

Optimal Stable Coresets for Geometric Median via Uniform Sampling

Amir Carmel 

Weizmann Institute of Science, Rehovot, Israel

Robert Krauthgamer 

The Harry Weinrebe Professorial Chair of Computer Science,
Weizmann Institute of Science, Rehovot, Israel

Nir Petruschka 

Weizmann Institute of Science, Rehovot, Israel

Abstract

The geometric median problem asks to find a point in $\mathbb{R}^d$ that minimizes the sum of Euclidean distances to an input set. It is a classical problem in computational geometry and appears as a subroutine in numerous optimization tasks, many of which require the solution to satisfy additional structural constraints. A common approach to reduce the input size is to construct a coreset, which is a small weighted subset that faithfully represents the input for a specific optimization problem. Strong coresets preserve the cost of every candidate solution but require linear time to construct; weak coresets admit sublinear construction, in fact by uniform sampling, but only preserve near-optimal solutions, which is insufficient when the solution is constrained. To address this, we focus instead on the recently introduced intermediate notion of a *stable coreset*, which simultaneously handles all constrained variants. Currently, there is a large gap between the known sample sizes for stable and weak coresets.

Our main result is that a uniform sample of size $O(\epsilon^{-2} \log \frac{1}{\epsilon})$ is a stable $(\epsilon, O(\epsilon))$ -coreset for the geometric median, with high constant probability, and this bound is tight up to the logarithmic factor. Our analysis adapts recent machinery of Carmel and Krauthgamer (ICLR 2026) for constructing stable coresets, which incurs an $O(\log d)$ factor. We show an iterative argument that progressively reduces the sample size, and eliminates this dependence on the dimension d . At a high level, this approach resembles the technique of iterative size reduction, which is applicable for strong coresets but not for weak coresets.

2012 ACM Subject Classification Theory of computation → Sketching and sampling; Theory of computation → Streaming models

Keywords and phrases clustering, geometric median, coresets, stable coresets, uniform sampling

Digital Object Identifier 10.4230/LIPIcs.APPROX/RANDOM.2026.30

Category APPROX

Related Version Full Version: <https://arxiv.org/abs/2606.30348>

Funding Robert Krauthgamer: Work partially supported by the Israel Science Foundation grant #1336/23.

1 Introduction

Clustering is a basic tool for understanding large collections of data, with applications throughout machine learning, statistics, and optimization. The core idea is to partition the data points into sets (called clusters) that share common characteristics, and a classical formulation is center-based clustering, where each cluster is represented by a single point. One of the most fundamental problems in this family is the *geometric median*, which asks for a point that minimizes the sum of distances to all the input points. Despite its



© Amir Carmel, Robert Krauthgamer, and Nir Petruschka;
licensed under Creative Commons License CC-BY 4.0

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2026).

Editors: Mohit Singh and Tom Gur; Article No. 30; pp. 30:1–30:14



Leibniz International Proceedings in Informatics

LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

apparent simplicity, the geometric median exhibits strong robustness properties and has broad applications, from robust statistics to facility location and machine learning. To define this problem formally, we view it as a special case of the *k-median problem*, where the input is a point set P in a metric space $(\mathcal{X}, \text{dist})$, and the goal is to find a set $C \subseteq \mathcal{X}$ of k centers that minimize the objective

$$\text{cost}(C, P) := \sum_{p \in P} \min_{c \in C} \text{dist}(p, c).$$

The *geometric median problem* (also called the *Fermat-Weber problem*) is the case $k = 1$ in the widely studied setting of Euclidean spaces, i.e., $\mathcal{X} = \mathbb{R}^d$ and $\text{dist}(x, y) = \|x - y\|_2$. Our main result addresses the geometric median problem, however the broader context is often useful, e.g., our proof connects this problem to 1-median in ℓ_1 spaces.

As datasets grow in size, data reduction techniques have become essential, as they enable algorithms to operate on small summaries rather than on the entire dataset. A coreset formalizes this idea by providing a small instance of the same problem, where solving the problem on the coreset immediately provides similar guarantees for the original dataset. Coresets usually offer a tunable tradeoff between size and quality, the latter measured as an approximation factor. A *strong ϵ -coreset* for k -median is a weighted subset $Q \subseteq P$ such that for every set of k centers $C \subset \mathbb{R}^d$, the cost is approximately preserved, namely,

$$\text{cost}(C, Q) \in (1 \pm \epsilon) \cdot \text{cost}(C, P). \quad (1)$$

Strong coresets for the classical k -median problem in Euclidean space, alongside numerous variants, have been studied extensively over the past two decades, see Section 1.3 for more details.

Many modern applications involve massive datasets, for which even linear-time processing might be prohibitive. Yet, constructing a strong coreset inherently requires reading each data point, which takes at least linear time $\Omega(nd)$, regardless of the coreset's size.¹ Therefore, to open the door for sublinear-time algorithms, one natural relaxation is the notion of a *weak coreset*. A weighted subset $Q \subseteq P$ is called a weak (ϵ, η) -coreset if every near-optimal solution for Q is also a near-optimal solution for P , namely, for every set of k centers $C \subset \mathbb{R}^d$,

$$\text{if } \text{cost}(C, Q) \leq (1 + \epsilon) \cdot \min_{|C'|=k} \text{cost}(C', Q), \text{ then } \text{cost}(C, P) \leq (1 + \eta) \cdot \min_{|C'|=k} \text{cost}(C', P). \quad (2)$$

It is worth noting that most constructions achieve $\eta = O(\epsilon)$, but covering the entire literature requires two separate parameters. Weak coresets can be constructed via uniform sampling, a method widely adopted in practice, often without theoretical guarantees, and naturally suited to streaming and distributed settings. See Table 1 for known bounds and references.

Despite its appeal, the weak coreset notion has a fundamental limitation: A strong coreset has guarantees for *every* candidate solution, but a weak coreset provides guarantees only for candidates that are near-optimal (on the coreset). This might be a severe limitation when the solutions of interest are constrained to a subset $\mathcal{X}' \subset \mathcal{X}$ (because an optimal solution in $\mathcal{X}'$ need not be optimal also in $\mathcal{X}$), and such constraints indeed arise from structural requirements or downstream tasks, as we discuss further below.

¹ Consider an instance P with a single extremely distant point. A strong coreset must include this point, which requires scanning the entire instance.

■ **Table 1** Sublinear-time constructions of coresets for the geometric median problem. In all of them, the sample complexity equals the (listed) coreset size; in fact, the coreset is simply the sample, except when marked by †, in which the sample is post-processed. Strong coresets are not listed as they require at least linear time.

Coreset type	Coreset size	Reference
weak $(0, \epsilon)$	$\tilde{O}(\epsilon^{-2}d)$	[1, 45]
	$\tilde{O}(\epsilon^{-4})$	[25]
weak $(\epsilon, O(\epsilon))$	$\tilde{O}(\epsilon^{-4})^\dagger$	[21]
	$\tilde{O}(\epsilon^{-3})$	[33]
	$\tilde{O}(\epsilon^{-2})$	[53]
stable $(\epsilon, O(\epsilon))$	$\tilde{O}(\epsilon^{-2} \log d)$	[13]
	$\tilde{O}(\epsilon^{-2})$	Theorem 1.1

The notion of a *stable coreset*, recently introduced by [13], is intermediate between the strong and weak coreset notions and addresses this issue, by preserving the relative quality of *all pairs* of candidate solutions. Formally, a weighted subset $Q \subseteq P$ is called a stable (ϵ, η) -coreset if for every pair of candidates $C, C' \subset \mathbb{R}^d$,

$$\text{if } \text{cost}(C, Q) \leq (1 + \epsilon) \cdot \text{cost}(C', Q), \text{ then } \text{cost}(C, P) \leq (1 + \eta) \cdot \text{cost}(C', P). \quad (3)$$

Crucially, since this guarantee holds simultaneously for all pairs of candidates, the same stable coreset can be reused across multiple constraints and downstream tasks, and its construction does not require a priori knowledge of the constraints.

The geometric median problem is used as a subroutine in numerous optimization tasks, and many of them require the solution to satisfy additional constraints, which essentially restrict it to a predetermined subset $\mathcal{X}' \subset \mathcal{X} = \mathbb{R}^d$. For instance, the solution may be required to lie in a low-dimensional linear subspace, or more generally on an algebraic variety, or to have a sparse coordinate structure (e.g., $\|c\|_0 \leq \alpha$).

We illustrate this with three concrete examples. First, in kernel-based methods, the input points are all images of a feature map φ and also the solution must be an image of φ [44, 47]; in this case, $\mathcal{X}'$ is simply the image set of φ . Second, the *discrete* geometric median problem requires the solution to be one of the input points in P [46, 30, 24]; in this case $\mathcal{X}' = P$ is discrete. Third, the *ultrametric 1-median* problem, in which the input distances form an ultrametric and the solution is a median terminal node, admits a $(1 + \epsilon)$ -approximation in $\tilde{O}(\epsilon^{-3})$ time [14]. Since finite ultrametrics embed isometrically into ℓ_2 [51], this is a special case of constrained ℓ_2 1-median.

In all these cases where we seek an optimal or near-optimal solution in $\mathcal{X}'$, a weak coreset provides no guarantees, while a stable coreset is applicable, in fact simultaneously for all possible restrictions $\mathcal{X}'$.

1.1 Our Results

A long line of work has progressively refined the size of weak coresets (see Table 1), however these results and techniques do not extend (at least not directly) to stable coresets, for which the known bound falls behind and even depends on the dimension d . Our main result is a stable coreset for geometric median that is constructed via uniform sampling (with no post-processing) and has near-optimal size and query complexity. In our result, a uniform sample is one drawn without replacement.

► **Theorem 1.1.** *Let $\epsilon \in (0, 1/4)$ and let $P \subset \mathbb{R}^d$ be finite. Then a uniform sample of size $O(\epsilon^{-2} \log \frac{1}{\epsilon})$ from P is a stable $(\epsilon, 17\epsilon)$ -coreset for 1-median in ℓ_2^d with probability at least $4/5$.*

This result improves over the stable coreset of [13], which is also via uniform sampling and has size $\tilde{O}(\epsilon^{-2} \log d)$. Our bound completely eliminates the dependence on the dimension. Technically, their main result is for ℓ_1 metrics, and the Euclidean case follows by Dvoretzky's theorem [29, 49]. Previously, a similar coreset size $\tilde{O}(\epsilon^{-2})$, also via uniform sampling, was known only for weak coreset [53], which is strictly less powerful. Technically, it is obtained as a corollary of a result for multiple ℓ_p regression, whose proof leverages ℓ_p Lewis weight sampling. In contrast, our proof refines the framework for stable coresets recently developed by [13], and employs an iterative argument to progressively reduce the sample size, see Section 1.2.

The dependence on ϵ in our bound is near-tight, as an $\Omega(\epsilon^{-2})$ lower bound follows from more general results of [16, 48]. For completeness, we provide a short self-contained proof in Appendix A.

► **Remark 1.2.** We consider only the case $k = 1$ in order to obtain clean theoretical guarantees. It is well-known that sublinear algorithms face a major challenge when $k > 1$, because small clusters can be missed by sublinear methods. Prior work by [33] (see also references therein) has formalized this issue through a *balancedness parameter* $\beta \in (0, 1]$, which measures the smallest cluster size in an optimal solution, modifying the coreset definition to consider only β -balanced clusterings of the coreset.

1.2 Technical Overview

We now overview the proof of our main result, Theorem 1.1. Our starting point is a result of [13, Corollary D.2], which establishes that a uniform sample of size $O(\epsilon^{-2} \log(d/\epsilon))$ is a stable $(\epsilon, O(\epsilon))$ -coreset for 1-median in ℓ_2^d with high constant probability. Their proof exploits the fact that ℓ_2^d embeds into $\ell_1^{O(d)}$, by Dvoretzky's theorem, and applies their analysis for the 1-median problem in ℓ_1 . We make a simple observation, that a folklore terminal embedding (Theorem 2.3) reduces the problem to the case where $d \leq n$, and therefore a uniform sample Q of size $O(\epsilon^{-2} \log(n/\epsilon))$ suffices for such a coreset, regardless of the ambient dimension d of our dataset P .

Building on this simple observation, our high-level idea is to apply iterative size reduction, a technique introduced by [9] for strong coresets, hence we need to adapt it to stable coresets. Suppose we have a uniform sample $Q \subset P$ of size $\tilde{O}(\epsilon^{-2} \log n)$, and then we pick from it a uniform sample $R \subseteq Q$ of size $\tilde{O}(\epsilon^{-2} \log |Q|) = \tilde{O}(\epsilon^{-2} \log \log n)$. By our preceding discussion, with constant high probability, Q is a stable coreset of P , and also R is a stable coreset of Q . Now if coresets can be composed (i.e., a coreset of a coreset is a coreset), which is indeed the case for strong coresets, then R is also a coreset of our initial dataset P . The crux is to apply this argument iteratively until the size of the sample no longer decreases, and by solving $|Q| \leq \epsilon^{-2} \log(|Q|/\epsilon)$ we can easily see that the final sample size is $\tilde{O}(\epsilon^{-2})$. Importantly, while this analysis goes through cumbersome iterations, the algorithm is simple: A uniform sample R from Q , which itself is a uniform sample from P , amounts to a uniform sample R taken directly from P . Therefore, the final coreset is just a uniform sample of size $\tilde{O}(\epsilon^{-2})$ from the input P .

Unfortunately, the above strategy has considerable gaps. One issue is that each iteration might fail with a constant probability, which is too large because the number of iterations is typically $O(\log^* n)$. One can decrease this probability by tuning parameters differently

(standard amplification by majority over repetitions would deviate from uniform sampling), but the final coreset size will depend on n , missing our intended goal. Another issue is that the errors accumulate over the iterations. Prior work [9, 21, 37] addressed this by setting a different approximation parameter ϵ_i for each iteration i . Typically, one starts with a small value, say $\epsilon_1 = 1/\log n$, and increases it exponentially fast until some final value ϵ_{final} , which determines the final coreset size. For strong coresets, the accumulated error is $\prod_i (1 + \epsilon_i) \leq 1 + \sum_i O(\epsilon_i) \leq 1 + O(\epsilon_{\text{final}})$, however for stable coresets the errors compound completely differently. Suppose that each iteration yields a stable $(\epsilon, 2\epsilon)$ -coreset, then we must set $\epsilon_{i+1} = \epsilon_i/2 = \epsilon_1/2^i$, which results overall in a stable $(\epsilon_{\text{final}}, \epsilon_1)$ -coreset. As the number of iterations grows with n , albeit slowly like $O(\log^* n)$, this guarantee is rather poor and certainly misses our intended goal.

Since iterative size reduction cannot be applied in a black-box manner, we need to implement the iterations in a more specialized manner. We identify that (a slight adaptation of) a notion called *Relative Cost-Difference Approximation* (RCDA), that was introduced by [13], avoids the core issue in the scheme above. Informally, the RCDA notion asserts that the difference in costs between any two candidate centers is approximately the same when measured on the sample versus on the full dataset. We compose this RCDA notion over the iterations, which requires some technical adaptations. It also introduces some error terms and failure probabilities, and tracking these carefully over the iterations is somewhat intricate. Our main technical lemma (Lemma 4.2) shows that the final sample is an RCDA for P , which, in turn, implies that it forms a stable coreset.

1.3 Related Work

Strong coresets, which are often simply referred to as coresets, have been extensively studied for the k -median and k -means problems in Euclidean metrics. After a long line of works, which includes [31, 15, 28, 52, 39, 9, 22], the current state-of-the-art constructions for k -median and k -means match the known lower bound of $\Omega(k\epsilon^{-2})$, up to polylogarithmic factors [20, 37]. For the ℓ_1 metric, a coreset of size $\text{poly}(k/\epsilon)$ follows from [40] by going through ℓ_2 -squared. The case of $k = 1$ has received special attention. In low-dimensional Euclidean spaces [32] obtain a strong coreset of size $\tilde{O}(\sqrt{d}/\epsilon)$, which was subsequently improved by [2] to $\tilde{O}(\epsilon^{-d/(d+1)})$ for any constant dimension d .

Many coreset variants of the k -median and k -means problems have been studied as well. Closest to our work is the line of weak coresets for 1-median in Euclidean spaces, where the construction can be performed in sublinear time; see Table 1 for a detailed comparison of previous results. Additionally, coresets have been studied for other variants of the geometric median problem and the more general k -median problem in Euclidean metrics, including capacitated k -median [7, 36], fair clustering [50, 35, 18], robust clustering [34, 41, 27], fuzzy clustering [6], ordered weighted clustering [8], and time-series clustering [38].

Another closely related problem is the 1-center, also known as the minimum enclosing ball problem. Strong coresets for this problem in Euclidean metrics require exponential dependence on the dimension [3]. Weak coresets, in contrast, were studied by [11, 10], who show a tight size bound of $\lceil 1/\epsilon \rceil$. For the ℓ_1 metric, [12] studied both strong and weak coresets, and showed that the coreset size must depend on the dimension.

Lastly, a separate line of work studies sublinear approximation algorithms for the geometric median (or the related 1-means objective), rather than constructing a coreset [19, 5].

2 Preliminaries

In the 1-median problem on the metric space $\ell_q^d = (\mathbb{R}^d, \|\cdot\|_q)$, we are given a set $P \subset \mathbb{R}^d$ and the goal is to find a point $x \in \mathbb{R}^d$ that minimizes the sum of distances to P , that is, to minimize the cost function

$$\text{cost}_q(x, P) = \frac{1}{|P|} \sum_{p \in P} \|x - p\|_q.$$

We denote by μ^P an optimal solution to the median problem on P , $\mu^P = \text{argmin}_{x \in \mathbb{R}^d} \text{cost}_q(x, P)$.

In this paper, we focus on the 1-median problem in the Euclidean metric, that is, $q = 2$. Throughout our analysis, we leverage metric embeddings to translate results between ℓ_2 and ℓ_1 .

The following theorem is a classical consequence of Gordon's refinement of Dvoretzky's theorem [29, 49].

► **Theorem 2.1** (Dvoretzky's theorem). *For every $\gamma \in (0, 1)$ and natural number d , there exists a linear map $f : \mathbb{R}^d \rightarrow \mathbb{R}^m$ with $m = O(\gamma^{-2}d)$ such that*

$$\forall x \in \mathbb{R}^d, \quad \|f(x)\|_1 \in (1 \pm \gamma) \|x\|_2.$$

Given such embedding f , it immediately follows that the cost function is also preserved up to $1 \pm \gamma$. For a set $P \subset \mathbb{R}^d$, we denote by $f(P) = \{f(p) : p \in P\}$ the image of P under f .

► **Corollary 2.2.** *Let $f : \mathbb{R}^d \rightarrow \mathbb{R}^m$ be as in Theorem 2.1 with parameter γ and let $P \subset \mathbb{R}^d$. Then*

$$\forall x \in \mathbb{R}^d, \quad \text{cost}_1(f(x), f(P)) \in (1 \pm \gamma) \text{cost}_2(x, P).$$

We will need the following folklore isometric terminal embedding in ℓ_2 . Although variants of it appear in [26, 39], we include a proof for completeness.

► **Theorem 2.3** (Folklore). *Let $P \subset \mathbb{R}^d$ of size n . Then there exists an isometric terminal embedding $f : \mathbb{R}^d \rightarrow \mathbb{R}^n$, that is, for all $x \in \mathbb{R}^d$ and $p \in P$, $\|f(x) - f(p)\|_2 = \|x - p\|_2$.*

Proof. Let $P = \{p_1, \dots, p_n\} \subset \mathbb{R}^d$ be a terminal set. By shifting the points, we can assume w.l.o.g. that p_n lies at the origin. Consider the $(n-1)$ -dimensional subspace V spanned by the vectors $p_1, \dots, p_{n-1}$. As every d -dimensional subspace of ℓ_2 is isometric to $\mathbb{R}^d$, we can also assume w.l.o.g. that $V = \mathbb{R}^{n-1}$. Let $\Pi : \mathbb{R}^d \rightarrow V$ be the orthogonal projection map onto V and define the terminal embedding $f : \mathbb{R}^d \rightarrow V \times \mathbb{R} = \mathbb{R}^n$ by $f(z) = (\Pi(z), \|\Pi(z) - z\|_2)$. Using the Pythagorean theorem, we conclude that for every $p \in P \subset V$ and $x \in \mathbb{R}^d$

$$\begin{aligned} \|f(p) - f(x)\|_2^2 &= \|\Pi(p) - \Pi(x)\|_2^2 + (\|\Pi(x) - x\|_2 - 0)^2 \\ &= \|p - \Pi(x)\|_2^2 + \|\Pi(x) - x\|_2^2 = \|p - x\|_2^2. \end{aligned}$$

◀

3 RCDA in ℓ_2 via Dimension Reduction

The RCDA property was formulated in [13]. Here we introduce a slightly modified definition.

► **Definition 3.1** (Relative Cost-Difference Approximation). *We say that $Q \subseteq P$ is an ϵ -RCDA $_q$ of P if*

$$\begin{aligned} \forall x, y \in \mathbb{R}^d, \quad & \left| [\text{cost}_q(x, P) - \text{cost}_q(y, P)] - [\text{cost}_q(x, Q) - \text{cost}_q(y, Q)] \right| \\ & \leq \epsilon \cdot (\text{cost}_q(x, P) + \text{cost}_q(y, P)). \end{aligned} \quad (4)$$

We will denote the left-hand side by $\text{RCDA}_q(x, y, Q, P)$ for brevity.

Our definition differs from that in [13] in that (4) quantifies over all pairs $x, y \in \mathbb{R}^d$, whereas the original property was defined with respect to x and the median μ^P . Specifically, the original definition required

$$\forall x \in \mathbb{R}^d, \quad \text{RCDA}_q(x, \mu^P, Q, P) \leq \epsilon \text{cost}_q(x, P). \quad (5)$$

However, the two definitions are essentially equivalent. Indeed, if for every $x \in \mathbb{R}^d$, $\text{RCDA}_q(x, \mu^P, Q, P) \leq \epsilon \cdot \text{cost}_q(x, P)$, then for every $x, y \in \mathbb{R}^d$,

$\text{RCDA}_q(x, y, Q, P) \leq \text{RCDA}_q(x, \mu^P, Q, P) + \text{RCDA}_q(\mu^P, y, Q, P) \leq \epsilon \cdot (\text{cost}_q(x, P) + \text{cost}_q(y, P))$, where the first inequality follows from the triangle inequality. Conversely, if for every $x, y \in \mathbb{R}^d$, we have $\text{RCDA}_q(x, y, Q, P) \leq \epsilon \cdot (\text{cost}_q(x, P) + \text{cost}_q(y, P))$, then for every $x \in X$,

$$\text{RCDA}_q(x, \mu^P, Q, P) \leq \epsilon \cdot (\text{cost}_q(x, P) + \text{cost}_q(\mu^P, P)) \leq 2\epsilon \cdot \text{cost}_q(x, P).$$

The main technical result from [13] establishes the RCDA property for a uniform sample in ℓ_1^d .

► **Lemma 3.2** (Lemma 4.6 in [13]). *Let $\epsilon \in (0, 1)$ and let $P \subset \mathbb{R}^d$ of size n . Then a uniform sample $Q \subseteq P$ of size $O(\epsilon^{-2} \log \frac{d}{\delta})$ is an ϵ -RCDA₁ of P in ℓ_1^d with probability at least $1 - \delta$.*

► **Remark 3.3.** While [13] does not explicitly state whether the uniform sample is taken with or without replacement, their result holds in both cases. The only step in their proof that depends on the sampling scheme is a theorem from [42] on ϵ -approximations, which applies equally to sampling with and without replacement [23].

A direct consequence of the RCDA property, is a uniform bound over all costs.

► **Proposition 3.4.** *Let $\epsilon \in (0, 1/2)$ and $\alpha \geq 1$. If Q is an ϵ -RCDA_q of P and $\text{cost}_q(\mu^P, Q) \leq \alpha \text{cost}_q(\mu^P, P)$, then,*

$$\forall x \in \mathbb{R}^d, \quad \text{cost}_q(x, Q) \leq 2\alpha \text{cost}_q(x, P).$$

Proof. By Definition 3.1, we have $\text{RCDA}_q(x, \mu^P, Q, P) \leq \epsilon \cdot (\text{cost}_q(x, P) + \text{cost}_q(\mu^P, P))$. Therefore:

$$\begin{aligned} \text{cost}_q(x, Q) &\leq \text{cost}_q(x, P) + \text{cost}_q(\mu^P, Q) - \text{cost}_q(\mu^P, P) + \epsilon(\text{cost}_q(x, P) + \text{cost}_q(\mu^P, P)) \\ &\leq (1 + \epsilon) \text{cost}_q(x, P) + (\alpha - 1 + \epsilon) \text{cost}_q(\mu^P, P) \\ &\leq (1 + \epsilon) \text{cost}_q(x, P) + (\alpha - 1 + \epsilon) \text{cost}_q(x, P) && (\mu^P \text{ is optimal}) \\ &= (\alpha + 2\epsilon) \text{cost}_q(x, P) \leq 2\alpha \text{cost}_q(x, P). && (2\epsilon \leq 1 \leq \alpha) \end{aligned}$$

◀

We now establish that uniform sampling yields RCDA in ℓ_2^d by reducing to the ℓ_1 case.

► **Lemma 3.5.** *Let $\epsilon \in (0, 1)$, $\alpha \geq 1$, and let $P \subset \mathbb{R}^d$ of size n . Then, with probability at least $1 - 1/\alpha - \delta$, a uniform sample $Q \subseteq P$ of size $O(\epsilon^{-2} \log \frac{\alpha \cdot n}{\epsilon \cdot \delta})$ satisfies:*

1. Q is an ϵ -RCDA₂ of P in ℓ_2^d , and,
2. For every $x \in \mathbb{R}^d$, $\text{cost}_2(x, Q) \leq 3\alpha \text{cost}_2(x, P)$.

Proof. By Theorem 2.3, we may assume without loss of generality that $d = n$.

Let $\gamma = \frac{\epsilon}{10\alpha}$. By Theorem 2.1, there exists $m = O(\alpha^2 \epsilon^{-2} n)$ and a linear map $f : \mathbb{R}^d \rightarrow \mathbb{R}^m$ such that for all $x \in \mathbb{R}^d$, $\|f(x)\|_1 \in (1 \pm \gamma) \|x\|_2$.

30:8 Optimal Stable Coresets for Geometric Median via Uniform Sampling

By Lemma 3.2, a uniform sample $Q \subseteq P$ of size $O(\epsilon^{-2} \log \frac{m}{\delta}) = O(\epsilon^{-2} \log \frac{\alpha n}{\epsilon \delta})$ is a $\frac{\epsilon}{2}$ -RCDA₁ of $f(P)$ in ℓ_1^m with probability at least $1 - \delta$. That is,

$$\begin{aligned} \forall x, y \in \mathbb{R}^d, \quad \text{RCDA}_1(f(x), f(y), f(Q), f(P)) \\ \leq \frac{\epsilon}{2} (\text{cost}_1(f(x), f(P)) + \text{cost}_1(f(y), f(P))). \end{aligned} \quad (6)$$

Let μ^P be the optimal median of P in ℓ_2^d . Since $\mathbb{E}_Q[\text{cost}_1(f(\mu^P), f(Q))] = \text{cost}_1(f(\mu^P), f(P))$, by Markov's inequality, with probability at least $1 - 1/\alpha$,

$$\text{cost}_1(f(\mu^P), f(Q)) \leq \alpha \text{cost}_1(f(\mu^P), f(P)). \quad (7)$$

In this event, by Proposition 3.4 applied in ℓ_1^m (noting $\epsilon/2 < 1/2$),

$$\forall x \in \mathbb{R}^d, \quad \text{cost}_1(f(x), f(Q)) \leq 2\alpha \text{cost}_1(f(x), f(P)). \quad (8)$$

Using Corollary 2.2, for all $x \in \mathbb{R}^d$,

$$\begin{aligned} \text{cost}_2(x, Q) &\leq (1 + \gamma) \text{cost}_1(f(x), f(Q)) \leq (1 + \gamma) \cdot 2\alpha \text{cost}_1(f(x), f(P)) \\ &\leq \frac{2\alpha(1 + \gamma)}{1 - \gamma} \text{cost}_2(x, P) \leq 3\alpha \text{cost}_2(x, P), \end{aligned}$$

where the last inequality uses $\gamma = \epsilon/(10\alpha) \leq 1/10$, establishing property (2).

To show property (1), let $x, y \in \mathbb{R}^d$. Again using Corollary 2.2 applied to each of the four cost terms and the triangle inequality:

$$\begin{aligned} \text{RCDA}_2(x, y, Q, P) &\leq \text{RCDA}_1(f(x), f(y), f(Q), f(P)) + \gamma(\text{cost}_1(f(x), f(P)) + \text{cost}_1(f(y), f(P)) \\ &\quad + \text{cost}_1(f(x), f(Q)) + \text{cost}_1(f(y), f(Q))) \end{aligned}$$

By the union bound (6) and (8) holds with probability at least $1 - 1/\alpha - \delta$. We get:

$$\begin{aligned} \text{RCDA}_2(x, y, Q, P) &\leq \frac{\epsilon}{2} (\text{cost}_1(f(x), f(P)) + \text{cost}_1(f(y), f(P))) \\ &\quad + \gamma(2\alpha + 1)(\text{cost}_1(f(x), f(P)) + \text{cost}_1(f(y), f(P))) \\ &\leq \frac{1}{1 - \gamma} \left(\frac{\epsilon}{2} + \gamma(2\alpha + 1) \right) (\text{cost}_2(x, P) + \text{cost}_2(y, P)) \\ &\leq \frac{10}{9} \left(\frac{\epsilon}{2} + \frac{\epsilon(2\alpha + 1)}{10\alpha} \right) (\text{cost}_2(x, P) + \text{cost}_2(y, P)) \\ &\leq \frac{10}{9} \left(\frac{\epsilon}{2} + \frac{3\epsilon}{10} \right) (\text{cost}_2(x, P) + \text{cost}_2(y, P)) \leq \epsilon(\text{cost}_2(x, P) + \text{cost}_2(y, P)). \blacktriangleleft \end{aligned}$$

4 Upper Bound (Proof of Theorem 1.1)

To apply the iterative size reduction technique, we first establish the following composition property for RCDA, which allows errors to be tracked across iterations.

► **Proposition 4.1.** *Let $Q'' \subseteq Q' \subseteq Q$ with $\epsilon, \epsilon' \in (0, 1)$ and $\beta \geq 1$. Suppose Q' is an ϵ -RCDA_q of Q and Q'' is an ϵ' -RCDA_q of Q' . If for every $x \in \mathbb{R}^d$ we have $\text{cost}_q(x, Q') \leq \beta \text{cost}_q(x, Q)$, then, Q'' is an $(\epsilon + \epsilon'\beta)$ -RCDA_q of Q .*

Proof. For $x, y \in \mathbb{R}^d$, using the triangle inequality,

$$\begin{aligned} \text{RCDA}_q(x, y, Q, Q'') &\leq \text{RCDA}_q(x, y, Q, Q') + \text{RCDA}_q(x, y, Q', Q'') \\ &\leq \epsilon(\text{cost}_q(x, Q) + \text{cost}_q(y, Q)) + \epsilon'(\text{cost}_q(x, Q') + \text{cost}_q(y, Q')) \\ &\leq (\epsilon + \epsilon'\beta)(\text{cost}_q(x, Q) + \text{cost}_q(y, Q)). \blacktriangleleft \end{aligned}$$

We are now ready to prove our main technical lemma. The proof involves balancing the errors with the guaranteed size reduction. To streamline the presentation, we use explicit constants and while we attempt to avoid huge constants, they are probably far from optimal.

► **Lemma 4.2.** *Let $\epsilon \in (0, \frac{1}{4})$ and let $P \subset \mathbb{R}^d$ of size n . Then, with probability at least $1 - \frac{1}{30}$, a uniform sample $Q \subseteq P$ of size $O(\epsilon^{-2} \log \frac{1}{\epsilon})$ is an ϵ -RCDA₂ of P in ℓ_2^d .*

Proof. By Theorem 2.3, we may assume $d = n$. Denote the i -th iterated logarithm (with base 2) by $\log^{(i)}$, i.e., $\log^{(i+1)} n = \log(\log^{(i)} n)$ where $\log^{(0)} n = n$. Let $t \geq 1$ be the smallest integer such that $\log^{(t-1)} n \leq 2048\epsilon^{-2} \log \epsilon^{-1}$; then it also satisfies $\log^{(t-1)} n > \log(2048\epsilon^{-2} \log \epsilon^{-1}) \geq 16$. For $i = 1, \dots, t$, define

$$\epsilon_i = \frac{\epsilon}{1440(\log^{(i)} n)^{\frac{1}{2}}}, \quad \alpha_i = 180(\log^{(i)} n)^{\frac{1}{4}}, \quad \delta_i = 1/\alpha_i.$$

We construct $Q_t \subseteq Q_{t-1} \subseteq \dots \subseteq Q_1 \subseteq Q_0 = P$, where each Q_{i+1} is obtained from Q_i by uniform sampling of size

$$O\left(\epsilon_{i+1}^{-2} \log\left(\frac{\alpha_{i+1} n_i}{\epsilon_{i+1} \delta_{i+1}}\right)\right)$$

with $n_i = |Q_i|$, as guaranteed by Lemma 3.5. We next prove the following claim.

► **Claim 4.3.** For every $i \in \{1, \dots, t\}$, with probability at least $1 - 3 \sum_{j=1}^i \delta_j$,

$$Q_i \text{ is an } \epsilon \sum_{j=1}^i \frac{1}{4(\log^{(j)} n)^{\frac{1}{4}}} \text{-RCDA}_2 \text{ of } P.$$

Proof. The proof proceeds by induction on $i \in \{1, \dots, t\}$.

Base case. By Lemma 3.5 with parameters $\epsilon_1, \alpha_1, \delta_1$, it follows that Q_1 is an ϵ_1 -RCDA₂ of P with probability at least $1 - 2\delta_1$. The base case holds as $\epsilon_1 = \epsilon/1440(\log^{(1)} n)^{\frac{1}{2}} \leq \epsilon/4(\log^{(1)} n)^{\frac{1}{4}}$.

Inductive step. Since Q_i is a uniform sample of P , by Markov's inequality, with probability at least $1 - 1/\alpha_{i+1}$,

$$\text{cost}_2(\mu^P, Q_i) \leq \alpha_{i+1} \text{cost}_2(\mu^P, P).$$

By Proposition 3.4, for every $x \in \mathbb{R}^d$, $\text{cost}_2(x, Q_i) \leq 2\alpha_{i+1} \text{cost}_2(x, P)$. Thus, applying Lemma 3.5 to sampling Q_{i+1} from Q_i gives that, with probability at least $1 - 2\delta_{i+1}$, Q_{i+1} is an ϵ_{i+1} -RCDA₂ of Q_i . As by the induction hypothesis, Q_i is an $\epsilon \sum_{j=1}^i \frac{1}{4(\log^{(j)} n)^{\frac{1}{4}}}$ -RCDA₂ of P with probability at least $1 - 3 \sum_{j=1}^i \delta_j$, both events hold with probability at least $1 - 3 \sum_{j=1}^{i+1} \delta_j$ by a union bound. Assuming they both hold, we can now apply Proposition 4.1 and get,

$$Q_{i+1} \text{ is an } \left(\epsilon \sum_{j=1}^i \frac{1}{4(\log^{(j)} n)^{\frac{1}{4}}} + \epsilon_{i+1} \cdot 2\alpha_{i+1} \right) \text{-RCDA}_2 \text{ of } P.$$

Since $\epsilon_{i+1} \cdot 2\alpha_{i+1} = \frac{\epsilon}{1440(\log^{(i+1)} n)^{\frac{1}{2}}} \cdot 360(\log^{(i+1)} n)^{\frac{1}{4}} = \frac{\epsilon}{4(\log^{(i+1)} n)^{\frac{1}{4}}}$, we have,

$$\epsilon \sum_{j=1}^i \frac{1}{4(\log^{(j)} n)^{\frac{1}{4}}} + \epsilon_{i+1} \cdot 2\alpha_{i+1} \leq \epsilon \sum_{j=1}^{i+1} \frac{1}{4(\log^{(j)} n)^{\frac{1}{4}}},$$

concluding the induction step, and proving the claim. $\triangleleft$

30:10 Optimal Stable Coresets for Geometric Median via Uniform Sampling

We proceed with the proof of Lemma 4.2. By choice of t , for $1 \leq j \leq t-2$, we have $(\log^{(j+1)} n)^{\frac{1}{4}} \leq \frac{1}{8}(\log^{(j)} n)^{\frac{1}{4}}$ since $\log^{(j)} n \geq 2^{16}$. Thus,

$$\epsilon \sum_{j=1}^{t-1} \frac{1}{4(\log^{(j)} n)^{\frac{1}{4}}} \leq \frac{\epsilon}{4(\log^{(t-1)} n)^{\frac{1}{4}}} \sum_{j=0}^{t-2} 8^{-j} \leq \frac{2\epsilon}{7(\log^{(t-1)} n)^{\frac{1}{4}}},$$

where the last inequality uses the standard bound over a sum of a geometric series. Additionally, since $\log^{(t-1)} n \geq 16$,

$$\epsilon \sum_{j=1}^t \frac{1}{4(\log^{(j)} n)^{\frac{1}{4}}} \leq \frac{2\epsilon}{7(\log^{(t-1)} n)^{\frac{1}{4}}} + \frac{\epsilon}{4(\log^{(t)} n)^{\frac{1}{4}}} \leq \frac{\epsilon}{2},$$

and similarly,

$$3 \sum_{j=1}^t \delta_j \leq 3 \sum_{j=1}^{t-1} \delta_j + 3\delta_t \leq \frac{24}{7} \delta_{t-1} + 3\delta_t \leq \frac{24}{7 \cdot 180(\log^{(t-1)} n)^{\frac{1}{4}}} + \frac{3}{180(\log^{(t)} n)^{\frac{1}{4}}} \leq \frac{1}{105} + \frac{1}{60} \leq \frac{3}{100},$$

where overall, we get that Q_t is ϵ -RCDA₂ of P with probability at least $1 - \frac{3}{100}$.

Sample size. It is easy to verify the invariant $n_i \leq C \epsilon^{-2} \log \epsilon^{-1} (\log^{(i)} n)^2$ for all $i \in \{1, \dots, t\}$, for a sufficiently large constant C . Thus, at iteration t , since $\log^{(t-1)} n = O(\epsilon^{-2} \log \epsilon^{-1})$, we have that $n_t = O(\epsilon^{-2} \log \epsilon^{-1} \cdot (\log(\epsilon^{-2} \log \epsilon^{-1}))^2) = O(\epsilon^{-2} \log^3 \frac{1}{\epsilon})$.

To further improve the sample size to $O(\epsilon^{-2} \log \frac{1}{\epsilon})$, we apply one additional iteration with constant parameters. Set $\epsilon_f = \frac{\epsilon}{4000}$, $\alpha_f = 1000$, and $\delta_f = 1/\alpha_f$, and uniformly sample $Q \subseteq Q_t$ of size

$$O\left(\epsilon_f^{-2} \log \frac{\alpha_f n_t}{\epsilon_f \delta_f}\right) = O(\epsilon^{-2} \log \frac{1}{\epsilon}),$$

where note that, since sampling is uniform at every step, Q is also a uniform sample of P .

Finally, by applying the same arguments as in the inductive step, we get that Q is an

$$\left(\frac{\epsilon}{2} + \epsilon_f \cdot 2\alpha_f\right)\text{-RCDA}_2 = \epsilon\text{-RCDA}_2$$

uniform sample of P of size $O(\epsilon^{-2} \log \frac{1}{\epsilon})$ with probability at least $1 - \frac{3}{100} - (2/\alpha_f + \delta_f) \geq 1 - \frac{1}{30}$. This concludes the proof of Lemma 4.2. $\blacktriangleleft$

We can now prove our main result, by plugging our notion of RCDA into the framework of [13].

► **Theorem 1.1.** *Let $\epsilon \in (0, 1/4)$ and let $P \subset \mathbb{R}^d$ be finite. Then a uniform sample of size $O(\epsilon^{-2} \log \frac{1}{\epsilon})$ from P is a stable $(\epsilon, 17\epsilon)$ -coreset for 1-median in ℓ_2^d with probability at least $4/5$.*

Proof. By Lemma 4.2, a uniform sample Q of size $O(\epsilon^{-2} \log \epsilon^{-1})$ is $\frac{\epsilon}{2}$ -RCDA₂ of P in ℓ_2^d with probability at least $29/30$. By Markov's inequality, with probability at least $5/6$, $\text{cost}_2(\mu^P, Q) \leq 6 \text{cost}_2(\mu^P, P)$. By the union bound, both events occur with probability at least $4/5$. Now, let $x, y \in \mathbb{R}^d$ such that $\text{cost}_2(x, Q) \leq (1 + \epsilon) \text{cost}_2(y, Q)$, then

$$\begin{aligned}
\text{cost}_2(x, P) &\leq \frac{\epsilon}{2} [\text{cost}_2(x, P) + \text{cost}_2(y, P)] + [\text{cost}_2(x, Q) - \text{cost}_2(y, Q)] + \text{cost}_2(y, P) \\
&\quad (Q \text{ is } \frac{\epsilon}{2}\text{-RCDA}_2 \text{ of } P) \\
&\leq \frac{\epsilon}{2} \text{cost}_2(x, P) + (1 + \frac{\epsilon}{2}) \text{cost}_2(y, P) + \epsilon \text{cost}_2(y, Q) \\
&\quad (\text{cost}_2(x, Q) \leq (1 + \epsilon) \text{cost}_2(x, Q)) \\
&\leq \frac{\epsilon}{2} \text{cost}_2(x, P) + (1 + \frac{\epsilon}{2} + 12\epsilon) \text{cost}_2(y, P) \\
&\quad (\text{by Proposition 3.4, } \text{cost}_2(y, Q) \leq 12 \text{cost}_2(y, P)) \\
&\leq (1 + \epsilon)(1 + 12.5\epsilon) \text{cost}_2(y, P) \leq (1 + 17\epsilon) \text{cost}_2(y, P).
\end{aligned}$$

This establishes that Q is a stable $(\epsilon, 17\epsilon)$ -coreset of P . ◀

References

- 1 Marcel R. Ackermann, Johannes Blömer, and Christian Sohler. Clustering for metric and nonmetric distance measures. *ACM Trans. Algorithms*, 6(4):59:1–59:26, 2010. doi:10.1145/1824777.1824779.
- 2 Peyman Afshani and Chris Schwiegelshohn. Optimal coresets for low-dimensional geometric median. In *ICML, Proceedings of Machine Learning Research*, pages 262–270. PMLR / OpenReview.net, 2024. URL: <https://proceedings.mlr.press/v235/afshani24a.html>.
- 3 Pankaj K. Agarwal, Sarel Har-Peled, and Kasturi R. Varadarajan. Geometric approximation via coresets. In *Combinatorial and computational geometry*, volume 52 of *MSRI Publications*, chapter 1, pages 1–30. Cambridge University Press, 2005. URL: <https://library2.msri.org/books/Book52/contents.html>.
- 4 Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.
- 5 Beatrice Bertolotti, Matteo Russo, Chris Schwiegelshohn, and Sudarshan Shyam. Simple and optimal sublinear algorithms for mean estimation. In *NeurIPS*, 2025. URL: <https://openreview.net/forum?id=9MLfWJZXTn>.
- 6 Johannes Blömer, Sascha Brauer, and Kathrin Bujna. Coresets for fuzzy k-means with applications. In *ISAAC*, volume 123 of *LIPIcs*, pages 46:1–46:12. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018. doi:10.4230/LIPIcs.ISAAC.2018.46.
- 7 Vladimir Braverman, Vincent Cohen-Addad, Shaofeng H.-C. Jiang, Robert Krauthgamer, Chris Schwiegelshohn, Mads Bech Tofttrup, and Xuan Wu. The power of uniform sampling for coresets. In *FOCS*, pages 462–473. IEEE, 2022. doi:10.1109/FOCS54457.2022.00051.
- 8 Vladimir Braverman, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for ordered weighted clustering. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 744–753. PMLR, 2019. URL: <http://proceedings.mlr.press/v97/braverman19a.html>.
- 9 Vladimir Braverman, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for clustering in excluded-minor graphs and beyond. In *SODA*, pages 2679–2696. SIAM, 2021. doi:10.1137/1.9781611976465.159.
- 10 Mihai Bădoiu and Kenneth L. Clarkson. Optimal core-sets for balls. *Comput. Geom.*, 40(1):14–22, 2008. doi:10.1016/j.comgeo.2007.04.002.
- 11 Mihai Bădoiu, Sarel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In *STOC*, pages 250–257. ACM, 2002. doi:10.1145/509907.509947.
- 12 Amir Carmel, Chengzhi Guo, Shaofeng H.-C. Jiang, and Robert Krauthgamer. Coresets for 1-center in ℓ_1 metrics. In *ITCS*, volume 325 of *LIPIcs*, pages 28:1–28:20. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025. doi:10.4230/LIPIcs.ITCS.2025.28.
- 13 Amir Carmel and Robert Krauthgamer. Stable coresets: Unleashing the power of uniform sampling. In *14th International Conference on Learning Representations, ICLR 2026*, 2026. URL: <https://openreview.net/forum?id=sOpAa8iR0A>.

- 14 Ching-Lueh Chang. On ultrametric 1-median selection. *Theoretical Computer Science*, 828-829:65–69, 2020. doi:10.1016/j.tcs.2020.04.021.
- 15 Ke Chen. On coresets for k -median and k -means clustering in metric and Euclidean spaces and their applications. *SIAM J. Comput.*, 39(3):923–947, 2009. doi:10.1137/070699007.
- 16 Xue Chen and Michal Derezhinski. Query complexity of least absolute deviation regression via robust uniform convergence. In *COLT*, Proceedings of Machine Learning Research, pages 1144–1179. PMLR, 2021. URL: <http://proceedings.mlr.press/v134/chen21d.html>.
- 17 Herman Chernoff. *Sequential analysis and optimal design*. SIAM, 1972.
- 18 Rachit Chhaya, Anirban Dasgupta, Jayesh Choudhari, and Supratim Shit. On coresets for fair regression and individually fair clustering. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 9603–9625. PMLR, 2022. URL: <https://proceedings.mlr.press/v151/chhaya22a.html>.
- 19 Michael B. Cohen, Yin Tat Lee, Gary L. Miller, Jakub Pachocki, and Aaron Sidford. Geometric median in nearly linear time. In *STOC*, pages 9–21. ACM, 2016. doi:10.1145/2897518.2897647.
- 20 Vincent Cohen-Addad, Kasper Green Larsen, David Saulpic, and Chris Schwiegelshohn. Towards optimal lower bounds for k -median and k -means coresets. In *STOC*, pages 1038–1051. ACM, 2022. doi:10.1145/3519935.3519946.
- 21 Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. Improved coresets and sublinear algorithms for power means in Euclidean spaces. In *NeurIPS*, pages 21085–21098. Curran Associates, Inc., 2021. URL: <https://proceedings.neurips.cc/paper/2021/hash/b035d6563a2adac9f822940c145263ce-Abstract.html>.
- 22 Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. A new coreset framework for clustering. In *STOC*, pages 169–182. ACM, 2021. doi:10.1145/3406325.3451022.
- 23 Mónika Csikós and Nabil H. Mustafa. Optimal approximations made easy. *Inf. Process. Lett.*, 176:106250, 2022. doi:10.1016/J.IPL.2022.106250.
- 24 Ovidiu Daescu and Ka Yaw Teo. Fast deterministic approximation of medoid in $\mathbb{R}^d$. In *CCCG*, pages 83–90, 2022.
- 25 Matan Danos. Coresets for clustering by uniform sampling and generalized rank aggregation. Master’s thesis, Weizmann Institute of Science, Rehovot, Israel, 2021. URL: https://www.wisdom.weizmann.ac.il/~robi/files/MatanDanos-MScThesis-2021_11.pdf.
- 26 Michael Elkin, Arnold Filtser, and Ofer Neiman. Terminal embeddings. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX-/RANDOM*, volume 40 of *LIPIcs*, pages 242–264. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2015. doi:10.4230/LIPIcs.APPROX-RANDOM.2015.242.
- 27 Ziyi Fang, Lingxiao Huang, and Runkai Yang. Coreset for robust geometric median: Eliminating size dependency on outliers. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025. URL: <https://openreview.net/forum?id=Yhfx12Azz1>.
- 28 Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *STOC*, pages 569–578. ACM, 2011. doi:10.1145/1993636.1993712.
- 29 Yehoram Gordon. Gaussian processes and almost spherical sections of convex bodies. *The Annals of Probability*, 16(1):180–188, 1988.
- 30 Sariel Har-Peled, Mitchell Jones, and Saladi Rahul. Active-learning a convex body in low dimensions. *Algorithmica*, 83(6):1885–1917, 2021. doi:10.1007/S00453-021-00807-W.
- 31 Sariel Har-Peled and Soham Mazumdar. On coresets for k -means and k -median clustering. In *STOC*, pages 291–300. ACM, 2004. doi:10.1145/1007352.1007400.
- 32 Lingxiao Huang, Ruiyuan Huang, Zengfeng Huang, and Xuan Wu. On coresets for clustering in small dimensional euclidean spaces. In *ICML*, Proceedings of Machine Learning Research, pages 13891–13915. PMLR, 2023. URL: <https://proceedings.mlr.press/v202/huang23h.html>.
- 33 Lingxiao Huang, Shaofeng H.-C. Jiang, and Jianing Lou. The power of uniform sampling for k -median. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 13933–13956. PMLR, 2023. URL: <https://proceedings.mlr.press/v202/huang23j.html>.

- 34 Lingxiao Huang, Shaofeng H.-C. Jiang, Jianing Lou, and Xuan Wu. Near-optimal coresets for robust clustering. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net, 2023. URL: <https://openreview.net/forum?id=Nc1ZkRW8Vde>.
- 35 Lingxiao Huang, Shaofeng H.-C. Jiang, and Nisheeth K. Vishnoi. Coresets for clustering with fairness constraints. In *NeurIPS*, pages 7587–7598, 2019. URL: <https://proceedings.neurips.cc/paper/2019/hash/810dfbbebb17302018ae903e9cb7a483-Abstract.html>.
- 36 Lingxiao Huang, Jian Li, Pinyan Lu, and Xuan Wu. Coresets for constrained clustering: General assignment constraints and improved size bounds. In *SODA*, pages 4732–4782. SIAM, 2025. doi:10.1137/1.9781611978322.161.
- 37 Lingxiao Huang, Jian Li, and Xuan Wu. On optimal coreset construction for Euclidean (k, z) -clustering. In *STOC*, pages 1594–1604. ACM, 2024. doi:10.1145/3618260.3649707.
- 38 Lingxiao Huang, K. Sudhir, and Nisheeth K. Vishnoi. Coresets for time series clustering. In *NeurIPS*, pages 22849–22862, 2021. URL: <https://proceedings.neurips.cc/paper/2021/hash/c115ba9e04ab27fbbb664f932112246d-Abstract.html>.
- 39 Lingxiao Huang and Nisheeth K. Vishnoi. Coresets for clustering in Euclidean spaces: importance sampling is nearly optimal. In *STOC*, pages 1416–1429. ACM, 2020. doi:10.1145/3357713.3384296.
- 40 Shaofeng H.-C. Jiang, Robert Krauthgamer, Jianing Lou, and Yubo Zhang. Coresets for kernel clustering. *Mach. Learn.*, 113(8):5891–5906, 2024. doi:10.1007/S10994-024-06540-Z.
- 41 Shaofeng H.-C. Jiang and Jianing Lou. Coresets for robust clustering via black-box reductions to vanilla case. In *ICALP*, volume 334 of *LIPIcs*, pages 101:1–101:18. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025. doi:10.4230/LIPIcs.ICALP.2025.101.
- 42 Yi Li, Philip M. Long, and Aravind Srinivasan. Improved bounds on the sample complexity of learning. *J. Comput. Syst. Sci.*, 62(3):516–527, 2001. doi:10.1006/jcss.2000.1741.
- 43 Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *J. Mach. Learn. Res.*, 5:623–648, 2004. URL: <https://jmlr.org/papers/volume5/mannor04b/mannor04b.pdf>.
- 44 Stanislav Minsker, Sanvesh Srivastava, Lizhen Lin, and David B. Dunson. Scalable and robust bayesian inference via the median posterior. In *ICML, JMLR Workshop and Conference Proceedings*, pages 1656–1664. JMLR.org, 2014. URL: <http://proceedings.mlr.press/v32/minsker14.html>.
- 45 Alexander Munteanu and Chris Schwiegelshohn. Coresets-methods and history: A theoreticians design pattern for approximation and streaming algorithms. *Künstliche Intell.*, 32(1):37–53, 2018. doi:10.1007/S13218-017-0519-3.
- 46 James Newling and François Fleuret. A sub-quadratic exact medoid algorithm. In *AISTATS*, Proceedings of Machine Learning Research, pages 185–193. PMLR, 2017. URL: <http://proceedings.mlr.press/v54/newling17a.html>.
- 47 Andreas Nienkötter and Xiaoyi Jiang. Kernel-based generalized median computation for consensus learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(5):5872–5888, 2023. doi:10.1109/TPAMI.2022.3202565.
- 48 Aditya Parulekar, Advait Parulekar, and Eric Price. L1 regression with lewis weights subsampling. In *APPROX-RANDOM*, LIPIcs, pages 49:1–49:21. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021. doi:10.4230/LIPIcs.APPROX/RANDOM.2021.49.
- 49 Gideon Schechtman. A remark concerning the dependence on ϵ in Dvoretzky’s theorem. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 1987–88*, pages 274–277. Springer, 2006.
- 50 Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. Fair coresets and streaming algorithms for fair k-means. In *WAOA*, Lecture Notes in Computer Science, pages 232–251. Springer, 2019. doi:10.1007/978-3-030-39479-0_16.
- 51 S.A Shkarin. Isometric embedding of finite ultrametric spaces in banach spaces. *Topology and its Applications*, 142(1):13–17, 2004. doi:10.1016/j.topol.2003.12.002.

- 52 Christian Sohler and David P. Woodruff. Strong coresets for k -median and subspace approximation: Goodbye dimension. In *FOCS*, pages 802–813. IEEE Computer Society, 2018. doi:10.1109/FOCS.2018.00081.
- 53 David P. Woodruff and Taisuke Yasuda. Coresets for multiple ℓ_p regression. In *ICML*, Proceedings of Machine Learning Research, pages 53202–53233. PMLR, 2024. URL: <https://proceedings.mlr.press/v235/woodruff24a.html>.

A Lower Bound

► **Theorem A.1.** *For every $\epsilon \in (0, 1/4)$ and every metric space $(\mathcal{X}, \text{dist})$ containing at least two distinct points, there exists a finite point set $P \subseteq \mathcal{X}$, such that every algorithm producing a weak $(0, \epsilon)$ -coreset (and in particular weak $(\epsilon/2, \epsilon)$ -coreset) for 1-median with probability at least $4/5$, has query complexity $\Omega(\epsilon^{-2})$.*

Proof. Without loss of generality assume $\mathcal{X}$ contains two distinct points x, y of distance 1. For simplicity, assume $\epsilon/2$ divides n .

Consider two multisets P_0 and P_1 . P_0 contains $n/2$ copies of x and $n/2$ copies of y , consequently, $\text{cost}(x, P_0) = \text{cost}(y, P_0) = 1/2$. P_1 contains $n(1 - \epsilon)/2$ copies of x and $n(1 + \epsilon)/2$ copies of y . For P_1 , $\text{cost}(x, P_1) = (1 + \epsilon)/2$ and $\text{cost}(y, P_1) = (1 - \epsilon)/2$, so $\text{cost}(x, P_1) > (1 + 2\epsilon) \text{cost}(y, P_1)$.

Let Q be a multiset containing s copies of x and $m - s$ copies of y . Then $\text{cost}(x, Q)/\text{cost}(y, Q) = (m - s)/s$. If $s \geq m/2$, then $\text{cost}(x, Q) \leq \text{cost}(y, Q)$. Therefore, producing a weak coreset requires distinguishing whether Q was sampled from P_0 or P_1 . As a weak $(0, \epsilon)$ -coreset of P_1 must have y as an optimal solution. Under uniform sampling, $s \sim \text{Binomial}(m, 1/2)$ for P_0 and $s \sim \text{Binomial}(m, (1 - \epsilon)/2)$ for P_1 . By an information-theoretic lower bound, distinguishing between two binomial distributions whose biases differ by ϵ with constant probability requires $m = \Omega(\epsilon^{-2})$ samples [17, 43, 4]. ◀