Optimal Coresets for Low-Dimensional Geometric Median

Peyman Afshani [1] **Chris Schwiegelshohn** [1]

Abstract

We investigate coresets for approximating the cost with respect to median queries. In this problem, we are given a set of points $P \subset \mathbb{R}^d$ and median queries are $\sum_{p \in P} \|p - c\|$ for any point $c \in \mathbb{R}^d$. Our goal is to compute a small weighted summary $S \subset P$ such that the cost of any median query is approximated within a multiplicative $(1 \pm \varepsilon)$ factor. We provide matching upper and lower bounds on the number of points contained in S of the order $\tilde{\Theta}\left(\varepsilon^{-d/(d+1)}\right)$.

1. Introduction

Large data sets pose a considerable challenge for the modern data analyst. Polynomial time algorithms are not automatically considered efficient and space requirements, previously perhaps of secondary concern, have become paramount. In addition, many novel computational models have emerged with the goal to design algorithms that scale to even the most massive data sets. Examples include streaming, where we read the data set once while storing a very small fraction of the input, and distributed models such as the MPC model, where scalability is achieved by spreading out the computation over multiple servers.

In this backdrop, coresets have emerged as a preeminent paradigm for big data analysis. Informally, given a set of queries $f \in \mathcal{F}$, a coreset of a data set $\mathcal{P}$ is a small summary S such that $f(\mathcal{P}) \approx_\varepsilon f(S)$ for all queries f, where ε denotes the desired precision. If one has the option of efficiently computing a coreset, all of the issues raised above can be addressed. Algorithms, when run on the summary S, become significantly faster, and the coreset by its very nature has excellent space requirements.

Moreover, for many problems encountered in machine learning, a query $f \in \mathcal{F}$ corresponds to a composable loss function, that is $f(S) := \sum_{p \in P} L(f, p)$, where $L(f, p)$ is the loss incurred for point p when evaluated with the query f. In this case coresets are composable, that given coresets S_1 of $\mathcal{P}_1$ and S_2 of $\mathcal{P}_2$, $S_1 \cup S_2$ is a coreset of $\mathcal{P}_1 \cup \mathcal{P}_2$. This immediately makes a coreset computation embarrassingly parallel and also yields streaming algorithms with roughly the space requirements proportional to the size of a coreset via a reduction by way of the Merge-&-Reduce framework (Bentley & Saxe, 1980).

Among the most studied coreset problems are coresets for center-based clustering problems. For k-clustering objectives such as Euclidean k-median, a coreset of size $\tilde{O}(k \cdot \varepsilon^{-2} \cdot \min(\sqrt[3]{k}, \varepsilon^{-1}))$ have been recently discovered (Cohen-Addad et al., 2022b) and this bound is tight up to polylog factors for certain ranges of k and ε (Huang et al., 2022). Notably, these coreset sizes are independent of the number of input points.

The situation for $k = 1$, often referred to as the geometric median and the Fermat-Weber problem, merits special consideration. Here, $\mathcal{P}$ is a point set in $\mathbb{R}^d$ and a point set S is an ε-coreset if for all points $c \in \mathbb{R}^d$

$$\left| \sum_{p \in P} \|p - c\| - \sum_{p \in S} w_p \|p - c\| \right| \leq \varepsilon \cdot \sum_{p \in P} \|p - c\|, \quad (1)$$

where $\|.\|$ denotes the Euclidean norm and w_p is a non-negative weight associated to p in S. The primary measure of goodness for a coreset S is the distinct number of points in S, henceforth called the coreset size.

Perhaps surprisingly, despite coreset research being very active since their inception over 20 years ago, even the basic geometric median problem is still not fully understood. In high dimensions, a coreset size of $\tilde{O}(\varepsilon^{-2})$ was discovered before the extension to k centers was known (Cohen-Addad et al., 2021). This bound is tight up to logarithmic factors if the coreset points are a subset of the input points (Cohen-Addad et al., 2021). Subsequently, (Braverman et al., 2022), showed that improvements are possible in low dimensions. Specifically, they gave a coreset construction consisting of $\tilde{O}(\varepsilon^{-2d/(d+1)})$ points. For the Euclidean plane, the construction yielded a coreset of size $\tilde{O}(\varepsilon^{-1.5})$. This was further improved recently by (Huang et al., 2023) to $\tilde{O}(\sqrt{d} \cdot \varepsilon^{-1})$. This result recovers the $\tilde{O}(\varepsilon^{-2})$ bound of (Cohen-Addad et al., 2021) in high dimensions, as using dimension reduction techniques one can always assume

[1]Department of Computer Science, Aarhus University, Denmark. Correspondence to: Peyman Afshani <peyman@cs.au.dk>, Chris Schwiegelshohn <schwiegelshohn@cs.au.dk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

$d \in O(\varepsilon^{-2} \log \varepsilon^{-1})$, while being a significant improvement in low dimensions. (Huang et al., 2023) also provided an $\tilde{O}(\sqrt{\varepsilon^{-1}})$ sized coreset for the line metric. The latter result admits a tight lower bound (Baker et al., 2020) which is also to date the best known lower bound in low dimensions.

1.1. Our Contribution

In this paper, we settle the coreset problem in low-dimensional Euclidean space. Specifically, we prove the following theorems.

Theorem 1.1. *Let $P \subset \mathbb{R}^d$ be a set of points, let $\varepsilon > 0$ a and let $\eta > 0$ be an absolute constant. Then there exists a deterministic algorithm computing an ε-median coreset $S \subseteq P$ of size $\tilde{O}(2^{\eta \cdot d} \varepsilon^{-d/(d+1)})$, that is Equation 1 holds for all $c \in \mathbb{R}^d$.*

This bound recovers the optimal $\tilde{O}(\sqrt{\varepsilon^{-1}})$ bound of (Huang et al., 2023) for line metric and improves over all prior constructions in the low-dimensional case. Notably, the coreset size is always sublinear in ε^{-1}. The weights are non-negative, that is the weight is 0 if a point is contained in the coreset and a positive number otherwise. We complement this result via the following lower bound.

Theorem 1.2. *There exists a set of points $P \subset \mathbb{R}^d$ such that any ε-median coreset $S \subseteq P$ must have size at least $\Omega(\varepsilon^{-d/(d+1)})$.*

The lower bound assumes that the coreset consists of input points. Extending this to arbitrary coresets is an interesting open question, see the discussion at the end of the paper.

1.2. Related Work

Corsets for clustering problems have been widely studied. In the earliest days of coreset research, the low dimensional case was the most widely studied setting for clustering problems (Har-Peled & Mazumdar, 2004; Har-Peled & Kushal, 2007). In particular, the result by (Har-Peled & Kushal, 2007) implied a coreset of size $O(\varepsilon^{-1})$ for coresets on the line. In a landmark result, Chen (Chen, 2009) showed that techniques from learning theory could be used to obtain coresets in a high dimensional setting. This connection was subsequently further explored (Langberg & Schulman, 2010) and most notably the seminal paper by Feldman and Langberg (Feldman & Langberg, 2011) who gave coresets of size $\tilde{O}(d\varepsilon^{-2})$. Subsequently, most research has focussed on dimension reduction techniques for the better part of a decade (Becchetti et al., 2019; Cohen-Addad et al., 2021; Feldman et al., 2013; Feng et al., 2021; Huang & Vishnoi, 2020; Sohler & Woodruff, 2018). Only recently did more advanced techniques commonly encountered when bounding Rademacher complexities for generalization bounds enter the picture, resulting in the to-date best high dimensional coreset bound of $\tilde{O}(\varepsilon^{-2})$ (Cohen-Addad et al., 2021; 2022a).

A weaker notion of coresets has also received some attention. In this relaxation, we only require that S preserves the optimum, that is any sufficiently good solution computed on S also yields a good solution for the entire point set, see (Munteanu & Schwiegelshohn, 2018) for an overview. Notable examples are coresets of size $O(\varepsilon^{-1]})$ for the minmum enclosing ball problem (Badoiu & Clarkson, 2008), which compares favourably to an $\exp(d)$ space bound for the more general, stronger corset guarantee (Agarwal et al., 2005). Similar results yielding weak coresets of size $O(\varepsilon^{-2})$ are known for the geometric median (Cohen et al., 2016), though these bounds no longer offer an improvement over the subsequently discovered coresets of size $\tilde{O}(\varepsilon^{-2})$ (Cohen-Addad et al., 2021).

1.3. Notations

We use $\mathrm{d}(p, q) = \|p - q\|$ to denote the Euclidean distance between two points p, q in $\mathbb{R}^d$. Also, in order to simplify the presentation of the bounds, we will use the notation $O_d(\cdot)$ to ignore exponential factors of d, i.e., factors of the size $2^{O(d)}$. To be precise, $f = O_d(g)$ implies the existence of a constant c such that $f \leq 2^{cd} g$. The notation $\Omega_d(\cdot)$ is also defined similarly. To be precise, $f = \Omega_d(g)$ implies the existence of a constant c such that $f \geq \frac{g}{2^{cd}}$. We will use the notation $\tilde{O}(f)$ to ignore polylogarithmic factors, i.e., factors of the size $\log^{O(1)} f$. The notations $\tilde{O}_d(\cdot), \tilde{\Omega}_d(\cdot)$ are also defined similarly.

1.4. Organization and a Summary of Our Techniques

In Section 2, we prove an integral lemma for our coreset construction, based on a "polynomial technique". Roughly speaking, we show that given a point set P inside the cube $[-1, 1]^d$, if P contains sufficiently many points, one can assign weights between -1 and 1 to the points of P such that the weighted sum of the distances from any point q to the pointset P drops very quickly as the function of $\|q\|$ (to be specific, as $\frac{1}{\|q\|^c}$ for any fixed constant c; the constant c here is an increasing function of $|P|$ and a decreasing function of d). In Section 3, we use this to build our coreset. We observe that we can use a modified quad-tree construction and exploit the "low-dimensional" nature of the problem. We observe that a quad-tree will not be able to have too many "large" cells close to any point $q \in \mathbb{R}^d$, meaning, many of the cells of the quad-tree will be far away. Consequently, they will contribute very little to the error. Finally, in Section 4 we prove a lower bound that shows our construction is best possible (up to small polylogarithmic factors of ε^{-1}). This is done by borrowing tools developed by (Alexander, 1990) for his analytical discrepancy lower bound for halfspaces in $\mathbb{R}^d$.

2. Minimizing Additive Error

In this section we will prove a lemma that later will be used in the construction of the coreset. The input is a set $P \subset \mathbb{R}^d$ of n points. We allow the points to have identical positions which can be modelled by allowing a point $p \in P$ to have some multiplicity $\mu(p)$. For the most part, we will assume $\mu(p) = 1$ but all the arguments generalize effortlessly to the case when points have larger multiplicities. Given a point set $P \subset \mathbb{R}^d$, a *weight assignment W to P* is a function $W : P \to \mathbb{R}$. A *δ-suitable weight assignment*, for a constant $0 < \delta < 1$, is one such that at least $\delta|P|$ of the points are assigned weight 1 and the remaining weights are real-values between -1 and 1. Given a weight assignment W to P, and a point $q \in \mathbb{R}^d$, we define the *additive error at* q, denoted by $\mathrm{Err}_P(q)$, as

$$\mathrm{Err}_P(q) = \sum_{p \in P} W(p)||q - p||. \tag{2}$$

Here, we prove the following lemma which will be the main tool used by our coreset. It enables us to build a coreset of a point set inside a given cube such that the additive error falls very rapidly as we move away from the pointset.

Lemma 2.1. *For any given parameter $c > 1$, there exists an integer t such that the following holds. For any given point set P with $|P| \geq t$, and such that P is contained inside the cube $\mathcal{Q} = [-1, 1]^d$ in $\mathbb{R}^d$, there exists a $\frac{1}{3}$-suitable weight assignment to P such that for any point q with $|q| > 3$, we have $|Err_P(q)| = O_c(|P| \, \|q\|^{-c})$.*

Proof. First, we consider a single point $p_i \in P$. The first idea is to use the standard Taylor expansion to estimate the distance between p_i and q. Observe that as $\|q\| > 3$, and $\|p_i\| \leq 1$ for all $p_i \in P$, it follows that $\|q\|^2 > \|p_i\|^2 - 2\vec{p_i} \cdot \vec{q}$. Thus, we can use the Taylor expansion below.

$$\begin{aligned}\mathrm{d}(p_i, q) &= \|p_i - q\| = \sqrt{\|q\|^2 + \|p_i\|^2 - 2\vec{p_i} \cdot \vec{q}} \\ &= \|q\| \left(1 + \sum_{j=1}^{\infty} \alpha_i \left(\frac{\|p_i\|^2 - 2\vec{p_i} \cdot \vec{q}}{\|q\|^2}\right)^j\right)\end{aligned} \tag{3}$$

where α_i's are constants that come from the Taylor expansion of the function $\sqrt{1 + x}$. Their exact value is not important but we simply note that $\alpha_i = O(2^{O(i)}) = O_i(1)$.

We split the terms in the Taylor expansion of $\mathrm{d}(p_i, q)$ into two parts; we consider those with $j = 1$ to $j = r$ and then those $j > r$, for some constant r. In other words, we write (3) $= A_i + B_i$ where

$$A_i = \|q\| \left(1 + \sum_{j=1}^{r} \alpha_i \left(\frac{\|p_i\|^2 - 2\vec{p_i} \cdot \vec{q}}{\|q\|^2}\right)^j\right)$$

and by properties of Taylor series, we have

$$\begin{aligned}B_i &= O_r \left(\|q\| \left(\frac{|\,\|p_i\|^2 - 2\vec{p_i} \cdot \vec{q}|}{\|q\|^2}\right)^r\right) \\ &= O_r \left(\|q\| \left(\frac{1 + 2\,\|q\|}{\|q\|^2}\right)^r\right) = O_r \left(\|q\|^{-r+1}\right).\end{aligned} \tag{4}$$

Let $q = (q_1, \cdots, q_d)$. Define $X_i(q) = \|p_i\|^2 - 2\vec{p_i} \cdot \vec{q}$ and observe that $X_i(q)$ is a d-variate polynomial over d indeterminates $q_1, \cdots, q_d$. Also, the coefficients of $X_i(q)$ depend on the coordinates of p_i. Fix a constant c' and define the d-variate polynomial $Q_i(q) = \sum_{j=0}^{c'} (X_i(q))^j$. Observe that $Q_i(q)$ is essentially A_i, after removing $\|q\|$ factors. Next, define the polynomial

$$Q(q) = \sum_{p_i \in P} W(p_i) Q_i(q). \tag{5}$$

Observe that Q is defined on the d indeterminates $q_1, \cdots, q_d$ and its coefficients depend on the coordinates (and multiplicities) of the points in P and more crucially, also linearly on the weight assignment. In addition, Q has degree r which implies Q has at most $F := \binom{d+r}{r}$ coefficients.

We now set $t = 3F$ and let $p_1, \cdots, p_m$ be the points of P and thus by our assumptions $m \geq t$. Observe that each coefficient of Q can be seen as a linear combination of the m weights, $W(p_1), \cdots, W(p_m)$. As mentioned, there are at most F coefficients but they might not all be linearly independent. We would like to set all the F coefficients of $Q(q)$ to zero; setting them to zero defines some F'-dimensional linear subspace, h, of $\mathbb{R}^m$, in the parameter space defined by the weights $W(p_1), \cdots, W(p_m)$, for some parameter $F' \leq F$. Next, consider the boundary Δ of the cube $[-1, 1]^m$ which is a convex polytope that contains the origin and thus Δ can be decomposed into j-dimensional faces, for $j = 0, \cdots, m - 1$. Consider the $m - F'$-dimensional faces of Δ. It thus follows that h intersects at least some of these $m - F'$-dimensional faces. Observe that every point on each such face has at least $m - F'$ coordinates equal to -1 or $+1$ and the magnitude of the other coordinates is at most one. Thus, the intersection of h and the $m - F'$-dimensional faces of $\mathcal{Q}$ will also have this property. The intersection point also gives us a weight assignement that sets all the coefficients of polynomial Q to zero; also, note that we can assume that at least $\frac{m-F'}{2} \geq \frac{m-m/3}{2} \geq \frac{m}{3}$ of the weights are equal to $+1$ since if not, we can just negate all the values. Consequently, this implies that there exists a $\frac{1}{3}$-suitable weight assignment that guarantees $Q(q) \equiv 0$. From this, it follows that $\sum_{p_i \in P} w_i A_i \equiv 0$. We now set $r = c + 1$ and the lemma follows from summing up eq 4 for all the points $p_i \in P$. $\square$

3. Building the Coreset

We begin with a “base coreset” in which we are allowed to take a large fraction of the input points as coreset points. Then, we will bootstrap the construction to build a general coreset.

3.1. A Base Coreset

We use our Lemma 2.1 in combination with a simple quad-tree construction. In this subsection, we assume we are given a pointset $P \subset \mathbb{R}^d$ containing n points and we are allowed to take $(1-\delta)n$ coreset points for some parameter $0 < \delta < 1$ that we will choose later.

A modified quad-tree. In Lemma 2.1, we pick $c = d+2$ and in the rest of this subsection, let t be the constant that is guaranteed to exist by the Lemma 2.1. We build a modified quad-tree T as follows. Every node v of T is associated with the subset of P, denoted by P_v and it is also enclosed in the minimum enclosing cube Q_v. v will be considered a leaf in three cases: if the depth of v is larger than $C\log|P|$ for a large enough constant C, then we treat v as a single point with multiplicity $|P_v|$ (i.e., we assume all the points of P_v are on top of each other; we will later show that this will only add a negligible amount of error to the final coreset). In this case we call v a *dense* leaf. Thus, assume v has depth less than $C\log|P|$. In this case, if P_v contains fewer than t points, then v is a *sparse* leaf, but if it contains between t and $C2^{d+2}t\log n$ points, then it is an *ordinary leaf*. If none of these cases hold, then v is not a leaf and thus Q_v is divided into 2^d congruent sub-cubes of half the side length and the construction continues recursively with respect to the points inside the sub-cubes.

Construction of the coreset. First, as the size of the coreset is allowed to be almost as large as the input size, we can assume that at most $\frac{n}{2}$ of the points of P are in dense leaves. If this is not the case, we are done. Specifically, after removing dense nodes, we might have $o(n)$ nodes left in the tree. But in this case, there must have been more than $n/2$ points removed from the dense nodes which means we can just take all the remaining points as the coreset. Thus, in the rest of the construction, we will assume that at least $\frac{n}{2}$ points are in sparse or ordinary leaves.

Consider a leaf v of T. If v is a sparse leaf, then all the points of P_v are added to the coreset with weight 1. We do the same thing when v is a dense leaf (note that in this case P_v has only one point but it's multiplicity is larger than one). Finally, consider the case when v is an ordinary leaf. Let us assume that there are m ordinary leaves. Sort the orindary leaves by the size of their enclosugin cubes in decreasing order. Thus, let $v_1, \cdots, v_m$ be the orindary leaves with corresponding enclusing cubes with side lengths $\ell_1, \cdots, \ell_m$ sorted in decreasing order. We add all the points P_{v_i} for $1 \le i \le m/2$ to the coreset. For any index $i > m/2$ we use Lemma 2.1; recall that Lemma 2.1 does a weight assignment where a point $p \in P_{v_i}$ is assigned a weight $W(p)$ which is between -1 and 1 and at least a fraction of the weights are guaranteed to be 1. In our coreset, we assign the weight $1 - W(p)$ to the point p. Crucially, if $W(p) = 1$, then the point is not added to the coreset. By Lemma 2.1, at most two third of the points in P_v are added to the coreset.

The next lemma shows that our construction ensures that only a fraction of the points will be in the sparse or dense leaves.

Lemma 3.1. *If at least $\frac{n}{2}$ points are in sparse or ordinary leaves, then at least $\frac{n}{4}$ points of T are in ordinary leaves.*

Proof. Consider a dense or sparse leaf v and let u be its parent. By construction, P_u contains more than $C2^{d+2}t\log n$ points and its depth is at most $C\log n$. We charge the points of P_v to the points of P_u, meaning, we distribute a total charge of $|P_v|$ among the points in P_u and thus each point of P_u receives a charge of $\frac{|P_v|}{|P_u|} < \frac{t}{C2^{d+2}t\log n}$.

We now look at how many charges a point can receive. The points of P_u can receive charges from all of its 2^d children. In addition, the points of P_u can also appear in the descendants of u to receive additional charges; however, the maximum depth of u is at most $C\log n$ and thus the average charge each point can receive is at most

$$\frac{t}{C2^{d+2}t\log n} \cdot 2^d \cdot C\log n < \frac{1}{4}. \tag{6}$$

Let n' be the number of points in sparse leaves. In our charging scheme, we are discharging n' charges from such points. As there are n points in total, and each point receives an average charge of $\frac{1}{4}$, it follows that $n' < n$. Thus, there are at least $\frac{3n}{4}$ points that are either in ordinary leaves or dense leaves but the number of latter types of points is $\frac{n}{2}$ which proves the lemma.

□

Corollary 3.2. *The size of the coreset is at most $\left(1 - \Omega_d\left(\frac{1}{\log n}\right)\right)n$.*

Proof. If more than $n/2$ points are in dense leaves, then the size of the coreset is upper bounded by $\frac{n}{2}$ plus the number of dense nodes which is at most $\frac{n}{C2^{d+2}t\log n} < \frac{n}{4}$, meaning, in this case, the lemma trivially holds. Thus, assume otherwise. By Lemma 3.1, at least $\frac{n}{4}$ points are in ordinary leaves. As the number of ordinary leaves is m and each ordinary leaf contains between t and $C2^{d+2}t\log n$ points, it follows that the ordinary leaves $v_{m/2+1}, \cdots, v_m$ have at least

$$\frac{t}{C2^{d+2}t\log n}n = \Omega_d\left(\frac{n}{\log n}\right) \tag{7}$$

points. The lemma then follows from Lemma 2.1 which guarantees the existence of $\frac{1}{3}$-suitable weight assignment; each weight that is assigned 1 is not placed in the coreset and thus Eq. 7 asymptotically bounds the number of points we are "saving", i.e., not placing in the coreset, proving the lemma. $\square$

3.1.1. Bounding the additive error

Next, we bound the additive error of our coreset. To facilitate this, we adopt the following notation. We use the notation U to refer to the multiset of the input points together with the coreset points. The input points are included with a positive weight that is equal to their multiplicity and the coreset points are included with a negative weight that is the product of their multiplicity by their weight in the coreset. Given a point $q \in \mathbb{R}^d$, the additive error at q is $\mathrm{Err}_U(q)$ and where the notion is defined by Eq. 2.

First we consider the dense leaves. Observe that if C is chosen large enough, then each dense leaf v is associated with a bounding cube which has side length at most $\ell_h 2^{-C\log n} \leq \frac{\ell_h}{n^2}$. Consider a dense leaf v. Changing the multiplicity of one point of P_v to $|P_v|$ and treating it as a single point corresonds to moving the points of P_v to lie on top of the chosen point (and thus increasing its multiplicity). By triangle inequality and since the weights assigned to the points are withint $[-1, 1]$, each movement of a point changes the additive error by at most $\frac{\ell_h}{n^2}$ and thus the total increase in the additive error is bounded by $\frac{\ell_h}{n}$ which will be asymptotically absorbed in the final additive error. In addition, observe that the points in sparse leaves also contribute nothing to the additive error as they are fully incluced in the coreset (i.e., they have been included twice in U, once with weight $+1$ and another time with weight -1).

W.l.o.g, assume that the unit cube is the smallest bounding cube of the input points P. It remains to consider the ordinary leaves. Recall that we have assumed that there are m ordinary leaves with side lengths $\ell_1, \cdots, \ell_m$ sorted in decreasing order. Let $h = \frac{m}{2}$ and consider ℓ_h. By construction, the points $\ell_1, \cdots, \ell_h$ do not contribute to the additive error. To bound the contributions of the rest of the points to additive error we will need the following simple lemmas.

Lemma 3.3. *The number of ordinary leaves v such that (i) the smallest distance between Q_v and q is at most f, for a parameter f, and (ii) the side-length of Q_v is between ℓ and 2ℓ is bounded by $O_d\left(\left(\frac{O(f+\ell)}{\ell}\right)^d\right)$.*

Proof. The proof follows from a simple packing argument and the observation that for two different ordinary leaves v and u, their minimum enclosing cubes Q_v and Q_u are disjoint; in particular, we simply need to divide the volume of a sphere of radius $O(f+\ell)$ by the volume of a sphere of radius f. $\square$

Lemma 3.4. *The total contribution of the ordinary points to the additive error is upper bounded by $O_d(\ell_h t\log n)$.*

Proof. Consider an ordinary leaf v with its bounding cube Q_v of side length ℓ and let f be the minimum distance of q to Q_v (f will be zero if q inside Q_v). To estimate the contribution of P_v to the additive error, we consider two cases. The first case is when $f \leq 3\ell$. In this case, we trivially upper bound the error by $C2^{d+2}t\log nf = O(\ell 2^{d+1}t\log n) = O_d(\ell_h t\log n)$. Also, by Lemma 3.3, the number of such cells if $O_d(1)$. The second and the more interesting case is when $f > 3\ell$. Here, we use Lemma 2.1. To use the lemma though, we need to scale by a factor ℓ, apply the lemma and then rescale. This yields the bound $O\left(\ell\cdot C2^{d+2}t\log n\left(\frac{\ell}{f}\right)^{d+2}\right)$. By Lemma 3.3, we bound the number of cells that have side-lengths between ℓ and 2ℓ and lie at distance between f and $2f$ of q. Combining these, we get that the total contribution such cells v is

$$\begin{aligned} &O_d\left(\ell t\log n\left(\frac{\ell}{f}\right)^{d+2}\cdot\left(\left(\frac{O(f+\ell)}{\ell}\right)^d\right)\right)\\ &= O_d\left(\ell t\log n\cdot\left(\frac{\ell}{f}\right)^2\right). \end{aligned} \tag{8}$$

Observe that summing the above expression over f ranging from 3ℓ and increasing powers of two will yield a geometric series and thus the sum is bounded by the largest term which is obtained by setting $f = 3\ell$. Then, we can do the same thing with respect to ℓ and observe that $\ell \leq \ell_h$ by our construction. This concludes the proof of the lemma. $\square$

Combining the above lemma with the other cases covered earlier we get that $\mathrm{Err}_U(q) = O_d(\ell_h t\log n)$.

Finally, to briefly remark on removing negative weights: We may increase all w_i's by 1 to get the 1 median with a small error: In other words, $\sum_{p_i\in P}(1+w_i)\,\mathrm{d}(p_i,q) = E + \sum_{p_i\in P}\mathrm{d}(p_i,q)$ so $1+w_i$ can be taken as the weights of the points in the coreset.

3.1.2. The multiplicative error

To be able to estimate the multiplicative error, we need to have an estimate of sum of total distances of q to the points of P. Luckily, this is relatively simple using Lemma 3.3. Observe that $t = O_d(1)$ and $C2^{d+2}t\log n = \tilde{O}_d(1)$ and thus

$h = \tilde{\Omega}_d(n)$. By Lemma 3.3, it follows that at least $\frac{h}{4}$ of the ordinary leaf nodes $v_1, \cdots, v_h$ lie at distance $\Omega_d\left(\ell_h n^{\frac{1}{d}}\right)$ away from h. Thus, sum of distances from q to the points in T is at least $\tilde{\Omega}_d\left(\ell_h n^{\frac{d+1}{d}}\right)$. Combining this with the additive distance yields that the resulting coreset has multiplicative error of $1 + \tilde{O}_d\left(n^{-\frac{d+1}{d}}\right)$.

3.2. A General Coreset

In this section, we generalize the construction from the previous subsection to be able to produce coresets with multiplicative $1+\varepsilon$ approximation for an arbitrary parameter $\varepsilon > 0$. To that, we first set $\varepsilon' = \varepsilon^2$ and then build a coreset $\mathcal{S}$ of size $\tilde{O}(\varepsilon'^{-2})$. There exist several examples in literature that do so, as mentioned in related work above. It is not important which one we use and it will be simple to modify the calculations even when using a coreset of size poly(ε^{-1}). In what follows, we assume that the initial size of the point set is $n = \tilde{O}(\varepsilon'^{-2})$.

Next, we apply the construction from the previous section repeatedly to create increasingly finer coresets. Suppose, we wish to compute a coreset of size $n/2$. We invoke the construction from the last section i times. In each iteration, we reduce the number of points by a multiplicative factor $1 - \Omega_d\left(\frac{1}{\log n}\right)$ as per Corollary 3.2. Thus, setting $i = O_d\left(\log n \cdot \log \varepsilon^{-1}\right)$, the coreset has size $\left(1 - \Omega_d\left(\frac{1}{\log n}\right)\right)^i \cdot n \le n/2$. The error after each coreset construction is $\left(1 + \tilde{O}_d\left(n^{-\frac{d+1}{d}}\right)\right)$ as per the discussion in Section 3.1.2. Thus, the accumulated error is

$$
\begin{aligned}
&\left(1 + \tilde{O}_d\left(n^{-\frac{d+1}{d}}\right)\right)^i \\
&\le \exp\left(\ln\left(1 + \tilde{O}_d\left(n^{-\frac{d+1}{d}}\right)\right) \cdot i\right) \\
&\le \exp\left(i \cdot \tilde{O}_d\left(n^{-\frac{d+1}{d}}\right)\right) \\
&\le \exp\left(\ln\left(1 + 2i \cdot \tilde{O}_d\left(n^{-\frac{d+1}{d}}\right)\right)\right) \\
&\le \left(1 + 2i \cdot \tilde{O}_d\left(n^{-\frac{d+1}{d}}\right)\right),
\end{aligned}
$$

where we used the inequalities $x/2 \le \ln(1+x) \le x$ for $x \le 1$ that immediately follow from the Mercator series. Observe that the error after this sequence of reductions is still $\left(1 + \tilde{O}_d\left(n^{-\frac{d+1}{d}}\right)\right)$, as i gets absorbed by the higher order terms.

Subsequently, repeating this series of compressions that reduce the coreset size by a factor of $1/2$ roughly $\log 1/\varepsilon$ times, each time and invoking the same calculations as above. At the end and after rearranging terms, we have an error of $\left(1 + \tilde{O}_d\left(\varepsilon\right)\right)$, with a corresponding size of $\tilde{O}\left(\varepsilon^{-d/(d+1)}\right)$. Obtaining the desired $(1+\varepsilon)$ error is now merely a matter of rescaling.

4. Lower Bound

In this section we prove an optimal lower bound for the size of coresets for 1-medians, under some assumptions; the assumptions can be weakened but we have avoided this to keep proof simple. We assume the coreset $\mathcal{S}$ is allowed to pick half of the original input points with weight 2 assigned to them. To prove the lower bound, we use the analytic techniques first employed by (Alexander, 1990) to lower bound the discrepancy of the set system defined by n points in $\mathbb{R}^d$ and the set of halfspaces.

Let us try to establish an intuitive relationship between a coreset lower bound and discrepancy theory. We would like to show that there exists a point in $\mathbb{R}^d$ that has high additive error, and in particular, at least one input will have high additive error. This will show a lower bound on the error which can be translated to a lower bound on the size. However, this view of a coreset lower bound relates naturally to the discrepancy theory. Below, we will quickly overview this and also give a very rough sketch of Alexander's proof.

4.1. An Overview of Alexander's Technique

The main result proven by Alexander is a discrepancy lower bound. Let P be a set of n points in $\mathbb{R}^d$. For a function $f : P \to \{-1, +1\}$ that assigns -1 or $+1$ to each point of P, the discrepancy of a halfspace h is defined as $\mathrm{Disc}_f(h, P) = |\sum_{p \in h \cap P} f(p)|$ and the discrepancy of f is defined as the $\mathrm{Disc}_f(P) = \max_{h \in \mathcal{H}} \mathrm{Disc}_f(h, P)$ where $\mathcal{H}$ is the set of all halfspaces. Finally, $\mathrm{Disc}(P) = \min_{f \in \mathcal{F}} \mathrm{Disc}_f(P)$ where $\mathcal{F}$ is the set of all functions that assign $+1$ and -1 to P. To show a lower bound for discrepancy, one often needs to construct a special point set P and then show that for every function $f \in \mathcal{F}$, there exists at least one halfspace with high discrepancy.

In Alexander's proof (Alexander, 1990), the starting point for proving the lower bound is a known result from measure theory that shows the existence of a translation invariant measure μ on hyperplanes. For two points $p, q \in R^d$, let $h(p, q)$ be the set of hyperplanes that intersect the line segment $\overline{pq}$. With proper normalization, we get that

$$\mu\left(h(p,q)\right) = \frac{\mathrm{d}(p,q)}{2}. \tag{9}$$

Next, the concept of $+1$ and -1 coloring are generalized to signed mass distributions; let ν be a signed mass distribution in $\mathbb{R}^d$ with total mass of 0 (i.e., the magnitude of the negative masses and the positives masses are equal); we call such a mass distribution *suitable*. Note that any function $f : P \to \{-1, +1\}$ can be turned to such a mass distribution by just adding $\mathrm{Disc}_f(\mathbb{R}^d, P)$ dummy points and assigning them -1 or $+1$ appropriately; this changes the discrepancy of any other halfspace by the same amount and thus at most doubles the discrepancy of the entire coloring.

For a hyperplane h, let h^+ and h^- be the halfspaces above and below h, respectively. For a suitable mass distribution, observe that $\nu(h^+) = -\nu(h^-)$ but also observe that if ν is obtained from a function $f \in \mathcal{F}$, then $|\nu(h^+)| = |\nu(h^-)| = \mathrm{Disc}_f(h, P)$, meaning, $\nu(h^+)\nu(h^-) = -\mathrm{Disc}_f^2(h, P)$. The foundational mathematical property that Alexander uses for his discrepancy lower bound the following observation, that follows from Eq. 9. Define $I(P) := \int\int \mathrm{d}(p, q) d\nu(p) d\nu(q)$. Then

$$I(P) = -\int \mathrm{Disc}_f^2(h, P) d\mu(h). \tag{10}$$

The above equivalence has significant implications for the method used by Alexander and in particular, it enables him to prove certain "convexity" properties for the left hand side of Eq. 10. As we shall see soon, the left hand side will enable us to prove our lower bound for the coresets.

Nonetheless, the right hand side of Eq. 10 in some sense is the "average" discrepancy of a hyperplane, however, this needs normalization since both sides of the above equality are scale dependent. To figure out that scaling factor, one needs to consider the hyperplanes that contribute to the right hand side. Observe that any hyperplane h that does not intersect the convex hull of P, contributes zero to the right hand side. Next, note that if P is a point set with diameter D, then it can be enclosed in a square Q of side length $2D$ and thus any hyperplane h that intersects P must also intersect one of the sides of Q. Each side of Q has length $2D$ and thus by Eq. 9, the total mass of the hyperplanes that intersect it is D and the total mass of hyperplanes that intersect Q is $O(d2^dD)$, as Q has $O(d2^d)$ edges. Thus,

$$-I(P) = \int \mathrm{Disc}_f^2(h, P) d\mu(h) = O(d2^d D \mathrm{Disc}_f^2(P)). \tag{11}$$

Using a number of additional properties, Alexander proves a lower bound for the left hand side. In particular, if the minimum distance between every two points in P is 1, he shows that the left hand side is lower bounded by $\Omega(n)$. Assuming d is a constant, when P is a grid of side length $n^{1/d}$, we have $D = O(n^{1/d})$ and thus one obtains a lower bound of $\mathrm{Disc}_f(P) = \Omega\left(n^{\frac{1}{2}-\frac{1}{2d}}\right)$.

4.2. A Lower Bound for Coresets

Consider a point set P of n points and consider a 1-median coreset $\mathcal{S}$ of size $\frac{n}{2}$. As discussed, we assume the coreset $\mathcal{S}$ is allowed to pick half of the original input points with weight 2 assigned to them. We would like to get a lower bound for the error assuming such a coreset. We can turn $\mathcal{S}$ and P into a mass distribution f. Each point $p_i \in P$ is assigned the weight $+1$ and each point $q_i \in \mathcal{S}$ is assigned the weight -1. This gives us a weight assignment. For a point $q \in P$, define

$$\mathrm{MedianDisc}_{\mathcal{S}}(q, P) = |\mathrm{Err}_P(q)| = |\sum_{p \in P} W(p)\, \mathrm{d}(q, p)|.$$

The crucial observation that ties our problem to the techniques developed by Alexander is the following:

$$-\sum_{q \in P} \mathrm{Err}_P(q) W(q) = -I(P). \tag{12}$$

Using the convexity properties of Eq. 10 in combination with operations such as convolution, Alexander proves the following main theorem in $\mathbb{R}^d$; here, we have slightly altered the presentation to ignore constant values that depend exponentially on the dimension.

Theorem 4.1. *(Alexander, 1990) Let ν be an atomic measure of total mass 0 concentrated on a pointset $X \subset \mathbb{R}^d$. If the minimum distance between any two points of X is at least one, then*

$$\sum_{p,q \in X} \nu(p)\nu(q) = \Omega_d\left(\sum_{p \in X} (\nu(p))^2\right).$$

We consider a point set P obtained via grid of side length $n^{1/d}$, composed of n grid cells of side length one. Under our weight assignment, and using Theorem 4.1 it follows that $\sum_{p,q \in X} \nu(p)\nu(q) = \Omega_d(n)$. By Eq. 12, we get that $\sum_{q \in P} \mathrm{Err}_P(q) W(q) = \Omega_d(n)$. Observe that for each point $q \in P$, we have $|W(q)| = 1$. Let $p_1, \cdots, p_n$ be the points in P. Thus, we get that the sum of n values, $\mathrm{Err}_P(p_1)W(p_1) + \cdots + \mathrm{Err}_P(p_n)W(p_n) = \Omega_d(n)$. As there are n terms in the sum, it thus follows that the absolute value of one of them must be $\Omega_d(1)$ (in fact, we get that this holds on average). This gives a lower bound on the additive error of the coreset. To calculate the multiplicative error, observe that the total sum of distances from any point $q \in P$ in our grid construction is lower bounded by $\sum_{p \in P} \mathrm{d}(q, p) = O(n^{1+1/d})$. This consequently implies that the coreset $\mathcal{S}$ has multiplicative error of $1 + \Omega_d(n^{-1-1/d})$.

To remove the assumption that the weights sum up to n, consider the case the weights sum up to some value n'. If $|I(P)| \in \Omega(n)$, the proof still goes through, so we assume that $I(P) \leq \eta \cdot n$ for some sufficiently small constant η. In this case, we add another point D with weight equal to $n - n'$ inside the grid, with the requirement that it is well separated from the remaining points, yielding a point set

$P' = P \cup \{D\}$. By the Theorem 4.1 $I(P') \in \Omega(n)$ which implies $W(D) \cdot \mathrm{Err}_P(D_n) \in \Omega(n)$, which implies that D $\Omega_d(1)$ additive error. The remaining steps are the same.

5. Conclusion and Open Problems

We presented a coreset construction for the geometric median problem consisting of $\tilde{O}_d(\varepsilon^{-d/(d+1)})$ points. This bound is sublinear in ε^{-1} which was previously only known for the line metric.

The dependency on d is substantial in our construction for high dimensions. We did not attempt to optimize with respect to this dependency, but using the quad-tree based approach, we will always incur a factor 2^d in the coreset size. It is not clear whether this is avoidable, as for some (large) value of d, a coreset using input points is required to consist of ε^{-2} many points. Obtaining a best of both worlds dependency between the previous high dimensional constructions of (Cohen-Addad et al., 2021; Huang et al., 2023) is an interesting open problem.

Another open problem is to remove the lower bound requirement that points must be selected from the input. All coreset construction we are aware of for the geometric median have this property or can be modified to have this property. While we presented the results for the geometric median, most of the results also work for powers of distance and in particular for squared distances. For squared distances, there exist a 0-coreset consisting of $O(1)$ points, if the coreset points are not required to be part of the input. Showing that one can improve our upper bounds using non-input points, or that this is not possible is interesting and likely challenging.

Acknowledgements

The authors thank the reviewers for helpful comments. Peyman Afshani is supported by a DFF (Det Frie Forskningsråd) project 1, grant number: 10.46540/3103-00334B. Chris Schwiegelshohn is partially supported by the Independent Research Fund Denmark (DFF) under a Sapere Aude Research Leader grant No 1051-00106B.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agarwal, P. K., Har-Peled, S., and Varadarajan, K. R. Geometric approximation via coresets. In *Combinatorial and computational geometry, MSRI*, pp. 1–30. University Press, 2005.

Alexander, R. Geometric methods in the study of irregularities of distribution. *Combinatorica*, 10:115–136, 1990.

Badoiu, M. and Clarkson, K. L. Optimal core-sets for balls. *Comput. Geom.*, 40(1):14–22, 2008. doi: 10.1016/j.comgeo.2007.04.002. URL http://dx.doi.org/10.1016/j.comgeo.2007.04.002.

Baker, D. N., Braverman, V., Huang, L., Jiang, S. H., Krauthgamer, R., and Wu, X. Coresets for clustering in graphs of bounded treewidth. In *International Conference on Machine Learning, ICML*, pp. 569–579, 2020. URL http://proceedings.mlr.press/v119/baker20a.html.

Becchetti, L., Bury, M., Cohen-Addad, V., Grandoni, F., and Schwiegelshohn, C. Oblivious dimension reduction for k-means: beyond subspaces and the johnson-lindenstrauss lemma. In *Symposium on Theory of Computing, STOC*, pp. 1039–1050, 2019. URL https://doi.org/10.1145/3313276.3316318.

Bentley, J. L. and Saxe, J. B. Decomposable searching problems i: Static-to-dynamic transformation. *J. Algorithms*, 1(4):301–358, 1980.

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 *Symposium on Foundations of Computer Science, FOCS*, pp. 462–473, 2022. doi: 10.1109/FOCS54457.2022.00051. URL https://doi.org/10.1109/FOCS54457.2022.00051.

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.

Cohen, M. B., Lee, Y. T., Miller, G. L., Pachocki, J., and Sidford, A. Geometric median in nearly linear time. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC*, 2016.

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

Cohen-Addad, V., Larsen, K. G., Saulpic, D., and Schwiegelshohn, C. Towards optimal lower bounds for k-median and k-means coresets. In *Symposium on Theory of Computing, STOC*, pp. 1038–1051, 2022a. doi: 10.1145/3519935.3519946. URL https://doi.org/10.1145/3519935.3519946.

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

Feldman, D. and Langberg, M. A unified framework for approximating and clustering data. In *Symposium on Theory of Computing, STOC*, pp. 569–578, 2011.

Feldman, D., Schmidt, M., and Sohler, C. Turning big data into tiny data: Constant-size coresets for *k*-means, PCA and projective clustering. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013, New Orleans, Louisiana, USA, January 6-8, 2013*, pp. 1434–1453, 2013.

Feng, Z., Kacham, P., and Woodruff, D. P. Dimensionality reduction for the sum-of-distances metric. In *International Conference on Machine Learning, ICML*, pp. 3220–3229, 2021. URL http://proceedings.mlr.press/v139/feng21a.html.

Har-Peled, S. and Kushal, A. Smaller coresets for k-median and k-means clustering. *Discrete & Computational Geometry*, 37(1):3–19, 2007.

Har-Peled, S. and Mazumdar, S. On coresets for k-means and k-median clustering. In *Symposium on Theory of Computing, STOC*, pp. 291–300, 2004.

Huang, L. and Vishnoi, N. K. Coresets for clustering in euclidean spaces: importance sampling is nearly optimal. In *Symposium on Theory of Computing, STOC 2020*, 2020. doi: 10.1145/3357713.3384296. URL https://doi.org/10.1145/3357713.3384296.

Huang, L., Li, J., and Wu, X. Towards optimal coreset construction for (k, z)-clustering: Breaking the quadratic dependency on k. *CoRR*, abs/2211.11923, 2022. doi: 10.48550/arXiv.2211.11923. URL https://doi.org/10.48550/arXiv.2211.11923.

Huang, L., Huang, R., Huang, Z., and Wu, X. On coresets for clustering in small dimensional euclidean spaces. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13891–13915. PMLR, 2023. URL https://proceedings.mlr.press/v202/huang23h.html.

Langberg, M. and Schulman, L. J. Universal ε-approximators for integrals. In *Symposium on Discrete Algorithms, SODA*, pp. 598–607, 2010. doi: 10.1137/1.9781611973075.50. URL http://dx.doi.org/10.1137/1.9781611973075.50.

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.

Sohler, C. and Woodruff, D. P. Strong coresets for k-median and subspace approximation: Goodbye dimension. In *Symposium on Foundations of Computer Science, FOCS*, pp. 802–813, 2018. doi: 10.1109/FOCS.2018.00081. URL https://doi.org/10.1109/FOCS.2018.00081.