The Power of Uniform Sampling for k-Median

Lingxiao Huang [1] **Shaofeng H.-C. Jiang** [2] **Jianing Lou** [2]

Abstract

We study the power of uniform sampling for k-MEDIAN in various metric spaces. We relate the query complexity for approximating k-MEDIAN, to a key parameter of the dataset, called the balancedness $\beta \in (0, 1]$ (with 1 being perfectly balanced). We show that any algorithm must make $\Omega(1/\beta)$ queries to the point set in order to achieve $O(1)$-approximation for k-MEDIAN. This particularly implies existing constructions of coresets, a popular data reduction technique, cannot be query-efficient. On the other hand, we show a simple uniform sample of $\mathrm{poly}(k\varepsilon^{-1}\beta^{-1})$ points suffices for $(1+\varepsilon)$-approximation for k-MEDIAN for various metric spaces, which nearly matches the lower bound. We conduct experiments to verify that in many real datasets, the balancedness parameter is usually well bounded, and that the uniform sampling performs consistently well even for the case with moderately large balancedness, which justifies that uniform sampling is indeed a viable approach for solving k-MEDIAN.

1. Introduction

We investigate the power of uniform sampling in data reduction for k-MEDIAN, which is a fundamental machine learning problem that has wide applications. Given a metric space $(\mathcal{X}, \mathrm{dist})$, k-MEDIAN takes an n-point dataset $X \subseteq \mathcal{X}$ and an integer parameter $k \geq 1$ as inputs, and the goal is to find a k-point center set $C \subseteq \mathcal{X}$ that minimizes the objective

$$\mathrm{cost}(X, C) := \sum_{x \in X} \mathrm{dist}(x, C),$$

where $\mathrm{dist}(x, C) := \min_{c \in C} \mathrm{dist}(x, c)$ is the distance to the closest center.

Data reduction is a powerful way for dealing with clustering problems, and a popular method called coreset (Har-Peled & Mazumdar, 2004) has been extensively studied during the last decades. Roughly, an ε-coreset aims to find a tiny but accurate proxy of the data, so that an ε-approximate center set C can be found by running existing algorithms on top of it. Specifically, coresets for k-MEDIAN in Euclidean $\mathbb{R}^d$ has been studied in a series of works (Har-Peled & Mazumdar, 2004; Har-Peled & Kushal, 2007; Chen, 2009; Feldman & Langberg, 2011; Feldman et al., 2020; Sohler & Woodruff, 2018; Huang & Vishnoi, 2020; Braverman et al., 2021b; Cohen-Addad et al., 2021b; Braverman et al., 2022), and coresets of size $\mathrm{poly}(k\varepsilon^{-1})$ were obtained, which is independent of the dimension d and the data size. In addition to speeding up existing algorithms, coresets can also be used to derive algorithms in sublinear modesl such as streaming (Har-Peled & Mazumdar, 2004), distributed (Balcan et al., 2013) and dynamic algorithms (Henzinger & Kale, 2020).

Despite the progress on the size bounds, an outstanding issue of coresets is that known coreset *constructions* are usually not query-efficient, i.e., it needs to access $\Omega(n)$ data points (even in sublinear models such as streaming). In fact, it is not hard to see that this limitation cannot be avoided, and in the worst case, any algorithm must make $\Omega(n)$ queries to the identity of data points in order to construct a coreset (see Theorem 1.2 which we discuss later). Technically, existing coreset constructions are often based on non-uniform sampling which is not data-oblivious, and this inherently requires to read the entire dataset. This also makes it more difficult to efficiently implement in practice due to the sophisticated sampling procedure. Hence, in order to achieve a sublinear query complexity, one must consider other methods than the coreset.

To this end, we consider uniform sampling as a natural alternative data reduction approach for clustering. Clearly, uniform sampling is data-oblivious and hence has a great

Authors are listed in alphabetical order. [1]State Key Laboratory of Novel Software Technology, Nanjing University, Jiangsu, China [2]Center on Frontiers of Computing Studies, Peking University, Beijing, China. Correspondence to: Lingxiao Huang <huanglingxiao1990@126.com>, Shaofeng H.-C. Jiang <shaofeng.jiang@pku.edu.cn>, Jianing Lou <loujn@pku.edu.cn>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

potential to achieve sublinear query complexity. Moreover, it often yields near-optimal solutions with only a few samples in practice, as was demonstrated by various experiments on real datasets in recent works on coresets for variants of clustering where uniform sampling is considered as a baseline (e.g., Marom & Feldman 2019; Jubran et al. 2020; Baker et al. 2020; Braverman et al. 2021a; Huang et al. 2023; Lu et al. 2023), even though it is known that uniform sampling cannot yield a coreset in the worst case.

Thus, the focus of this paper is to understand the sampling complexity of uniform sampling for k-MEDIAN and to justify its performance in practice.

1.1. Our Results

We first give a hardness result (Theorem 1.2) that even for $k = 2$ and 1D line, any algorithm, not only the uniform sampling, must make $\Omega(n)$ queries to the identity of points (which is the coordinate in 1D), in order to be $O(1)$-approximate to k-MEDIAN. In addition, Theorem 1.2 further states that the number of queries must depend on a parameter $\beta \in (0, 1]$, called *balancedness* (Definition 1.1), which is a property of the dataset. Intuitively, β measures how balance the optimal solution is, and precisely, it requires the size of the smallest cluster in an optimal solution is at least β times of the average cluster size $\frac{|X|}{k}$.

Definition 1.1 (Balancedness)**.** Given a dataset $X \subseteq \mathcal{X}$, the balancedness $\beta \in (0, 1]$ of X is the largest number such that there is an optimal solution of k-MEDIAN on X satisfying that every cluster[1] has at least $\frac{\beta|X|}{k}$ points.

The same notion of balancedness was considered in Meyerson et al. (2004) which also studied the complexity of uniform sampling for k-MEDIAN but achieved a weaker bound (which we will discuss later). This balancedness was also enforced as a constraint to clustering problems (Bradley et al., 2000), and more generally in capacitated clustering (Charikar et al., 2002) and fair clustering (Chierichetti et al., 2017). A similar idea of introducing the "balancedness" has been investigated in the context of coresets for logistic problems (Mai et al., 2021; Munteanu et al., 2021; 2022; Lu et al., 2023; Munteanu et al., 2023; Woodruff & Yasuda, 2023). However, due to the substantial differences between logistic regression and clustering, their findings can not be directly applied in clustering problem.

Theorem 1.2. *Any $O(1)$-approximate algorithm for 2-MEDIAN with success probability at least $3/4$ must query the identify of data points in X for $\Omega(1/\beta)$ times, where β is the balancedness of the dataset X, even if the queried points have free access to the distance function.*

[1]For a center set $C = \{c_i\}_{i=1}^k$, each c_i defines a cluster $X_i \subseteq X$ that consists of points whose nearest neighbor in C is c_i.

Intuitively, the hard instance in Theorem 1.2 has two clusters, one has only one point and the other has many points with a small diameter, and the two clusters are very far away. Then the uniform sampling fails to pick any point from singleton cluster with high probability, which incurs a big error. The detailed proof can be found in Appendix E.

Theorem 1.2 suggests that the balancedness β of the dataset may be a fundamental parameter that determines the necessary size of uniform sampling. In our main result, stated in Theorem 1.3, we give a nearly-matching upper bound (with respect to β) for k-MEDIAN in Euclidean $\mathbb{R}^d$, which helps to justify this parameter is indeed fundamental. This bound breaks the $\Omega(n)$ query complexity barrier of coresets, and it readily yields sublinear-time algorithms for k-MEDIAN. The theorem may also be interpreted as beyond worst case analysis for uniform sampling, where the parameter β provides a refined description to the structure of the dataset.

Theorem 1.3 (Informal version of Theorem 3.1)**.** *Given a dataset $X \subset \mathbb{R}^d$ with balancedness $\beta \in (0, 1]$, $\varepsilon \in (0, 0.5)$ and integer $k \geq 1$, let S be $\tilde{O}(\frac{k^2}{\beta\varepsilon^3})$ uniform samples[2] from X, then with probability 0.9, one can find a $(1 + \varepsilon)$-approximation for k-MEDIAN on X only using S.*

As mentioned, the dependence of β in Theorem 1.3 is nearly optimal. Furthermore, the dependence of k and ε^{-1} is only low-degree polynomial, and it also matches the known coreset size bounds. Another feature of our bound is that it does not have a dependence in the Euclidean dimension d, thus it is very useful for dealing with high dimensional and/or sparse datasets.

In addition to the Euclidean case, we show a more general version (Theorem 3.1) that relates the sampling complexity to a notion of covering number (Definition 3.5) which measures the complexity of the underlying metric space $(\mathcal{X}, \mathrm{dist})$. By bounding the mentioned covering number (see Section 3.4), we also obtain similar bounds of $\mathrm{poly}(k\varepsilon^{-1}\beta^{-1})$ for various other metric spaces such as doubling metrics and shortest-path metrics of graphs with bounded treewidth. For general finite metric spaces, we obtain a bound of $\mathrm{poly}(k\varepsilon^{-1}\beta^{-1}\log|\mathcal{X}|)$.

Compared with the notion of coreset (Har-Peled & Mazumdar, 2004), the uniform sample S may be interpreted as a coreset in a *weaker* sense. Specifically, coresets usually require the clustering cost be preserved for *all* center sets $C \subseteq \mathcal{X}$, while in Theorem 1.3 we only guarantee the cost on *some* specific solutions, yet we still ensure that it suffices for finding good approximation efficiently by running algorithms on S. Another important difference is that for coresets, the size does not need to have a dependence in β which we have, but as mentioned earlier, this size bound of coreset cannot be realized by query-efficient algorithms,

[2]Throughout, $\tilde{O}(f) = O(f \,\mathrm{poly}\log f)$.

which the uniform sampling does.

Indeed, similar idea of using uniform sampling was previously studied in the literature, but the focus was mostly on the Euclidean case, and our bounds for doubling metrics and graph metrics are new. Even for Euclidean spaces, only the special case of $k = 1$ was studied, and previous bounds either depend on d (Munteanu & Schwiegelshohn, 2018) or have a worse ε^{-4} dependence than our ε^{-3} (Cohen-Addad et al., 2021a; Danos, 2021) (noting that $\beta = 1$ if $k = 1$). For general metrics, Meyerson et al. (2004) gave a very similar size bound as in ours (which also depends on β^{-1}), but it only achieves $O(1)$ error instead of our ε. Somewhat less related, Mishra et al. (2001); Ben-David (2007); Czumaj & Sohler (2007) gave uniform sampling bounds for the additive error guarantee, which is incomparable to ours. Recently, Braverman et al. (2022) proposes a coreset construction algorithm based on uniform sampling. However, it should be noted that their overall sampling distribution is actually non-uniform. Specifically, they partition the dataset into multiple parts and only apply uniform sampling in each part. As a result, their approach is not query-efficient. In contrast, our approach focuses on a "weaker" coreset and relies solely on uniform sampling, thereby achieving sublinear query complexity.

Finally, even though Theorem 1.3 implies a small uniform sample S suffices for a sublinear-time algorithm for k-MEDIAN, to actually find the near-optimal solution on the sample S can be tricky. A natural idea is to find the optimal k-MEDIAN on S as the approximate solution for X, but we show that this does not work, even when the dataset is balanced. In particular, if one uses the optimal k-MEDIAN on S as the approximate solution, then it still requires $\Omega(n)$ samples even for a balanced dataset (see Lemma F.1).

This motivates us to consider a variant of k-MEDIAN, called (k, β)-MEDIAN, which aims to find the optimal center set C subject to the constraint that C is β-balanced (i.e., every cluster induced by C has size at least $\beta \frac{|X|}{k}$; see Definition 2.1). The balancedness constraint in (k, β)-MEDIAN is intuitive, since if the dataset is already balanced, then by definition there must be a balanced optimal solution, which (k, β)-MEDIAN can find. We show in Theorem 3.1 (i.e., the full statement of Theorem 1.3) that an α-approximate (k, β)-MEDIAN on S is $O(\alpha(1 + \varepsilon))$-approximate k-MEDIAN on X with constant probability.

Experiments Our experiments focus on validating the performance of running an algorithm for k-MEDIAN (instead of (k, β)-MEDIAN) on top of the uniform S, since it is arguably the most natural approach and is likely to be used in practice. Our experiments were conducted on various real datasets of different types of metric spaces including Euclidean $\mathbb{R}^d$ and shortest-path metrics. We find that the solution returned by the k-MEDIAN algorithm is fairly balanced, which effectively enforces the balancedness constraint of (k, β)-MEDIAN. Moreover, we find that these datasets are mostly balanced, especially when the number of clusters k is small. Even when k is relatively large and the balancedness becomes worse, the factor of $1/\beta$ in our upper bound is still reasonably bounded (~100), and we also observe that the practical performance is not significantly worse than that of the small-k case, if at all. All these findings, together with our Theorem 1.3, justifies the effectiveness of uniform sampling in real datasets.

1.2. Technical Overview

Our proof of Theorem 1.3 builds on a structural lemma (Lemma 3.7), which states that if a center set C is "bad", i.e., its cost is larger than $(1 + \varepsilon)\,\mathrm{OPT}$, then this badness carries on to the sample. Specifically, let $C^\star$ be the optimal center set, if a center set C satisfies $\mathrm{cost}(X, C) \geq (1 + \varepsilon)\,\mathrm{cost}(X, C^\star)$, then we have $\mathrm{cost}(S, C) - \mathrm{cost}(S, C^\star) \geq \frac{O(\varepsilon)|S|}{n}\,\mathrm{cost}(X, C^\star)$ with probability $1 - \exp(-O(|S|))$ (the big O hides dependence in other parameters such as k and ε). Intuitively, conditioning on this event, one can conclude that any "good" center found in S is also good in X, which implies the main theorem.

The special case $k = 1$ of the structural lemma was proved in Munteanu & Schwiegelshohn (2018), but the analysis does not seem to generalize our case $k \geq 2$. Specifically, to bound $\mathrm{cost}(S, C) - \mathrm{cost}(S, C^\star)$, it suffices to bound $\mathrm{dist}(x, C) - \mathrm{dist}(x, C^\star)$ for $x \in S$, and in Munteanu & Schwiegelshohn (2018) they can use triangle inequality $\mathrm{dist}(x, C) - \mathrm{dist}(x, C^\star) \leq \mathrm{dist}(C, C^\star)$ since $|C| = |C^\star| = 1$, and the remaining analysis is on $\mathrm{dist}(C, C^\star)$ which does not depend on the variable x. However, when $k \geq 2$ such a triangle inequality no longer holds, and this forces us to use an alternative bound $\mathrm{dist}(x, C) - \mathrm{dist}(x, C^\star) \leq \max_{c \in C} \mathrm{dist}(c, C^\star) + \max_{c^\star \in C^\star} \mathrm{dist}(c^\star, C)$. To analyze this, we crucially use a new observation of balancedness: if C is a balanced center set, then C and $C^\star$ must be "close" which depends on $1/\beta$ (see Definition 3.3 and Lemma 3.4). This implies that both $\max_{c \in C} \mathrm{dist}(c, C^\star)$ and $\max_{c^\star \in C^\star} \mathrm{dist}(c^\star, C)$ are small enough, and this bound eventually allows us to apply a concentration inequality to bound $\mathrm{cost}(S, C) - \mathrm{cost}(S, C^\star)$. Indeed, the use of balancedness is a fundamental difference to Munteanu & Schwiegelshohn (2018).

Once the structural lemma is established, a natural next step is to apply it with a union bound on all centers that are close to C^*. While these centers can still be infinitely many, one can apply standard discretization techniques, such as ρ-nets in $\mathbb{R}^d$, to reduce the number of events in the union bound.

Removing Dependence on d However, for Euclidean $\mathbb{R}^d$, a naive application of the net argument only leads to a size bound that depends on d. To remove the dependence on d for the Euclidean case, we need a better union bound. To this end, we use an alternative interpretation of the structural lemma. In particular, we change the "variable" in the lemma from the center set C to a vector $v^C := (\mathrm{dist}(x, C) - \mathrm{dist}(x, C^\star))_{x \in X} \in \mathbb{R}^X$ which represents all the relevant costs that C induce. Then the structural lemma is equivalently stated as: if some $v \in \mathbb{R}^X$ satisfies $\|v\|_1 \geq \varepsilon \cdot \mathrm{cost}(X, C^\star)$, then $\|v|_S\|_1 \geq \frac{O(\varepsilon)|S|}{n} \mathrm{cost}(X, C^\star)$ with high probability, where $v|_S$ means restricting v to the uniform sample S. Now, we try to apply the union bound on a discretization of the cost vectors $\{v^C\}_C$, instead of the space of all center points C.

Specifically, for a fixed sample S, we need to find a discretized set U such that for every v, U contains a vector v' with $\|v|_S\|_1 \approx \|v'|_S\|_1$ (recalling that we do not need to preserve $\|v\|_1$). Note that this U needs not be a subset of $\{v^C\}_C$. In other words, a vector $u \in U$ may not correspond to a center set $C \subset \mathcal{X}$, and they can be any real vector in $\mathbb{R}^X$. This turns out to be great freedom compared with discretization of center sets. Indeed, for Euclidean $\mathbb{R}^d$, to preserve $\|v|_S\|$, we can map $S \cup C^\star$ into a low dimensional space of dimension $d' := O(\log(|S \cup C^\star|)/\varepsilon^2)$ using a terminal version of the Johnson-Lindenstrauss transform Narayanan & Nelson 2019, and discretize only in this lower dimensional space. This removes the dependence in d and replaces it with d'. In addition, we use a chaining argument (Talagrand, 1996) which was also used in several recent works about coresets (Cohen-Addad et al., 2021a; 2022a; Huang et al., 2022), to further save an ε^{-1} factor and obtain ε^{-3} dependence in the final bound. Compared with a closely related work Cohen-Addad et al. (2021a), our chaining argument is applied on the entire X, while theirs is applied on $O(\varepsilon^{-2})$ pieces of X separately which results in an addition ε^{-1} factor more than ours. Finally, we note that it may cause randomness issues since we assumed fixed sample S before finding U. Luckily, we manage to fix this by relating it with a Gaussian process.

Beyond Euclidean Spaces In fact, the above argument is useful not only for removing the dependence on d for Euclidean spaces, but also for obtaining bounds for other metric spaces. We show it suffices to bound the size of U in order to obtain a bound for uniform sampling. We formulate this minimum size of U as the covering number (see Definition 3.5), and we derive covering number bounds for several types of metric spaces. Previously, A similar notion of covering number as well as its use to bound the coreset size was also considered in the coreset literature (Cohen-Addad et al., 2022a; Huang et al., 2022). We give a more detailed comparison in Section 3.2.

1.3. Related Work

Stemming from Har-Peled & Mazumdar (2004), the study of size bounds for coresets has been very fruitful. For k-MEDIAN in Euclidean $\mathbb{R}^d$, a series of works improves the size from $O(\mathrm{poly}(k)\varepsilon^{-d}\log n)$ all the way to $\mathrm{poly}(k\varepsilon^{-1})$ (Har-Peled & Mazumdar, 2004; Har-Peled & Kushal, 2007; Chen, 2009; Feldman & Langberg, 2011; Feldman et al., 2020; Sohler & Woodruff, 2018; Cohen-Addad et al., 2021a; 2022a). Recent works focus on deriving tight bounds for k and ε^{-1}. The state-of-the-art coresets for k-MEDIAN in $\mathbb{R}^d$ achieves a size of $\tilde{O}(\min\{k\varepsilon^{-3}, k^{\frac{4}{3}}\varepsilon^{-2}\})$ (Cohen-Addad et al., 2022a; Huang et al., 2022), which nearly matches a lower bound of $\Omega(k\varepsilon^{-2})$ (Huang & Vishnoi, 2020; Cohen-Addad et al., 2022a). Beyond Euclidean spaces, coresets for k-MEDIAN were obtained in doubling metrics (Huang et al., 2018), shortest-path metrics of graphs with bounded treewidth (Baker et al., 2020) and graphs that exclude a fixed minor (Braverman et al., 2021b). In addition, coresets for variants of clustering have also been studied, notably capacitated clustering and the highly related fair clustering (Schmidt et al., 2019; Huang et al., 2019; Bandyapadhyay et al., 2021; Braverman et al., 2022), and robust clustering (Feldman & Schulman, 2012; Huang et al., 2023).

2. Preliminaries

We define (k, β)-MEDIAN problem in Definition 2.2 which depends on the notion of balanced center sets (Definition 2.1). The notion of weak coreset (Definition 2.3) captures the main guarantee of Theorem 3.1.

Definition 2.1 (Balanced Center Set)**.** Given a dataset $X \subseteq \mathcal{X}$, an integer $k \geq 1$ and $\beta \in (0, 1)$, we say a center set $C \subseteq \mathcal{X}$ is β-balanced if for every cluster X_i ($i \in [k]$) induced by C contains at least $\beta|X|/k$ points. Let $\mathcal{C}_\beta(X)$ denote the collection of all β-balanced center sets on X.

Recall that if the balancedness of X is β, there exists an optimal solution of k-MEDIAN on X that is β-balanced. Then we naturally define the following (k, β)-MEDIAN problem that aims to find optimal solutions within $C_\beta(X)$.

Definition 2.2 ((k, β)-MEDIAN)**.** Given a dataset $X \subseteq \mathcal{X}$, an integer $k \geq 1$ and $\beta \in (0, 1]$, the goal of the (k, β)-MEDIAN problem is to find a k-point set $C \in \mathcal{C}_\beta(X)$ that minimizes $\mathrm{cost}(X, C)$. Let $\mathrm{OPT}_\beta(X) := \mathrm{cost}(X, C^\star)$ where $C^\star$ is an optimal solution for (k, β)-MEDIAN on X.

Again, if the balancedness of X is β, solving (k, β)-MEDIAN also solves k-MEDIAN and $\mathrm{OPT}_\beta(X)$ equals the optimal k-MEDIAN cost on X. A similar notion of (k, β)-MEDIAN also appears in the literature, e.g., Bradley et al. (2000); Malinen & Fränti (2014); Costa et al. (2017); Lin et al. (2019); Ding (2020). The main difference is that they allow points being assigned to a non-closest center for

achieving a balanced dataset partition, and hence, consider all possible k points sets instead of $\mathcal{C}_\beta(X)$.

Definition 2.3 (Weak Coreset for (k,β)-MEDIAN)**.** Given a dataset $X \subseteq \mathcal{X}$, an integer $k \geq 1$ and $\beta, \varepsilon \in (0,1]$, an ε-weak coreset for (k,β)-MEDIAN on X is a subset $S \subseteq X$ such that for every k-point set $C \in \mathcal{C}_{\beta/2}(S)$ with $\sum_{x\in S} \operatorname{dist}(x,C) \leq (1+\varepsilon)\operatorname{OPT}_{\beta/2}(S)$, it holds that $\sum_{x\in X} \operatorname{dist}(x,C) \leq (1+O(\varepsilon))\operatorname{OPT}_\beta(X)$.

Intuitively, the above definition requires that any near-optimal solution for $(k,\beta/2)$-MEDIAN on a weak coreset S is a near-optimal solution for (k,β)-MEDIAN on X. Consequently, solving $(k,\beta/2)$-MEDIAN on S leads to a near-optimal solution for k-MEDIAN on X if the balancedness of X is β. The points are unweighted in our weak coreset, as opposed to the weighted points considered in (strong) coresets (Har-Peled & Mazumdar, 2004). This is natural since we consider uniform sampling.

Note that an optimal solution $C^\star$ for (k,β)-MEDIAN on X may not be β-balanced on S due to the small size of S, and hence, we consider a relaxed balancedness $\beta/2$ instead of β for S such that the considered collection $\mathcal{C}_{\beta/2}(S)$ is likely to include $C^\star$.

It is noteworthy that existing literature usually defines a weak coreset as one that preserves the cost for all near-optimal solutions, and thus allows any algorithms to run on it. In contrast, the weak coreset we define in Definition 2.3 requires a $(k,\beta/2)$-MEDIAN algorithm. Nonetheless, as stated in Section 1.1, this relaxation is necessary, provided that only uniform sampling is used and we have excluded the possibility of a set of $o(n)$ uniform samples even preserving the cost for the optimal solution. (see Lemma F.1).

3. Uniform Sampling Yields Weak Coreset

Theorem 3.1 (Main Theorem)**.** *Let $(\mathcal{X},\operatorname{dist})$ be a metric space and $X \subseteq \mathcal{X}$ be a dataset. Given an integer $k \geq 1$ and real numbers $\beta \in (0,1], \varepsilon \in (0,0.5)$, let integer m satisfy that*

$$m \geq O\left(\frac{k}{\beta\varepsilon^2}\left(\sum_{i=1}^{\log\varepsilon^{-1}} \sqrt{2^{-i}\log N_X^{2^{-i}}(m)}\right)^2\right) \tag{1}$$

where $N_X^\alpha(m)$ is the covering number defined in Definition 3.5. Then, a set S of m uniform samples from X is an ε-weak coreset for (k,β)-MEDIAN on X with probability at least 0.9.

The factor $\sum_{i=1}^{\log\varepsilon^{-1}} \sqrt{2^{-i}\log N_X^{2^{-i}}(m)}$ plays a similar role as the entropy integral (or Dudley integral) that are commonly used in the chaining argument (see e.g., Corollary 5.25 of van Handel 2016). A more concise sufficient condition for (1) is

$$m \geq O\left(\frac{k}{\beta\varepsilon^2}\cdot\log N_X^\varepsilon(m)\right) \tag{2}$$

which is derived directly by the monotonicity of the covering number, i.e., $N_X^{2^{-i}}(m) \leq N_X^\varepsilon(m)$ for all $i \leq \log\varepsilon^{-1}$. However, we still need to use (1) in order to obtain better sample complexity, especially for Euclidean spaces. We give bounds for this term for various metrics in Section 3.4.

Proof Overview To utilize the balancedness, we consider *candidate center sets* $\mathcal{C}'(X)$ (Definition 3.3) as a collection of center sets C that are "close" to $C^\star$. We show that any near-optimal center set for $(k,\beta/2)$-Median on S is likely to be a candidate (Lemma 3.4). Then it suffices to show that all center sets $C \in \mathcal{C}'(X)$ with $\operatorname{cost}(X,C) \geq (1+O(\varepsilon))\operatorname{OPT}_\beta(X)$, denoted as $\mathcal{C}^{\text{bad}}(X)$, are likely to have a large cost on S, i.e., $\operatorname{cost}(S,C) > (1+\varepsilon)\operatorname{OPT}_{\beta/2}(S)$. In other words, we need a uniform convergence guarantee on $\mathcal{C}^{\text{bad}}(X)$. To this end, we need to bound the "complexity" of $\mathcal{C}^{\text{bad}}(X)$, by considering the notion of *covering*, which may be viewed as a set of representatives, and the *covering number* which measures the complexity of the covering (Definition 3.5). In the last steps of the proof (Section 3.3), we reduce the above requirement for $\mathcal{C}^{\text{bad}}(X)$ to a Gaussian process and applies a chaining argument based on the covering.

3.1. Good Event and Candidate Center Sets

We first introduce some useful notations. Let $\lambda > 1000$ be a constant throughout this section. For a center set $C \subseteq \mathcal{X}$ and $x \in \mathcal{X}$, denote by $C(x) = \arg\min_{c\in C}\operatorname{dist}(c,x)$ the closest center in C to x (breaking ties arbitrarily). Let $C^\star$ denote an optimal center set of X for (k,β)-MEDIAN. For a subset $A \subseteq X$, and real numbers $\eta \in (0,1), \alpha > 0$, denote by $\mathcal{F}_\eta^{(\alpha)}(A) := \{C \in \mathcal{X}^k : \operatorname{cost}(A,C) \leq (1+\alpha)\operatorname{OPT}_\eta(A)\}$ the collection of all k-points whose cost is at most $(1+\alpha)\operatorname{OPT}_\eta(A)$ and let $\overline{\mathcal{F}}_\eta^{(\alpha)}(A) = \mathcal{X}^k \setminus \mathcal{F}_\eta^{(\alpha)}(A)$. By definition, $\mathcal{C}_S := \mathcal{C}_{\beta/2}(S) \cap \mathcal{F}_{\beta/2}^{(\varepsilon)}(S)$ is the collection of all $(1+\varepsilon)$-approximate k-point set for $(k,\beta/2)$-MEDIAN on S, and thus we know that S is an ε-coreset weak coreset for (k,β)-MEDIAN on X if and only if

$$\mathcal{C}_S \cap \overline{\mathcal{F}}_\beta^{(O(\varepsilon))}(X) = \emptyset. \tag{3}$$

Let $\mathcal{P}^\star = \{X_1^\star, ..., X_k^\star\}$ be the partition of X induced by $C^\star$. We denote by ξ_S the event that

$$\begin{aligned} &\frac{1}{m}\operatorname{cost}(S,C^\star) \leq \lambda\cdot\frac{1}{n}\operatorname{OPT}_\beta(X) \\ \wedge\quad &\forall i\in[k], \frac{|S\cap X_i^\star|}{|S|} \in (1\pm\frac{1}{2})\frac{|X_i^\star|}{|X|}, \end{aligned} \tag{4}$$

where the first condition requires that the average k-MEDIAN cost of S to $C^\star$ is not too large compared to that of X, and the second condition requires that the ratio of sampled points in every cluster $X_i^\star$ is close to the underlying one. The following lemma claims that ξ_S happens with high probability, and hence, we can condition on ξ_S in the analysis. The proof can be found in Appendix A.1.

Lemma 3.2. *ξ_S happens with probability at least 0.99.*

Definition 3.3 (Candidate Center Sets)**.** We say a k-point set $C \subseteq \mathcal{X}$ is a *candidate center set* if we have

$$\frac{1}{n}\sum_{x\in X}\operatorname{dist}(C^\star(x), C) \le \frac{6\lambda}{n}\operatorname{OPT}_\beta(X), \tag{5}$$

and for every $x \in X$,

$$|\operatorname{dist}(x, C) - \operatorname{dist}(x, C^\star)| \le \frac{6\lambda k}{\beta n}\operatorname{OPT}_\beta(X). \tag{6}$$

Let $\mathcal{C}'(X) := \{C \subseteq \mathcal{X} : |C| = k, C \text{ satisfies (5),(6)}\}$ denote the collection of all candidate center sets on X.

Intuitively, a candidate center set C is "close" to $C^\star$. (5) means that the average distance from every $C^\star(x)$ to C is not too large, and (6) states that all distance differences $|\operatorname{dist}(x, C) - \operatorname{dist}(x, C^\star)|$ are small. We note that the definition of candidate center sets is independent of S, which is useful for the probability analysis on S.

The following lemma states that all $(1+\varepsilon)$-approximate center sets for $(k, \beta/2)$-MEDIAN on S are candidate center sets conditioning on ξ_S. Recall that we want to prove (3), hence by Lemma 3.4, it remains to prove $\mathcal{C}_S \cap \left(\mathcal{C}'(X) \cap \overline{\mathcal{F}}_\beta^{(O(\varepsilon))}(X)\right) = \emptyset$, which is easier to handle due to good properties of $\mathcal{C}'(X)$.

Lemma 3.4. *$\mathcal{C}_S \subseteq \mathcal{C}'(X)$ holds conditioning on ξ_S.*

3.2. Covering and Covering Number

The notion of covering and covering number, defined in Definition 3.5, plays a crucial role in our analysis. We start with giving the definition, and then discuss how the several relevant parameters are chosen for our application.

Definition 3.5 (Covering and Covering Number)**.** Given a dataset $X \subseteq \mathcal{X}$ and a subset $S \subseteq X$, a set of vectors $V \subset \mathbb{R}^X$, an error function $\operatorname{err} : X \times V \to \mathbb{R}$ and real numbers $\alpha, \gamma > 0$, we say $U \subset \mathbb{R}^X$ is a γ-bounded α-covering of V w.r.t. (S, err) if the following holds:

1. (Bounded Covering Error) for every $v \in V$, there exists a vector $u \in U$ such that

$$\forall x \in S, \quad |v_x - u_x| \le \alpha \cdot \operatorname{err}(x, v)$$

2. (Bounded L_∞ Norm) for every $u \in U$, $\|u\|_\infty \le \gamma$.

Define $N^{\alpha,\gamma}(S, V, \operatorname{err})$ to be the minimum cardinality $|U|$ of any γ-bounded α-covering U of V w.r.t. (S, err). Moreover, let $\mathcal{S} \subseteq 2^X$ be a collection of subsets and define the γ-bounded α-covering number of V w.r.t. $(\mathcal{S}, \operatorname{err})$ to be

$$N_X^{\alpha,\gamma}(\mathcal{S}, V, \operatorname{err}) := \max_{S\in\mathcal{S}} N^{\alpha,\gamma}(S, V, \operatorname{err})$$

Explanation of Definition 3.5 The idea of ε-covering has also been used in the coreset literature, e.g., Cohen-Addad et al. (2022a;b); Huang et al. (2022). Intuitively, the covering may be viewed as a discretization/representative of V, and the covering number measures its complexity. The parameter α together with the error function err controls the granularity of the discretization of V, and the covering number $N^{\alpha,\gamma}(S, V, \operatorname{err})$ increases as α becomes larger. The relative errors $\operatorname{err}(x, v)$ should often be tailored to the application (e.g., Cohen-Addad et al. 2022a;b; Huang et al. 2022) and we need to use a specific definition of it. Compared with a standard definition of covering, we additionally require $\|u\|_\infty$ for all $u \in U$ bounded (by parameter γ). This requirement is useful for bounding the variance of a Gaussian process in our analysis, which plays a similar role as excluding huge subsets as in the definition of covering in Huang et al. (2022) (their Definition 3.2). A natural choice of $\mathcal{S}$ is the collection of all $S \subseteq X$ with a fixed cardinality but in our case we need additional constraints on $\mathcal{S}$ to bound the overall covering errors.

Specifying V, γ, err and $\mathcal{S}$ For a center set $C \subseteq \mathcal{X}$, define $v^C \in \mathbb{R}^X$ to be a cost vector such that

$$\forall x \in X, \quad v_x^C = \operatorname{dist}(x, C) - \operatorname{dist}(x, C^\star),$$

and this is motivated by (6) which considers the difference of the distances. Since our proof is focusing on $\mathcal{C}'(X)$, we consider the following V on candidate center sets:

$$V = \{v^C : C \in \mathcal{C}'(X)\}.$$

As (6) implies $\|v^C\|_\infty \le \frac{6\lambda k}{\beta n}\operatorname{OPT}_\beta(X)$, we select

$$\gamma = \frac{12\lambda k}{\beta n}\operatorname{OPT}_\beta(X).$$

Now we define function $\operatorname{err} : X \times V \to \mathbb{R}$. For every $x \in X$ and $v^C \in V$,

$$\begin{aligned} &\operatorname{err}(x, v^C) \\ =\ & v_x^C + 2\operatorname{dist}(x, C^\star) + \frac{1}{n}\operatorname{OPT}_\beta(X) \\ =\ & \operatorname{dist}(x, C) + \operatorname{dist}(x, C^\star) + \frac{1}{n}\operatorname{OPT}_\beta(X). \end{aligned}$$

The term $\frac{1}{n}\operatorname{OPT}_\beta(X)$ is consistent with the selection of γ. The term $\operatorname{dist}(x, C) + \operatorname{dist}(x, C^\star)$ is mainly designed

for obtaining a dimension-independent covering number in Euclidean spaces; see Lemma D.2 for details.

Finally, we specify $\mathcal{S}$ with an additional restriction ξ_S.

$$\mathcal{S}(m) = \{S \subseteq X : |S| \le m, \xi_S\}.$$

We shorten the notation of the covering number by

$$N_X^{\alpha}(m) := N_X^{\alpha,\gamma}\left(\mathcal{S}(m), V, \mathrm{err}\right),$$

and call it the α-covering number. We have the following lemma showing that ξ_S leads to a bound of the total covering error, which is helpful for bounding the variance of our Gaussian process. The proof can be found in Appendix B.

Lemma 3.6 (ξ_S Implies Bounded Covering Error)**.** *For S such that ξ_S happens, we have for every $C \in \mathcal{C}'(X)$,*

$$\sum_{x\in S} \mathrm{err}(x, v^C) \le \frac{15\lambda m}{n} \mathrm{OPT}_\beta(X).$$

3.3. Proof of Main Theorem: Theorem 3.1

Omitted proofs can be found in Appendix C. Conditioning on ξ_S, for every $C \in \mathcal{C}_S$, it holds that

$$\begin{aligned}\sum_{x\in S} \mathrm{dist}(x, C) &\le (1+\varepsilon) \sum_{x\in S} \mathrm{dist}(x, C^\star) \\ &\le \sum_{x\in S} \mathrm{dist}(x, C^\star) + \frac{\lambda\varepsilon m}{n} \mathrm{OPT}_\beta(X),\end{aligned} \tag{7}$$

where the second inequality is due to (4). Let

$$\mathcal{C}^{\mathrm{bad}}(X) := \mathcal{C}'(X) \cap \overline{\mathcal{F}}_\beta^{(10\lambda^2\varepsilon)}(X)$$

denote the collection of all candidate solutions C which are bad on X, i.e., $\mathrm{cost}(X, C) \ge (1 + 10\lambda^2\varepsilon)\,\mathrm{OPT}_\beta(X)$. To show S is an ε-weak coreset, let ϕ_S denotes the event that for every $C \in \mathcal{C}^{\mathrm{bad}}(X)$, (7) is far from being satisfied, that is, $\sum_{x\in S} \mathrm{dist}(x, C) \ge \sum_{x\in S} \mathrm{dist}(x, C^\star) + \frac{\lambda^2\varepsilon m}{n} \mathrm{OPT}_\beta(X)$. Then it suffices to bound the following probability

$$\begin{aligned}&\Pr_S[\mathcal{C}_S \cap \overline{\mathcal{F}}_\beta^{(10\lambda^2\varepsilon)}(X) \ne \emptyset] \\ &= \Pr_S[\mathcal{C}_S \cap \overline{\mathcal{F}}_\beta^{(10\lambda^2\varepsilon)}(X) \ne \emptyset \wedge \phi_S] \\ &\quad + \Pr_S[\mathcal{C}_S \cap \overline{\mathcal{F}}_\beta^{(10\lambda^2\varepsilon)}(X) \ne \emptyset \wedge \neg\phi_S] \\ &\le \Pr_S[\mathcal{C}_S \cap \overline{\mathcal{F}}_\beta^{(10\lambda^2\varepsilon)}(X) \ne \emptyset \wedge \phi_S] + \Pr_S[\neg\phi_S].\end{aligned}$$

For the first term, we have

$$\begin{aligned}&\Pr_S[\mathcal{C}_S \cap \overline{\mathcal{F}}_\beta^{(10\lambda^2\varepsilon)}(X) \ne \emptyset \wedge \phi_S] \\ &= \Pr_S[\mathcal{C}_S \cap \overline{\mathcal{F}}_\beta^{(10\lambda^2\varepsilon)}(X) \ne \emptyset \wedge \phi_S \mid \xi_S] \Pr[\xi_S] \\ &\quad + \Pr_S[\mathcal{C}_S \cap \overline{\mathcal{F}}_\beta^{(10\lambda^2\varepsilon)}(X) \ne \emptyset \wedge \phi_S \mid \neg\xi_S] \Pr_S[\neg\xi_S] \\ &\le \Pr_S[\mathcal{C}_S \cap \overline{\mathcal{F}}_\beta^{(10\lambda^2\varepsilon)}(X) \ne \emptyset \wedge \phi_S \mid \xi_S] + \Pr_S[\neg\xi_S]\end{aligned}$$

By the discussion above, conditioning on ξ_S, it holds that $\mathcal{C}_S \subseteq \mathcal{C}'(X)$. So $\mathcal{C}_S \cap \bar{\mathcal{F}}_0^{(10\lambda^2\varepsilon)}(X) \ne \emptyset$ is equivalent to $\mathcal{C}_S \cap \mathcal{C}^{\mathrm{bad}}(X) \ne \emptyset$, which implies that there exists $C \in \mathcal{C}^{\mathrm{bad}}(X)$ such that (7) holds. We know this contradicts ϕ_S. Therefore, we have

$$\Pr_S[\mathcal{C}_S \cap \overline{\mathcal{F}}_\beta^{(10\lambda^2\varepsilon)}(X) \ne \emptyset \wedge \phi_S \mid \xi_S] = 0,$$

and thus

$$\Pr_S[\mathcal{C}_S \cap \overline{\mathcal{F}}_\beta^{(10\lambda^2\varepsilon)}(X) \ne \emptyset \wedge \phi_S] \le \Pr_S[\neg\xi_S] \le 0.01,$$

where the second inequality follows from Lemma 3.2.

It remains to bound $\Pr_S[\neg\phi_S] \le 0.09$. Recall that for any $C \in \mathcal{C}^{\mathrm{bad}}(X)$, $\|v^C\|_1 = \sum_{x\in X}(\mathrm{dist}(x, C) - \mathrm{dist}(x, C^\star)) \ge 10\lambda^2\varepsilon\,\mathrm{OPT}_\beta(X)$. $\Pr[\neg\phi_S]$ can be regarded as a uniformly convergence guarantee for all $v^C \in V$ and, let $\mu := \frac{\varepsilon m}{n}\mathrm{OPT}_\beta(X)$, it is equivalent to bound

$$\Pr_S\left[\inf_{C\in\mathcal{C}^{\mathrm{bad}}(X)} \sum_{x\in S} v_x^C \le \lambda^2\mu\right] \le 0.09 \tag{8}$$

To bound (8), our plan is to set up a Gaussian process and apply the chaining argument. The following lemma provides a convergence guarantee for a single C.

Lemma 3.7. *For any $C \in \mathcal{C}^{\mathrm{bad}}(X)$, the following holds:* $\Pr_S\left[\sum_{x\in S} v_x^C \le 5\lambda^2\mu\right] \le 2\exp\left(-\Theta\left(\varepsilon^2\beta m/k\right)\right)$.

By Lemma 3.7, it remains to bound the "complexity" of $\mathcal{C}^{\mathrm{bad}}(X)$. To this end, we reduce to a Gaussian process where we only need to consider the complexity of coverings with respect to the sample set S. This idea is formalized in the following lemma.

Lemma 3.8 (Reduction to Gaussian Process)**.** (8) *holds if the following holds:*

$$\mathbb{E}_S\left[\mathbb{E}_{g_i}\left[\sup_{C\in\mathcal{C}^{\mathrm{bad}}(X)} \frac{1}{\mu}\left|\sum_{i=1}^m g_i u_{s_i}^C\right|\right] \mid \xi_S\right] \le \lambda. \tag{9}$$

Here, $g_1, \ldots, g_m$ are the independent standard Gaussian random variables, U_ε is an ε-covering of V w.r.t. a random set $S = \{s_1, \ldots, s_m\} \subseteq X$, and $u^C \in U_\varepsilon$ denotes the ε-covering of v^C for $C \in \mathcal{C}^{\mathrm{bad}}(X)$,

Now it suffices to prove (9). The main idea is to apply a chaining argument. For every $C \in \mathcal{C}^{\mathrm{bad}}(X)$, let $v^{C,h} \in U_{2^{-h}}$ denote the 2^{-h}-covering of v^C, then we can rewrite u^C as a telescoping sum $u^C = \sum_{h=1}^{\log \varepsilon^{-1}} (v^{C,h} - v^{C,h-1})$. Hence, it suffices to prove the following lemma which implies (9), and this completes the proof of Theorem 3.1.

Lemma 3.9 (Bounding Error in the Chaining Argument)**.** *Conditioning on ξ_S, the following holds:*

$$\sum_{h=1}^{\log \varepsilon^{-1}} \mathbb{E}_{g_i} \left[\sup_{C \in \mathcal{C}^{\mathrm{bad}}(X)} \frac{1}{\mu} \left| \sum_{i=1}^{m} g_i (v_{s_i}^{C,h} - v_{s_i}^{C,h-1}) \right| \right] \leq \lambda.$$

The proof of Lemma 3.9 heavily relies on our definition of covering. In particular, it allows us to bound the difference $|v_x^{C,h} - v_x^{C,h-1}|$ either by $2^{-h+2} \cdot \mathrm{err}(x, v^C)$, or by an absolute value 2γ. This eventually guarantees that each Gaussian variable $\sum_{i=1}^{m} g_i (v_{s_i}^{C,h} - v_{s_i}^{C,h-1})$ has a well bounded variance.

3.4. Weak Coresets in Various Metric Spaces

We apply Theorem 3.1 to various metric spaces and obtain the following theorem, by analyzing their covering number. The details can be found in Appendix D.

Theorem 3.10. *For a metric space $M = (\mathcal{X}, \mathrm{dist})$ and a dataset $X \subseteq \mathcal{X}$, an integer $k \geq 1$ and real numbers $\beta, \varepsilon \in (0, 1)$, let S be a set of uniform samples with size*

- $O\left(\frac{k^2}{\beta\varepsilon^3} \cdot \log^2 \frac{k}{\beta\varepsilon} \cdot \log^2 \frac{1}{\varepsilon}\right)$ *if M is Euclidean $\mathbb{R}^d$;*
- $O\left(\frac{k^2}{\beta\varepsilon^2} \cdot \mathrm{ddim} \cdot \log \frac{k}{\beta\varepsilon}\right)$ *if M has doubling dimension* ddim*;*
- $O\left(\frac{k^2}{\beta\varepsilon^2} \cdot \log |\mathcal{X}| \cdot \log \frac{k}{\beta\varepsilon}\right)$ *if M is a finite metric;*
- $O\left(\frac{k^2}{\beta\varepsilon^2} \cdot \mathrm{tw} \cdot \log \frac{k}{\beta\varepsilon}\right)$ *if M is the shortest-path metric of a graph with treewidth* tw.

*Then S is an ε-weak coreset for (k, β)-*MEDIAN *on X with probability* 0.9.

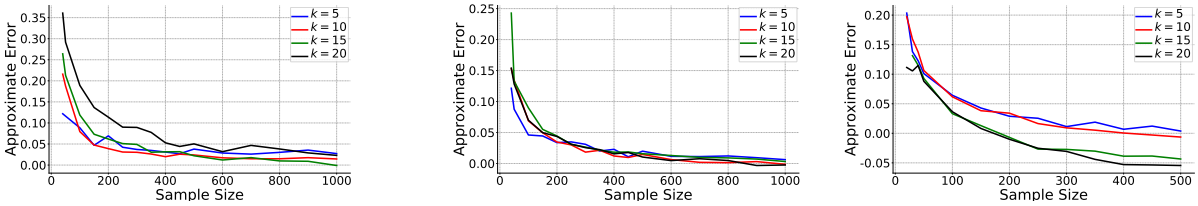
4. Experiments

Experiment Setup Our experiments are conducted on 3 real datasets: Twitter, Census1990 and NY. For Twitter (Chan et al., 2018) and Census1990 (Meek et al., 1990), we select numerical features and normalize them into Euclidean vectors in $\mathbb{R}^d$, with $n = 21040936$ data points and $d = 2$ for Twitter, and $n = 2458285, d = 68$ for Census1990. The NY dataset is a weighted graph with $|V| = 2276360$ vertices and $|E| = 2974516$ edges, representing the largest connected component of the road network of New York State, extracted from OpenStreetMap (OpenStreetMap contributors, 2017). All experiments are conducted on a PC with Intel Core i7 CPU and 16 GB memory, and algorithms are implemented in C++ 11.

Error Measure To measure the error of a center set C (which is a solution to k-MEDIAN), ideally one should compare $\mathrm{cost}(X, C)$ with OPT. However, since the exact value of OPT is NP-hard to find, we turn to compare it with an approximate solution C^{apx} instead. Hence, our measure of error for a center set C is defined as the following $\hat{\varepsilon}$ which is a signed relative error compared with $\mathrm{cost}(X, C^{\mathrm{apx}})$: $\hat{\varepsilon}(C) := \frac{\mathrm{cost}(X,C) - \mathrm{cost}(X,C^{\mathrm{apx}})}{\mathrm{cost}(X,C^{\mathrm{apx}})}$. The reason to keep the sign is that C^{apx} is not the optimal solution, hence it is possible that $\mathrm{cost}(X, C)$ is better than $\mathrm{cost}(X, C^{\mathrm{apx}})$ which leads to a negative sign. To compute C^{apx}, we first construct a coreset (e.g., using Feldman & Langberg (2011)) to reduce the size of the dataset, and then apply a standard approximation algorithm for k-MEDIAN (Arya et al., 2001) on top of it. The use of coreset is to ensure the local search can run for sufficiently large number of rounds to converge in a reasonable time.

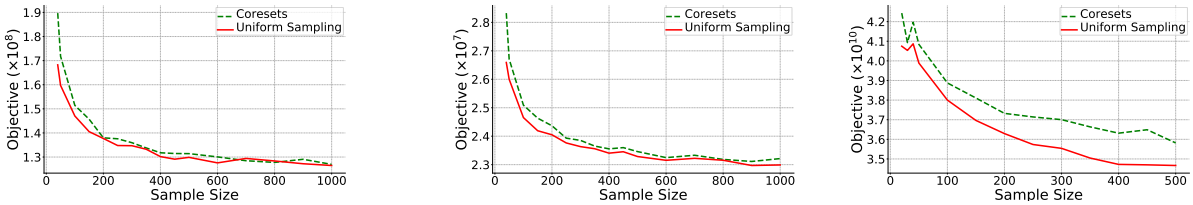
Experiment: Size-error Tradeoff and Balancedness Given $k \geq 1$, we take m uniform samples S from the dataset with varying m, and run a local search algorithm (Arya et al., 2001) on S to find a center set C_S for k-MEDIAN. We evaluate the error $\hat{\varepsilon}(C_S)$ and plot the size-error curves in Figure 1 for various choices of k. In fact, it turns out that the choice of k also affect the balancedness parameter β. Hence, for each dataset and value of k, we evaluate the balancedness β of the abovementioned C^{apx} on the original dataset, and we also take $m = 500$ uniform samples S and evaluate the balancedness β' of C_S on S. The resulted β and β' are reported in Table 1. To make the measurement stable, we repeat the sampling and local search 20 times independently and report the average statistics.

As can be seen from Figure 1, uniform sampling shows an outstanding performance, and admits a similar convergence of the error curve regardless of the choice of k and the consequent balancedness of datasets. In NY dataset, it even achieves negative error when $k = 20$, which means it performs even better than the baseline. We also observe in Table 1 that the datasets are mostly balanced even when k is relatively large, and thus the factor of $1/\beta$ is reasonably bounded (i.e., no greater than 100), and even for the less balanced scenario, the error is still very small. Moreover, the center set computed on S, although computed using a vanilla local search for k-MEDIAN, actually satisfies the balancedness constraint of (k, β)-MEDIAN problem, that is, $\beta' \geq \beta/2$ for every choice of k and every dataset. This suggests that it is not necessary to run the more sophisticated (k, β)-MEDIAN at all in practice. These findings help to justify why it is often seen in practice that a few uniform samples are enough to compute a good approximation.



Twitter dataset Census1990 dataset NY dataset

Figure 1: The size-error tradeoff of the uniform sampling.



Twitter dataset Census1990 dataset NY dataset

Figure 2: The size-objective tradeoff, compared with that of coresets.

Experiment: Comparison to Coreset Since uniform sampling is usually very efficient, we expect to see an advantage in running time compared with coreset constructions. Our second experiment aims to demonstrate this advantage, particularly to compare with the coresets constructed by importance sampling (Feldman & Langberg, 2011). We set $k = 20$, and vary the sample size m for both uniform sampling and importance sampling. Let S and S' be the subsets constructed by uniform sampling and importance sampling, respectively. We run a local search algorithms on each of S and S', and compute the objective value of the output center set on the original dataset. In Figure 2, we plot the objective-size curves, and observe that uniform sampling is comparable to coreset on all these datasets. However, for the running time, it takes 391s/114s/298s to compute the coreset on Twitter/Census1990/NY dataset, while the runtime of the local step is 16s/42s/203s. Hence, the coreset construction is even more costly than the local search, and the coreset actually becomes the bottleneck of the total running time. On the other hand, the uniform sampling takes merely 10^{-3}s to sample 500 points from the dataset which is significantly more efficient. This also justifies the practical benefit of using uniform sampling than methods like importance sampling.

Table 1: Balancedness under different values of k.

k	TWITTER		CENSUS1990		NY	
	β	β'	β	β'	β	β'
5	0.549	0.523	0.302	0.292	0.036	0.127
10	0.107	0.273	0.359	0.326	0.053	0.093
15	0.138	0.192	0.120	0.249	0.073	0.121
20	0.202	0.184	0.045	0.216	0.072	0.120

ACKNOWLEDGMENTS

Research is partially supported by a national key R&D program of China No. 2021YFA1000900, a startup fund from Peking University, and the Advanced Institute of Information Technology, Peking University.

References

Arya, V., Garg, N., Khandekar, R., Meyerson, A., Munagala, K., and Pandit, V. Local search heuristic for k-median and facility location problems. In *STOC*, pp. 21–29. ACM, 2001.

Assouad, P. Plongements lipschitziens dans $\mathbb{R}^n$. *Bulletin de la Société Mathématique de France*, 111:429–448, 1983.

Baker, D. N., Braverman, V., Huang, L., Jiang, S. H., Krauthgamer, R., and Wu, X. Coresets for clustering in graphs of bounded treewidth. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 569–579. PMLR, 2020.

Balcan, M., Ehrlich, S., and Liang, Y. Distributed k-means and k-median clustering on general communication topologies. In *NIPS*, pp. 1995–2003, 2013.

Bandyapadhyay, S., Fomin, F. V., and Simonov, K. On

coresets for fair clustering in metric and Euclidean spaces and their applications. In *ICALP*, volume 198 of *LIPIcs*, pp. 23:1–23:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.

Ben-David, S. A framework for statistical clustering with constant time approximation algorithms for k-median and k-means clustering. *Mach. Learn.*, 66(2-3):243–257, 2007.

Bradley, P. S., Bennett, K. P., and Demiriz, A. Constrained k-means clustering. *Microsoft Research, Redmond*, 20 (0):0, 2000.

Braverman, V., Jiang, S. H., Krauthgamer, R., and Wu, X. Coresets for clustering with missing values. In *NeurIPS*, pp. 17360–17372, 2021a.

Braverman, V., Jiang, S. H., Krauthgamer, R., and Wu, X. Coresets for clustering in excluded-minor graphs and beyond. In *SODA*, pp. 2679–2696. SIAM, 2021b.

Braverman, V., Cohen-Addad, V., Jiang, S. H., Krauthgamer, R., Schwiegelshohn, C., Toftrup, M. B., and Wu, X. The power of uniform sampling for coresets. In *FOCS*, pp. 462–473. IEEE, 2022.

Chan, T.-H. H., Guerquin, A., and Sozio, M. Twitter data set, 2018. https://github.com/fe6Bc5R4JvLkFkSeExHM/k-center.

Charikar, M., Guha, S., Tardos, É., and Shmoys, D. B. A constant-factor approximation algorithm for the k-median problem. *J. Comput. Syst. Sci.*, 65(1):129–149, 2002.

Chen, K. 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.

Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii, S. Fair clustering through fairlets. In *NIPS*, pp. 5029–5037, 2017.

Cohen-Addad, V., Saulpic, D., and Schwiegelshohn, C. Improved coresets and sublinear algorithms for power means in Euclidean spaces. In *NeurIPS*, pp. 21085–21098, 2021a.

Cohen-Addad, V., Saulpic, D., and Schwiegelshohn, C. A new coreset framework for clustering. In *STOC*, pp. 169–182. ACM, 2021b.

Cohen-Addad, V., Larsen, K. G., Saulpic, D., and Schwiegelshohn, C. Towards optimal lower bounds for k-median and k-means coresets. In *STOC*, pp. 1038–1051. ACM, 2022a.

Cohen-Addad, V., Larsen, K. G., Saulpic, D., Schwiegelshohn, C., and Sheikh-Omar, O. A. Improved coresets for euclidean k-means. In *NeurIPS*, pp. 2679–2694, 2022b.

Costa, L. R., Aloise, D., and Mladenovic, N. Less is more: basic variable neighborhood search heuristic for balanced minimum sum-of-squares clustering. *Inf. Sci.*, 415:247–253, 2017.

Czumaj, A. and Sohler, C. Sublinear-time approximation algorithms for clustering via random sampling. *Random Struct. Algorithms*, 30(1-2):226–256, 2007.

Danos, M. Coresets for clustering by uniform sampling and generalized rank aggregation. Master's thesis, Weizmann Institute of Science, 2021.

Ding, H. Faster balanced clusterings in high dimension. *Theor. Comput. Sci.*, 842:28–40, 2020.

Feldman, D. and Langberg, M. A unified framework for approximating and clustering data. In *STOC*, pp. 569–578. ACM, 2011.

Feldman, D. and Schulman, L. J. Data reduction for weighted and outlier-resistant clustering. In *SODA*, pp. 1343–1354. SIAM, 2012.

Feldman, D., Schmidt, M., and Sohler, C. Turning big data into tiny data: Constant-size coresets for k-means, pca, and projective clustering. *SIAM J. Comput.*, 49(3):601–657, 2020.

Gupta, A., Krauthgamer, R., and Lee, J. R. Bounded geometries, fractals, and low-distortion embeddings. In *FOCS*, pp. 534–543. IEEE Computer Society, 2003.

Har-Peled, S. and Kushal, A. Smaller coresets for k-median and k-means clustering. *Discret. Comput. Geom.*, 37(1):3–19, 2007.

Har-Peled, S. and Mazumdar, S. On coresets for k-means and k-median clustering. In *STOC*, pp. 291–300. ACM, 2004.

Henzinger, M. and Kale, S. Fully-dynamic coresets. In *ESA*, volume 173 of *LIPIcs*, pp. 57:1–57:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

Huang, L. and Vishnoi, N. K. Coresets for clustering in Euclidean spaces: importance sampling is nearly optimal. In *STOC*, pp. 1416–1429. ACM, 2020.

Huang, L., Jiang, S. H., Li, J., and Wu, X. Epsilon-coresets for clustering (with outliers) in doubling metrics. In *FOCS*, pp. 814–825. IEEE Computer Society, 2018.

Huang, L., Jiang, S. H., and Vishnoi, N. K. Coresets for clustering with fairness constraints. In *NeurIPS*, pp. 7587–7598, 2019.

Huang, L., Li, J., and Wu, X. On optimal coreset construction for euclidean (k, z)-clustering. *CoRR*, abs/2211.11923, 2022.

Huang, L., Jiang, S. H., Lou, J., and Wu, X. Near-optimal coresets for robust clustering. In *ICLR*, 2023. To appear.

Jubran, I., Tukan, M., Maalouf, A., and Feldman, D. Sets clustering. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4994–5005. PMLR, 2020.

Lin, W., He, Z., and Xiao, M. Balanced clustering: A uniform model and fast algorithm. In *IJCAI*, pp. 2987–2993. ijcai.org, 2019.

Lu, F., Raff, E., and Holt, J. A coreset learning reality check. In *AAAI*, 2023. To appear.

Mai, T., Musco, C., and Rao, A. Coresets for classification - simplified and strengthened. In *NeurIPS*, pp. 11643–11654, 2021.

Malinen, M. I. and Fränti, P. Balanced k-means for clustering. In *S+SSPR*, volume 8621 of *Lecture Notes in Computer Science*, pp. 32–41. Springer, 2014.

Marom, Y. and Feldman, D. k-Means clustering of lines for big data. In *NeurIPS*, pp. 12797–12806, 2019.

Massart, P. *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer, 2007.

Meek, C., Thiesson, B., and Heckerman, D. Uci machine learning repository, 1990. http://archive.ics.uci.edu/ml/datasets/US+Census+Data+(1990).

Meyerson, A., O'Callaghan, L., and Plotkin, S. A. A k-median algorithm with running time independent of data size. *Mach. Learn.*, 56(1-3):61–87, 2004.

Mishra, N., Oblinger, D., and Pitt, L. Sublinear time approximate clustering. In *SODA*, pp. 439–447. ACM/SIAM, 2001.

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

Munteanu, A., Omlor, S., and Woodruff, D. P. Oblivious sketching for logistic regression. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7861–7871. PMLR, 2021.

Munteanu, A., Omlor, S., and Peters, C. p-generalized probit regression and scalable maximum likelihood estimation via sketching and coresets. In *AISTATS*, volume 151 of *Proceedings of Machine Learning Research*, pp. 2073–2100. PMLR, 2022.

Munteanu, A., Omlor, S., and Woodruff, D. P. Almost linear constant-factor sketching for l_1 and logistic regression. In *ICLR*, 2023. To appear.

Narayanan, S. and Nelson, J. Optimal terminal dimensionality reduction in Euclidean space. In *STOC*, pp. 1064–1069. ACM, 2019.

OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org, 2017.

Schmidt, M., Schwiegelshohn, C., and Sohler, C. Fair coresets and streaming algorithms for fair k-means. In *WAOA*, volume 11926 of *Lecture Notes in Computer Science*, pp. 232–251. Springer, 2019.

Sohler, C. and Woodruff, D. P. Strong coresets for k-median and subspace approximation: Goodbye dimension. In *FOCS*, pp. 802–813. IEEE Computer Society, 2018.

Talagrand, M. Majorizing measures: the generic chaining. *The Annals of Probability*, 24(3):1049–1103, 1996.

van Handel, R. *Probability in High Dimension*. 2016.

Woodruff, D. P. and Yasuda, T. Online lewis weight sampling. In *SODA*, pp. 4622–4666. SIAM, 2023.

Yao, A. C. Lower bounds by probabilistic arguments (extended abstract). In *FOCS*, pp. 420–428. IEEE Computer Society, 1983.

A. Missing Proofs in Section 3.1

A.1. Proof of Lemma 3.2

Lemma A.1 (Restatement of Lemma 3.2)**.** *ξ_S happens with probability at least 0.99.*

Proof. Since $C^\star$ is a β-balanced center set on X, we have $|X_i^\star| \geq \frac{\beta n}{k}$ for every $i \in [k]$. Recall that S is a set of uniform samples, therefore by Chernoff bound, we have

$$\begin{aligned}\Pr\left[\frac{|S \cap X_i^\star|}{|S|} \notin \left(1 \pm \frac{1}{2}\right)\frac{|X_i^\star|}{|X|}\right] &\leq 2\exp\left(-\frac{\beta m}{12k}\right)\\ &\leq 0.001/k.\end{aligned}$$

By the union bound, we have

$$\Pr\left[\forall i \in [k], \frac{|S \cap X_i^\star|}{|S|} \in \left(1 \pm \frac{1}{2}\right)\frac{|X_i^\star|}{|X|}\right] \geq 0.999.$$

Also note that $\mathbb{E}_S[\mathrm{cost}(S, C^\star)] = \frac{m}{n}\,\mathrm{OPT}_\beta(X)$. Then by the Markov inequality, with probability at least 0.999, we have $\mathrm{cost}(S, C^\star) \leq \lambda \cdot \frac{m}{n}\,\mathrm{OPT}_\beta(X)$ since $\lambda > 1000$. This completes the proof. □

A.2. Proof of Lemma 3.4

Lemma A.2 (Restatement of Lemma 3.4)**.** *$\mathcal{C}_S \subseteq \mathcal{C}'(X)$ holds conditioning on ξ_S.*

Proof. Conditioning on ξ_S, we have $C^\star \in \mathcal{C}_{\beta/2}(S)$ and for every $C \in \mathcal{C}_S$, it holds that

$$\begin{aligned}&\sum_{x\in S}\mathrm{dist}(C^\star(x), C)\\ \leq\ &\sum_{x\in S}(\mathrm{dist}(x, C^\star) + \mathrm{dist}(x, C))\\ \leq\ &(2+\varepsilon)\sum_{x\in S}\mathrm{dist}(x, C^\star)\\ \leq\ &\frac{3\lambda m}{n}\,\mathrm{OPT}_\beta(X),\end{aligned} \tag{10}$$

where the first derivation is due to the triangle inequality, the second derivation is because $\sum_{x\in S}\mathrm{dist}(x, C) \leq (1+\varepsilon)\,\mathrm{OPT}_{\beta/2}(S) \leq (1+\varepsilon)\sum_{x\in S}\mathrm{dist}(x, C^\star)$, and the last derivation is due to (4).

We can rewrite $\sum_{x\in S}\mathrm{dist}(C^\star(x), C)$ as $\sum_{c_i^\star \in C^\star}\mathrm{dist}(c_i^\star, C)\cdot|S \cap X_i^\star|$, where $c_i^\star$ denotes the center of cluster $X_i^\star$. Therefore, we have by (4),

$$\begin{aligned}\sum_{x\in S}\mathrm{dist}(C^\star(X), C) \geq\ &\frac{m}{2n}\sum_{c_i^\star \in C^\star}\mathrm{dist}(c_i^\star, C)\cdot|X_i^\star|\\ =\ &\frac{m}{2n}\sum_{x\in X}\mathrm{dist}(C^\star(x), C),\end{aligned}$$

which combining with (10) completes the proof of (5).

To prove (6), we observe that for every $x \in X$, it holds that

$$\begin{aligned}&|\mathrm{dist}(x, C) - \mathrm{dist}(x, C^\star)|\\ \leq\ &\max\{\mathrm{dist}(C(x), C^\star), \mathrm{dist}(C^\star(x), C)\}.\end{aligned}$$

We only upper bound $\mathrm{dist}(c, C^\star)$ for every $c \in C$, and bounding $\mathrm{dist}(c^\star, C)$ for every $c^\star \in C^\star$ is almost the same. Since

$C \in \mathcal{C}_{\beta/2}(S)$, for every $c \in C$, we have

$$\begin{aligned}
\operatorname{dist}(c, C^\star) &\leq \sum_{c \in C} \operatorname{dist}(c, C^\star) \\
&\leq \frac{2k}{\beta} \cdot \frac{1}{m} \sum_{c \in C} \frac{\beta m}{2k} \operatorname{dist}(c, C^\star) \\
&\leq \frac{2k}{\beta} \cdot \frac{1}{m} \sum_{x \in S} \operatorname{dist}(C(x), C^\star) \\
&\leq \frac{2k}{\beta} \cdot \frac{1}{m} \sum_{x \in S} (\operatorname{dist}(x, C) + \operatorname{dist}(x, C^\star)) \\
&\leq \frac{2k}{\beta} \cdot \frac{2+\varepsilon}{m} \sum_{x \in S} \operatorname{dist}(x, C^\star) \\
&\leq \frac{6\lambda k}{\beta n} \operatorname{OPT}_\beta(X),
\end{aligned}$$

where the third derivation is because $\sum_{x \in S} \operatorname{dist}(C(x), C^\star) = \sum_{c_i \in C} \operatorname{dist}(c_i, C^\star) \cdot |X_i|$ for $\{X_1, \ldots, X_k\}$ being the partition of X induced by C, and $|X_i| \geq \frac{\beta m}{2k}$ since $C \in \mathcal{C}_{\beta/2}(S)$. □

B. Proof of Lemma 3.6

Lemma B.1 (Restatement of Lemma 3.6)**.** *For S such that ξ_S happens, we have for every $C \in \mathcal{C}'(X)$,*

$$\sum_{x \in S} \operatorname{err}(x, v^C) \leq \frac{15\lambda m}{n} \operatorname{OPT}_\beta(X).$$

Proof. ξ_S implies that $\operatorname{cost}(S, C^\star) \leq \lambda \cdot \frac{m}{n} \operatorname{OPT}_\beta(X)$ and

$$\begin{aligned}
\sum_{x \in S} \operatorname{dist}(C^\star(x), C) &= \sum_{c_i^\star \in C^\star} \operatorname{dist}(c_i^\star, C) \cdot |S \cap X_i^\star| \\
&\leq \sum_{c_i^\star \in C^\star} \operatorname{dist}(c_i^\star, C) \cdot 2|X_i^\star| \cdot \frac{m}{n} \\
&\leq \frac{2m}{n} \sum_{x \in X} \operatorname{dist}(C^\star(x), C) \\
&\leq \frac{12\lambda m}{n} \operatorname{OPT}_\beta(X).
\end{aligned}$$

where the second derivation is due to (4), and the forth derivation is due to (5). Therefore for every $C \in \mathcal{C}'(X)$, it holds that

$$\begin{aligned}
\sum_{x \in S} \operatorname{err}(x, v^C) &= \sum_{x \in S} \left(\operatorname{dist}(x, C) + \operatorname{dist}(x, C^\star) + \frac{1}{n} \operatorname{OPT}_\beta(X) \right) \\
&\leq \frac{m}{n} \operatorname{OPT}_\beta(X) + \sum_{x \in S} (2 \operatorname{dist}(x, C^\star) + \operatorname{dist}(C^\star(x), C)) \\
&\leq \frac{15\lambda m}{n} \operatorname{OPT}_\beta(X)
\end{aligned}$$

□

C. Missing Proofs in Section 3.3

Lemma C.1 (Restatement of Lemma 3.7)**.** *For any $C \in \mathcal{C}^{\mathrm{bad}}(X)$, the following holds:*

$$\Pr_S\left[\sum_{x\in S} v_x^C \le 5\lambda^2\mu\right] \le 2\exp\left(-\Theta\left(\frac{\varepsilon^2\beta m}{k}\right)\right),$$

Proof. Since $\mathcal{C}^{\mathrm{bad}}(X) \subseteq \mathcal{C}'(X)$, due to (6), for every $C \in \mathcal{C}^{\mathrm{bad}}(X)$, it holds that

$$\|v^C\|_\infty \le \gamma = \frac{12\lambda k}{\beta n}\,\mathrm{OPT}_\beta(X).$$

Since S is a set of m uniform samples, and $\mathbb{E}\sum_{x\in S} v_x^C = \frac{m}{n}\|v^C\|_1 \ge \frac{10\lambda^2\varepsilon m}{n}\,\mathrm{OPT}_\beta(X)$, we have

$$\Pr_S\left[\sum_{x\in S} v_x^C \le \frac{5\lambda^2\varepsilon m}{n}\,\mathrm{OPT}_\beta(X)\right] \le \Pr_S\left[\left|\sum_{x\in S} v_x^C - \mathbb{E}\sum_{x\in S} v_x^C\right| \ge \frac{5\lambda^2\varepsilon m}{n}\,\mathrm{OPT}_\beta(X)\right]$$

Here we apply Bernstein inequality to finish the proof. To this end, we should also bound the variance of $\sum_{x\in S} v_x^C$, which is

$$\begin{aligned}
&\le m\cdot\frac{1}{n}\sum_{x\in X}(v_x^C)^2\\
&\le \gamma\cdot\frac{m}{n}\sum_{x\in X}|v_x^C|\\
&\le \gamma\cdot\frac{m}{n}\sum_{x\in X}(\mathrm{dist}(x,C)+\mathrm{dist}(x,C^\star))\\
&\le \gamma\cdot\frac{m}{n}\sum_{x\in X}(2\,\mathrm{dist}(x,C^\star)+\mathrm{dist}(C^\star(x),C))\\
&\le O\left(\frac{km}{\beta n^2}(\mathrm{OPT}_\beta(X))^2\right).
\end{aligned}$$

where the last derivation is due to (5). Therefore, by Bernstein inequality, we have

$$\Pr_S\left[\left|\sum_{x\in S} v_x^C - \mathbb{E}\sum_{x\in S} v_x^C\right| \ge \frac{5\lambda^2\varepsilon m}{n}\,\mathrm{OPT}_\beta(X)\right] \le 2\exp\left(-\Theta\left(\frac{\varepsilon^2\beta m}{k}\right)\right)$$

which completes the proof. □

Lemma C.2 (Restatement of Lemma 3.8)**.** (8) *holds if the following holds:*

$$\mathbb{E}_S\left[\mathbb{E}_{g_i}\left[\sup_{C\in\mathcal{C}^{\mathrm{bad}}(X)}\frac{1}{\mu}\left|\sum_{i=1}^m g_i u_{s_i}^C\right|\right] \mid \xi_S\right] \le \lambda. \tag{11}$$

Here, $g_1,\ldots,g_m$ are the independent standard Gaussian random variables, U_ε is an ε-covering of V w.r.t. a random set $S=\{s_1,\ldots,s_m\}\subseteq X$, and $u^C \in U_\varepsilon$ denotes the ε-covering of v^C for $C\in\mathcal{C}^{\mathrm{bad}}(X)$.

Proof. Let $S' = \{s'_1,\ldots,s'_m\}$ be another set of independent uniform samples from X that is independent of S. We first use the symmetrization trick to show that

$$\Pr_S\left[\inf_{C\in\mathcal{C}^{\mathrm{bad}}(X)}\sum_{i=1}^m v_{s_i}^C \le \lambda^2\mu\right] \le 2\Pr_{S,S'}\left[\sup_{C\in\mathcal{C}^{\mathrm{bad}}(X)}\left|\sum_{i=1}^m (v_{s_i}^C - v_{s'_i}^C)\right| \ge 4\lambda^2\mu\right].$$

To see this, we assume for some S, the event ϕ_S, i.e., $\inf_{C\in\mathcal{C}^{\mathrm{bad}}(X)}\sum_{i=1}^m v_{s_i}^C \le \lambda^2\mu$ happens, then take an arbitrarily $C_S \in \mathcal{C}^{\mathrm{bad}}(X)$ such that $\sum_{i=1}^m v_{s_i}^{C_S} \le \lambda^2\mu$ (if ξ_S does not happen, we let C_S be an arbitrarily center set). If the event $\sum_{i=1}^m v_{s'_i}^{C_S} \ge 5\lambda^2\mu$, denoted by $\varphi_{C_S,S'}$, happens, then it holds that

$$\left|\sum_{i=1}^m (v_{s_i}^{C_S} - v_{s'_i}^{C_S})\right| \ge 4\lambda^2\mu.$$

Note that $\varphi_{C_S,S'}$ is a convergence guarantee for $C_S \in \mathcal{C}^{\text{bad}}(X)$, and Lemma 3.7 gives a lower bound of $1/2$ to $\Pr_{S'}[\varphi_{C_S,S'}]$. Therefore, the following holds.

$$\begin{aligned}
&\Pr_{S,S'}\left[\sup_{C\in\mathcal{C}^{\text{bad}}(X)}\left|\sum_{i=1}^{m}(v_{s_i}^C - v_{s_i'}^C)\right| \geq 4\lambda^2\mu\right] \\
&\geq \Pr_{S,S'}\left[\phi_S \wedge \varphi_{C_S,S'}\right] \\
&= \Pr_{S}[\phi_S]\Pr_{S,S'}[\varphi_{C_S,S'} \mid \phi_S] \\
&= \Pr_{S}[\phi_S]\,\mathbb{E}_S\left[\Pr_{S'}[\varphi_{C_S,S'}] \mid \phi_S\right] \\
&\geq \frac{1}{2}\Pr_{S}[\phi_S],
\end{aligned}$$

Let $r_1, ..., r_m$ be independent Rademacher random variables[3]. We have

$$\begin{aligned}
&\Pr_{S,S'}\left[\sup_{C\in\mathcal{C}^{\text{bad}}(X)}\left|\sum_{i=1}^{m}(v_{s_i}^C - v_{s_i'}^C)\right| \geq 4\lambda^2\mu\right] \\
&= \Pr_{S,S',r_i}\left[\sup_{C\in\mathcal{C}^{\text{bad}}(X)}\left|\sum_{i=1}^{m}r_i(v_{s_i}^C - v_{s_i'}^C)\right| \geq 4\lambda^2\mu\right] \\
&\leq \Pr_{S,S',r_i}\left[\sup_{C\in\mathcal{C}^{\text{bad}}(X)}\left(\left|\sum_{i=1}^{m}r_i v_{s_i}^C\right| + \left|\sum_{i=1}^{m}r_i v_{s_i'}^C\right|\right) \geq 4\lambda^2\mu\right]
\end{aligned}$$

where the first derivation is because $v_{s_i}^C - v_{s_i'}^C$ is symmetric and thus is distributed identically to $r_i(v_{s_i}^C - v_{s_i'}^C)$, the second derivation is due to the triangle inequality. If $\sup_{C\in\mathcal{C}^{\text{bad}}(X)}\left(\left|\sum_{i=1}^m r_i v_{s_i}^C\right| + \left|\sum_{i=1}^m r_i v_{s_i'}^C\right|\right) \geq 4\lambda^2\mu$ holds, then either $\sup_{C\in\mathcal{C}^{\text{bad}}(X)}\left|\sum_{i=1}^m r_i v_{s_i}^C\right| \geq 2\lambda^2\mu$ or $\sup_{C\in\mathcal{C}^{\text{bad}}(X)}\left|\sum_{i=1}^m r_i v_{s_i'}^C\right| \geq 2\lambda^2\mu$ holds. Since S and S' are distributed identically, by union bound, we have

$$\begin{aligned}
&\Pr_{S,S',r_i}\left[\sup_{C\in\mathcal{C}^{\text{bad}}(X)}\left(\left|\sum_{i=1}^{m}r_i v_{s_i}^C\right| + \left|\sum_{i=1}^{m}r_i v_{s_i'}^C\right|\right) \geq 4\lambda^2\mu\right] \\
&\leq 2\Pr_{S,r_i}\left[\sup_{C\in\mathcal{C}^{\text{bad}}(X)}\left|\sum_{i=1}^{m}r_i v_{s_i}^C\right| \geq 2\lambda^2\mu\right] \\
&= 2\Pr_{S,r_i}\left[\sup_{C\in\mathcal{C}^{\text{bad}}(X)}\left|\sum_{i=1}^{m}r_i v_{s_i}^C\right| \geq 2\lambda^2\mu \mid \xi_S\right]\Pr[\xi_S] \\
&\quad + 2\Pr_{S,r_i}\left[\sup_{C\in\mathcal{C}^{\text{bad}}(X)}\left|\sum_{i=1}^{m}r_i v_{s_i}^C\right| \geq 2\lambda^2\mu \mid \neg\xi_S\right]\Pr[\neg\xi_S] \\
&\leq 2\Pr_{S,r_i}\left[\sup_{C\in\mathcal{C}^{\text{bad}}(X)}\left|\sum_{i=1}^{m}r_i v_{s_i}^C\right| \geq 2\lambda^2\mu \mid \xi_S\right] + 0.02,
\end{aligned} \tag{12}$$

It suffices to prove that

$$\mathbb{E}_S\left[\mathbb{E}_{r_i}\sup_{C\in\mathcal{C}^{\text{bad}}(X)}\left|\sum_{i=1}^{m}r_i v_{s_i}^C\right| \mid \xi_S\right] \leq 20\lambda\mu. \tag{13}$$

We can thus apply Markov inequality to bound the probability in (12) by 0.01, which leads to $\Pr[\neg\phi_S] \leq 0.08$ and completes the proof of (8). It remains to show how to derive (13).

[3] A Rademacher random variable r takes value -1 with probability $1/2$ and takes value 1 with probability $1/2$.

Recall that U_ε denotes an ε-covering of V w.r.t. random subset S. We next replace the cost vector v^C with its ε-covering $u^C \in U_\varepsilon$, which satisfies that $|u_x^C - v_x^C| \le \varepsilon \cdot \mathrm{err}(x, v^C)$ for every $x \in S$.

$$\begin{aligned}
&\sup_{C\in\mathcal{C}^{\mathrm{bad}}(X)} \left|\sum_{i=1}^m r_i v_{s_i}^C\right| \\
\le &\sup_{C\in\mathcal{C}^{\mathrm{bad}}(X)} \left|\sum_{i=1}^m r_i u_{s_i}^C + r_i(v_{s_i}^C - u_{s_i}^C)\right| \\
\le &\sup_{C\in\mathcal{C}^{\mathrm{bad}}(X)} \left(\left|\sum_{i=1}^m r_i u_{s_i}^C\right| + \varepsilon \sum_{i=1}^m \mathrm{err}(s_i, v^C)\right) \\
\le &\sup_{C\in\mathcal{C}^{\mathrm{bad}}(X)} \left|\sum_{i=1}^m r_i u_{s_i}^C\right| + \varepsilon \cdot \frac{15\lambda m}{n} \mathrm{OPT}_\beta(X) \\
\le &\sup_{C\in\mathcal{C}^{\mathrm{bad}}(X)} \left|\sum_{i=1}^m r_i u_{s_i}^C\right| + 15\lambda\mu
\end{aligned} \tag{14}$$

where the second derivation is due to the triangle inequality, and the third derivation is due to Lemma 3.6 and $|r_i| = 1$ for every $i \in [m]$. Then it suffices to prove

$$\mathbb{E}_S\left[\mathbb{E}_{r_i} \sup_{C\in\mathcal{C}^{\mathrm{bad}}(X)} \left|\sum_{i=1}^m r_i u_{s_i}^C\right| \mid \xi_S\right] \le 5\lambda\mu$$

which is equivalent to

$$\mathbb{E}_S\left[\mathbb{E}_{r_i} \sup_{C\in\mathcal{C}^{\mathrm{bad}}(X)} \frac{1}{\mu}\left|\sum_{i=1}^m r_i u_{s_i}^C\right| \mid \xi_S\right] \le 5\lambda \tag{15}$$

Finally, we replace the Rademacher random variables with standard Gaussian random variables.

Lemma C.3 (Lemma 7.4 of van Handel (2016))**.** *For $r_1, \ldots, r_m$ are Rademacher random variables, let $g_1, \ldots, g_m$ be the independent standard Gaussian random variables, it holds that*

$$\begin{aligned}
&\mathbb{E}_{r_i}\left[\sup_{C\in\mathcal{C}^{\mathrm{bad}}(X)} \frac{1}{\mu}\left|\sum_{i=1}^m r_i u_{s_i}^C\right|\right] \\
\le &\sqrt{\frac{\pi}{2}}\, \mathbb{E}_{g_i}\left[\sup_{C\in\mathcal{C}^{\mathrm{bad}}(X)} \frac{1}{\mu}\left|\sum_{i=1}^m g_i u_{s_i}^C\right|\right]
\end{aligned}$$

It suffices to prove

$$\mathbb{E}_S\left[\mathbb{E}_{g_i} \sup_{C\in\mathcal{C}^{\mathrm{bad}}(X)} \frac{1}{\mu}\left|\sum_{i=1}^m g_i u_{s_i}^C\right| \mid \xi_S\right] \le \lambda$$

which leads to (15) by Lemma C.3, and completes the proof. □

Lemma C.4 (Restatement of Lemma 3.9)**.** *Conditioning on ξ_S, the following holds:*

$$\sum_{h=1}^{\log\varepsilon^{-1}} \mathbb{E}_{g_i}\left[\sup_{C\in\mathcal{C}^{\mathrm{bad}}(X)} \frac{1}{\mu}\left|\sum_{i=1}^m g_i(v_{s_i}^{C,h} - v_{s_i}^{C,h-1})\right|\right] \le \lambda.$$

Proof. For every $h \in [\log\varepsilon^{-1}]$, let

$$E_h := \mathbb{E}_{g_i}\left[\sup_{C\in\mathcal{C}^{\mathrm{bad}}(X)} \frac{1}{\mu}\left|\sum_{i=1}^m g_i(v_{s_i}^{C,h} - v_{s_i}^{C,h-1})\right|\right].$$

For every $C \in \mathcal{C}^{\mathrm{bad}}(X)$, $\sum_{i=1}^{m} \frac{g_i}{\mu}(v_{s_i}^{C,h} - v_{s_i}^{C,h-1})$ is Gaussian with zero mean and variance

$$\begin{aligned}
&\sum_{i=1}^{m} \left(\frac{1}{\mu}(v_{s_i}^{C,h} - v_{s_i}^{C,h-1})\right)^2 \\
&\leq \frac{\|v^{C,h}\|_\infty + \|v^{C,h-1}\|_\infty}{\mu} \sum_{i=1}^{m} \frac{|v_{s_i}^{C,h} - v_{s_i}^{C,h-1}|}{\mu} \\
&\leq \frac{24\lambda k}{\beta\varepsilon m} \sum_{i=1}^{m} \frac{|v_{s_i}^{C,h} - v_{s_i}^{C} + v_{s_i}^{C} - v_{s_i}^{C,h-1}|}{\mu} \\
&\leq \frac{24\lambda k}{\beta\varepsilon m} \sum_{x\in S} \frac{2^{-h+2}\mathrm{err}(x, v^C)}{\mu} \\
&\leq O\left(\frac{2^{-h+6}k}{\beta\varepsilon^2 m}\right)
\end{aligned}$$

where the second derivation is due to $\|v^{C,h}\|_\infty \leq \gamma$ for all $h \in [\log \varepsilon^{-1}]$ and the last derivation is due to Lemma 3.6. The following lemma demonstrates that an upper bound of the variance of each $\sum_{i=1}^{m} \frac{g_i}{\mu}(v_{s_i}^{C,h} - v_{s_i}^{C,h-1})$ leads to an upper bound of E_h.

Lemma C.5 (Lemma 2.3 of Massart (2007))**.** *Let $g_i \sim N(0, \sigma_i^2)$ for $i \in [m]$ be Gaussian random variables (not need to be independent) and let $\sigma = \max_{i\in m} \sigma_i$, then it holds that*

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

The number of distinct difference vector $v^{C,h} - v^{C,h-1}$ is at most $|U_{2^{-h}} \times U_{2^{-h+1}}| \leq N_X^{2^{-h}}(m) \cdot N_X^{2^{-h+1}}(m)$. Therefore, we have

$$\begin{aligned}
E_h &\leq \sqrt{8\log(|U_{2^{-h}}| \cdot |U_{2^{-h+1}}|) \cdot O\left(\frac{2^{-h+6}k}{\beta\varepsilon^2 m}\right)} \\
&\leq \sqrt{O(1) \cdot \log N_X^{2^{-h}}(m) \cdot \frac{2^{-h}k}{\beta\varepsilon^2 m}}
\end{aligned}$$

Plug in $\sum_{h=1}^{\log \varepsilon^{-1}} E_h$, we have

$$\begin{aligned}
&\sum_{h=1}^{\log \varepsilon'^{-1}} E_h \\
&\leq O(1) \cdot \sqrt{\frac{k}{\beta\varepsilon^2 m}} \sum_{h=1}^{\log \varepsilon^{-1}} \sqrt{\log N_X^{2^{-h}}(m)}.
\end{aligned} \tag{16}$$

Since $m \geq t \cdot \frac{k}{\beta\varepsilon^2}\left(\sum_{i=1}^{\log \varepsilon^{-1}} \sqrt{2^{-i}\log N_X^{2^{-i}}(m)}\right)^2$ for sufficiently large constant t, we can bound $\sum_{h=1}^{\log \varepsilon^{-1}} E_h$ by λ. which completes the proof. □

D. Proof of Theorem 3.10: Sample Sizes in Various Metrics

Theorem D.1 (Restatement of Theorem 3.10)**.** *For a metric space $M = (\mathcal{X}, \mathrm{dist})$ and a dataset $X \subseteq \mathcal{X}$, an integer $k \geq 1$ and real numbers $\beta, \varepsilon \in (0, 1)$, let S be a set of uniform samples with size*

- $O\left(\frac{k^2}{\beta\varepsilon^3} \cdot \log^2 \frac{k}{\beta\varepsilon} \cdot \log^2 \frac{1}{\varepsilon}\right)$ *if M is Euclidean $\mathbb{R}^d$;*

- $O\left(\frac{k^2}{\beta\varepsilon^2}\cdot \mathrm{ddim}\cdot\log\frac{k}{\beta\varepsilon}\right)$ *if M has doubling dimension* ddim*;*
- $O\left(\frac{k^2}{\beta\varepsilon^2}\cdot \log|\mathcal{X}|\cdot\log\frac{k}{\beta\varepsilon}\right)$ *if M is a finite metric;*
- $O\left(\frac{k^2}{\beta\varepsilon^2}\cdot \mathrm{tw}\cdot\log\frac{k}{\beta\varepsilon}\right)$ *if M is the shortest-path metric of a graph with treewidth* tw.

*Then S is an ε-weak coreset for (k,β)-*MEDIAN *on X with probability* 0.9.

We prove Theorem D.1 by bounding the covering number in each metric space respectively.

D.1. Euclidean Space

Lemma D.2 (Covering Number in Euclidean Space)**.** *In Euclidean space $(\mathbb{R}^d,\mathrm{dist})$, for integer $m>0$ and real number $0<\alpha<1/2$, it holds that*

$$\log|N_X^\alpha(m)|\le O\left(k\alpha^{-2}\log(m+k)\log\frac{k}{\beta\alpha}\right)$$

Proof. By Definition 3.5 of the covering number, it suffices to construct an α-covering of V w.r.t (S,err) for every $S\in\mathcal{S}(m)$. We need *terminal embedding* as stated in the following theorem:

Theorem D.3 (Terminal Johnson-Lindenstrauss Lemma (Narayanan & Nelson, 2019))**.** *For every $\varepsilon\in(0,1/2)$ and finite set $Y\subseteq\mathbb{R}^d$, there exists an embedding $g:\mathbb{R}^d\to\mathbb{R}^t$ for $t=O(\varepsilon^{-2}\log|Y|)$ such that*

$$\forall x\in Y,\forall c\in\mathbb{R}^d,\quad \mathrm{dist}(x,c)\le\mathrm{dist}(g(x),g(c))\le(1+\varepsilon)\,\mathrm{dist}(x,c).$$

We call g is an ε-terminal embedding for Y.

For every $S\in\mathcal{S}(m)$, let $Y:=S\cup C^\star$, by Theorem D.3, there exists an α-terminal embedding g for Y with target dimension $t=O(\alpha^{-2}\log|Y|)=O(\alpha^{-2}\log(m+k))$. For a set $A\subset\mathbb{R}^d$, we denote by $g(A):=\{g(x):x\in A\}$ the set of images of all $x\in A$. By the definition of $\mathcal{C}'(X)$, for every $C\in\mathcal{C}'(X)$, $c\in C$,

$$\begin{aligned}\mathrm{dist}(g(c),g(C^\star))&\le\mathrm{dist}(g(c),g(C^\star(c)))\\&\le 2\,\mathrm{dist}(c,C^\star)\\&\le\frac{12\lambda k}{\beta n}\,\mathrm{OPT}(X)\end{aligned}$$

We denote by $B_A(c,r):=\{x\in A:\mathrm{dist}(x,c)\le r\}$ the ball centered at c of radius r for $A\subseteq\mathcal{X}$. Thus we have for every $C\in\mathcal{C}'(X)$,

$$g(C)\subseteq\bigcup_{c^\star\in C^\star}B_{\mathbb{R}^t}\left(g(c^\star),\frac{12\lambda k}{\beta n}\,\mathrm{OPT}_\beta(X)\right)$$

To construct a covering for the set of cost vectors, we discretize each ball via a classical *covering* of a point set.

Definition D.4 (Covering of a Point Set)**.** For a metric space $(\mathcal{X},\mathrm{dist})$, a point set $A\subseteq\mathcal{X}$ and real number $0\le\alpha<1$, we say $T\subseteq\mathcal{X}$ is an α-covering of A if for every $x\in A$, there exists $y\in T$ such that $\mathrm{dist}(x,y)\le\alpha$.

Note that the above definition of covering is different from Definition 3.5, since they are for different objects. The following lemma bounds the cardinality of α-covering of an Euclidean ball.

Lemma D.5 (Covering of a Euclidean Ball)**.** *For $\alpha>0$ and an Euclidean ball $B\subset\mathbb{R}^t$ of radius $\Delta>0$, there exists an α-covering $T\subseteq B$ of size at most* $\exp\left(O(t\log(\Delta/\alpha))\right)$.

For every $c^\star\in C^\star$, let $T_{c^\star}$ be such an $\frac{\alpha\,\mathrm{OPT}_\beta(X)}{n}$-covering of $B_{\mathbb{R}^t}\left(g(c^\star),\frac{12\lambda k}{\beta}\cdot\frac{1}{n}\,\mathrm{OPT}_\beta(X)\right)$, and let $T_{C^\star}:=\bigcup_{c^\star\in C^\star}T_{c^\star}$. By Lemma D.5, we have

$$|T_{C^\star}|\le k\cdot\exp\left(O\left(t\log\frac{k}{\alpha\beta}\right)\right)$$

Let $g' : \mathbb{R}^d \to T_{C^\star}$ be a function satisfying that

$$g'(x) = \arg\min_{y \in T_{C^\star}} \operatorname{dist}(g(x), y),$$

Construct U is $\{\tilde{v}^C \in \mathbb{R}^X : C \in \mathcal{C}'(X)\}$, where $\tilde{v}^C$ is a cost function defined as: for every $x \in X$, $\tilde{v}^C_x := \operatorname{dist}(g(x), g'(C)) - \operatorname{dist}(g(x), g(C^\star))$. Observe that

$$|U| \le |T_{C^\star}|^k \le \exp\left(O\left(kt \log \frac{k}{\alpha\beta}\right)\right),$$

which implies that $\log |U| \le O(kt \log \frac{k}{\beta\alpha}) = O(k\alpha^{-2} \log(m+k) \log \frac{k}{\beta\alpha})$, it remains to show that U is the desired α-covering of V.

Bounded Covering Error By the definition of g', we have that for every $C \in \mathcal{C}'(X)$ and $c \in C$,

$$\operatorname{dist}(g(c), g'(C)) \le \operatorname{dist}(g(c), g'(c)) \le \frac{\alpha}{n} \operatorname{OPT}_\beta(X),$$

and

$$\operatorname{dist}(g'(c), g(C)) \le \operatorname{dist}(g'(c), g(c)) \le \frac{\alpha}{n} \operatorname{OPT}_\beta(X),$$

and thus for every $x \in S$, let c_1 denote the closest center in $g(C)$ to $g(x)$, and c_2 denote the closest center in $g'(C)$ to $g(x)$, then it holds that

$$|\operatorname{dist}(g(x), g(C)) - \operatorname{dist}(g(x), g'(C))| \le \max\{\operatorname{dist}(c_1, g'(C)), \operatorname{dist}(c_2, g(C))\} \le \frac{\alpha}{n} \operatorname{OPT}_\beta(X). \tag{17}$$

The error incurred by the covering is

$$\begin{aligned}
|v^C_x - \tilde{v}^C_x| &= |\operatorname{dist}(x, C) - \operatorname{dist}(x, C^\star) - \operatorname{dist}(g(x), g'(C)) + \operatorname{dist}(g(x), g(C^\star))| \\
&\le |\operatorname{dist}(x, C) - \operatorname{dist}(g(x), g'(C))| + |\operatorname{dist}(x, C^\star) - \operatorname{dist}(g(x), g(C^\star))| \\
&\le |\operatorname{dist}(x, C) - \operatorname{dist}(g(x), g(C))| + |\operatorname{dist}(x, C^\star) - \operatorname{dist}(g(x), g(C^\star))| + \frac{\alpha}{n} \operatorname{OPT}_\beta(X) \\
&\le \alpha \operatorname{dist}(x, C) + \alpha \operatorname{dist}(x, C^\star) + \frac{\alpha}{n} \operatorname{OPT}_\beta(X) \\
&= \alpha \cdot \operatorname{err}(x, v^C)
\end{aligned}$$

where the second derivation is due to triangle inequality, the third derivation is due to (17), and the forth derivation is due to the terminal embedding guarantee.

Bounded L_∞ Norm For every $\tilde{v}^C \in U$,

$$\begin{aligned}
\|\tilde{v}^C\|_\infty &= \max_{x \in X} |\operatorname{dist}(g(x), g'(C)) - \operatorname{dist}(g(x), g(C^\star))| \\
&\le \max_{x \in X} |\operatorname{dist}(g(x), g(C)) - \operatorname{dist}(g(x), g(C^\star))| + \frac{\alpha}{n} \operatorname{OPT}_\beta(X) \\
&\le \max\left\{\max_{c \in C} \operatorname{dist}(g(c), g(C^\star)), \max_{c^\star \in C^\star} \operatorname{dist}(g(c^\star), g(C))\right\} + \frac{\alpha}{n} \operatorname{OPT}_\beta(X) \\
&\le (1+\alpha) \max\left\{\max_{c \in C} \operatorname{dist}(c, C^\star), \max_{c^\star \in C^\star} \operatorname{dist}(c^\star, C)\right\} + \frac{\alpha}{n} \operatorname{OPT}_\beta(X) \\
&\le \frac{12\lambda k}{\beta n} \operatorname{OPT}_\beta(X),
\end{aligned}$$

where the second derivation is due to (17), the third derivation is due to the triangle inequality, and the last derivation follows from a similar argument of Lemma 3.4. Thus we can construct α-covering of V with respect to any $S \in \mathcal{S}(m)$, which concludes the proof. □

Proof of Theorem D.1 in Euclidean space. We only need to bound the summation in (1).

$$\begin{aligned}
&\sum_{i=1}^{\log \varepsilon^{-1}} \sqrt{2^{-i} \log N_X^{2^{-i}}(m)} \\
\leq &\sum_{i=1}^{\log \varepsilon^{-1}} O(1) \cdot \sqrt{2^{-i} \cdot k \cdot 2^{2i} \log m \log \frac{k}{\beta 2^{-i}}} \\
\leq &\, O(1) \cdot \log \varepsilon^{-1} \cdot \sqrt{\varepsilon^{-1} \cdot k \log m \log \frac{k}{\beta \varepsilon}}.
\end{aligned}$$

Therefore, it suffices to set $m = O\left(\frac{k^2}{\beta\varepsilon^3} \cdot \log^2 \frac{k}{\beta\varepsilon} \cdot \log^2 \frac{1}{\varepsilon}\right)$. □

D.2. Doubling Metric Space

Definition D.6 (Doubling Dimension (Assouad, 1983; Gupta et al., 2003))**.** The doubling dimension of a metric space $M = (\mathcal{X}, \mathrm{dist})$, denoted as $\mathrm{ddim}(M)$, is the smallest integer d such that any ball can be covered by at most 2^d balls of half the radius.

Lemma D.7 (Covering Number in Doubling Metric Space)**.** *In a metric space $(\mathcal{X}, \mathrm{dist})$ with doubling dimension ddim, for integer $m > 0$ and real number $0 < \alpha < 1/2$, it holds that*

$$\log |N_X^{\alpha}(m)| \leq O(k \cdot \mathrm{ddim} \cdot \log(k/\alpha\beta)).$$

Proof. It suffices to construct α-covering with respect to S for every $S \in \mathcal{S}(m)$. Different from the Euclidean case (Lemma D.2), we directly apply the covering of point set without using terminal embedding. The following lemma bounds the cardinality of a covering of point set in doubling metrics.

Lemma D.8 (Gupta et al. 2003)**.** *For $\alpha > 0$ and a metric space $(\mathcal{X}, \mathrm{dist})$ with doubling dimension ddim and diameter Δ, there exists an α-covering T of $\mathcal{X}$ with $|T| \leq \exp\left(O(\mathrm{ddim} \cdot \log(\Delta/\alpha))\right)$.*

By definition of $\mathcal{C}'(X)$, for every $C \in \mathcal{C}'(X)$, $c \in C$, it holds that $\mathrm{dist}(c, C^\star) \leq \frac{6\lambda k}{\beta} \cdot \frac{1}{n} \mathrm{OPT}(X)$. Hence, consider an $\frac{\alpha \mathrm{OPT}(X)}{n}$-covering $T_{c^\star}$ of $B_{\mathcal{X}}\left(c^\star, \frac{6\lambda k}{\beta} \cdot \frac{1}{n} \mathrm{OPT}(X)\right)$ for every $c^\star \in C^\star$ and let $T_{C^\star} := \bigcup_{c^\star \in C^\star} T_{c^\star}$. By Lemma D.8, we have

$$|T_{C^\star}| \leq k \cdot \exp\left(O\left(\mathrm{ddim} \cdot \log \frac{k}{\alpha\beta}\right)\right)$$

We define $g : V \to V$ such that for every $c \in V$,

$$g(c) = \arg\min_{c' \in T_{C^\star}} \mathrm{dist}(c, c').$$

Construct U is $\{\tilde{v}^C \in \mathbb{R}^X : C \in \mathcal{C}'(X)\}$, where $\tilde{v}^C$ is a cost function defined as: for every $x \in X$, $\tilde{v}_x^C := \mathrm{dist}(x, g(C)) - \mathrm{dist}(x, C^\star)$. Then a similar analysis for bounding the covering error and L_∞ norm as in Euclidean case(Lemma D.2) certifies that U is an α-covering of V with respect to any $S \in \mathcal{S}(m)$, and $\log |U| \leq O(k \cdot \mathrm{ddim} \cdot \log \frac{k}{\alpha\beta})$. □

Proof of Theorem D.1 in doubling metric space and general discrete metric space. Directly plugging the covering number in Lemma D.7 into (2) concludes the proof in doubling metric space. For general discrete metric space $M = (\mathcal{X}, \mathrm{dist})$, we know that the doubling dimension is $O(\log |\mathcal{X}|)$, and thus directly applying the result of doubling metric space completes the proof. □

D.3. Shortest-path Metric of a Graph with Bounded Treewidth

Definition D.9 (Tree Decomposition and Treewidth)**.** A tree decomposition of a graph $G = (V, E)$ is a tree $\mathcal{T} = (\mathcal{V}, \mathcal{E})$, where each node in $\mathcal{V}$, called a bag, is a subset of vertices in V, such that the following holds.

- $\bigcup_{S\in\mathcal{V}} S = V$.
- $\forall u \in V$, the nodes in $\mathcal{V}$ that contain u form a connected component in $\mathcal{T}$.
- $\forall (u,v) \in E$, there exists $S \in \mathcal{V}$ such that $\{u,v\} \subseteq S$.

The treewidth of G, denoted as $\mathrm{tw}(G)$, is the smallest integer t such that there is a tree decomposition of G with maximum bag size $t+1$.

Lemma D.10 (Covering Number in Shortest-path Metric of a Graph)**.** *In shortest-path metric $M = (\mathcal{X}, \mathrm{dist})$ of a graph with bounded treewidth* tw*, for integer $m > 0$ and real number $0 < \alpha < 1/2$, it holds that*

$$\log |N_X^\alpha(m)| \le O(k \cdot \mathrm{tw} \cdot \log(k/\alpha\beta + m/\alpha)).$$

Proof. To bound the covering number in graph metrics, it suffices to construct an α-covering of V with respect S for every $S \in \mathcal{S}(m)$. The proof of Lemma D.10 relies on the following structural lemma, proposed by (Baker et al., 2020).

Lemma D.11 (Structural Lemma, (Baker et al., 2020))**.** *Given graph $G = (\mathcal{X}, E)$ with treewidth* tw*, and $A \subseteq \mathcal{X}$, there exists a collection $\mathcal{T}_A$ of subsets of $\mathcal{X}$, such that the following holds.*

1. $\bigcup_{T\in\mathcal{T}_A} T = \mathcal{X}$.
2. $|\mathcal{T}_A| \le \mathrm{poly}(|A|)$.
3. *For every $T \in \mathcal{T}_A$, $|T \cap A| \le O(\mathrm{tw})$ and there exists $P_T \subseteq \mathcal{X}$ with $|P_T| \le O(\mathrm{tw})$ such that there is no edge in E between T and $\mathcal{X} \setminus (T \cup P_T)$.*

Recall that any $S \in \mathcal{S}(m)$ makes ξ_S happens, i.e., S satisfies (4). Thus for every $C \in \mathcal{C}'(X)$ and every $x \in S$,

$$\mathrm{dist}(x, C^\star) \le \sum_{x\in S} \mathrm{dist}(x, C^\star) \le \frac{\lambda m}{n} \mathrm{OPT}(X),$$

and thus

$$\mathrm{dist}(x, C) \le \mathrm{dist}(C(x), C^\star) + \mathrm{dist}(x, C^\star) \le \left(\frac{6k}{\beta} + m\right) \cdot \frac{\lambda}{n} \mathrm{OPT}(X).$$

Let $\mathcal{T}_S$ be a collection of subsets asserted by Lemma D.11. We can construct the covering as follows:

For every $T \in \mathcal{T}$, and $c \in T$, let $P_T \subseteq \mathcal{X}$ be a set asserted in condition 3 in Lemma D.11. By Lemma D.11, we have for every $x \in \mathcal{X} \setminus T$, it holds that

$$\mathrm{dist}(x, c) = \min_{x \ne y \in P_T} \{\mathrm{dist}(x, y) + \mathrm{dist}(y, c)\},$$

since the shortest path between x and c must pass by some vertices in P_T. Let $I_T := (S \cap T) \cup P_T$ denote the *important vertices* of T, we define a *rounded distance function* $\mathrm{dist}'_T : T \times X \to \mathbb{R}_+$ w.r.t. T as follows:

- $\forall c \in T, x \in I_T$, $\mathrm{dist}'_T(c, x)$ is the closest multiple of $\frac{\alpha}{n}\mathrm{OPT}(X)$ to $\mathrm{dist}(x, c)$ no greater than $\left(\frac{6k}{\beta} + m\right) \cdot \frac{\lambda}{n}\mathrm{OPT}(X)$.
- $\forall c \in T, x \in X \setminus I_T$, $\mathrm{dist}'_T(c, x) = \min_{y\in P_T} \left\{\mathrm{dist}'_T(c, y) + \mathrm{dist}(y, x)\right\}$.

Note that for $x \in T \setminus S$ or for $x \in \mathcal{X}$ that satisfies $\mathrm{dist}(c, x) > \left(\frac{6k}{\beta} + m\right) \cdot \frac{\lambda}{n}\mathrm{OPT}(X)$, the rounded distance $\mathrm{dist}'_T(c, x)$ may be distorted badly. However, as discussed above, we only case about $x \in S$ which is ensured that the distance to c is not that far.

Notice that for a fixed $c \in T$, the rounded distance function $\mathrm{dist}(c, \cdot)$ is determined by the values of $\mathrm{dist}'_T(c, x)$ for all $x \in I_T$, and each of them is among $\frac{6\lambda k}{\alpha\beta} + \frac{\lambda m}{\alpha}$ possible values. Therefore the number of distinct rounded distance functions $|\{\mathrm{dist}'_T(c, \cdot) : c \in T\}|$ w.r.t. T is

$$\le \left(\frac{2\lambda k}{\alpha\beta} + \frac{\lambda m}{\alpha}\right)^{|I_T|} \le \left(\frac{k}{\alpha\beta} + \frac{m}{\alpha}\right)^{O(\mathrm{tw})}$$

We define $\mathrm{dist}' : \mathcal{X} \times X \to \mathbb{R}_+$ as: for every $c \in \mathcal{X}$, let $T_c \in \mathcal{T}$ that contains c, it holds that $\mathrm{dist}'(c, x) = \mathrm{dist}'_{T_c}(c, x)$. Then we have

$$|\{\mathrm{dist}'(c, \cdot) : c \in \mathcal{X}\}| \leq |\mathcal{T}_S| \cdot \left(\frac{k}{\alpha\beta} + \frac{m}{\alpha}\right)^{O(\mathrm{tw})} \leq \mathrm{poly}(|S|) \cdot \left(\frac{k}{\alpha\beta} + \frac{m}{\alpha}\right)^{O(\mathrm{tw})}.$$

For every $C \in \mathcal{C}'(X)$, we denote by $\tilde{v}^C \in \mathbb{R}^X$ the *rounded coset vector* defined as follows: for every $x \in X$, $\tilde{v}^C_x := \min_{c \in C} \mathrm{dist}'(c, x) - \mathrm{dist}(x, C^\star)$. Let $\tilde{V}$ denote the set $\{\tilde{v}^C : C \in C'(X)\}$, we construct U as follows: for every $\tilde{v} \in \tilde{V}$, U contains one v^C for some $C \in \mathcal{C}'(X)$ such that $\tilde{v}^C = \tilde{v}$.

We next show that U is the desired α-covering.

Cardinality of U Observe that the cardinality of U is upper bounded by the number of distinct rounded distance functions $|\{\min_{c \in C} \mathrm{dist}'(c, \cdot) : C \in \mathcal{X}^k\}|$, which is at most

$$\left(\mathrm{poly}(|S|) \cdot \left(\frac{k}{\alpha\beta} + \frac{m}{\alpha}\right)\right)^{O(\mathrm{tw}) \cdot k} \leq \left(\frac{k}{\alpha\beta} + \frac{m}{\alpha}\right)^{O(k \cdot \mathrm{tw})}$$

Thus it holds that $\log |U| \leq O(k \cdot \mathrm{tw} \cdot \log(\frac{k}{\alpha\beta} + \frac{m}{\alpha}))$.

Bouneded Covering Error For every $C \in \mathcal{C}'(X)$, and $x \in S$, we have $\mathrm{dist}(x, C) \leq \left(\frac{6k}{\beta} + m\right) \cdot \frac{\lambda}{n} \mathrm{OPT}(X)$. Hence, the error incurred by rounding, i.e., $|\mathrm{dist}'(c, x) - \mathrm{dist}(c, x)|$ is at most $\frac{\alpha}{n} \mathrm{OPT}(X)$. Let $\tilde{v}^C$ be the rounded cost vector with respect to v^C, by definition, there exists $C' \in \mathcal{C}'(X)$ such that $v^{C'} \in U$ and $\tilde{v}^{C'} = \tilde{v}^C$. The covering error is

$$\begin{aligned}
\left|v^C_x - v^{C'}_x\right| &\leq \left|v^C_x - \tilde{v}^C + \tilde{v}^{C'} - v^{C'}_x\right| \\
&\leq \left|v^C_x - \tilde{v}^C\right| + \left|\tilde{v}^{C'} - v^{C'}_x\right| \\
&\leq \frac{2\alpha}{n} \mathrm{OPT}(X) \\
&\leq 2\alpha \cdot \mathrm{err}(x, v^C).
\end{aligned}$$

It suffices to rescale α.

Bounded L_∞ Norm For every $v^C \in U$, we have $v^C \in V$ also, thus the L_∞ norm is bounded because the L_∞ norm of cost vectors in V is bounded. □

Proof of Theorem D.1 in graph metric space. Similar to doubling metric case, we plug the covering number in Lemma D.10 into (2) to concludes the proof. □

E. Proof of Theorem 1.2: An $\Omega(1/\beta)$ Query Complexity Lower Bound for Any Algorithm

Theorem E.1 (Restatement of Theorem 1.2)**.** *There exists a family of datasets $X \subset \mathbb{R}$ with balancedness β such that any (randomized) $O(1)$-approximate algorithm for* 2-MEDIAN *with success probability at least $3/4$ must query the identify of data points in X for $\Omega(1/\beta)$ times (provided that queried points have free access to distance function).*

To prove our lower bound, we first apply the Yao's principle (Yao, 1983) and derive the following lemma that reduce to proving lower bounds for deterministic algorithms with respect to some input distribution.

Lemma E.2. *For real number $\alpha > 1$, let D be a distribution over a family of datasets $X \subset \mathbb{R}$, if any deterministic algorithm must query $\Omega(1/\beta)$ times to computes an α-approximate center set for* 2-MEDIAN *on X sampled from D with success probability at least $3/4$, then any randomized α-approximate algorithm for* 2-MEDIAN *with success probability at least $3/4$ must query $\Omega(1/\beta)$ times.*

Proof. For an algorithm $\mathcal{A}$, we denote by $\mathcal{A}(X)$ the output of A when running on X. For any randomized algorithm $\mathcal{R}$ which makes at most $o(1/\beta)$ queries, it can be seen as a distribution over some deterministic algorithms $\mathcal{R}_1, \ldots, \mathcal{R}_s$. By assumption, for every $i \in [s]$, it holds that

$$\Pr_{X \sim D}\left[\mathrm{cost}(X, \mathcal{R}_i(X)) > \alpha\, \mathrm{OPT}(X)\right] > 1/4.$$

Therefore, averaging over all deterministic algorithms, we have

$$\Pr_{X \sim D}\left[\mathrm{cost}(X, \mathcal{R}(X)) > \alpha\, \mathrm{OPT}(X)\right] > 1/4.$$

where the randomness is from the choice of X and the randomness of $\mathcal{R}$. Hence, there exists an instance $X \in \mathrm{supp}(D)$ such that

$$\Pr\left[\mathrm{cost}(X, \mathcal{R}(X)) > \alpha\, \mathrm{OPT}(X)\right] > 1/4.$$

which completes the proof. □

Proof of Theorem E.1. For an integer $n \geq 10/\beta$, let $m = \beta n/2$, we construct a family of n-point datasets as follows. For a set $\Pi \in [n]$ and a real number t, let $X^{\Pi,t} := \{x_1^{\Pi,t}, \ldots, x_n^{\Pi,t}\}$ denotes an ordered n-point set in 1-dimensional space such that,

$$\forall i \in [n], \quad x_i^{\Pi,t} = \begin{cases} t & i \in \Pi \\ 0 & i \notin \Pi \end{cases}.$$

Let $\mathcal{X} := \{X^{\Pi,t}; |\Pi| = m, t \in \{1,2\}\}$ be a family of n-point datasets. Clearly, every $X \in \mathcal{X}$ is of balancedness β for 2-MEDIAN, and the optimal objective value of X for 2-MEDIAN is 0. Let D be a uniform distribution over $\mathcal{X}$.

By Lemma E.2, it suffices to prove that for any $\alpha > 1$ and any deterministic algorithm $\mathcal{A}$, which makes at most $\frac{1}{2\beta}$ queries, will compute a 2-point center set C such that $\mathrm{cost}(X, C) > \alpha \cdot \mathrm{OPT}(X)$ for X sampled from D with probability at least $1/4$. We assume $\mathcal{A}$ queries with index $i_1, \ldots, i_l$ with $l \leq \frac{1}{2\beta}$, and receives $x_{i_1}, \ldots, x_{i_l}$ from the oracle. We observe that $\mathcal{A}$ finds a 2-center set C such that $\mathrm{cost}(X, C) > \alpha \cdot \mathrm{OPT}(X)$ if and only if $\mathcal{A}$ finds the exact optimal solution (otherwise, we say $\mathcal{A}$ fails). Therefore, we have

$$\begin{aligned}
&\Pr_{X \sim D}\left[\mathrm{cost}(X, \mathcal{A}(X)) > \alpha \cdot \mathrm{OPT}(X)\right] \\
=\ &\Pr_{X \sim D}[\mathcal{A} \text{ fails}] \\
=\ &\Pr_{X \sim D}\left[\mathcal{A} \text{ fails} \mid \forall j \in [l], X_{i_j} = 0\right] \Pr_{X \sim D}\left[\forall j \in [l], x_{i_j} = 0\right] \\
&+ \Pr_{X \sim D}\left[\mathcal{A} \text{ fails} \mid \exists j \in [l], X_{i_j} \neq 0\right] \Pr_{X \sim D}\left[\exists j \in [l], x_{i_j} \neq 0\right] \\
\geq\ &\Pr_{X \sim D}\left[\mathcal{A} \text{ fails} \mid \forall j \in [l], X_{i_j} = 0\right] \Pr_{X \sim D}\left[\forall j \in [l], x_{i_j} = 0\right]
\end{aligned} \tag{18}$$

Recall that $\mathcal{A}$ makes queries with index $i_1, \ldots, i_l$, and thus the output $\mathcal{A}(X)$ depends only on x_{i_j} for all $j \in [l]$. For X sampled uniformly from $\mathcal{X}$, its optimal center set is $\{0, 1\}$ with probability 0.5 and $\{0, 2\}$ with probability 0.5. Therefore, condition on $\forall j \in [l], X_{i_j} = 0$, $\mathcal{A}$ will fail with probability at least 0.5, i.e.,

$$\Pr[\mathcal{A} \text{ fails} \mid \forall j \in [l], X_{i_j} = 0] \geq \frac{1}{2} \tag{19}$$

since it can not determine the center other than 0. It remains to bound $\Pr_{X\sim D}[\forall j \in [l], X_{i_j} = 0]$.

$$\begin{aligned}
\Pr_{X\sim D}[\forall j \in [l], X_{i_j} = 0] &= \binom{n-l}{m} \cdot \binom{n}{m}^{-1} \\
&= \frac{(n-l)!(n-m)!}{(n-l-m)!n!} \\
&= \frac{(n-l-m+1)(n-l-m+2)\cdots(n-l)}{(n-m+1)(n-m+2)\cdots n} \\
&= \left(1-\frac{l}{n-m+1}\right)\left(1-\frac{l}{n-m+2}\right)\cdots\left(1-\frac{l}{n}\right) \\
&\geq \left(1-\frac{l}{n-m+1}\right)^m \\
&\geq 1-\frac{lm}{n-m+1} \\
&\geq 1-\frac{1}{2\beta}\cdot\frac{\beta n}{2}\cdot\left(\left(1-\frac{\beta}{2}\right)n+1\right)^{-1} \\
&\geq \frac{1}{2}
\end{aligned} \tag{20}$$

where the sixth derivation is due to Bernoulli's inequality. Therefore, plugging (19) and (20) into (18), we have

$$\Pr_{X\sim D}\left[\mathrm{cost}(X, \mathcal{A}(X)) > \alpha \cdot \mathrm{OPT}(X)\right] \geq \frac{1}{4}$$

which completes the proof. □

F. Vanilla k-MEDIAN on Uniform Sample Can Incur Big Error

Lemma F.1. *There exists a family of* 0.5*-balanced datasets* $X_n \subset \mathbb{R}$ *with* $|X_n| = n$ *for any integer* $n \geq 1$*, such that letting* S_n *be a set of* $o(n)$ *uniform samples from* X_n*, with probability at least* $1/4$*, it holds that* $\mathrm{cost}(X_n, C_n^\star) \geq 1.01 \cdot \mathrm{OPT}(X_n)$*, where* $C_n^\star$ *is an optimal center set for* 3-MEDIAN *on* S_n*.*

Proof. Our proof is constructive. For $n \geq 10$, assume S_n has $f(n)$ points sampled uniformly and independently from X_n, where $3 \leq f(n) < n/2$. To simplify the proof, we construct X_n with size of $O(n)$ (instead of n) as follows. We place n points at 0 and n points at 1. Also, we place n points at w, $\frac{n}{1.01\cdot f(n)}$ points at $w + f(n)$, and let $w \to \infty$. Clearly, the optimal 3-point center set $C_n^\star$ is $\{0, 1, w\}$, and the objective value is $n/1.01$.

Recall that S_n is a set of $f(n)$ uniform samples from X_n, we have the probability of that $w + f(n) \in S_n$ is

$$\begin{aligned}
\Pr[w + f(n) \in S_n] &= 1 - \left(1 - \frac{n/(1.01\cdot f(n))}{3n + n/(1.01\cdot f(n))}\right)^{f(n)} \\
&\geq 1 - \exp\left(-\frac{n/1.01}{3n + n/(1.01\cdot f(n))}\right) \\
&\geq 1 - \exp\left(-\frac{1}{3.03 + 1/3}\right) \\
&\geq 0.25
\end{aligned}$$

where the third derivation is due to $f(n) \geq 3$. Condition on $w + f(n) \in S_n$, we observe that S_n contains at most $f(n)$ points at 0, at most $f(n)$ points at 1, at most $f(n)$ points at w and at least 1 point at $w + f(n)$. In this case, the optimal 3-point center set C' must contain $w + f(n)$. On X_n, the optimal center set that contains $w + f(n)$ is either $\{0, w, w + f(n)\}$ or $\{1, w, w + f(n)\}$, both of which have an objective value $n \geq 1.01 \cdot \mathrm{OPT}(X_n)$. Therefore, $f(n)$ must be greater than $n/2$, which completes the proof. □