notations.

2.1 Partition into Rings and Groups

Let $A^* = \{a_1^*, \ldots, a_k^* \in X\} \in X^k$ be a constant approximation for the $k$-MEDIAN problem. Let $P_i = \{x \in P : \arg \min_{j \in [k]} d(x, a_j) = i\}$ be the $i$-th cluster induced by $A^*$ (breaking ties arbitrarily). For each $i \in [k]$, denote $\Delta_i := \frac{1}{|P_i|} \text{cost}(P_i, A^*)$ to be the average cost of $P_i$ to $A^*$.

---

2Their NeurIPS version was made public on Nov 01 2022 and their arXiv version ([https://arxiv.org/abs/2211.08184](https://arxiv.org/abs/2211.08184)) was uploaded on Nov 15 2022.

7
Ring structure and group structure. Following [17, 14], we first partition the points into a collection of rings, and then group some rings into the group structure, based on $P_i$s and $A^\ast$. We first partition each clusters $P_i$ into rings according to the ratio $\frac{d(p,A^\ast)}{\Delta_i}$ for $p \in P_i$.

**Definition 2.1 (Ring structure [17, 14])** For each $i \in [k]$ and $j \in \mathbb{Z}$, define ring $R_{ij} := \{p \in P_i : 2^i \Delta_i \leq d(p, A^\ast) < 2^{i+1} \Delta_i\}$ to be the set of points in $P_i$ whose distances to $A^\ast$ is within $[2^i \Delta_i, 2^{i+1} \Delta_i)$. For each $i \in \mathbb{Z}$, define $R(j) := \bigcup_{i \in [k]} R_{ij}$ to be the union of all $j$-th level rings $R_{ij}$. For each $P_i$, we divide rings in the following way:

- If $j \leq \log \varepsilon$, we call $R_{ij}$ an inner ring, where each point $p \in R_{ij}$ satisfies $d(p, A^\ast) \leq \varepsilon \Delta_i$.
- Let $R_i^{(o)} = \bigcup_{j \geq 2 \log \varepsilon^{-1}} R_{ij}$ denote an outer ring, where each point $p \in R_{ij}$ satisfies $d(p, A^\ast) \geq \varepsilon^{-2} \Delta_i$. Let $R_i^{(o)} = \bigcup_{i \in [k]} R_i^{(o)}$ denote the collection of all points in outer rings.
- If $\log \varepsilon < j < 2 \log \varepsilon^{-1}$, we call $R_{ij}$ a main ring.

By merging rings $R$ at the same level from different clusters with similar $\text{cost}(R,A^\ast)$, we propose the following group structure.

**Definition 2.2 (Group structure)** For each integer $\log \varepsilon < j < 2 \log \varepsilon^{-1}$ and $b \leq 0$, we denote the interval $I(j, b) = [2^b \cdot \text{cost}(R(j), A^\ast), 2^{b+1} \cdot \text{cost}(R(j), A^\ast)]$ and define a main group

$$G^{(m)}(j, b) := \bigcup_{i \in [k] : \text{cost}(R_{ij}, A^\ast) \in I(j, b)} R_{ij}$$

as the collection of main rings $R_{ij}$ with $2^b \cdot \text{cost}(R(j), A^\ast) \leq \text{cost}(R_{ij}, A^\ast) < 2^{b+1} \cdot \text{cost}(R(j), A^\ast)$. Let $G^{(m)}(j) := \{G^{(m)}(j, b) : \log(\varepsilon/4) - \log k < b \leq 0\}$ be the collection of $j$-th level main groups. Let $G^{(m)} := \{G \in G^{(m)}(j) : \log \varepsilon < j < 2 \log \varepsilon^{-1}\}$ be the collection of all main groups.

Similarly, for each integer $b \leq 0$, we define $I(o, b) = [2^b \cdot \text{cost}(R^{(o)}(j), A^\ast), 2^{b+1} \cdot \text{cost}(R^{(o)}(j), A^\ast)]$ and an outer group

$$G^{(o)}(b) := \bigcup_{i \in [k] : \text{cost}(R_i^{(o)}, A^\ast) \in I(o, b)} R_i^{(o)}$$

as the collection of outer rings $R_i^{(o)}$ with $2^b \cdot \text{cost}(R^{(o)}(j), A^\ast) \leq \text{cost}(R_i^{(o)}, A^\ast) < 2^{b+1} \cdot \text{cost}(R^{(o)}(j), A^\ast)$. Let $G^{(o)} := \{G^{(o)}(b) : \log(\varepsilon/4) - \log k < b \leq 0\}$ be the collection of outer groups.

Let $G := G^{(m)} \cup G^{(o)}$ be the collection of all groups.

By definition, we know that all groups in $G$, including main groups $G^{(m)}(j, b)$s and outer groups $G^{(o)}(b)$s, are pairwise disjoint. For main groups, we provide an illustration in Figure 1. We also have the following observation that lower bounds $\frac{d(p,A^\ast)}{\text{cost}(G,A^\ast)}$ for main groups $G \in G^{(m)}$.

**Observation 2.3 (Main group cost [14])** Let $G \in G^{(m)}$ be a main group. Let $i \in [k]$ be an integer satisfying that $P_i \cap G \neq \emptyset$. For any $p \in P_i \cap G$, we have

$$\text{cost}(G, A^\ast) \leq 2k \cdot \text{cost}(P_i \cap G, A^\ast) \leq 4k \cdot |P_i \cap G| \cdot d(p, A^\ast), \text{ and } |P_i \cap G| \cdot d(p, A^\ast) \leq 2\text{cost}(P_i \cap G).$$
Note that $\mathcal{G}$ may not contain all points in $P$ – actually, we discard some “light” rings in Definition 2.2. We will see that we only need to take samples from groups in $\mathcal{G}$ for coreset construction and the remaining points can be “represented by” points in $A^*$ with a small estimation error. Our group structure is slightly different from that in [17, 14]: they gather groups $G^{(m)}(j, b)/G^{(o)}(b)$ with $-\log k \leq b \leq 0$ as an entirety and sample from them together. Our group structure increases the number of groups in $\mathcal{G}$ by a polylog factor $O(\log(\varepsilon^{-1}) \log k)$ (summarized as Lemma 2.4); however, this minor change is essential for reducing the coreset size.

**Lemma 2.4 (Group number)** There exist at most $O(\log(\varepsilon^{-1}) \log k)$ groups in $\mathcal{G}$.

We also have the following lemma for the construction time of $\mathcal{G}$.

**Lemma 2.5 (Construction time of $\mathcal{G}$)** Given a constant approximation $A^* \in \mathcal{X}^k$, it takes $O(nk)$ time to construct $\mathcal{G}$.

**Proof:** Firstly, it takes $O(nk)$ time to compute all distances $d(p, A^*)$ for $p \in P$ and construct clusters $P_i$s. Then it takes $O(n)$ time to compute all $\Delta_i$s and construct all rings and their corresponding cost $\text{cost}(R_{ij}, A^*)$ and $\text{cost}(R^{(o)}_i, A^*)$. Finally, it takes at most $O(n)$ time to construct all groups $G^{(m)}(j, b)$ and $G^{(o)}(b)$. Overall, we complete the proof. \qed

### 2.2 A Coreset Construction Algorithm for k-Median [17, 14]

Now we are ready to present our coreset construction algorithm for $k$-Median (see Algorithm 1). We use the same importance sampling procedure proposed in [17, 14]. The key is to perform an importance sampling procedure for each group $G \in \mathcal{G}$ with a carefully selected number of samples $\Gamma_G$ (Line 1). For $k$-Median, $\Gamma_G$ is defined in Theorem 3.5. We also weigh each center $a^*_i \in A^*$
by the number of remaining points in $P_i \setminus \mathcal{G}$ (We slightly abuse the notation $P_i \setminus \mathcal{G}$ to denote $P_i \setminus (\cup_{G \in \mathcal{G}} G)$. See Line 2). Algorithm 1 can be easily extended to general $(k, z)$-CLUSTERING; The details can be found in Appendix A.

Algorithm 1 Coreset Construction Algorithm for k-Median [17, 14]

Input: a metric space $(\mathcal{X}, d)$, a dataset $P \subseteq \mathcal{X}$, integer $k \geq 1$; A constant approximation $A^* \in \mathcal{X}^k$ for $P$, and the partition $P_1, \ldots, P_k$ of $P$ according to $A^*$; The collection $\mathcal{G}$ of groups as in Definition 2.2 together with the number of samples $\Gamma_G$ for each $G \in \mathcal{G}$.

Output: a weighted subset $(S, w)$

1: For each group $G \in \mathcal{G}$, sample a collection $S_G$ of $\Gamma_G$ points from $G$, where each sample $p \in G$ is selected with probability $\frac{d(p, A^*)}{\text{cost}(G, A^*)}$ and weighted by $w(p) = \frac{\text{cost}(G, A^*)}{\text{cost}(G, d(p, A^*))}$.
2: For each $i \in [k]$, set the weight of $a_i^* \in A^*$ as $w(a_i^*) = |P_i \setminus \mathcal{G}|$.
3: return $S := A^* \cup (\bigcup_{G \in \mathcal{G}} S_G)$ together with the weights $w$.

3 Analysis of Algorithm 1

In this section, we provide a new analysis for Algorithm 1 (Theorem 3.5).

3.1 $\alpha$-Covering and Covering Number of Groups

We first define $\alpha$-covering of groups in $\mathcal{G}$. The new definition of $\alpha$-covering for main groups is crucial for improving the coreset size.

Coverings of main groups. Recall that $A^* = \{a_1^*, \ldots, a_k^* \in \mathcal{X}\} \in \mathcal{X}^k$ is a constant approximation for the $k$-MEDIAN problem. We define the following huge subset of main groups $G \in \mathcal{G}^{(m)}(j)$ w.r.t. a $k$-center set $C \in \mathcal{X}^k$:

$H(G, C) := \{p \in R_{ij} \cap G : i \in [k], d(a_i^*, C) \geq 2^{j+4} \varepsilon^{-1} \Delta_i\}$,

i.e., the collection of rings $R_{ij} \in G$ with $d(a_i^*, C) \geq 2^{j+4} \varepsilon^{-1} \cdot 2^j \Delta_i$. By construction, for all points $p \in H(G, C)$, the distances $d(p, C)$ are "close" to each other, which is an important property. We summarize the property in the following observation.

Observation 3.1 For a $k$-center set $C \in \mathcal{X}^k$, $i \in [k]$ and $p \in P_i \cap H(G, C)$, we have $d(p, C) \in (1 \pm \frac{\varepsilon}{4}) \cdot d(a_i^*, C)$.

By this observation, we can use a single distance $d(a_i^*, C)$ to approximate $d(p, C)$ with a small error for all points $p \in H(G, C)$. For points in $G \setminus H(G, C)$, we define the following notions of coverings and covering numbers of main groups.

Definition 3.2 ($\alpha$-Coverings of main groups) Let $G \in \mathcal{G}^{(m)}$ be a main group. Let $S \subseteq G$ be a subset and $\alpha > 0$. We say a set $V \subseteq \mathbb{R}^{|S|}$ of cost vectors is an $\alpha$-covering of $S$ if for each $C \in \mathcal{X}^k$, there exists a cost vector $v \in V$ such that for any $i \in [k]$ and $p \in P_i \cap S \setminus H(G, C)$, the following holds:

$$|d(p, C) - d(a_i^*, C) - v_p| \leq \alpha \cdot \text{err}(p, C),$$

(4)
where \( \text{err}(p, C) \) is called the relative covering error of \( p \) to \( C \) defined as follows

\[
\text{err}(p, C) := \left( \sqrt{d(p, C) \cdot d(p, A^*)} + d(p, A^*) \right) \cdot \sqrt{\frac{\text{cost}(G, C + A^*)}{\text{cost}(G, A^*)}},
\]

where we use \( \text{cost}(G, C + A^*) \) as a shorthand notation of \( \text{cost}(G, C) + \text{cost}(G, A^*) \) throughout.

**Definition 3.3 (Covering numbers of main groups)** Define \( \mathcal{N}^{(m)}(S, \alpha) \) to be the minimum cardinality \( |V| \) of any \( \alpha \)-covering \( V \) of \( S \). Given an integer \( \Gamma \geq 1 \), define the \((\Gamma, \alpha)\)-covering number of \( G \) to be

\[
\mathcal{N}_G^{(m)}(\Gamma, \alpha) := \max_{S \subseteq G, |S| \leq \Gamma} \mathcal{N}^{(m)}(S, \alpha),
\]

i.e., the maximum cardinality \( \mathcal{N}^{(m)}(S, \alpha) \) over all possible subsets \( S \subseteq G \) of size at most \( \Gamma \).

Intuitively, the covering number is a complexity measure of the set of the all possible distance difference vectors \( \{d(p, C) - d(a^*_i, C)\}_{p \in G} \). As \( \mathcal{N}_G^{(m)}(\Gamma, \alpha) \) becomes larger, all possible center sets in \( \mathcal{X}^k \) can induce more types of such vectors (up to the relative covering error).

**Remark:** We remark that [14, Definition 3] also introduces a discretization of \( \mathcal{X}^k \). Our definition has the following differences from theirs, which are crucial for improving the coreset sizes for several metric spaces:

1. Our covering is constructed on the distance difference \( d(p, C) - d(a^*_i, C) \) instead of \( d(p, C) \), which is upper bounded by \( d(p, a^*_i) \) by the triangle inequality. For comparison, the covering error defined in [14, Definition 3] is \( d(p, C) + d(p, A^*) \). Our covering error \( \text{err}(p, C) \) is typically smaller (especially when \( d(p, C) \gg d(p, A^*) \)). This smaller error is essential for the improvement of coreset sizes, since its scale affects the variance of our sampling scheme (Lemma 3.14) and the variance bound directly appears in the coreset size (Lemma 3.13).

2. We consider coverings for subsets \( S \subseteq G \) of different sizes \( \Gamma \) instead of the entire dataset \( G \). This allows us to design a one-stage sampling framework for general discrete metric space and shortest path metric. On the other hand, the analysis in [17] assumes that their algorithm is run over a coreset found by another existing coreset algorithm, which is a two-stage algorithm.

**Coverings of outer groups.** Similarly, we define coverings of an outer group \( G \in \mathcal{G}^{(o)} \). The same as [14], we define the following far subset of \( G \):

\[
F(G, C) := \left\{ p \in R_i^{(o)} \cap G : i \in [k], \exists q \in R_i^{(o)} \cap G \text{ with } d(q, C) \geq 4 \cdot d(q, A^*) \right\},
\]

i.e., the collection of rings \( R_i^{(o)} \) in \( G \) that contain at least one point \( q \) with \( d(q, C) \geq 4 \cdot d(q, A^*) \). Note that for any \( p \in G \setminus F(G, C) \), we have \( \frac{d(p, C)}{d(p, A^*)} \leq 4 \). Again, we formally define coverings and covering numbers of outer groups, which is very similar to those for main groups.

**Definition 3.4 (Coverings and covering numbers of outer groups)** Let \( G \in \mathcal{G}^{(o)} \) be an outer group. Let \( S \subseteq G \) be a subset and \( \alpha > 0 \). We say a set \( V \subset \mathbb{R}^{|S|} \) of cost vectors is an \( \alpha \)-covering
of $S$ if for each $C \in \mathcal{X}^k$, there exists a cost vector $v \in V$ such that for any $p \in S \setminus F(G, C)$, the following inequality holds:

$$|d(p, C) - v_p| \leq \alpha \cdot (d(p, C) + d(p, A^*)).$$

Define $\mathcal{N}(\gamma)(S, \alpha)$ to be the minimum cardinality $|V|$ of an arbitrary $\alpha$-covering $V$ of $S$. Given an integer $\Gamma \geq 1$, define the $(\Gamma, \alpha)$-covering number of $G$ to be

$$\mathcal{N}_G(\Gamma, \alpha) := \max_{S \subseteq G: |S| \leq \Gamma} \mathcal{N}(\gamma)(S, \alpha),$$

i.e., the maximum cardinality $\mathcal{N}(\gamma)(S, \alpha)$ over all possible subsets $S \subseteq G$ of size at most $\Gamma$.

Our definition for outer groups is almost the same as [14, Definition 3], except that we consider coverings for subsets $S \subseteq G$ instead of $G$, which enables us to design a one-stage sampling algorithm.

### 3.2 The Main Theorem for Algorithm 1

Now we are ready to state the main theorem for Algorithm 1.

**Theorem 3.5 (Coreset for $k$-Median)** Let $(\mathcal{X}, d)$ be a metric space and $P \subseteq \mathcal{X}$ be a set of $n$ points. Let integer $k \geq 1$ and $\varepsilon \in (0, 1)$ be the precision parameter. We define the number $\Gamma_G$ of samples for group $G$ (in Algorithm 1) as follows:

1. For each main group $G \in \mathcal{G}^{(m)}$, let $\Gamma_G$ be the smallest integer satisfying that

$$\Gamma_G \geq \Omega \left( \varepsilon^{-2} \left( \int_0^1 \sqrt{\log \mathcal{N}_G^{(m)}(\Gamma_G, \alpha)} d\alpha \right)^2 + k \varepsilon^{-2} \log(k \varepsilon^{-1}) \right); \quad (5)$$

2. For each outer group $G \in \mathcal{G}^{(o)}$, let $\Gamma_G$ be the smallest integer satisfying that

$$\Gamma_G \geq \Omega \left( \varepsilon^{-2} \left( \int_0^1 \sqrt{\log \mathcal{N}_G^{(o)}(\Gamma_G, \alpha)} d\alpha \right)^2 + k \varepsilon^{-2} \log(k \varepsilon^{-1}) \right); \quad (6)$$

With probability at least 0.9, Algorithm 1 outputs an $O(\varepsilon)$-coreset $(S, w)$ of size $k + \sum_{G \in G} \Gamma_G$, and the running time is $O(nk)$.

$\Gamma_G$ appears in the familiar form of the entropy integral (or Dudley integral) commonly used in the chaining argument (see e.g., [39, Corollary 5.25]). Hence, the remaining task is to upper bound the entropy integrals. Moreover, combining with Lemma 2.5, we know that the coreset construction time is at most $O(nk)$ time (for building the rings and groups), given a constant factor approximation $A^*$. For general $z \geq 1$, we can obtain the following theorem, and the proof can be found in Appendix D.

**Theorem 3.6 (Coreset for $(k, z)$-Clustering in general metrics)** Let $(\mathcal{X}, d)$ be a metric space and $P \subseteq \mathcal{X}$ be a set of $n$ points. Let integers $k \geq 1$, constant $z \geq 1$, and $\varepsilon \in (0, 1)$ be the precision parameter. Let the number of samples $\Gamma_G$ for group $G$ be defined as in (5) and (6). With probability at least 0.9, Algorithm 1 outputs an $\varepsilon$-coreset $(S, w)$ for $(k, z)$-Clustering of size $k + 2^{O(z)} \sum_{G \in G} \Gamma_G$, and the running time is $O(nk)$.

---

3The definition of covering numbers $\mathcal{N}_G^{(m)}(\Gamma_G, \alpha)$ and $\mathcal{N}_G^{(o)}(\Gamma_G, \alpha)$ are extended to constant $z \geq 1$ in a straightforward way; see details in Section D.1.
Compared to Theorem 3.5, the coreset size for general $z \geq 1$ has an additional multiplicative factor $2^O(z)$. Typically, we treat $z$ as a constant and thus this additional factor is also a constant.

### 3.3 Proof of Theorem 3.5 (Main Theorem for $k$-Median)

In this section, we prove Theorem 3.5 for $k$-Median. The coreset size is directly from Line 1 of Algorithm 1. Thus, it remains to verify that $S$ is an $O(\varepsilon)$-coreset.

The key is the following lemma that summarizes the estimation error induced by each group $G \in \mathcal{G}$ and the remaining points $P \setminus \mathcal{G}$. For a group $G \in \mathcal{G}$, we define $P_G := \{p \in P : \exists i \in [k], p \in P_i \text{ and } P_i \cap G \neq \emptyset\}$ to be the union of all clusters $P_i$ that intersects with $G$. Recall that $\text{cost}(G, C + A^*) = \text{cost}(G, C) + \text{cost}(G, A^*)$.

**Lemma 3.7 (Error analysis of groups and remaining points)** We have the followings:

1. For each main group $G \in \mathcal{G}^{(m)}$, we have
   \[
   
   \mathbb{E}_{S_G} \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}(G, C + A^*)} \cdot \sum_{p \in S_G} w(p) \cdot d(p, C) - \text{cost}(G, C) \right] \leq 3\varepsilon.
   \]

2. For each outer group $G \in \mathcal{G}^{(o)}$, we have
   \[
   
   \mathbb{E}_{S_G} \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}(P_G, C + A^*)} \cdot \sum_{p \in S_G} w(p) \cdot d(p, C) - \text{cost}(G, C) \right] \leq 2\varepsilon.
   \]

3. For any center set $C \in \mathcal{X}^k$, we have
   \[
   
   \left| \text{cost}(P \setminus \mathcal{G}, C) - \sum_{i \in [k]} w(a_i^{*}) \cdot d(a_i^{*}, C) \right| \leq \varepsilon \cdot \text{cost}(P, C + A^*).
   \]

Here, the expectations above are taken over the randomness of the sample set $S_G$.

Note that the normalization term of a main group $G \in \mathcal{G}^{(m)}$ is $\frac{1}{\text{cost}(G, C + A^*)}$ that only relates to $G$, but the normalization term of an outer group $G \in \mathcal{G}^{(o)}$ is $\frac{1}{\text{cost}(P_G, C + A^*)}$ that depends on a larger subset $P_G$. This is because “remote” points in outer groups $G \in \mathcal{G}^{(o)}$ introduce an even larger empirical error than $\text{cost}(G, C + A^*)$, but the total number of these remote points is small compared to $|P_G|$, which enables us to upper bound the empirical error by $\varepsilon \cdot (\text{cost}(P_G, C + A^*))$.

The proof of Lemma 3.7 requires the chaining argument and the improved bound of variance and is deferred to Section 3.4. Theorem 3.5 is a direct corollary of Lemma 3.7. The argument is almost the same as the proof of Theorem 4 in [14]. Intuitively, we can verify that for any center set $C \in \mathcal{X}^k$, the expected estimation error of $(S, w)$ is small, say

\[

\mathbb{E}_S \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}(P, C + A^*)} \cdot \sum_{p \in S} w(p) \cdot d(p, C) - \text{cost}(P, C) \right] \leq O(\varepsilon).

\]

Then we only need to apply the Markov inequality and use the fact that $A^*$ is an $O(1)$-approximate solution. For completeness, we provide the proof of Theorem 3.5 from Lemma 3.7 in Section B.
Lemma 3.11 (Estimation error of \( \xi_G \) found in Section C.)

For \( x \), we have the following lemma whose proof idea is from that in [14, Lemma 15] and can be found in Section C.

\[
\sum_{p \in P_i \cap S_G} w(p) = \sum_{p \in P_i \cap S_G} \frac{\text{cost}(G, A^\ast)}{\Gamma_G \cdot d(p, A^\ast)} \in (1 \pm \varepsilon) \cdot |P_i \cap G|.
\]

Event \( \xi_G \) implies that for each cluster \( P_i \), the total sample weight of \( P_i \cap S_G \) is close to the cardinality \( |P_i \cap G| \). It is not difficult to show that \( \xi_G \) happens with high probability via standard concentration inequality.

**Definition 3.9 (Cost vectors of main groups)** Given a \( k \)-center set \( C \in \mathcal{X}^k \), we define the following cost vectors:

- Let \( u^C \in \mathbb{R}_{\geq 0}^{|G|} \) be a cost vector satisfying that 1) for any \( p \in H(G, C) \), \( u^C_p := d(p, C) \); 2) for any \( p \in G \setminus H(G, C) \), \( u^C_p := 0 \).
- Let \( x^C \in \mathbb{R}_{\geq 0}^{|G|} \) be a cost vector satisfying that 1) for any \( i \in [k] \) and \( p \in P_i \cap G \setminus H(G, C) \), \( x^C_p := d(a^*_i, C) \); 2) for any \( p \in H(G, C) \), \( x^C_p := 0 \).
- Let \( y^C \in \mathbb{R}^{|G|} \) be a cost vector satisfying that 1) for any \( i \in [k] \) and \( p \in P_i \cap G \setminus H(G, C) \), \( y^C_p := d(p, C) - d(a^*_i, C) \); 2) for any \( p \in H(G, C) \), \( y^C_p := 0 \).

The main difference from [14] is that we create two cost vectors \( x^C \) and \( y^C \) for \( G \setminus H(G, C) \) for variance reduction. Note that \( \text{cost}(G, C) = \|u^C\|_1 + \|x^C + y^C\|_1 \). Hence, to prove Item 1, it suffices to upper bound the estimation error of \( u^C \), \( x^C \) and \( y^C \) by \( \varepsilon \) respectively and we give the corresponding lemmas (Lemmas 3.10 to 3.12) in the following.

For \( u^C \), we have the following lemma from [14] since \( \Gamma_G \geq O(k \varepsilon^{-2} \log k) \).

**Lemma 3.10 (Estimation error of \( u^C \) [14, Lemma 15])** The following inequality holds:

\[
\mathbb{E}_{S_G} \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}(G, C + A^\ast)} \sum_{p \in S_G} w(p) \cdot u^C_p - \|u^C\|_1 \right] \leq \varepsilon.
\]

For \( x^C \), we have the following lemma whose proof idea is from that in [14, Lemma 15] and can be found in Section C.

**Lemma 3.11 (Estimation error of \( x^C \))** The following inequality holds:

\[
\mathbb{E}_{S_G} \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}(G, C + A^\ast)} \sum_{p \in S_G} w(p) \cdot x^C_p - \|x^C\|_1 \right] \leq \varepsilon.
\]
The main difficulty is to upper bound the estimation error for \( y^C \) by \( \varepsilon \) as in the following lemma.

**Lemma 3.12 (Estimation error of \( y^C \))** The following inequality holds:

\[
\mathbb{E}_{S_G} \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}(G, C + A^*)} \left| \sum_{p \in S_G} w(p) \cdot y^C_p - \|y^C\|_1 \right| \right] \leq \varepsilon.
\]

Adding the inequalities in the above lemmas, we directly conclude Item 1 since for any \( C \in \mathcal{X}^k \),

\[
\left| \sum_{p \in S_G} w(p) \cdot d(p, C) - \text{cost}(G, C) \right| \leq \left| \sum_{p \in S_G} w(p) \cdot u^C_p - \|u^C\|_1 \right| + \left| \sum_{p \in S_G} w(p) \cdot x^C_p - \|x^C\|_1 \right| + \left| \sum_{p \in S_G} w(p) \cdot y^C_p - \|y^C\|_1 \right|,
\]

which implies the first item of Lemma 3.7:

\[
\mathbb{E}_{S_G} \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}(G, C + A^*)} \cdot \left| \sum_{p \in S_G} w(p) \cdot d(p, C) - \text{cost}(G, C) \right| \right] \leq 3\varepsilon.
\]

Hence, it remains to prove Lemma 3.12.

**Proof:** [of Lemma 3.12]

We first use symmetrization trick to reduce the left hand side to a Gaussian process, and then apply a chaining argument.

**Reduction to a Gaussian process.** Let \( \xi = \{\xi_p \sim N(0, 1) : p \in S_G\} \) be a collection of \( \Gamma_G \) independent standard Gaussian random variables. Note that \( S_G \) is drawn from the unbiased importance sampling algorithm (Line 1 of Algorithm 1), which implies that \( \mathbb{E}_{S_G} \left[ \sum_{p \in S_G} w(p) \cdot y^C_p \right] = \|y^C\|_1 \) holds for any \( C \in \mathcal{X}^k \). By [39, Lemma 7.4], we have

\[
\mathbb{E}_{S_G} \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}(G, C + A^*)} \left| \sum_{p \in S_G} w(p) \cdot y^C_p - \|y^C\|_1 \right| \right] \leq \sqrt{2\pi} \cdot \mathbb{E}_{S_G, \xi} \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}(G, C + A^*)} \left| \sum_{p \in S_G} \xi_p \cdot w(p) \cdot y^C_p \right| \right],
\]

where \( \xi_p \)s are i.i.d. standard Gaussian variables. Thus, it suffices to prove that for any (multi-)set \( S_G \subseteq G \) of size \( \Gamma_G \),

\[
\mathbb{E}_\xi \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}(G, C + A^*)} \left| \sum_{p \in S_G} \xi_p \cdot w(p) \cdot y^C_p \right| \right] \leq \frac{\varepsilon}{6}. \quad (8)
\]
Note that we fix $S_G$ and the only randomness is from $\xi$ (hence we get a Gaussian process).

By the triangle inequality, we directly have $|d(p, C) - d(a_i^*, C)| \leq d(p, a_i^*)$ for any $C \in X^k$, $i \in [k]$ and $p \in P_i \cap G \setminus H(G, C)$. Hence, for any $\alpha \geq 1$, $N_G^{(m)}(\Gamma, \alpha) = 1$ always holds by Definition 3.2 and we can trivially use $0^{|S_G|}$ as an $\alpha$-covering of $S_G$. Consequently, we only need to consider the range $0 < \alpha \leq 1$.

For integer $h \geq 0$, let $V_h \subset \mathbb{R}^G$ be a $2^{-h}$-covering of $S_G$. For any $C \in X^k$, define $v^{C,h} \in V_h$ to be a vector satisfying that for each $p \in S_G$,

$$|y_p^C - v_p^{C,h}| \leq 2^{-h} \cdot \text{err}(p, C) = 2^{-h} \cdot \left( \sqrt{d(p, C) \cdot d(p, A^*)} + d(p, A^*) \right) \cdot \sqrt{\frac{\text{cost}(G, C + A^*)}{\text{cost}(G, A^*)}}.$$

Now we consider the following estimator

$$X_{C,h} := \frac{1}{\text{cost}(G, C + A^*)} \sum_{p \in H(S_G; C)} \xi_p \cdot w(p) \cdot (v_p^{C,h+1} - v_p^{C,h}),$$

and let $X_C := \sum_{h \geq 0} X_{C,h}$. Note that for any $p \in S_G \setminus H(S_G; C)$, we have $y_p^C = 0$ by definition. Thus, Inequality (8) is equivalent to the following inequality:

$$\mathbb{E}_\xi \sup_{C \in X^k} |X_C| \leq \frac{\varepsilon}{6},$$

since $X_C$ is a separable process [39, Definition 5.22] with $\lim_{h \to +\infty} v_p^{C,h} = y_p^C$. Since $X_C \coloneqq \sum_{h \geq 0} X_{C,h}$, it suffices to prove

$$\sum_{h \geq 0} \mathbb{E}_\xi \sup_{C \in X^k} |X_{C,h}| \leq \frac{\varepsilon}{6} \quad (9)$$

In the following, we focus on proving Inequality (9).

**A chaining argument for Inequality (9).** Fix an integer $h \geq 0$ and we first show how to upper bound $\mathbb{E}_\xi \sup_{C \in X^k} |X_{C,h}|$ by a chaining argument. Note that each $X_{C,h}$ is a Gaussian variable. The main idea is to upper bound the variance of each $X_{C,h}$, which leads to an upper bound for $\mathbb{E}_\xi \sup_{C \in X^k} |X_{C,h}|$ by the following lemma.

**Lemma 3.13 ([34, Lemma 2.3])** Let $g_i \sim N(0, \sigma_i^2)$ for each $i \in [n]$ be Gaussian random variables (not need to be independent) and suppose $\sigma = \max_{i \in [n]} \sigma_i$. Then

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

Note that each $X_{C,h}$ is a Gaussian variable with mean 0. Hence, the key is upper bounding the variance of $X_{C,h}$ as in the following lemma.

**Lemma 3.14 (Variance of $X_{C,h}$)** Fix a $k$-center set $C \in X^k$ and an integer $h \geq 0$. The variance of $X_{C,h}$ is always at most

$$\text{Var} [X_{C,h}] = \sum_{p \in S_G \setminus H(S_G; C)} \left( \frac{w(p) \cdot (v_p^{C,h+1} - v_p^{C,h})}{\text{cost}(G, C + A^*)} \right)^2 \leq \frac{2^{-2h+4k}}{\Gamma_G}.$$
Moreover, conditioned on event $\xi_G$ (Inequality (7)), the variance of $X_{C,h}$ is always at most

$$\text{Var}[X_{C,h} \mid \xi_G] = \sum_{p \in S_G \setminus H(S_G, C)} \left( \frac{w(p) \cdot (v_p^{C,h+1} - v_p^{C,h})}{\text{cost}(G, C + A^*)} \right)^2 \leq \frac{2^{-2h+6}}{\Gamma_G}.$$  

The proof and analysis of the above lemma can be found in Section 3.5. Compared with [14, Lemma 23], the above lemma provides a tighter variance bound, saving a min $\{\varepsilon^{-1}, k\}$ factor.

Now we are ready to upper bound $\mathbb{E}_\xi \sup_{C \in \mathcal{X}^k} |X_{C,h}|$. By law of total expectation, we have

$$\mathbb{E}_\xi \sup_{C \in \mathcal{X}^k} |X_{C,h}| = \mathbb{E}_\xi \left[ \sup_{C \in \mathcal{X}^k} |X_{C,h}| \mid \xi_G \right] \cdot \Pr[\xi_G] + \mathbb{E}_\xi \left[ \sup_{C \in \mathcal{X}^k} |X_{C,h}| \mid \bar{\xi}_G \right] \cdot \Pr[\bar{\xi}_G]. \quad (10)$$

For each $C \in \mathcal{X}^k$, we define $\psi^{(C,h)} = v^{C,h+1} - v^{C,h}$ as a certificate of $C$. Note that there are at most $|V_{h+1}| \cdot |V_h|$ different certificates $\psi^{(C,h)}$s and let $\Psi_h = \{\psi^{(C,h)} : C \in \mathcal{X}^k\}$ be the collection of all possible certificates $\psi^{(C,h)}$s. For each certificate $\psi \in \Psi$, we select a $k$-center set $C_\psi \in \mathcal{X}^k$ with

$$C_\psi := \arg \min_{C \in \mathcal{X}^k, \psi^{(C,h)} = \psi} \text{cost}(G, C),$$

to be the representative center set of $\psi$. By definition, we note that for any $C \in \mathcal{X}^k$ with $\psi^{(C,h)} = \psi$,

$$\left| \frac{1}{\text{cost}(G, C + A^*)} \sum_{p \in S_G \setminus H(S_G, C)} \xi_p \cdot w(p) \cdot (v_p^{C,h+1} - v_p^{C,h}) \right| \leq \left| \frac{1}{\text{cost}(G, C_\psi + A^*)} \sum_{p \in S_G \setminus H(S_G, C_\psi)} \xi_p \cdot w(p) \cdot (v_p^{C_\psi,h+1} - v_p^{C_\psi,h}) \right|. \quad (11)$$

Let $\mathcal{C}_h = \{C_\psi : \psi \in \Psi\}$ be the collection of all such $C_\psi$s. By construction, we have $|\mathcal{C}_h| \leq |V_{h+1}| \cdot |V_h|$. Now we have

$$\mathbb{E}_\xi \left[ \sup_{C \in \mathcal{X}^k} |X_{C,h}| \mid \xi_G \right] \cdot \Pr[\xi_G] \leq \mathbb{E}_\xi \left[ \sup_{C \in \mathcal{C}_h} \left| \frac{1}{\text{cost}(G, C + A^*)} \sum_{p \in S_G \setminus H(S_G, C)} \xi_p \cdot w(p) \cdot (v_p^{C,h+1} - v_p^{C,h}) \right| \mid \xi_G \right]$$

$$= \mathbb{E}_\xi \left[ \sup_{C \in \mathcal{C}_h} \left| \frac{1}{\text{cost}(G, C + A^*)} \sum_{p \in S_G \setminus H(S_G, C)} \xi_p \cdot w(p) \cdot (v_p^{C,h+1} - v_p^{C,h}) \right| \mid \xi_G \right]$$

$$\leq 2 \sup_{C \in \mathcal{C}_h} \text{Var}[X_{C,h} \mid \xi_G] \cdot \sqrt{2 \ln(|V_{h+1}| \cdot |V_h|)}$$

$$\leq \frac{96^{-h}}{\sqrt{\Gamma_G}} \cdot \sqrt{|V_{h+1}|},$$

17
where the first inequality follows from the definition of $X_{C,h}$, the first equality holds since the supreme must achieve at some center set $C \in C_h$ due to Inequality (11), the second inequality follows from Lemma 3.13, and the last inequality holds due to Lemma 3.14. Furthermore, we have

$$\mathbb{E}_\xi \left[ \sup_{C \in \mathcal{X}^k} |X_{C,h}| \right] \cdot \Pr [\xi_G] \leq 2 \sum_{C \in \mathcal{C}_h} \left( \frac{w(p) \cdot (v_p^{C,h+1} - v_p^{C,h})}{\cos(G, C + A^*)} \right)^2 \cdot \sqrt{2 \ln(|V_{h+1}| \cdot |V_h|)} \cdot \Pr [\xi_G]$$

where the first inequality follows from Inequality (10), the second inequality from Lemma 3.14, and the last inequality from Lemma 3.8. Using Eq. (10) and Ineqs. (12) and (13), we can conclude that

$$\sum_{h \geq 0} \mathbb{E}_\xi \sup_{C \in \mathcal{X}^k} |X_{C,h}| \leq \frac{2^{7-h}}{\sqrt{\Gamma_G}} \cdot \sqrt{\ln |V_{h+1}|} \cdot \Pr [\xi_G]$$

(Eq. (10), Ineqs. (12) and (13))

$$\leq \frac{\varepsilon}{16},$$

(Defn. of $\Gamma_G$)

$$\leq \frac{\varepsilon}{6},$$

(Defn. of $V_h$)

i.e., Inequality (9) holds. Hence, we have completed the proof of Lemma 3.12. \qed

**Item 2.** For Item 2 of Lemma 3.7, the proof is almost the same as that of [14, Lemma 13]. The only difference is that we consider the coverings on $S_G$ instead of $G$.

**Item 3.** Item 3 of Lemma 3.7 has been proved in [17, 14]; see e.g., [17, Lemma 4]. Overall, we complete the proof of Lemma 3.7.

### 3.5 Proof of Lemma 3.14: A Tighter Variance Bound

We first present the proof of Lemma 3.14, and then show that our relative covering error $\text{err}(p, C)$ leads to almost tight variance upper bound.

**Proof:** [of Lemma 3.14] Let $M_G \subseteq [k]$ denote the collection of $i \in [k]$ with $P_i \cap S_G \setminus H(S_G, C) \neq \emptyset$. By the definition of $H(S_G, C)$, we note that for any $i \in M_G$, $P_i \cap S_G \setminus H(S_G, C) = P_i \cap S_G$. For any $i \in [k]$ and point $p \in P_i \cap G$, let $q = \arg \min_{p' \in P_i \cap G} d(p', C)$. Suppose $p, q \in R_{ij}$ for some $j$. We
\[ d(p, C) \leq d(q, C) + d(p, A^*) + d(q, A^*) \] (triangle ineq.)
\[ \leq d(q, C) + 3d(p, A^*) \] (Defn. of \( R_{ij} \))
\[ \leq \frac{\text{cost}(P_i \cap G, C)}{|P_i \cap G|} + 3d(p, A^*) \] (Defn. of \( q \)).

Then we can bound the variance as follows:

\[
\sum_{p \in S_G \setminus H(S_G, C)} \left( \frac{w(p) \cdot (v_{p}^{C,h+1} - v_{p}^{C,h})}{\text{cost}(G, C + A^*)} \right)^2
\]
\[= \sum_{p \in S_G \setminus H(S_G, C)} \left( \frac{w(p) \cdot (v_{p}^{C,h+1} - y_{p}^{C} - y_{p}^{C} - v_{p}^{C,h})}{\text{cost}(G, C + A^*)} \right)^2
\]
\[\leq \frac{2^{-2h+2}}{\Gamma_G} \sum_{p \in S_G \setminus H(S_G, C)} \frac{w(p) \cdot (d(p, C) + d(p, A^*))}{\text{cost}(G, C + A^*)}
\]
\[\leq \frac{2^{-2h+2}}{\Gamma_G} \sum_{i \in M_C} \sum_{p \in P_i \cap S_G} \frac{w(p) \cdot \text{cost}(P_i \cap G, C) + \text{cost}(P_i \cap G, A^*)}{\text{cost}(G, C + A^*)} \cdot \sum_{p \in P_i \cap S_G} w(p)
\]

where the first inequality holds due to the Defn. of \( v_{p}^{C,h+1} \) and \( v_{p}^{C,h} \), the second follows from the Definition of \( w(p) \), the third from Inequality (14) and the definition of \( M_C \), and the last from Observation 2.3. Recall that event \( \xi_G \) is defined as

\[
\sum_{p \in P_i \cap S_G} w(p) = \sum_{p \in P_i \cap S_G} \frac{\text{cost}(G, A^*)}{\Gamma_G \cdot d(p, A^*)} \in (1 \pm \varepsilon) \cdot |P_i \cap G|. \tag{16}
\]

Hence, conditioning on \( \xi_G \) and continue the derivation, we can see that

\[
\sum_{p \in S_G \setminus H(S_G, C)} \left( \frac{w(p) \cdot (v_{p}^{C,h+1} - v_{p}^{C,h})}{\text{cost}(G, C + A^*)} \right)^2
\]
\[\leq \frac{2^{-2h+5}}{\Gamma_G} \sum_{i \in M_C} \frac{(1 + \varepsilon)|P_i \cap G|}{\text{cost}(G, C + A^*)} \left( \frac{\text{cost}(P_i \cap G, C) + \text{cost}(P_i \cap G, A^*)}{|P_i \cap G|} \right) \quad \text{(Eq. (16))}
\]
\[\leq \frac{2^{-2h+6}}{\Gamma_G} \sum_{i \in M_C} \frac{(\text{cost}(P_i \cap G, C) + \text{cost}(P_i \cap G, A^*))}{\text{cost}(G, C) + \text{cost}(G, A^*)} \quad (\varepsilon \in (0, 1))
\]
\[\leq \frac{2^{-2h+6}}{\Gamma_G} \cdot \frac{\text{cost}(G, C) + \text{cost}(G, A^*)}{\text{cost}(G, C + A^*)}
\]
\[\leq \frac{2^{-2h+6}}{\Gamma_G} .
\]
In the general case without conditioning on $\xi_G$, we also have

\[
\sum_{p \in S_G \setminus H(S_G, C)} \left( \frac{w(p) \cdot (v_p^{C,h+1} - v_p^{C,h})}{\text{cost}(G, C + A^*)} \right)^2 \leq \frac{2^{-2h+2}}{\Gamma_G} \sum_{i \in M_C} \sum_{p \in P_i \cap S_G} \frac{w(p)}{\text{cost}(G, C + A^*)} \cdot \left( \frac{\text{cost}(P_i \cap G, C)}{|P_i \cap G|} + 4d(p, A^*) \right)
\]

\[
= \frac{2^{-2h+2}}{\Gamma_G} \sum_{i \in M_C} \sum_{p \in P_i \cap S_G} \frac{\text{cost}(G, A^*)}{\Gamma_G \cdot d(p, A^*) \cdot \text{cost}(G, C + A^*)} \cdot \left( \frac{\text{cost}(P_i \cap G, C)}{|P_i \cap G|} + 4d(p, A^*) \right)
\]

\[
\leq \frac{2^{-2h+2}}{\Gamma_G} \sum_{i \in M_C} \sum_{p \in P_i \cap S_G} \frac{4k \cdot \text{cost}(P_i \cap G, C) + 4\text{cost}(G, A^*)}{\Gamma_G \cdot \text{cost}(G, C + A^*)}
\]

\[
\leq \frac{2^{-2h+4} \cdot k}{\Gamma_G},
\]

where the first inequality follows from Ineq. (15), the first equality from the Definition of $w(p)$, the second inequality from Observation 2.3, and the third inequality is due to the fact that $|P_i \cap S_G| \leq \Gamma_G$. This completes the proof. \qed

**Remark 3.15** Lemma 3.14 provides an upper bound of $O\left(\frac{2^{-2h}}{\Gamma_G}\right)$ for $\text{Var} \left[ X_{C,h} \ | \ \xi_G \right]$. One can verify that this bound is almost tight for a large collection of $C \in \mathcal{X}^k$. Suppose $H(G, C) = \emptyset$. For any $i \in [k]$ and point $p \in P_i \cap G$, suppose $d(p, C) \geq 4d(p, A^*)$, we have $d(p, C) = \Omega\left( \frac{\text{cost}(P_i \cap G, C)}{|P_i \cap G|} \right)$ by a reverse argument of Inequality (14). By reversing the argument of the proof of Lemma 3.14, we can verify that conditioning on $\xi_G$,

\[
\sum_{p \in S_G \setminus H(S_G, C)} \left( \frac{w(p) \cdot (v_p^{C,h+1} - v_p^{C,h})}{\text{cost}(G, C + A^*)} \right)^2 = \Omega\left( \frac{2^{-2h}}{\Gamma_G} \right).
\]

The improvement here is mainly due to the new relative covering error $\text{err}(p, C)$, which is smaller than that of [17, 14]. Choosing a smaller error may increase the covering number, however, the covering number may not increase proportionally. We choose the best tradeoff between variance and covering number.

### 4 Improved Coreset Size for Euclidean $(k, z)$-Clustering

In this section, we consider the coreset construction for Euclidean $(k, z)$-Clustering. Following the same reasoning as in [17, 14], we can make the following assumption without loss of generality.

**Assumption 1 (Assumptions on Euclidean datasets)** Let $\varepsilon \in (0, 1)$. We can assume that the given dataset $P$ satisfies
• The number of distinct points $\|P\|_0$ is at most $2^{O(z)} \cdot \text{poly}(k\varepsilon^{-1})$;

• The dimension $d = O(z^2 \varepsilon^{-2} \log \|P\|_0)$;

• $P$ is unweighted.

Now, we state our main theorem for Euclidean coreset as follows.

**Theorem 4.1 (Coreset for Euclidean $(k, z)$-Clustering)** Let $\mathcal{X} = \mathbb{R}^d$, $P \subset \mathbb{R}^d$, $\varepsilon \in (0, 1)$ and constant $z \geq 1$. For each $G \in \mathcal{G}$, let $\Gamma_G = O\left(k^{2/3} \varepsilon^{-2} \log(k\varepsilon^{-1}) \cdot d^{4/3} \varepsilon^{-2}\right)$. With probability at least $0.9$, Algorithm 1 outputs an $\varepsilon$-coreset of $P$ for Euclidean $(k, z)$-CLUSTERING of size

$$2^{O(z)} k^{2/3} \varepsilon^{-2} \log^2(k\varepsilon^{-1}) \log^5 \varepsilon^{-1} = \tilde{O}_z\left(k^{2/3} \varepsilon^{-2}\right).$$

Our coreset size breaks the dependence on $k^2$ in prior work [27, 17, 14] (when $k \leq \varepsilon^{-1}$). Specifically, for $z = 1$, our coreset size is $\tilde{O}(k^{4/3} \varepsilon^{-2})$, which improves the prior upper bound [14] by a factor $k^{2/3}$. For simplicity, we prove the case of $z = 1$ in the main text. The proof for general $z \geq 1$ can be found in Appendix E.

### 4.1 Proof of Theorem 4.1: Coreset for Euclidean $k$-Median

In light of Theorem 3.5, it remains to upper bound the covering number $\mathcal{N}_G^{(m)}(\Gamma_G, \alpha) \text{ and } \mathcal{N}_G^{(o)}(\Gamma_G, \alpha)$, as in the following lemma.

**Lemma 4.2 (Covering number in Euclidean metrics)** For each $G \in \mathcal{G}^{(m)}$ and $0 < \alpha \leq 1$,

$$\log \mathcal{N}_G^{(m)}(\Gamma_G, \alpha) = O\left(k \cdot \min\{k^{1/3} \alpha^{-2} \log \|P\|_0 \log \varepsilon^{-1}, d\} \cdot \log(\varepsilon^{-1} \alpha^{-1})\right).$$

For each $G \in \mathcal{G}^{(o)}$ and any $0 < \alpha \leq 1$,

$$\log \mathcal{N}_G^{(o)}(\Gamma_G, \alpha) = O(k \cdot \min\{\alpha^{-2} \log \|P\|_0, d\} \cdot \log \alpha^{-1}).$$

The proof can be found in Section 4.3. Theorem 4.1 is a direct corollary of the above lemma.

**Proof:** [of Theorem 4.1 for Euclidean $k$-MEDIAN] Fix a main group $G \in \mathcal{G}^{(m)}(j)$ for some integer $\log \varepsilon < j < 2 \log \varepsilon^{-1}$. To apply Theorem 3.5, the key is to upper bound the entropy integral $\int_0^1 \sqrt{\log \mathcal{N}_G^{(m)}(\Gamma_G, \alpha)} \, d\alpha$. We first have

$$\int_0^1 \sqrt{\log \mathcal{N}_G^{(m)}(\Gamma_G, \alpha)} \, d\alpha = \int_0^\varepsilon \sqrt{\log \mathcal{N}_G^{(m)}(\Gamma_G, \alpha)} \, d\alpha + \int_\varepsilon^1 \sqrt{\log \mathcal{N}_G^{(m)}(\Gamma_G, \alpha)} \, d\alpha. \quad (17)$$

Next, we apply Lemma 4.2 to upper bound the above two terms on the right side separately. For the first term, we have

$$\int_0^\varepsilon \sqrt{\log \mathcal{N}_G^{(m)}(\Gamma_G, \alpha)} \, d\alpha = O(1) \cdot \int_0^\varepsilon \sqrt{k d \log(\varepsilon^{-1} \alpha^{-1})} \, d\alpha \quad \text{(Lemma 4.2)}$$

$$= O\left(\sqrt{k \varepsilon^{-2} \log(k\varepsilon^{-1})}\right) \cdot \int_0^\varepsilon \sqrt{\log(\varepsilon^{-1} \alpha^{-1})} \, d\alpha \quad \text{(Asm. 1)} \quad (18)$$

$$= O\left(\sqrt{k \log(k\varepsilon^{-1}) \cdot \log(\varepsilon^{-1})}\right).$$
For the second term, we have the following upper bound

\[
\int_{\epsilon}^{1} \sqrt{\log \mathcal{N}_G^{(m)}(\Gamma_G, \alpha)} \, d\alpha = O \left( \int_{\epsilon}^{1} \sqrt{k^{4/3} \alpha^{-2} \log \| \mathcal{P} \|_0 \log \epsilon^{-1} \log (\epsilon^{-1} \alpha^{-1})} \, d\alpha \right) \quad \text{(Lemma 4.2)}
\]

\[
= O \left( \sqrt{k^{4/3} \log (k \epsilon^{-1}) \log \epsilon^{-1}} \cdot \int_{\epsilon}^{1} \sqrt{\alpha^{-2} \log (\epsilon^{-1} \alpha^{-1})} \, d\alpha \right) \quad \text{(letting } 2^\beta = k^{1/3} \text{ and Asm. 1)}
\]

\[
= O \left( \sqrt{k^{4/3} \log (k \epsilon^{-1}) \log \epsilon^{-1}} \cdot \int_{\epsilon}^{1} \sqrt{\alpha^{-2} \log (\epsilon^{-1})} \, d\alpha \right) \quad \text{Ineq. (18) and (19)}
\]

\[
= O \left( \sqrt{k^{4/3} \log (k \epsilon^{-1}) \log \epsilon^{-1}} \right) \quad \text{(19)}
\]

Now we are ready to upper bound \( \int_{0}^{1} \sqrt{\log \mathcal{N}_G^{(m)}(\Gamma_G, \alpha)} \, d\alpha \). Combining with Inequalities (17) to (19), we conclude that

\[
\int_{0}^{1} \sqrt{\log \mathcal{N}_G^{(m)}(\Gamma_G, \alpha)} \, d\alpha \leq O \left( \sqrt{k \log (k \epsilon^{-1}) \log \epsilon^{-1}} \right) + O \left( \sqrt{k^{4/3} \log (k \epsilon^{-1}) \log \epsilon^{-1}} \right)
\]

\[
\leq O \left( \sqrt{k^{4/3} \log (k \epsilon^{-1}) \log \epsilon^{-1}} \right).
\]

Similarly, we can prove that for an outer group \( G \in G^{(o)} \),

\[
\int_{0}^{1} \sqrt{\log \mathcal{N}_G^{(o)}(\Gamma_G, \alpha)} \, d\alpha = O \left( \sqrt{k \log (k \epsilon^{-1}) \log \epsilon^{-1}} \right).
\]

Consequently, we have that \( \Gamma_G = O \left( k^{4/3} \epsilon^{-2} \log (k \epsilon^{-1}) \log \epsilon^{-1} \right) \) satisfies Inequality (6). This completes the proof.

\[\square\]

### 4.2 Terminal Embedding with Additive Errors

Before proving Lemma 4.2, we introduce two types of terminal embeddings that are useful for dimension reduction.

**Terminal embedding.** Roughly speaking, a terminal embedding projects a point set \( X \subseteq \mathbb{R}^d \) to a low-dimensional space while approximately preserving all pairwise distances between \( X \) and \( \mathbb{R}^d \).

**Definition 4.3 (Terminal embedding)** Let \( \alpha \in (0, 1) \) and \( X \subseteq \mathbb{R}^d \) be a collection of \( n \) points. A mapping \( f : \mathbb{R}^d \to \mathbb{R}^m \) is called an \( \alpha \)-terminal embedding of \( X \) if for any \( p \in X \) and \( q \in \mathbb{R}^d \),

\[
d(p, q) \leq d(f(p), f(q)) \leq (1 + \alpha) \cdot d(p, q).
\]
We have the following recent result on terminal embedding.

**Theorem 4.4 (Optimal terminal embedding [35, 9])** Let $\alpha \in (0, 1)$ and $X \subseteq \mathbb{R}^d$ be a collection of $n$ points. There exists an $\alpha$-terminal embedding $f$ with a target dimension $O(\alpha^{-2} \log n)$. Specifically, $f$ is constructed as an extension of a Johnson-Lindenstrauss (JL) transform with the following properties:

1. Let $g : X \to \mathbb{R}^{m-1}$ be a JL transform.
2. For each $p \in X$, let $f(p) = (g(p), 0)$.
3. For each $q \in \mathbb{R}^d$, the mapping $f(q) \in \mathbb{R}^m$ satisfies that $d(p, q) \leq d(f(p), f(q)) \leq (1 + \alpha) \cdot d(p, q)$ for all $p \in X$.

Accordingly, if $\|X\|_0 = \text{poly}(k\varepsilon^{-1})$, there exists an $\alpha$-terminal embedding of target dimension $O(\alpha^{-2} \log(k\varepsilon^{-1}))$.

**Additive terminal embedding.** We introduce a new notion of terminal embedding with additive error, in order to handle the covering number $\Lambda_G^{(m)}(\Gamma_G, \alpha)$ of a main group. The new variant may be of independent interest.

**Definition 4.5 (Additive terminal embedding)** Let $\alpha \in (0, 1)$, $r > 0$, and $X \subseteq \mathbb{R}^d$ be a collection of $n$ points within a ball $B(0,r)$ with $0^d \in X$. A mapping $f : \mathbb{R}^d \to \mathbb{R}^m$ is called an $\alpha$-additive terminal embedding of $X$ if for any $p \in X$ and $q \in \mathbb{R}^d$, $|d(p, q) - d(f(p), f(q))| \leq \alpha \cdot r$.

The main difference of the above definition from the classic terminal embedding [35, 9] is that we consider an additive error $\alpha \cdot r$ instead of a multiplicative error $\alpha \cdot d(p, q)$. Our error is smaller for remote points $q \in \mathbb{R}^d \setminus B(0,2r)$. Since $p \in B(0,r)$, we know that $d(p, q) \geq r$ by the triangle inequality. Hence, $\alpha \cdot r$ is a smaller error compared to $\alpha \cdot d(p, q)$, which means that an $\alpha$-additive terminal embedding is also an $\alpha$-terminal embedding. We have the following theorem showing that the target dimension of an additive terminal embedding is the same as the multiplicative version, using the same mapping as in [33, 35].

**Theorem 4.6 (Additive terminal embedding)** Let $\alpha \in (0, 1)$, $r > 0$, and $X \subseteq \mathbb{R}^d$ be a collection of $n$ points within a ball $B(0,r)$ with $0^d \in X$. There exists an $\alpha$-additive terminal embedding with a target dimension $O(\alpha^{-2} \log n)$.

**Proof:** Let $m = O(\alpha^{-2} \log n)$. We first let $g : X \to \mathbb{R}^{m-1}$ be a JL transform and assume that the distortion of $g$ is $1 + \alpha$ over the points of $X$. For each $p \in X$, we let $f(p) = (g(p), 0)$. Then we show how to extend $f$ to an additive terminal embedding $f : \mathbb{R}^d \to \mathbb{R}^m$. For each $q \in B(0,2r)$, by Theorem 4.4, we know that there exists a mapping $f(q) \in \mathbb{R}^m$ such that for any $p \in X$,

$$d(p,q) \leq d(f(p),f(q)) \leq (1 + \alpha) \cdot d(p, q) \leq d(p,q) + 8\alpha r,$$

since $d(p, q) \leq 4r$.

Thus, we only need to focus on points $q \in \mathbb{R}^d \setminus B(0,2r)$. By the property of JL transform, we have the following lemma.
Lemma 4.7 (Restatement of [35, Lemma 3.1]) With probability at least 0.9, for any \( q \in \mathbb{R}^d \), there exists \( q' \in \mathbb{R}^{m-1} \) such that \( \|q'\|_2 \leq \|q\|_2 \) and
\[
\forall p \in X, \quad |\langle g(p), q' \rangle - \langle p, q \rangle| \leq \alpha \cdot \|p\|_2 \|q\|_2.
\]

By Lemma 4.7, there exists \( q' \in \mathbb{R}^{m-1} \) such that \( \|q'\| \leq \|q\|_2 \) and for any \( p \in X \),
\[
|\langle g(p), q' \rangle - \langle p, q \rangle| \leq \alpha \cdot \|p\|_2 \|q\|_2 \leq \alpha r \|q\|_2. \tag{20}
\]

Construct \( f(q) = (q', \sqrt{\|q\|_2 - \|q'\|_2}), \) which is the same mapping as in [33, 35]. The construction implies that \( d(f(0), f(q)) = \|f(q)\|_2 = \|q\|_2 = d(0, q) \). It remains to prove the correctness of \( |d(p, q) - d(f(p), f(q))| \leq \alpha \cdot r. \) By construction, we have
\[
d(p, q)^2 = \|p\|^2 + \|q\|^2 - 2\langle p, q \rangle, \tag{21}
\]
and
\[
d(f(p), f(q))^2 = \|f(p)\|^2 + \|f(q)\|^2 - 2\langle f(p), f(q) \rangle. \tag{22}
\]

Since \( \|f(q)\|_2 = \|q\|_2 \), we have
\[
|d(p, q)^2 - d(f(p), f(q))^2| \leq \|p\|^2_2 - \|f(p)\|^2_2 + 2|\langle g(p), q' \rangle - \langle p, q \rangle| \quad \text{(Ineqs. (21) and (22))}
\]
\[
\leq 4\alpha r^2 + 2|\langle g(p), q' \rangle - \langle p, q \rangle| \quad \text{ (p \in P and JL)} \tag{23}
\]
\[
\leq 4\alpha r^2 + 2\alpha r \|q\|_2 \quad \text{(Ineq. (20))}
\]

Hence, we obtain that
\[
|d(p, q) - d(f(p), f(q))| = \frac{|d(p, q)^2 - d(f(p), f(q))^2|}{d(p, q) + d(f(p), f(q))} \leq \frac{4\alpha r^2 + 2\alpha r \|q\|_2}{\|q\|_2 - \|p\|_2 + \|f(q)\|_2 - \|f(p)\|_2} \quad \text{(Ineq. (23) and triangle ineq.)}
\]
\[
= \frac{2\|q\|_2^2 - \|p\|_2^2 - \|f(q)\|_2^2 + \|f(p)\|_2^2}{2\|q\|_2 - 3r} \quad (\|f(q)\|_2 = \|q\|_2)
\]
\[
\leq \frac{4\alpha r^2 + 2\alpha r \|q\|_2}{2\|q\|_2 - 3r} \quad \text{(JL)}
\]
\[
\leq \frac{4\alpha r \|q\|_2}{0.5\|q\|_2} \quad (\|q\|_2 > 2r)
\]
\[
= 8\alpha r.
\]

Overall, we complete the proof. \qed

4.3 Proof of Lemma 4.2: Bounding the Covering Number in Euclidean Case

Recall that \( \Gamma_G \) is the number of samples for each main/outer group \( G \in \mathcal{G} \) (Line 1 of Algorithm 1). By [14, Lemma 21], we can obtain the following upper bound for \( \mathcal{N}^{(o)}_G(\Gamma_G, \alpha) \), which suffices for our purpose.
\[
\log \mathcal{N}^{(o)}_G(\Gamma_G, \alpha) = O \left( k \cdot \min \left\{ \alpha^{-2} \log \|P\|_0, d \right\} \cdot \log \alpha^{-1} \right).
\]
Now, we consider the main groups. Fix $\alpha \in (0,1)$ and let $G \in \mathcal{G}^{(m)}(j)$ for some $j$ be a main group. In the following, we focus on proving the upper bound for the covering number $\mathcal{N}^{(m)}_G(\Gamma_G, \alpha)$. We first have the following lemma that provides a common construction for coverings.

**Lemma 4.8 (α-Covering for Euclidean spaces)** Suppose $d \geq \log k$. Let $\alpha \in (0,1)$ and $v \geq 1$. Let $1 \leq t \leq k$ be an integer, $a_1, \ldots, a_t \in \mathbb{R}^d$ be $t$ centers and $r_1, \ldots, r_t > 0$ be $t$ radius. Let $X = X_1 \cup X_2 \cup \cdots \cup X_t \subset \mathbb{R}^d$ be a dataset that consists of $t$ disjoint subsets $X_1 \subset B(a_1, r_1), X_2 \subset B(a_2, r_2), \ldots, X_t \subset B(a_t, r_t)$. There exists an $\alpha$-covering $V \subset \mathbb{R}^{|X|}$ of $X$ with $\log |V| = Okd \log (u\alpha^{-1})$, i.e., for any $k$-center set $C \subset X$, there exists a cost vector $v \in V$ for any $i \in [t]$ with $d(a_i, C) \leq u \cdot r_i$ and $p \in X_i$ such that

$$|d(p, C) - d(a_i, C) - v_p| \leq \alpha \cdot r_i.$$  

**Proof:** For each $i \in [t]$, take an $\alpha \cdot r_i$-net of the Euclidean ball $B(a_i, 10ur_i)$. Since $d \geq \log k$, the union $\Lambda$ of these nets has size at most

$$k \cdot \exp\left(O\left(d \log (u\alpha^{-1})\right)\right) = \exp\left(O\left(d \cdot \log (u\alpha^{-1})\right)\right).$$

For any $S \subseteq G$, we then define an $\alpha$-covering $V \subset \mathbb{R}^{|S|}$ of $S$ as follows: for any $k$-center set $C \subseteq \Lambda$, we construct a vector $v \in V$ in which the entry corresponding to point $p$ (say $p \in R_{ij} \cap S$) is

$$v_p = d(p, C) - d(a_i, C).$$

Obviously, we have $\log |V| = O(kd \cdot \log (u\alpha^{-1}))$. It remains to verify that $V$ is indeed an $\alpha$-covering of $S$. For any $C = (c_1, \ldots, c_k) \in X^k$, let $C' = (c'_1, \ldots, c'_k) \in V$ such that $c'_i$ is the closest point of $c_i$ in $\Lambda (i \in [k])$. Then for any $i \in [t]$ with $d(a_i, C) \leq u \cdot r_i$ and $p \in X_i$, we have the following observation: for every $c_i \in C$,

1. If $d(p, c_i) \geq (u + 3) \cdot r_i$, then $d(p, c'_i) \geq (u + 2) \cdot r_i \geq d(p, C')$ since $d(p, C') \leq d(a_i, C) + d(p, a_i)$.

2. If $d(p, c) \leq (u + 3) \cdot r_i$, then $d(p, c') \in d(p, c) \pm \alpha \cdot r_i$ by the construction of $\Lambda$.

Consequently, we have

$$d(p, C') \in d(p, C) \pm \frac{\alpha}{2} \cdot r_i.$$

Similarly, we can prove that

$$d(a_i, C') \in d(a_i, C) \pm \frac{\alpha}{2} \cdot r_i.$$

The above two inequalities directly lead to the following conclusion:

$$|d(p, C) - d(a_i, C) - v_p| = |d(p, C) - d(a_i, C) - (d(p, C') - d(a_i, C'))| \leq 2\alpha \cdot r_i \leq \alpha,$$

which completes the proof. □

We go back to upper bound $\mathcal{N}^{(m)}_G(\Gamma_G, \alpha)$.

**The first upper bound for $\mathcal{N}^{(m)}_G(\Gamma_G, \alpha)$**. We first verify

$$\log \mathcal{N}^{(m)}_G(\Gamma_G, \alpha) = O(kd \log (\epsilon^{-1} \alpha^{-1})).$$

This is actually a direct corollary of Lemma 4.8 by letting $X = S$, $X_i = S \cap R_{ij}$, $r_i = 2^j + 1 \Delta_i$ for $i \in [k]$, and $u = 10\epsilon^{-1}$.  

25
Figure 2: An example of Definition 4.9

The second upper bound for $\mathcal{N}_{G}^{r}(\Gamma_{G}, \alpha)$. Next, we verify

$$
\log \mathcal{N}_{G}^{r}(\Gamma_{G}, \alpha) = O(k^{4/3}\alpha^{-2}\log \| P \|_{0} \log \varepsilon^{-1} \log (\varepsilon^{-1} \alpha^{-1})).
$$

For preparation, we propose the following notion of partitions of $G \setminus H(G, C)$ w.r.t. center sets $C \in \mathfrak{X}$ according to the ratio $\frac{d(a^{*}, C)}{\Delta_{i}}$.

**Definition 4.9 (A partition of $G \setminus H(G, C)$)** Let $G \in \mathcal{G}^{(m)}(j)$ for some integer $\log \varepsilon < j < 2 \log \varepsilon^{-1}$ be a main group. For a $k$-center set $C \in \mathfrak{X}$ and integer $\beta \geq 1$, we define

$$
R(G, C, \beta) := \left\{ p \in R_{ij} \cap G \setminus H(G, C) : i \in [k], 2^{\beta+j+1} \Delta_{i} \leq d(a^{*}, C) < 2^{\beta+j+2} \Delta_{i} \right\}.
$$

Also, define

$$
R(G, C, 0) := \left\{ p \in R_{ij} \cap G \setminus H(G, C) : i \in [k], d(a^{*}, C) < 2^{j+2} \Delta_{i} \right\}.
$$

For any subset $S \subseteq G$ and $k$-center set $C \in \mathfrak{X}$, we define $R(S, C, \beta) = S \cap R(G, C, \beta)$ for integer $\beta \geq 0$.

By the above definition, for a main group $G \in \mathcal{G}^{(m)}(j)$ and a given $k$-center set $C \in \mathfrak{X}$, $H(G, C)$ and all $R(G, C, \beta)s (0 \leq \beta \leq 2 + \log \varepsilon^{-1})$ are disjoint, and their union is exactly $G$; see Figure 2 for an example. Intuitively, for all points $p \in R(G, C, \beta)$, the fractions $\frac{d(p, C)}{d(p, A^{*})}$ are “close”, which is an important property for our coreset construction. More concretely, combining with Definition 2.2 and the triangle inequality, we have the following observation.

**Observation 4.10 (Relations between $d(p, C)$ and $d(p, A^{*})$ in partitions $R(G, C, \beta)$)** For a $k$-center set $C \in \mathfrak{X}$,

1. If $p \in R(G, C, 0)$, then $d(p, C) \leq 6d(p, A^{*})$;
2. If $p \in R(G, C, \beta)$ for some integer $1 \leq \beta \leq 2 + \log \varepsilon^{-1}$, then $(2^{\beta} - 1) \cdot d(p, A^{*}) \leq d(p, C) \leq (2^{\beta+2} + 1) \cdot d(p, A^{*})$. 

26
For each $C \in \mathcal{X}^k$, denote a certificate $\phi(C) \in \{0, 1, \ldots, 2 + \lfloor \log \frac{1}{\varepsilon} \rfloor, +\infty\}^{|G|}$ as follows: for each $i \in [k]$ with ring $R_{ij} \in G$,

- If there exists some integer $0 \leq \beta \leq 2 + \log \varepsilon^{-1}$ such that $R_{ij} \subseteq R(G, C, \beta)$, let $\phi_i(C) = \beta$;
- Otherwise, let $\phi_i(C) = +\infty$.

Intuitively, every entry $\phi_i(C)$ reflects the distance between ring $R_{ij}$ and center set $C$. As $C$ is far away from points in $R_{ij}$, $\phi_i(C)$ becomes larger. Note that there are at most $(3 + \log \varepsilon^{-1})^k$ possible $\phi(C)$'s over all $C \in \mathcal{X}^k$, since each $\phi_i(C)$ has at most $3 + \log \varepsilon^{-1}$ different choices. Fix an arbitrary subset $S \subseteq G$ of size at most $\Gamma_G$ and a vector $\phi \in \{0, 1, \ldots, 2 + \lfloor \log \varepsilon^{-1} \rfloor, +\infty\}^k$. In the following, we investigate the covering number induced by all center sets $C \in \mathcal{X}^k$ with $\phi(C) = \phi$. For each integer $0 \leq \beta \leq 2 + \log \varepsilon^{-1}$, let $S_{\phi, \beta}$ denote the collection of $R_{ij} \cap S$ with $\phi_i = \beta$. Note that $S_{\phi, \beta}$s are disjoint for all $\beta$, and the union of all $S_{\phi, \beta}$s and $S \cap H(G, C)$ is exactly $S$. Also note that for each $C \in \mathcal{X}^k$ with $\phi(C) = \phi$, we have $R(S, C, \beta) = S_{\phi, \beta}$ for any integer $0 \leq \beta \leq 2 + \log \varepsilon^{-1}$. We have the following lemma.

**Lemma 4.11 (An upper bound of the covering number for each $\beta$)** Let $G \in \mathcal{G}(m)(j)$ for some integer $\log \varepsilon < j < 2 \log \varepsilon^{-1}$ be a main group and let $S \subseteq G$ be a subset. Fix a vector $\phi \in \{0, 1, \ldots, 2 + \lfloor \log \varepsilon^{-1} \rfloor, +\infty\}^k$ and an integer $0 \leq \beta \leq 2 + \log \varepsilon^{-1}$. There exists an $\alpha$-covering $V_{\phi, \beta} \subset \mathbb{R}^{[S_{\phi, \beta}]}$ of $S_{\phi, \beta}$ satisfying that for each $C \in \mathcal{X}^k$ with $\phi(C) = \phi$, there exists a cost vector $v \in V_{\phi, \beta}$ such that for any $i \in [k]$ and $p \in R_{ij} \cap S_{\phi, \beta}$, Inequality (4) holds, i.e.,

$$|d(p, C) - d(a^*_i, C) - v_p| \leq \alpha \cdot (\sqrt{d(p, C) \cdot d(p, A^*)} + d(p, A^*)) \cdot \sqrt{\frac{\text{cost}(G, C + A^*)}{\text{cost}(G, A^*)}}.$$

Moreover, the size of $V_{\phi, \beta}$ satisfies

$$\log |V_{\phi, \beta}| = O(k \cdot \min \left\{ 2^\beta \alpha^{-2} \log \|P\|_0, (1 + k^2 2^{-2\beta}) \alpha^{-2} \log \|P\|_0 \right\} \cdot \log(\varepsilon^{-1} \alpha^{-1})).$$

By this lemma, we can directly construct an $\alpha$-covering $V$ of $S$ as follows:

1. For each $\phi \in \{0, 1, \ldots, 2 + \lfloor \log \varepsilon^{-1} \rfloor, +\infty\}^k$, construct $V_{\phi}$ to be the collection of cost vectors $v \in \mathbb{R}^{\Gamma_G}$ satisfying that 1) for each integer $0 \leq \beta \leq 2 + \log \varepsilon^{-1}$, $v|_{S_{\phi, \beta}} \in V_{\phi, \beta}$; 2) for $p \in S \cap H(G, C)$, $v_p = 0$.

2. Let $V$ be the union of all $V_{\phi}$s.

By construction, we can upper bound the size of $V$ by

$$\log |V| \leq \log(3 + \log \varepsilon^{-1})^k + \sum_{0 \leq \beta \leq 2 + \log \varepsilon^{-1}} \log |V_{\phi, \beta}| \leq O(k \log \varepsilon^{-1}) + \sum_{0 \leq \beta \leq 2 + \log \varepsilon^{-1}} O(k \cdot \min \left\{ 2^\beta \alpha^{-2} \log \|P\|_0, (1 + k^2 2^{-2\beta}) \alpha^{-2} \log \|P\|_0 \right\} \log(\varepsilon^{-1} \alpha^{-1}))$$

$$\leq O(k \log \varepsilon^{-1}) + \sum_{0 \leq \beta \leq 2 + \log \varepsilon^{-1}} O(k^{4/3} \alpha^{-2} \log \|P\|_0 \log(\varepsilon^{-1} \alpha^{-1}))$$

$$\leq O \left( k^{4/3} \alpha^{-2} \log \|P\|_0 \log(\varepsilon^{-1} \alpha^{-1}) \right),$$

27
where the second inequality holds due to Lemma 4.11 and the third inequality follows from letting $2^\beta = k^{1/3}$ and Assumption 1. This completes the proof of Inequality (25). Hence, it remains to prove Lemma 4.11.

**Proof:** [of Lemma 4.11]

**Case** $\beta = 0$. Let $f$ be an $\alpha$-terminal embedding of $A^* \cup S_{\phi,0}$ into $m = O(\alpha^{-2} \log ||P||_0)$ dimensions given by Theorem 4.4. Given a $k$-center set $C \in \mathcal{X}^k$ with $\phi(C) = \phi$, we have that for any $i \in [k]$ and $p \in R_{ij} \cap S_{\phi,0}$,

$$d(f(p), f(C)) = \min_{q \in C} d(f(p), f(q)) \in (1 \pm \alpha) \cdot \min_{q \in C} d(p, q) \quad \text{(Theorem 4.4)}$$

and

$$d(f(a^*_i), f(C)) \in d(a^*_i, C) \pm O(\alpha) \cdot d(p, A^*).$$

Moreover, by the definition of $S_{\phi,0}$, we know that $d(f(p), f(C)) \leq O(1) \cdot 2^i \Delta_i$.

By applying Lemma 4.8 with $d = m$, $X = S_{\phi,0}$, $X_i = R_{ij} \cap S_{\phi,0}$, $r_i = 2^{i+4} \Delta_i$ for $i \in [k]$, and $u = O(1)$, we can construct an $\alpha$-covering $V_{\phi,0} \subset \mathbb{R}^{||S_{\phi,0}||}$ of $f(S)$ satisfying that

1. $|V_{\phi,0}| = \exp \left(O(k m \log (u \alpha^{-1}))\right) = \exp \left(O(k \alpha^{-2} \log ||P||_0 \log (\alpha^{-1}))\right)$;

2. For any $C \in \mathcal{X}^k$ with $\phi(C) = \phi$, there exists $v \in V_{\phi,0}$ such that for any $i \in [k]$ and $p \in R_{ij} \cap S_{\phi,0}$,

$$|d(f(p), f(C)) - d(f(a^*_i), f(C)) - v_p| \leq \alpha \cdot r_i \leq O(\alpha) \cdot d(p, A^*). \quad \text{(Ineq. (26))}$$

It remains to verify that $V_{\phi,0}$ is an $O(\alpha)$-covering of $S$. For each $C \in \mathcal{X}^k$ with $\phi(C) = \phi$, we find a cost vector $v \in V_{\phi,0}$ satisfying Inequality (27). Consequently, we have for any $i \in [k]$ and $p \in R_{ij} \cap S_{\phi,0}$,

$$|d(p, C) - d(a^*_i, C) - v_p| \leq |d(f(p), f(C)) - d(f(a^*_i), f(C)) - v_p| + O(\alpha) \cdot d(p, A^*) \quad \text{(Ineq. (27))}$$

which completes the proof.

**Case** $\beta \geq 1$. We first construct an $\alpha$-covering $V_{\phi, \beta}$ of $S_{\phi, \beta}$ with

$$\log |V_{\phi, \beta}| = O(k^{2^\beta \alpha^{-2}} \log ||P||_0 \cdot \log (\varepsilon^{-1} \alpha^{-1})). \quad \text{(28)}$$

Let $f$ be an $2^{-\beta/2} \alpha$-terminal embedding of $A^* \cup S_{\phi, \beta}$ into $m = O(2^{\beta} \alpha^{-2} \log ||P||_0)$ dimensions given by Theorem 4.4. Given a $k$-center set $C \in \mathcal{X}^k$ with $\phi(C) = \phi$, we have that for any $i \in [k]$ and $p \in R_{ij} \cap S_{\phi, \beta}$,

$$d(f(p), f(C)) \in (1 \pm 2^{-\beta/2} \alpha) \cdot d(p, C) \quad \text{(Observation 4.10)}$$

$$\in d(p, C) \pm O(\alpha) \cdot \sqrt{d(p, C)} \cdot d(p, A^*), \quad \text{(Observation 4.10)}$$

28
and
\[ d(f(a_i^*), f(C)) \in d(a_i^*, C) \pm O(\alpha) \cdot \sqrt{d(p, C) \cdot d(p, A^*)}. \]
Moreover, by the definition of \( S_{\phi, \beta} \), we know that \( d(f(p), f(C)) \leq O(1) \cdot 2^{j+1} \Delta_i \).

By applying Lemma 4.8 with \( d = m \), \( X = S_{\phi, \beta}, X_i = R_{ij} \cap S_{\phi, \beta}, r_i = 2^{j+1} \Delta_i \) for \( i \in [k] \), and \( u = O(2^\beta) \), we can construct an \( \alpha \)-covering \( V_{\phi, \beta} \subset \mathbb{R}^{|S_{\phi, \beta}|} \) of \( f(S) \) with
\[ |V_{\phi, \beta}| = \exp \left( O(km \log(2^\beta \alpha^{-1})) \right) = \exp \left( O(k2^\beta \alpha^{-2} \log \|P\|_0 \cdot \log(e^{-1} \alpha^{-1})) \right); \]

and for any \( C \in \mathcal{X}^k \) with \( \phi(C) = \phi \), there exists a cost vector \( v \in V_{\phi, \beta} \) such that for any \( i \in [k] \) and \( p \in R_{ij} \cap S_{\phi, \beta}, \)
\[ |d(f(p), f(C)) - d(f(a_i^*), f(C)) - v_p| \leq \alpha \cdot r_i \leq O(\alpha) \cdot d(p, A^*). \]

Combining with Inequality (29), we conclude that \( V_{\phi, \beta} \) is also an \( O(\alpha) \)-covering of \( S_{\phi, \beta} \), which completes the proof of Inequality (28).

Finally, we construct an \( \alpha \)-covering \( V_{\phi, \beta} \) of \( S_{\phi, \beta} \) with
\[ \log |V_{\phi, \beta}| = O(k^2 2^{-2\beta} \alpha^{-2} \log \|P\|_0 \cdot \log(e^{-1} \alpha^{-1})). \tag{30} \]
We first note that for any \( C \in \mathcal{X}^k \) with \( \phi(C) = \phi \),
\[
\frac{\text{cost}(G, C)}{\text{cost}(G, A^*)} \geq \sum_{i \in B} \frac{\text{cost}(R_{ij} \cap G, C)}{\text{cost}(G, A^*)} \\
\geq \sum_{i \in B} \frac{\text{cost}(R_{ij} \cap G, C)}{2k \cdot \text{cost}(R_{ij} \cap G, A^*)} \quad \text{(Definition 2.2)} \\
\geq \sum_{i \in B} \frac{2^\beta - 1}{2k} \quad \text{(Observation 4.10)} \\
\geq \frac{|B|2^\beta - 2}{k}. \tag{31}
\]
Let \( B \) be the collection of \( i \in [k] \) with \( \phi_i = \beta \). For each \( i \in B \), let \( g_i \) be an \( \sqrt{|B|k^{-1}2^\beta} \alpha \)-additive terminal embedding of \( T_i = \{ p - a_i^*: p \in R_{ij} \cap S_{\phi, \beta} \text{ or } p = a_i^* \} \) into \( m = O(|B|k^{-1}2^{-2\beta} \alpha^{-2} \log \|P\|_0) \) dimensions given by Theorem 4.6. Here, we can assume \( \sqrt{|B|k^{-1}2^\beta} \alpha = \Omega(1) \). This is because we only need to consider the range of \( \alpha \) satisfying that there exists some \( p \in S_{\phi, \beta} \) and some \( C \in \mathcal{X}^k \) with \( \phi(C) = \phi \) such that \( \alpha \cdot \text{err}(p, C) \leq d(p, A^*) \), which implies \( \sqrt{|B|k^{-1}2^\beta} \alpha = \Omega(1) \) by the definition of \( \text{err}(p, C) \).

Let \( r = 2^{j+1} \Delta_i \). Next, we define a function \( f_i \) as follows: Recall that \( G \in G^{(m)}(j) \), and hence, \( R_{ij} \cap S_{\phi, \beta} \in B(a_i^*, r) \). Then by Theorem 4.6, we have that for any \( p \in R_{ij} \cap S_{\phi, \beta} \) or \( p = a_i^* \), and any \( q \in \mathbb{R}^d, \)
\[ |d(p, q) - d(f_i(p), f_i(q))| \leq \sqrt{|B|k^{-1}2^\beta} \alpha r. \]
Consequently, for any \( p \in R_{ij} \cap S_{\phi, \beta} \) or \( p = a^*_i \),
\[
|d(f_i(p), f_i(C)) - d(p, C)| \leq 2 \cdot \max_{c \in C} |d(f_i(p), f_i(c)) - d(p, c)|
\leq \sqrt{|B|}k^{-1}2^{\beta+1}\alpha r
\leq O(\alpha) \cdot \sqrt{d(p, C) \cdot d(p, A^*)} \cdot \sqrt{\frac{|B|2^{\beta-2}}{k}}
\leq O(\alpha) \cdot \sqrt{d(p, C) \cdot d(p, A^*)} \cdot \sqrt{\frac{\text{cost}(G, C)}{\text{cost}(G, A^*)}}.
\]

(32)

Moreover, by the definition of \( S_{\phi, \beta} \), we know that \( d(f_i(p), f_i(C)) \leq O(1) \cdot 2^{i+\beta} \Delta_i \). By Equation (24), we can construct an \( \alpha \)-covering \( V_i \subseteq \mathbb{R}^{|R_{ij} \cap S_{\phi, \beta}|} \) of \( f_i(R_{ij} \cap S_{\phi, \beta}) \) with
\[
|V_i| = \exp \left( O(km \log(\varepsilon^{-1}\alpha^{-1})) \right) = \exp \left( O(|B|^{-1}k^22^{-2\beta}2^{-2}\alpha^{-2} \log \|P\|_0 \cdot \log(\varepsilon^{-1}\alpha^{-1})) \right)
\]
such that for any \( C \in \mathcal{X}^k \) with \( \phi(C) = \phi \), there exists a cost vector \( v \in V_{\phi, \beta} \) such that for any \( i \in [k] \) and \( p \in R_{ij} \cap S_{\phi, \beta} \),
\[
|d(f_i(p), f_i(C)) - d(f_i(a^*_i), f_i(C)) - v_p| \leq O(\alpha) \cdot 2^i \Delta_i \leq O(\alpha) \cdot d(p, A^*).
\]

Combining with Inequality (32), we conclude that \( V_i \) is also an \( O(\alpha) \)-covering of \( R_{ij} \cap S_{\phi, \beta} \). Now we can construct an \( \alpha \)-covering \( V_{\phi, \beta} \subseteq \mathbb{R}^{|S_{\phi, \beta}|} \) of \( S_{\phi, \beta} \) as the Cartesian Product of all \( V_i \)
\[
V_{\phi, \beta} = \prod_{i \in B} V_i = \{(v_1, \ldots, v_{|B|}) : v_i \in V_i \text{ for all } i \in B\}.
\]

(33)

By the construction of \( V_{\phi, \beta} \), we know that for each \( C \in \mathcal{X}^k \) with \( \phi(C) = \phi \), there exists a cost vector \( v^{(i)} \in \mathbb{R}^{|R_{ij} \cap S_{\phi, \beta}|} \) such that for any \( p \in R_{ij} \cap S_{\phi, \beta} \), Inequality (4) holds. Since \( \prod_{i \in B} v^{(i)} \in V_{\phi, \beta} \) by construction, we have that \( V_{\phi, \beta} \) is indeed an \( \alpha \)-covering of \( S_{\phi, \beta} \). By Inequality (33) and the construction of \( V_{\phi, \beta} \), it is obvious that \( |V_{\phi, \beta}| \) satisfies Inequality (30). Overall, we complete the proof. \( \square \)

5 Optimal Coresets in Doubling and General Discrete Metrics

We first consider the coreset construction for \((k, z)\)-CLUSTERING in doubling metrics. The main theorem is as follows.

**Theorem 5.1 (Coreset for \((k, z)\)-CLUSTERING in doubling metrics)** Let \( \varepsilon \in (0, 1) \) and constant \( \beta \geq 1 \). Suppose the doubling dimension of \((\mathcal{X}, d)\) is \( \text{ddim} \) and \( P \subseteq \mathcal{X} \) is a set of \( n \) points. Let \( \Gamma_G = O(k \cdot \text{ddim} \cdot \varepsilon^{-2} \log(k\varepsilon^{-1})) \) for each \( G \in G \). With probability at least 0.9, Algorithm 1 outputs an \( \varepsilon \)-coreset of \( P \) for \((k, z)\)-CLUSTERING of size
\[
2^{|O\varepsilon\text{ddim} \cdot \varepsilon^{-2} \log(k\varepsilon^{-1}) \log \varepsilon^{-1}} = \tilde{O}_z(k \cdot \text{ddim} \cdot \varepsilon^{-2}).
\]

30
For constant $z \geq 1$, our coreset size matches the lower bound in [14] up to poly-log factors. Different from the Euclidean metrics that require a pre-processing of dimension reduction for Assumption 1, we can directly apply Algorithm 1 in the doubling metrics. Hence, our algorithm is a one-stage sampling algorithm for coreset construction in doubling metrics. On the other hand, the analysis in [17, Corollary 4] assumes that the algorithm is run on a coreset found by any other coreset algorithm, leading to a two-stage algorithm.

Theorem 5.1 is a consequence of Theorem 3.6 and the following lemma which upper bounds the covering number $N_G^{(m)}(\Gamma_G, \alpha)$.

**Lemma 5.2 (Covering number in doubling metrics)** For each main group $G \in G^{(m)}$, constant $0 < \alpha \leq 1$, and integer $0 \leq \beta < \log(z \varepsilon^{-1})$,

$$\log N_G^{(m)}(\Gamma_G, \alpha) = O(z k \cdot \text{ddim} \cdot \log(\varepsilon^{-1} \alpha^{-1}) + k \log k).$$

For each outer group $G \in G^{(o)}$ and any constant $0 < \alpha \leq 1$,

$$\log N_G^{(o)}(\Gamma_G, \alpha) = O(z k \cdot \text{ddim} \cdot \log(\alpha^{-1}) + k \log k).$$

**Proof:** [of Lemma 5.2] The proof idea is the same as that for the Euclidean case (Lemma E.1, which is a generalized version of Lemma 4.2 to constant $z \geq 1$). We just replace $d$ by $\text{ddim}$ and construct the $\alpha$-covering $V$. Due to the packing property of doubling metrics, we know that $|V|$ has size at most

$$k \cdot \exp \left( O(z \cdot \text{ddim} \cdot \log(\varepsilon^{-1} \alpha^{-1})) \right) = \exp \left( O(z \cdot \text{ddim} \cdot \log(\varepsilon^{-1} \alpha^{-1})) + \log k \right),$$

which implies the lemma. □

Theorem 5.1 is a direct corollary of the above lemma.

**Proof:** [of Theorem 5.1] Fix a main group $G \in G^{(m)}(j)$ for some integer $z \log(\varepsilon/z) < j < 2z \log(\varepsilon^{-1})$. We have

$$\int_0^1 \sqrt{\log N_G^{(m)}(\Gamma_G, \alpha)} d\alpha \leq O(\sqrt{z}) \cdot \int_0^1 \sqrt{k \cdot \text{ddim} \cdot \log(\varepsilon^{-1} \alpha^{-1}) + k \log k} d\alpha 
\leq O(\sqrt{k \cdot \text{ddim} \cdot \log(\varepsilon^{-1}) + k \log k}). \quad \text{(Ineq. (18))}$$

Consequently, we have that $\Gamma_G = O(\varepsilon^{-2} \cdot k \cdot \text{ddim} \cdot \log(\varepsilon^{-1}))$ satisfies

$$\Gamma_G \geq O \left( \varepsilon^{-2} \left( \int_0^1 \sqrt{\log N_G^{(m)}(\Gamma_G, \alpha)} d\alpha \right)^2 + k \varepsilon^{-2} \log(k \varepsilon^{-1}) \right).$$

This completes the proof. The proof for outer groups $G \in G^{(o)}$ is similar. □

We finish this section by proving Lemma 5.2.

---

4The following covering numbers $N_G^{(m)}(\Gamma_G, \alpha)$ and $N_G^{(o)}(\Gamma_G, \alpha)$ are defined in Section D.1 (see Definition D.2), which generalizes Definitions 3.2 and 3.4 to constant $z \geq 1$. 31
General discrete metric spaces. Suppose \((\mathcal{X}, d)\) is a general discrete metric space and \(P \subseteq \mathcal{X}\). Since the doubling dimension of \((\mathcal{X}, d)\) is \(O(\log |\mathcal{X}|)\), we have the following corollary from Theorem 5.1.

**Corollary 5.3 (Coreset for \((k, z)\)-Clustering in general discrete metrics) Let \(\varepsilon \in (0, 1)\) and constant \(z \geq 1\). Let \((\mathcal{X}, d)\) be a general discrete metric space and \(P \subseteq \mathcal{X}\). Let \(\Gamma_G = O\left(k \cdot \log |\mathcal{X}| \cdot \varepsilon^{-2} \log(ke^{-1})\right)\) for each \(G \in \mathcal{G}\). With probability at least 0.9, Algorithm 1 outputs an \(\varepsilon\)-coreset of \(P\) for \((k, z)\)-Clustering of size \(2^O(z)k \cdot \log |\mathcal{X}| \cdot \varepsilon^{-2} \log^2(ke^{-1}) \log \varepsilon^{-1}\).

For constant \(z \geq 1\), the coreset size is \(\tilde{O}(k \log |\mathcal{X}| \cdot \varepsilon^{-2})\), which matches the lower bound in [14].

### 6 Coreset for \((k, z)\)-Clustering in Shortest Path Metrics with Bounded Treewidth

In this section, we apply Theorem 3.5 to shortest path metrics defined by undirected graphs with bounded treewidth. Constructing coresets in such metrics has been studied in [2, 17]. We are given an edge-weighted undirected graph \(G = (\mathcal{X}, E)\) and a subset \(P \subseteq \mathcal{X}\), together with a distance function \(d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}\), where \(d(p, q)\) is the shortest path distance between \(p, q \in \mathcal{X}\). Note that \((\mathcal{X}, d)\) is a metric space. Suppose the treewidth of \(G\) is \(tw \geq 1\). The state-of-the-art coreset size for \((k, z)\)-Clustering w.r.t. \((\mathcal{X}, d)\) is \(\tilde{O}\left(k \cdot tw \cdot \varepsilon^{-2} \min\{k, 1 + \varepsilon^{-2}\}\right)\). We have the following theorem that removes the term \(\min\{k, 1 + \varepsilon^{-2}\}\) in the coreset size.

**Theorem 6.1 (Coreset for \((k, z)\)-Clustering in graphs with bounded treewidth) Let \(\varepsilon \in (0, 1)\) and constant \(z \geq 1\). Let \(G = (\mathcal{X}, E)\) be a given edge-weighted graph with treewidth at most \(tw \geq 1\), \((\mathcal{X}, d)\) be the shortest path metric on \(G\), and \(P \subseteq \mathcal{X}\). Let \(\Gamma_G = O\left(k \cdot \log \log(\alpha^{-1} \cdot \varepsilon^{-1})\right)\) for each \(G \in \mathcal{G}\). With probability at least 0.9, Algorithm 1 outputs an \(\varepsilon\)-coreset of \(P\) for \((k, z)\)-Clustering of size \(2^O(z)k \cdot \log(ke^{-1}tw) \log(ke^{-1}) \log \varepsilon^{-1}\).

For constant \(z \geq 1\), our coreset size is \(\tilde{O}(k \cdot \log \log(\alpha^{-1} \cdot \varepsilon^{-1}))\). The current known lower bound is \(O(k \cdot \log \log(\alpha^{-1} \cdot \varepsilon^{-1}))\) [2], which leaves an \(\varepsilon^{-1}\) gap. Moreover, for this metric, we obtain a one-stage sampling algorithm, which the analysis in [17, Corollary 5] requires two.

Again, we have the following lemma that upper bounds the covering numbers.

**Lemma 6.2 (Covering number for graphs with bounded treewidth) For each \(G \in \mathcal{G}^{(m)}\), \(0 < \alpha \leq 1\), and integer \(0 \leq \beta < \log(\varepsilon^{-1})\),
\[
\log \mathcal{N}_G^{(m)}(\Gamma_G, \alpha) = O\left(ztw \log(\varepsilon^{-1} \alpha^{-1}) + k \log(\Gamma_G + k)\right).
\]

For each \(G \in \mathcal{G}^{(o)}\) and any \(0 < \alpha \leq 1\),
\[
\log \mathcal{N}_G^{(o)}(\Gamma_G, \alpha) = O\left(ztw \log(\alpha^{-1}) + k \log(\Gamma_G + k)\right).
\]

Note that by selecting \(\Gamma_G\) as in Theorem 6.1, the term \(k \log \Gamma_G = \Omega(k \log(ktw\varepsilon^{-1}))\) in the above lemma. Thus, using similar argument as for doubling metrics, we can see that Theorem 6.1 is a direct corollary of the above lemma. It remains to prove Lemma 6.2.
Proof: [of Lemma 6.2] Again, we only need to prove the bound for main groups. Fix a main group $G \in \mathcal{G}^{(m)}(j)$ for some integer $z \log(\varepsilon/z) < j < 2z \log(\varepsilon z^{-1})$ and $0 \leq \alpha \leq 1$. Fix a subset $S \subseteq G$ of size $\Gamma_G$. For each $C \in \mathcal{X}^k$, let $M_C$ denote the collection of $i \in [k]$ with $R_{ij} \cap S \setminus H(G, C) \neq \emptyset$.

The proof follows almost the same lines as in [17, Lemma 18]. Fix a subset $S \subseteq G$ of size at most $\Gamma_G$. They can construct a covering $V' \subseteq \mathbb{R}^{|S|}$ such that for any $C \in \mathcal{X}^k$, there exists a cost vector $v' \in V'$ satisfying that for any $i \in [k]$ and $p \in R_{ij} \cap S \setminus H(G, C)$ or $p = a^*_i$,

$$|d(p, C) - v'_p| \leq \frac{\alpha}{z} 2^j \Delta_i,$$

which implies that

$$|d^x(p, C) - (v'_p)^x| \leq O(\alpha) \cdot (d^x(p, C) + 2^j \Delta_i).$$

They prove that the size $|V'| = \text{poly}(|S| + k) \cdot (\frac{z}{\varepsilon \alpha})^{O(tw)}$ [17, Lemma 18].

We only need to do a slight modification to their construction: Construct a covering $V \subseteq \mathbb{R}^{|S|+k}$ such that for any $C \in \mathcal{X}^k$, there exists a cost vector $v \in V$ satisfying that for any $i \in M_C$ and $p \in \{a^*_i\} \cup (R_{ij} \cap S \setminus H(G, C))$,

$$|d(p, C) - v_p| \leq \frac{\alpha 2^{j-z(\beta+6)}}{z} \Delta_i,$$

which implies that

$$|d^x(p, C) - d^x(a^*_i, C) - (v_p)^x + (v_{a^*_i})^x| \leq O(\alpha) \cdot (\sqrt{d^x(p, C) \cdot d^x(p, A^*)} + d^x(p, A^*)).$$

This is exactly what we want in Definition D.2. Since $z = O(1)$ and $0 \leq \beta < \log(\varepsilon z^{-1})$, the following size bound still holds:

$$|V|^k \leq \left(\text{poly}(|S| + k) \cdot (\frac{z}{\varepsilon \alpha})^{O(tw)}\right)^k \leq \left(\text{poly}(\Gamma_G + k) \cdot (\frac{z}{\varepsilon \alpha})^{O(tw)}\right)^k.$$

Thus, we know that

$$\log \mathcal{N}_G^{(m)}(\Gamma_G, \alpha) \leq \log(|V|^k) = O\left(\frac{z}{\varepsilon \alpha} \log(\varepsilon^{-1} \alpha^{-1}) + k \log(\Gamma_G + k)\right),$$

which completes the proof. \qed

References

[1] David Arthur and Sergei Vassilvitskii. $k$-means++: the advantages of careful seeding. In SODA, pages 1027–1035, 2007.

[2] Daniel N. Baker, Vladimir Braverman, Lingxiao Huang, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for clustering in graphs of bounded treewidth. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of Proceedings of Machine Learning Research, pages 569–579. PMLR, 2020.

5Note that [17, Lemma 18] suppose $\alpha = \varepsilon$. The above holds by using $\frac{2}{z} \cdot \text{dist}(x, A^*)$ as multiples instead of $\frac{z}{\varepsilon} \cdot \text{dist}(x, A^*)$ in their arguments.
[3] Johannes Blömer, Sascha Brauer, and Kathrin Bujna. Coresets for fuzzy $k$-means with applications. In Wen-Lian Hsu, Der-Tsai Lee, and Chung-Shou Liao, editors, 29th International Symposium on Algorithms and Computation, ISAAC 2018, December 16-19, 2018, Jiaoxi, Yilan, Taiwan, volume 123 of LIPIcs, pages 46:1–46:12. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018.

[4] Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near-optimal coresets for least-squares regression. IEEE transactions on information theory, 59(10):6880–6892, 2013.

[5] Vladimir Braverman, Vincent Cohen-Addad, Shaofeng Jiang, Robert Krauthgamer, Chris Schwiegelshohn, Mads Bech Toftrup, and Xuan Wu. The power of uniform sampling for coresets. In 62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2022. IEEE Computer Society, 2022.

[6] Vladimir Braverman, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for ordered weighted clustering. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of Proceedings of Machine Learning Research, pages 744–753. PMLR, 2019.

[7] Vladimir Braverman, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for clustering in excluded-minor graphs and beyond. In SODA, pages 2679–2696. SIAM, 2021.

[8] 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.

[9] Yeshwanth Cherapanamjeri and Jelani Nelson. Terminal embeddings in sublinear time. In 62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2021, Denver, CO, USA, February 7-10, 2022, pages 1209–1216. IEEE, 2021.

[10] Rachit Chhaya, Anirban Dasgupta, and Supratim Shit. On coresets for regularized regression. In International conference on machine learning, pages 1866–1876. PMLR, 2020.

[11] Adam Coates and Andrew Y. Ng. Learning feature representations with $k$-means. In Grégoire Montavon, Genevieve B. Orr, and Klaus-Robert Müller, editors, Neural Networks: Tricks of the Trade - Second Edition, volume 7700 of Lecture Notes in Computer Science, pages 561–580. Springer, 2012.

[12] Michael B Cohen, Yin Tat Lee, Cameron Musco, Christopher Musco, Richard Peng, and Aaron Sidford. Uniform sampling for matrix approximation. In Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science, pages 181–190. ACM, 2015.

[13] Michael B Cohen, Cameron Musco, and Christopher Musco. Input sparsity time low-rank approximation via ridge leverage score sampling. In Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 1758–1777. SIAM, 2017.

[14] Vincent Cohen-Addad, Kasper Green Larsen, David Saulpic, and Chris Schwiegelshohn. Towards optimal lower bounds for $k$-median and $k$-means coresets. In STOC ’22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 - 24, 2022, pages 1038–1051. ACM, 2022.
[15] Vincent Cohen-Addad, Kasper Green Larsen, David Saulpic, Chris Schwiegelshohn, and Omar Ali Sheikh-Omar. Improved coresets for euclidean $k$-means. In Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, 2022.

[16] 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.

[17] Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. A new coreset framework for clustering. In Samir Khuller and Virginia Vassilevska Williams, editors, STOC ’21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021, pages 169–182. ACM, 2021.

[18] Petros Drineas, Michael W Mahoney, and Shan Muthukrishnan. Sampling algorithms for $\ell_2$ regression and applications. In Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm, pages 1127–1136. Society for Industrial and Applied Mathematics, 2006.

[19] Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In Proceedings of the forty-third annual ACM symposium on Theory of computing, pages 569–578. ACM, 2011. https://arxiv.org/abs/1106.1379.

[20] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for $k$-means, pca and projective clustering. In Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 1434–1453. SIAM, 2013.

[21] Sariel Har-Peled and Soham Mazumdar. On coresets for $k$-means and $k$-median clustering. In 36th Annual ACM Symposium on Theory of Computing,, pages 291–300, 2004.

[22] Lingxiao Huang, Shaofeng H.-C. Jiang, Jian Li, and Xuan Wu. $\varepsilon$-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. IEEE Computer Society, 2018.

[23] Lingxiao Huang, Shaofeng H.-C. Jiang, Jianing Lou, and Xuan Wu. Near-optimal coresets for robust clustering. CoRR, abs/2210.10394, 2022.

[24] Lingxiao Huang, Shaofeng H.-C. Jiang, and Nisheeth K. Vishnoi. Coresets for clustering with fairness constraints. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, editors, Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 7587–7598, 2019.

[25] Lingxiao Huang, K Sudhir, and Nisheeth Vishnoi. Coresets for regressions with panel data. Advances in Neural Information Processing Systems, 33:325–337, 2020.
[26] Lingxiao Huang, K. Sudhir, and Nisheeth K. Vishnoi. Coresets for time series clustering. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 22849–22862, 2021.

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

[28] Ibrahim Jubran, Alaa Maalouf, and Dan Feldman. Fast and accurate least-mean-squares solvers. In *Advances in Neural Information Processing Systems*, pages 8305–8316, 2019.

[29] Michael Langberg and Leonard J Schulman. Universal $\varepsilon$-approximators for integrals. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 598–607. SIAM, 2010.

[30] Mu Li, Gary L Miller, and Richard Peng. Iterative row sampling. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 127–136. IEEE, 2013.

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

[32] Mario Lucic, Matthew Faulkner, Andreas Krause, and Dan Feldman. Training Gaussian mixture models at scale via coresets. *The Journal of Machine Learning Research*, 18(1):5885–5909, 2017.

[33] Sepideh Mahabadi, Konstantin Makarychev, Yury Makarychev, and Ilya Razenshteyn. Nonlinear dimension reduction via outer bi-lipschitz extensions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1088–1101, 2018.

[34] Pascal Massart, Jean Picard, and École d’été de probabilités de Saint-Flour. Concentration inequalities and model selection. 2007.

[35] Shyam Narayanan and Jelani Nelson. Optimal terminal dimensionality reduction in Euclidean space. *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, 2019.

[36] Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. Fair coresets and streaming algorithms for fair k-means. In Evripidis Bampis and Nicole Megow, editors, *Approximation and Online Algorithms - 17th International Workshop, WAOA 2019, Munich, Germany, September 12-13, 2019, Revised Selected Papers*, volume 11926 of *Lecture Notes in Computer Science*, pages 232–251. Springer, 2019.

[37] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, et al. Cluster analysis: basic concepts and algorithms. *Introduction to data mining*, 8:487–568, 2006.
[38] Murad Tukan, Xuan Wu, Samson Zhou, Vladimir Braverman, and Dan Feldman. New coresets for projective clustering and applications. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event, volume 151 of Proceedings of Machine Learning Research, pages 5391–5415. PMLR, 2022.

[39] Ramon van Handel. Probability in high dimension. 2014.
A  Extension of Algorithm 1 to Any Constant $z \geq 1$

We need to generalize the notations of rings and groups. Again, we let $A^* = \{a_1^*, \ldots, a_k^* \in X\} \in X^k$ be a constant approximation for the $(k, z)$-CLUSTERING problem.

Ring structure and group structure for general $z \geq 1$. We let $\Delta_i = \frac{\text{cost}_z(p, A^*)}{|P_i|}$ and define ring $R_{ij} := \{p \in P_i : 2^j \Delta_i \leq \delta_z(p, A^*) < 2^{j+1} \Delta_i\}$. The range of $j$ for main rings is generalized to be $z \log(\epsilon/z) < j < 2z \log(z \epsilon^{-1})$.

For the group structure, we generalize the range of $b$ to be $z \log(\epsilon/4z) - \log k < b \leq 0$. Observation 2.3 is generalized as follows.

Observation A.1 (Main group cost for general $z \geq 1$ [14]) Let $G \in \mathcal{G}^{(m)}$ be a main group. Let $i \in [k]$ be an integer satisfying that $P_i \cap G \neq \emptyset$. For any $p \in P_i \cap G$, we have

$$\text{cost}_z(G, A^*) \leq 2k \cdot \text{cost}_z(P_i \cap G, A^*) \leq 4k \cdot |P_i \cap G| \cdot \delta_z(p, A^*).$$

We also generalize Lemma 2.4.

Lemma A.2 (Group number for general $z \geq 1$) For general $z \geq 1$, there exist at most $O(z^2 \log(ke^{-1}) \log(ze^{-1}))$ groups in $\mathcal{G}$.

By using the above general group structure $\mathcal{G}$, we can generalize Algorithm 1 to any constant $z \geq 1$. The only difference is that in Line 1 of Algorithm 1, we let the probability be $\frac{\delta_z(p, A^*)}{\text{cost}_z(G, A^*)}$ and let weights $w(p) = \frac{\text{cost}_z(G, A^*)}{\Gamma_G \cdot \delta_z(p, A^*)}$.

B  Proof of Theorem 3.5

We complete the missing details of the proof of Theorem 3.5 from Lemma 3.7. in Section 3.3. The proof is almost the same as that in [14].
Proof: [of Theorem 3.5] We have

\[
\mathbb{E}_S \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}(P, C + A^*)} \cdot \sum_{p \in S} w(p) \cdot d(p, C) - \text{cost}(P, C) \right] \leq \mathbb{E}_S \sup_{C \in \mathcal{X}^k} \sum_{G \in \mathcal{G}} \left[ \frac{1}{\text{cost}(P, C + A^*)} \cdot \sum_{p \in S_G} w(p) \cdot d(p, C) - \text{cost}(G, C) \right] 
\]

\[
+ \frac{1}{\text{cost}(P, C + A^*)} \cdot \sum_{i \in [k]} w(a^*_i) \cdot d(a^*_i, C) - \text{cost}(P \setminus G, C) \right] \leq \mathbb{E}_S \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}(P, C + A^*)} \cdot \sum_{i \in [k]} w(a^*_i) \cdot d(a^*_i, C) - \text{cost}(P \setminus G, C) \right] 
\]

\[
\leq 4 \varepsilon. 
\]

By the Markov inequality, with probability at least 0.9, the following inequality holds

\[
\sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}(P, C + A^*)} \cdot \sum_{p \in S} w(p) \cdot d(p, C) - \text{cost}(P, C) \right] \leq 40 \varepsilon. 
\]

Then for any \( C \in \mathcal{X}^k \), we have

\[
\left| \sum_{p \in S} w(p) \cdot d(p, C) - \text{cost}(P, C) \right| \leq 40 \varepsilon \cdot (\text{cost}(P, C + A^*)) \leq O(\varepsilon) \cdot \text{cost}(P, C),
\]

since \( A^* \) is an \( O(1) \)-approximate solution. This completes the proof. \( \square \)
C Proof of Lemma 3.11: Error Analysis for $x^C$

Proof: The proof idea is from that in [14, Lemma 15]. We have

\[ \mathbb{E}_{S_G} \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}(G, C + A^*)} \sum_{p \in S_G} w(p) \cdot x^C_p - \|x^C\|_1 \right] = \mathbb{E}_{S_G} \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}(G, C + A^*)} \sum_{p \in S_G} w(p) \cdot x^C_p - \|x^C\|_1 \mid \xi_G \right] \cdot \Pr[\xi_G] \] (34)

For the first term on the right side, a trivial upper bound for $\Pr[\xi_G]$ is 1. For each $C \in \mathcal{X}^k$, let $M_C \subseteq [k]$ be the collection of $i$ with $P_i \cap H(G, C) = \emptyset$. Assuming $\xi_G$ holds, we have that

\[ \sum_{p \in S_G} w(p) \cdot x^C_p = \sum_{i \in M_C} \sum_{p \in P_i \cap S_G} w(p) \cdot x^C_p \]

\[ = \sum_{i \in M_C} d(a_i^*, C) \cdot \sum_{p \in P_i \cap S_G} w(p) \] (Definition 3.9)

\[ \in (1 \pm \varepsilon) \sum_{i \in M_C} d(a_i^*, C) \cdot |P_i \cap G| \] (Defn. of $\xi_G$)

which implies that

\[ \mathbb{E}_{S_G} \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}(G, C + A^*)} \sum_{p \in S_G} w(p) \cdot x^C_p - \|x^C\|_1 \mid \xi_G \right] \cdot \Pr[\xi_G] \]

\[ \leq \varepsilon \cdot \mathbb{E}_{S_G} \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}(G, C + A^*)} \|x^C\|_1 \right] \] (Ineq. (35))

\[ \leq \varepsilon. \] (triangle ineq.)

On the other hand, assuming $\xi_G$ does not hold, we have $\Pr[\xi_G^C] \leq k \cdot \exp(-\varepsilon^2 \Gamma_G/9k)$ by Lemma 3.8. Since $\Gamma_G \geq O(\varepsilon^{-2} \log k)$, we have $\Pr[\xi_G^C] \leq \varepsilon/4k$. Moreover,

\[ \sum_{p \in S_G} w(p) \cdot x^C_p = \sum_{i \in M_C} \sum_{p \in P_i \cap S_G} w(p) \cdot x^C_p \]

\[ = \sum_{i \in M_C} d(a_i^*, C) \cdot \sum_{p \in P_i \cap S_G} \frac{\text{cost}(G, A^*)}{\Gamma_G} \cdot d(p, A^*) \] (Defn. of $w(p)$)

\[ \leq \sum_{i \in M_C} d(a_i^*, C) \cdot \sum_{p \in P_i \cap S_G} \frac{4k|P_i \cap G|}{\Gamma_G} \] (Observation 2.3)

\[ \leq 4k \sum_{i \in M_C} d(a_i^*, C) \cdot |P_i \cap G| \] ($|P_i \cap S_G| \leq \Gamma_G$)

\[ = 4k\|x\|_1, \]
which implies that

\[
\mathbb{E}_{\mathcal{G}} \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}(G, C + A^*)} \left| \sum_{p \in \mathcal{S}_G} w(p) \cdot x_p^C - \|x_G^C\|_1 \right| \frac{x_G^C}{x_G^C} \right] \cdot \Pr \left[ \xi_G \right] \\
\leq 4k \cdot \mathbb{E}_{\mathcal{G}} \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}(G, C + A^*)} \|x_G^C\|_1 \right] \cdot \Pr \left[ \xi_G \right] \\
\leq 4k \cdot \Pr \left[ \xi_G \right] \\
\leq 4k \cdot \frac{\varepsilon}{4k} \\
= \varepsilon.
\]

The lemma is a direct corollary of Inequalities (34), (36) and (38).

\[\square\]

D Proof of Theorem 3.6: Analysis of Algorithm 1 for General \( z \geq 1 \)

We first show how to choose the number of samples \( \Gamma_G \) for group \( G \) in Theorem 3.6. For preparation, we generalize the definition of coverings.

D.1 Coverings for general \( z \geq 1 \)

For a main group \( G \in \mathcal{G}^{(m)}(j, b) \), we generalize the definition of \( H(G, C) \) to be

\[
H(G, C) := \left\{ p \in R_{ij} \cap G^{(m)}(j, b) : i \in [k], d(a^*_i, C) \geq 8z\varepsilon^{-1} \cdot 2^j \Delta_i \right\},
\]

such that we have the following observation.

Observation D.1 For a \( k \)-center set \( C \in \mathcal{X}^k \), \( i \in [k] \) and \( p \in P_i \cap H(G, C) \), we have \( d^z(p, C) \in (1 \pm \varepsilon) \cdot d^z(a^*_i, C) \).

We also generalize Definition 3.2 as follows.

Definition D.2 (Coverings and covering numbers of main groups for general \( z \geq 1 \)) Fix constant \( z \geq 1 \). Let \( G \in \mathcal{G}^{(m)} \) be a group. Let \( S \subseteq G \) be a subset, \( \alpha > 0 \) and \( 0 \leq \beta < \log(ze^{-1}) \) be an integer. We say a set \( V \subseteq \mathbb{R}^{|S|} \) of cost vectors is an \( \alpha \)-covering of \( S \) if for each \( C \in \mathcal{X}^k \), there exists a cost vector \( v \in V \) such that for any \( i \in [k] \) and \( p \in P_i \cap S \setminus H(G, C) \), the following inequality holds:

\[
|d^z(p, C) - d^z(a^*_i, C) - v_p| \leq \alpha \cdot (\sqrt{d^z(p, C) \cdot d^z(p, A^*)} + d^z(p, A^*)) \cdot \sqrt{\frac{\text{cost}_z(G, C + A^*)}{\text{cost}_z(G, A^*)}},
\]

where we use \( \text{cost}_z(G, C + A^*) \) as a shorthand notation of \( \text{cost}_z(G, C) + \text{cost}_z(G, A^*) \) throughout. Define \( \mathcal{N}^{(m)}(S, \alpha) \) to be the minimum cardinality \( |V| \) of an arbitrary \( \alpha \)-covering \( V \) of \( S \). The definition of \( (\Gamma, \alpha) \)-covering number \( \mathcal{N}^{(m)}(\Gamma, \alpha) \) of \( G \) remains the same.

For outer groups, Definition 3.4 remains the same except that we replace \( d \) by \( d^z \) for general \( z \geq 1 \).
D.2 Proof of Theorem 3.6

Fix constant $z \geq 1$. The proof is similar to that of Theorem 3.5. Again, we only need to verify the correctness. For preparation, we provide the following relaxed triangle inequality.

**Lemma D.3 (Relaxed triangle inequality)** Let $a, b, c \in \mathcal{X}$ and $z \geq 1$. For every $t \in (0, 1]$, the following inequalities hold:

$$d^z(a, b) \leq (1 + t)^{z-1}d^z(a, c) + (1 + \frac{1}{t})^{z-1}d^z(b, c).$$

We have the following lemma that generalizes Lemma 3.7.

**Lemma D.4 (Generalization of Lemma 3.7)** The followings hold

1. For each $G \in \mathcal{G}^{(m)}$, we have
   $$\mathbb{E}_{S_G} \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}_z(G, C + A^*)} \cdot \sum_{p \in S_G} w(p) \cdot d^z(p, C) - \text{cost}_z(G, C) \right] \leq \varepsilon.$$

2. For each $G \in \mathcal{G}^{(o)}$, we have
   $$\mathbb{E}_{S_G} \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}_z(P^G, C + A^*)} \cdot \sum_{p \in S_G} w(p) \cdot d^z(p, C) - \text{cost}_z(G, C) \right] \leq \varepsilon.$$

3. For any $C \in \mathcal{X}^k$, we have
   $$\left| \text{cost}_z(P \setminus G, C) - \sum_{i \in [k]} w(a_i^*) \cdot d^z(a_i^*, C) \right| \leq \varepsilon \cdot \text{cost}_z(P, C + A^*).$$

By the same argument as for Theorem 3.5, we know that Theorem 3.6 is a direct corollary of the above three lemmas. Hence, it remains to prove Lemma D.4.

**Proof:** [of Lemma D.4]

**Item 1.** The proof is almost the same as Item 1 of Lemma 3.7. Fix a main group $G \in \mathcal{G}^{(m)}$ and let $\Gamma_G' = 2^{O(z)} \Gamma_G \geq 2^{3^z} \Gamma_G$ be the sample number of $G$. The good event $\xi_G$ is generalized to be

$$\sum_{p \in P_i \cap S_G} w(p) = \sum_{p \in P_i \cap S_G} \frac{\text{cost}_z(G, A^*)}{\Gamma_G' \cdot d^z(p, A^*)} \in (1 \pm \varepsilon) \cdot |P_i \cap G|,$$

and Lemma 3.8 still holds. All $d(\cdot, \cdot)$ in Definition 3.9 generalizes to be $d^z(\cdot, \cdot)$. Then Lemmas 3.10 and 3.11 still hold by generalizing all cost$(\cdot, \cdot)$ to cost$_z(\cdot, \cdot)$. The arguments are the same – still relying on Lemma 3.8.

The main difference is the following generalization of Lemma 3.12.
Lemma D.5 (Estimation error of \(y^C\) for general \(z \geq 1\)) The following inequality holds:

\[
\mathbb{E}_{S_G} \sup_{C \in \mathcal{X}^k} \left[ \frac{1}{\text{cost}_{\xi}(G, C + A^*)} \left| \sum_{p \in S_G} w(p) \cdot y^C_p - \|y^C\|_1 \right| \right] \leq \varepsilon.
\]

Similarly, Item 1 of Lemma D.4 is a direct corollary. Thus, it suffices to prove Lemma D.5. We still reduce to a Gaussian process and consider the following generalized estimators:

\[
X_{C,h} := \frac{1}{\text{cost}_{\xi}(G, C + A^*)} \sum_{p \in S_G \setminus H(S_G, C)} \xi_p \cdot w(p) \cdot (v^C_{p,h+1} - v^C_{p,h}),
\]

The key difference is the variance of \(X_{C,h}\); summarized by the following generalization of Lemma 3.14.

Lemma D.6 (Variance of \(X_{C,h}\) for general \(z \geq 1\)) Fix a \(k\)-center set \(C \in \mathcal{X}^k\). The variance of \(X_{C,h}\) is always at most

\[
\sum_{p \in S_G \setminus H(S_G, C)} \left( \frac{w(p) \cdot (v^C_{p,h+1} - v^C_{p,h})}{\text{cost}_{\xi}(G, C + A^*)} \right)^2 \leq \frac{2^{-2h+4k}}{\Gamma_G}.
\]

Moreover, conditioned on event \(\xi_G\), the variance of \(X_{C,h}\) is always at most

\[
\sum_{p \in S_G \setminus H(S_G, C)} \left( \frac{w(p) \cdot (v^C_{p,h+1} - v^C_{p,h})}{\text{cost}_{\xi}(G, C + A^*)} \right)^2 \leq \frac{2^{-2h+5}}{\Gamma_G}.
\]

Since the variance of \(X_{C,h}\) has an additional factor \(2^{O(z)}\), we can upper bound

\[
\sum_{h \geq 0} \mathbb{E}_{\xi} \sup_{C \in \mathcal{X}^k} |X_{C,h}| \leq \frac{\varepsilon}{6},
\]

which completes the proof of Lemma D.5.

It remains to prove Lemma D.6. We still let \(M_C \subseteq [k]\) denote the collection of \(i \in [k]\) with \(P_i \cap S_G \setminus H(S_G, C) \neq \emptyset\). For any \(i \in [k]\) and point \(p \in P_i \cap G\), let \(q = \arg \min_{p' \in P \cap G} d(p', C)\). Firstly, Inequality (14) generalizes to be

\[
d^z(p, C) \leq (d(q, C) + d(p, A^*) + d(q, A^*))^z \quad \text{(triangle ineq.)}
\]

\[
\leq (d(q, C) + 3d(p, A^*))^z \quad \text{(Defn. of } R_{ij})
\]

\[
\leq 2^z d(q, C) + 6^z d^z(p, A^*) \quad \text{(Lemma D.3)}
\]

\[
\leq \frac{2^z \text{cost}_{\xi}(P_i \cap G, C)}{|P_i \cap G|} + 6^z d^z(p, A^*) \quad \text{(Defn. of } q).\]

Then conditioned on \(\xi_G\), we have

\[
\sum_{p \in S_G \setminus H(S_G, C)} \left( \frac{w(p) \cdot (v^C_{p,h+1} - v^C_{p,h})}{\text{cost}_{\xi}(G, C + A^*)} \right)^2 \leq \frac{2^{-2h+2}}{\Gamma_G} \cdot \sum_{i \in M_C} \sum_{p \in P_i \cap S_G} \frac{w(p)}{\text{cost}_{\xi}(G, C + A^*)} \left( \frac{2^z \text{cost}_{\xi}(P_i \cap G, C)}{|P_i \cap G|} + 6^z d^z(p, A^*) \right)
\]

\[
\leq \frac{2^{-2h+5}}{\Gamma_G}.
\]
In general, we also have

\[ \sum_{p \in S_G \setminus H(S_G, C)} \left( \frac{w(p) \cdot (v_{p,C,h+1} - v_{p,C,h})}{\text{cost}_z(G, C + A^*)} \right)^2 \leq \frac{2^{-2h+2}}{\Gamma_G} \cdot \sum_{i \in M_C} \sum_{p \in P_i \cap S_G} \frac{w(p)}{\text{cost}_z(G, C + A^*)} \cdot \left( \frac{2^2 \text{cost}_z(P_i \cap G, C)}{|P_i \cap G|} + 6^2 d^2(p, A^*) \right) \]

\[ \leq \frac{2^{-2h+5}}{\Gamma_G}. \]

Thus, we complete the proof of Lemma D.6, which completes the proof of Item 1.

**Item 2.** Again, the proof idea is almost the same as that of [14, Lemma 13]. The same as Item 2 of Lemma 3.7, we can consider the coverings on \( S_G \) instead of \( G \). Recall that [14, Lemma 18] selects the sample size \( \delta = \frac{z}{O(z)} \Gamma_G \) containing factor \( \frac{z}{O(z)} \Gamma_G \) instead of \( 2\frac{O(z)}{O(z)} \Gamma_G \). However, by the proof of [14, Lemma 18], we can see that the variance of \( X_{C,h} \) is actually upper bounded by \( 16 \frac{z^2}{\Gamma_G} \). Hence, the sample size \( \delta \) in [14, Lemma 18] can be modified to be \( 2\frac{z}{O(z)} \Gamma_G \), which matches Item 2 of Lemma D.4.

**Item 3.** Item 3 has been proved in [17, 14]; see e.g., [17, Lemma 4]. \( \square \)

## E Proof of Theorem 4.1 for general \( z \geq 1 \): Euclidean coresets

The key is to prove the following lemma that generalizes Lemma 4.2.

**Lemma E.1** (Covering number in Euclidean metrics for general \( z \geq 1 \)) For each \( G \in \mathcal{G}^{(m)} \), \( 0 < \alpha \leq 1 \), and integer \( 0 \leq \beta < \log(ze^{-1}) \),

\[ \log \mathcal{N}_G^{(m)}(\Gamma_G, \alpha) = 2^{O(z)}k \cdot \min \left\{ k^{\frac{1}{2}} \alpha^{-2} \log \|P\|_0 \log \varepsilon^{-1}, d \right\} \cdot \log(\varepsilon^{-1} \alpha^{-1}). \]

For each \( G \in \mathcal{G}^{(o)} \) and any \( 0 < \alpha \leq 1 \),

\[ \log \mathcal{N}_G^{(o)}(\Gamma_G, \alpha) = O(z^2 k \cdot \min \left\{ \alpha^{-2} \log \|P\|_0, d \right\} \cdot \log \alpha^{-1}). \]

**Proof:** By [14, Lemma 21], we can obtain the upper bound for \( \mathcal{N}_G^{(o)}(\Gamma_G, \alpha) \). Fix \( \alpha \in (0, 1) \), \( z \geq 1 \) and let \( G \in \mathcal{G}^{(m)}(j) \) be a main group. We generalize Definition 4.9, the partition of \( G \), by setting the range of \( \beta \) to be \( 0 \leq \beta < 2 + \log(ze^{-1}) \). Fix a subset \( S \subseteq G \) of size \( \Gamma_G \).

**Generalization of Lemma 4.8.** We first show how to generalize Lemma 4.8.

**Lemma E.2** (Covering construction for general \( z \geq 1 \) in Euclidean spaces) Suppose \( d \geq \log k \) and constant \( z \geq 1 \). Let \( X \subseteq \mathbb{R}^d \) be a dataset that consists of \( X_1 \subseteq B(a_1^*, r_1), X_1 \subseteq B(a_2^*, r_2), \ldots, X_t \subseteq B(a_t, r_t) \) for some \( 1 \leq t \leq k, a_1^*, \ldots, a_t^* \in \mathbb{R}^d \) and \( r_1, \ldots, r_t > 0 \). Let \( \alpha \in (0, 1) \).
and \( u \geq 1 \). There exists an \( \alpha \)-covering \( V \subset \mathbb{R}^{|X|} \) of \( X \) with \( \log |V| = O(\kappa d \log(\alpha^{-1})) \), i.e., for any \( k \)-center set \( C \in \mathcal{X}^k \), there exists a cost vector \( v \in V \) with \( d(a^*_i, C) \leq u \cdot r_i \) and \( p \in X_i \) such that

\[
|d^z(p, C) - d^z(a^*_i, C) - v_p| \leq \alpha \cdot r_i^z.
\]

**Proof:** For each \( i \in [t] \), take an \( \frac{\alpha}{10(u+2)^z} \cdot r_i \)-net of the Euclidean ball \( B(a^*_i, 10zur_i) \). Since \( d \geq \log k \) and \( z = O(1) \), the union \( \Lambda \) of these nets has size at most

\[
k \cdot \exp\left( O(zd \log(\alpha^{-1})) \right) = \exp\left( O(zd \cdot \log(zu^{-1})) \right).
\]

For any \( S \subseteq G \), we then define an \( \alpha \)-covering \( V \subset \mathbb{R}^{|S|} \) of \( S \) as follows: for any \( k \)-center set \( C \subseteq \Lambda \), we construct a vector \( v \in V \) in which the entry corresponding to point \( p \) (say \( p \in R_{ij} \cap S \)) is

\[
v_p = d^z(p, C) - d^z(a^*_i, C).
\]

Obviously, we have \( \log |V| = O(\kappa d \cdot \log(\alpha^{-1})) \). It remains to verify that \( V \) is indeed an \( \alpha \)-covering of \( S \). For any \( C = (c_1, \ldots, c_k) \in \mathcal{X}^k \), let \( C' = (c'_1, \ldots, c'_k) \in V \) such that \( c'_i \) is the closest point of \( c_i \) in \( \Lambda \) (\( i \in [k] \)). Then for any \( i \in [t] \) with \( d(a^*_i, C) \leq u \cdot r_i \) and \( p \in X_i \), we have the following observation: for every \( c_i \in C \),

1. If \( d(p, c_i) \geq 4u \cdot r_i^z \), then \( d(p, c'_i) \geq 2u \cdot r_i \) since \( d(p, C) \leq d(a^*_i, C) + d(p, a_i^*) \).
2. If \( d(p, c) \leq 4u \cdot r_i \), then \( d(p, c') \in d(p, c) \pm \frac{\alpha}{10(u+2)^z} \cdot r_i \) by the construction of \( \Lambda \).

Consequently, we have

\[
d(p, C') \in d(p, C) \pm \frac{\alpha}{10(u+2)^z} \cdot r_i,
\]

which implies that

\[
|d^z(p, C') - d^z(p, C)| \\
\leq |d(p, C') - d(p, C)| \cdot z(d^{z-1}(p, C') + d^{z-1}(p, C)) \\
\leq \frac{\alpha}{10(u+2)^z} \cdot r_i \cdot \left( (d(a^*_i, C') + r_i)^{z-1} + (d(a^*_i, C) + r_i)^{z-1} \right) \quad \text{(triangle ineq. and Ineq. (40))} \\
\leq \frac{\alpha}{10(u+2)^z} \cdot r_i \cdot 2(u+2)^{z-1}r_i^{z-1} \quad \text{(d(a^*_i, C) \leq u \cdot r_i)} \\
\leq \frac{\alpha}{2} \cdot r_i^z.
\]

The above two inequalities directly lead to the following conclusion:

\[
|d^z(p, C) - d^z(a^*_i, C) - v_p| = |d^z(p, C) - d^z(a^*_i, C) - (d^z(p, C') - d^z(a^*_i, C'))| \leq \alpha \cdot r_i^z,
\]

which completes the proof. \( \square \)
Generalization of Inequality (24). We first verify the following inequality that generalizes Inequality (24):

$$\log \mathcal{N}^m_G(\Gamma_G, \alpha) = O(zkd \log(\varepsilon^{-1} \alpha^{-1})).$$

(41)

This is actually a direct corollary of Lemma 4.8 by letting $X = S$, $X_i = S \cap R_{ij}$, $r_i = 2^{j+1} \Delta_i$ for $i \in [k]$, and $u = 10\varepsilon^{-1}$.

Generalization of Inequality (25). Next, we verify the following inequality that generalizes Inequality (25):

$$\log \mathcal{N}^m_G(\Gamma_G, \alpha) \leq 2^{O(z)} k^{\frac{2\alpha+2}{\varepsilon+2}} \alpha^{-2} \log \|P\|_0 \log \varepsilon^{-1} \log(\varepsilon^{-1} \alpha^{-1})).$$

(42)

We again partition $G \setminus H(G, C)$ as Definition 4.9. The only difference is that the range of $\beta$ changes to be $0 \leq \beta \leq 2 + \log(\varepsilon^{-1})$. The same as the proof in Section 4.3, we denote $\phi(C) \in \{0, 1, \ldots, 2 + \log(\varepsilon^{-1})\}$ for each $C \in \mathcal{X}^k$. There are at most $(3 + \log(\varepsilon^{-1}))^k$ distinct $\phi(C)$s. For each $\phi \in \{0, 1, \ldots, 2 + \log(\varepsilon^{-1})\}$, we again denote $S_{\phi, \beta}$ for each integer $0 \leq \beta \leq 2 + \log(\varepsilon^{-1})$. We have the following lemma that generalizes Lemma 4.11.

**Lemma E.3 (An upper bound of the covering number for each $\beta$ for general $\varepsilon \geq 1$)** Let $G \in \mathcal{G}^{(m)}(j)$ for some integer $\log \varepsilon < j < 2\log \varepsilon^{-1}$ be a main group and let $S \subseteq G$ be a subset. Fix a vector $\phi \in \{0, 1, \ldots, 2 + \log(\varepsilon^{-1})\}$ and an integer $0 \leq \beta \leq 2 + \log(\varepsilon^{-1})$. There exists an $\alpha$-covering $V_{\beta, \phi} \subset \mathbb{R}^{|S_{\phi, \beta}|}$ such that for each $e \in \mathcal{X}^k$, there exists a cost vector $v \in V_{\beta, \phi}$ such that for any $i \in [k]$ and $p \in R_{ij} \cap S_{\phi, \beta}$, Inequality (39) holds, i.e.,

$$|d^z(p, C) - d^z(a_i^*, C) - v_p| \leq \alpha \cdot (\sqrt{d^z(p, C) \cdot d^z(p, A^*)} + d^z(p, A^*)) \cdot \frac{\sqrt{\text{cost}_z(G, C + A^*)}}{\text{cost}_z(G, A^*)}.$$

Moreover, the size of $V_{\beta, \phi}$ satisfies

$$\log |V_{\beta, \phi}| \leq 2^{O(z)} k \cdot \min \left\{2^{\beta z} \alpha^{-2} \log(k \varepsilon^{-1}), (1 + k 2^{-2\beta}) \alpha^{-2} \log(k \varepsilon^{-1}) \right\} \cdot \log(\varepsilon^{-1} \alpha^{-1}).$$

By a similar argument as that in Section 4.3, we can prove that Lemma E.1 is a corollary of the above lemma. The only difference is that we should upper bound the term $\min \{2^{\beta z}, 1 + k 2^{-2\beta}\}$ instead of $\min \{2^{\beta z}, 1 + k 2^{-2\beta}\}$. Letting $2^{\beta} = k^{\frac{1}{2 + z}}$, we have

$$\min \{2^{\beta z}, 1 + k 2^{-2\beta}\} \leq k^{\frac{\beta}{2 + z}},$$

which implies Lemma E.1. Then it remains to prove Lemma E.3.

**Proof:** [of Lemma E.3]

46
Case $\beta = 0$ for general $z \geq 1$. Let $f$ be an $2^{-5z} \alpha$-terminal embedding of $A^* \cup S$ into $m = 2^{10z} \alpha^{-2} \log \|P\|_0$ dimensions given by Theorem 4.4. Given a $k$-center set $C \in \mathcal{X}^k$ with $\phi(C) = \phi$, we have that for any $p \in A \cup S_{\phi,0}$,

$$d^z(f(p), f(C)) = \min_{q \in C} d^z(f(p), f(q))$$

$$\in (1 + \alpha 2^{-5z})^z \cdot \min_{q \in C} d^z(p, q) \quad \text{(Theorem 4.4)}$$

$$\in (1 + \alpha 2^{-3z}) \cdot d^z(p, C)$$

$$\in d^z(p, C) \pm \alpha \cdot d^z(p, A^*). \quad \text{(Observation 4.10)}$$

The remaining argument is the same as in Lemma 4.2, say applying Lemma 4.8 with $d = m$, $X = S_{\phi,0}$, $X_i = R_{ij} \cap S_{\phi,0}$, $r_i = 2^{i+4} \Delta_i$ for $i \in [k]$, and $u = O(1)$ to construct an $\alpha$-covering of $f(S)$. Note that the target dimension is $m = 2^{O(z)} \alpha^{-2} \log \|P\|_0$, which completes the proof.

Case $\beta \geq 1$ for general $z \geq 1$. We first construct an $\alpha$-covering $V_{\phi,\beta}$ of $S_{\phi,\beta}$ with

$$\log |V_{\phi,\beta}| = 2^{O(z)} k 2^{\beta z} \alpha^{-2} \log \|P\|_0 \cdot \log(\varepsilon^{-1} \alpha^{-1}). \quad (44)$$

Let $f$ be an $2^{-5z-\beta z/2} \alpha$-terminal embedding of $A^* \cup S_{\phi,\beta}$ into $m = 2^{10z+\beta z -2} \alpha^{-2} \log \|P\|_0$ dimensions given by Theorem 4.4. Given a $k$-center set $C \in \mathcal{X}^k$ with $\phi(C) = \phi$, we have that for any $i \in [k]$ and $p \in R_{ij} \cap S_{\phi,\beta}$,

$$d^z(f(p), f(C)) = \min_{q \in C} d^z(f(p), f(q))$$

$$\in (1 + \alpha 2^{-5z-\beta z/2})^z \cdot \min_{q \in C} d^z(p, q) \quad \text{(Theorem 4.4)}$$

$$\in (1 + \alpha 2^{-3z-\beta z/2}) \cdot d^z(p, C)$$

$$\in d^z(p, C) \pm O(\alpha) \cdot \sqrt{d^z(p, C) \cdot d^z(p, A^*)}, \quad \text{(Observation 4.10)}$$

and

$$d^z(f(a_i^*), f(C)) \in d^z(a_i^*, C) \pm O(\alpha) \cdot \sqrt{d^z(p, C) \cdot d^z(p, A^*)}.$$}

The remaining argument is the same as in Lemma 4.11 by applying Lemma 4.8 with $d = m$, $X = S_{\phi,0}$, $X_i = R_{ij} \cap S_{\phi,0}$, $r_i = 2^{i+4} \Delta_i$ for $i \in [k]$, and $u = 2^{10z+10}$ to construct an $\alpha$-covering of $f(S)$. Since the target dimension is $m = 2^{O(z)-\beta z} \alpha^{-2} \log \|P\|_0$, we complete the proof of Inequality (44).

Finally, we construct an $\alpha$-covering $V_{\phi,\beta}$ of $S_{\phi,\beta}$ with

$$\log |V_{\phi,\beta}| = 2^{O(z)} (k 2^{\beta z} \alpha^{-2} \log \|P\|_0 \cdot \log(\varepsilon^{-1} \alpha^{-1})). \quad (45)$$

We again let $B \subseteq [k]$ be the collection of $i \in [k]$ with $R_{ij} \cap S \subseteq S_{\phi,\beta}$. Similar to Inequality (31), we note that for any $C \in \mathcal{X}^k$ with $\phi(C) = \phi$,

$$\frac{\text{cost}_z(G, C)}{\text{cost}_z(G, A^*)} \geq \frac{|B| 2^{(\beta z -1)}}{k}. \quad (46)$$

For each $i \in B$, let $g_i$ be an $\sqrt{|B| k^{-1} 2^{\beta z}} \alpha$-additive terminal embedding of $T_i = \{p - a_i^*: p \in R_{ij} \cap S\}$ into $m = O(|B|^{-1} k 2^{10z-3z} \alpha^{-2} \log \|P\|_0)$ dimensions given by Theorem 4.6.
Let \( r = 2^{j+1} \Delta_i \). We again define a function \( f_i \) as follows: \( f_i(p) = g_i(p - a_i^*) \) for any \( p \in \mathbb{R}^d \). Recall that \( G \in \mathcal{G}(m)(j) \), and hence, \( R_{ij} \cap S \in B(a_i^*, r) \). Then by Theorem 4.6, we have that for any \( p \in R_{ij} \cap S \) or \( p = a_i^* \) and \( q \in \mathbb{R}^d \), \(|d(p, q) - d(f_i(p), f_i(q))| \leq \sqrt{|B|k^{-1}2^{\beta - 5z} \alpha r} \). Given a \( k \)-center set \( C \in \mathcal{X}_k \) with \( \phi(C) = \phi \), we have that for any \( p \in R_{ij} \cap S \), \( d(a_i^*, C) \geq 2^{3j+2} \Delta_i \geq 4r \). Consequently, we know that \( C \subset \mathbb{R}^d \setminus B(a_i^*, 2r) \). For any \( p \in R_{ij} \cap S_{\phi, \beta} \) and \( c_p = \arg \min_{c \in C} d(p, C) \in \mathbb{R}^d \), we have

\[
\begin{align*}
&d^2(f_i(p), f_i(c_p)) \\
&\in (d(p, c) \pm \sqrt{|B|k^{-1}2^{\beta - 5z} \alpha r})^2 \\
&\in (1 \pm 2^{1-5z} \sqrt{|B|k^{-1} \alpha})^2 d^2(p, c_p) \\
&\in (1 \pm 2^{-4z} \sqrt{|B|k^{-1} \alpha})^2 d^2(p, c_p) \\
&\in d^2(p, c_p) \pm \alpha \cdot d^2(p, c_p) \cdot d^2(p, A^*) \cdot \frac{\text{cost}(G, C)}{\text{cost}(G, A^*)} \\
&\in d^2(p, C) \pm \alpha \cdot d^2(p, C) \cdot d^2(p, A^*) \cdot \frac{\text{cost}(G, C)}{\text{cost}(G, A^*)} \\
&\in d^2(p, c_p) \pm \alpha \cdot d^2(p, c_p) \cdot d^2(p, A^*) \cdot \frac{\text{cost}(G, C)}{\text{cost}(G, A^*)} \\
&\in d^2(p, C) \pm \alpha \cdot d^2(p, C) \cdot d^2(p, A^*) \cdot \frac{\text{cost}(G, C)}{\text{cost}(G, A^*)}
\end{align*}
\]

Moreover, we have for any \( p \in R_{ij} \cap S_{\phi, \beta} \) and \( c_p = \arg \min_{c \in C} d(p, C) \in \mathbb{R}^d \),

\[
|d^2(f_i(a_i^*), f_i(C)) - d^2(a_i^*, C)| \leq O(\alpha) \cdot \sqrt{d^2(p, C) \cdot d^2(p, A^*)} \cdot \sqrt{\frac{\text{cost}(G, C)}{\text{cost}(G, A^*)}}.
\]

Combining the above two inequalities, we directly have

\[
|(d^2(f_i(p), f_i(C)) - d^2(f_i(a_i^*), f_i(C))) - (d^2(p, C) - d^2(a_i^*, C))| \\
\leq O(\alpha) \cdot \sqrt{d^2(p, C) \cdot d^2(p, A^*)} \cdot \sqrt{\frac{\text{cost}(G, C)}{\text{cost}(G, A^*)}}.
\]

The remaining argument is the same as in Lemma 4.11. Since the target dimension is \( m = 2^{O(\alpha)} |B|^{-1} k 2^{-2\beta} \alpha^{-2} \log \|P\|_0 \), we complete the proof of Inequality (45).

Overall, we complete the proof of Lemma E.1.

By the same argument as for \( z = 1 \), Theorem 4.1 is a direct corollary of Lemma E.1.