Approximation and Complexity of the Capacitated Geometric Median Problem

Vladimir Shenmaier(✉)

Sobolev Institute of Mathematics, Novosibirsk, Russia

Abstract. In the Capacitated Geometric Median problem, we are given n points in d-dimensional real space and an integer m, the goal is to locate a new point in space (center) and choose m of the input points to minimize the sum of Euclidean distances from the center to the chosen points. We show that this problem admits an "almost exact" polynomial-time algorithm in the case of fixed d and an approximation scheme PTAS in high dimensions. On the other hand, we prove that, if the dimension of space is not fixed, Capacitated Geometric Median is strongly NP-hard and does not admit a scheme FPTAS unless P = NP.

Keywords: Facility location · Geometric median · Outliers · Euclidean distances · Approximation scheme · NP-hardness

1 Introduction

We study the question of the polynomial-time solvability and approximability of the Capacitated Geometric Median problem, which is formulated as follows.

Capacitated Geometric Median. Let X be a set of n points in space $\mathbb{R}^d$, m be a positive integer, and $\|.\|$ denote the Euclidean norm. Find a center $c \in \mathbb{R}^d$ and an m-element subset $S \subseteq X$ to minimize the value of

$$cost(S, c) = \sum_{x \in S} \|x - c\|.$$

This problem may also be referred to as *Geometric Median with outliers*. In fact, it consists of finding a center $c \in \mathbb{R}^d$ minimizing the total distance from c to m nearest input points.

In an equivalent version, we need to find a center $c \in \mathbb{R}^d$ and a subset $S \subseteq X$ of the maximun cardinality for which the value of $cost(S, c)$ does not exceed a given upper bound. It is easy to see that this "inverse" version is reduced to a series of instances of the original Capacitated Geometric Median problem, with

The study was carried out within the framework of the state contract of the Sobolev Institute of Mathematics (project 0314-2019-0014).

R. Santhanam and D. Musatov (Eds.): CSR 2021, LNCS 12730, pp. 422–434, 2021.
https://doi.org/10.1007/978-3-030-79416-3_26

different values of m. On the other hand, the original problem is reduced to a series of instances of the inverse version, with different cost bounds.

As in the usual Geometric Median problem, where $m = n$, the input points represent clients or demand points, whereas the desired center represents a location for placing a facility to serve the clients. The $n-m$ clients which are removed from this service in the solution are called *outliers*. The problem with outliers arises naturally in the following situations.

- The facility to be placed at the center we are looking for has a limited capacity and may not serve all the demand points.
- There is an upper limit for the transportation cost, so we need to remove a minimum possible number of clients from the service to satisfy this limit.
- The data contains noise and errors. In this case, a few most distant clients may exert a disproportionately strong influence over the final solution and correspond to the least robust input points.
- The discovered outliers do not fit the rest of the data and they are worthy of further investigation. In particular, once identified, they can be used to discover anomalies in the data.

Related Work. Besides the practical considerations, the problem we study is theoretically interesting. Strictly speaking, no polynomial-time algorithms are known even for the usual Geometric Median problem, where we find the best center for the whole set X, i.e., the *geometric median* of X. Finding this center is complicated by the fact that, even for 5-element sets, the geometric median is not expressible by radicals over the rationals [4]. However, one can say that the usual Geometric Median problem is polynomially solvable "almost exactly" since, e.g., the randomized algorithm from [10] computes a $(1+\varepsilon)$-approximate solution of this problem in time $\mathcal{O}\big(dn \log^3(n/\varepsilon)\big)$. Moreover, by using constructions based on random sampling, a $(1+\varepsilon)$-approximate geometric median can be found in time almost or completely independent of n [5,10,17].

In the case of arbitrary $m \le n$, the Capacitated Geometric Median problem becomes much harder due to the exponential number of m-subsets $S \subseteq X$. Applying random sampling to this problem seems to be effective only if m is sufficiently large. In the general case, when m may be arbitrarily small, any bounded number of random samples may "miss" good m-subsets.

ElGindy and Keil [12] consider the mentioned above version of the problem where it is required to find a maximum-cardinality subset satisfying a given upper bound for the cost value. They suggest an $\mathcal{O}(n^{2.5} \log^4 n)$-time exact algorithm for the 2-dimensional case with rectilinear distances.

Two well-known single location problems closest to Capacitated Geometric Median are Smallest m-Enclosing Ball and m-Variance. The first consists of finding m input points minimizing the radius of the Euclidean ball enclosing these points. In the second, we find m input points minimizing the sum of squared distances from these points to their mean. In high dimensions, both problems are strongly NP-hard [15,21,22] but admit approximation schemes PTAS with running time $\mathcal{O}\big(dn^{\lceil 1/\varepsilon \rceil}\big)$ [1] and $\mathcal{O}\big(dn^{\lceil 2/\varepsilon \rceil + 1}\big)$ [20], respectively.

Another close problem is Geometric 2-Median, which consists of finding two centers in space $\mathbb{R}^d$ minimizing the total Euclidean distance from the input points to nearest centers. In high dimensions, this problem admits fast approximation schemes based on random sampling and coresets [5,6,9,17]. However, the computational complexity of Geometric 2-Median is an open question.

A natural generalization of Capacitated Geometric Median is the problem of finding k disjoint clusters of total cardinality m in an n-element input set and selecting centers of these clusters to minimize the total distance from the cluster centers to cluster elements. The discrete version of this problem, in which all the centers must be selected from a given finite set, and also other similar problems are considered in [7,8,11,16].

Our Contributions. Both algorithmic and complexity results for Capacitated Geometric Median are given. First, we describe a randomized algorithm which, with constant probability, computes a $(1+\varepsilon)$-approximate solution of the problem in time $\mathcal{O}\big(dn^{\lfloor (d+1)/2 \rfloor} m^{\lceil (d+3)/2 \rceil} \log^3(m/\varepsilon)\big)$. Thus, in the case of fixed d, the problem is solvable "almost exactly" in polynomial time. Next, we show that, in high dimensions, Capacitated Geometric Median admits an approximation scheme PTAS with running time $\mathcal{O}\big(dn^{\lceil \log(2/\varepsilon)/\varepsilon \rceil + 1}\big)$.

On the other hand, we prove that, if the dimension of space is not fixed, Capacitated Geometric Median is strongly NP-hard and does not admit an approximation scheme FPTAS unless P = NP. In fact, it is proved for the special case when $m = n/2$. The proof is done by a reduction from the Maximum Bisection problem in a 3-regular graph.

2 Algorithms

In this section, we present two polynomial-time algorithms for finding close-to-optimal solutions of the Capacitated Geometric Median problem, in fixed and in high dimensions.

2.1 Algorithm for Fixed Dimensions

The first algorithm is based on the property that an optimal subset consists of m input points nearest to some point in space. It allows to solve the problem by enumerating the cells of the m-order Voronoi diagram of the input set and finding geometric medians of the subsets defining these cells. A similar idea is used in [2,23] for solving a number of related vector-subset problems.

Definition 1. *Given an n-element set $X \subset \mathbb{R}^d$ and a non-empty subset $S \subset X$, the* Voronoi cell *of S is the set*

$$V(S,X) = \big\{ z \in \mathbb{R}^d \,\big|\, \|z-x\| < \|z-y\| \text{ for all } x \in S,\, y \in X \setminus S \big\}.$$

*Given an integer $m \in \{1,\dots,n-1\}$, the m-*order Voronoi diagram *of X is the collection $V_m(X)$ of all the non-empty Voronoi cells $V(S,X)$, where $S \subset X$ and $|S| = m$, labeled by S.*

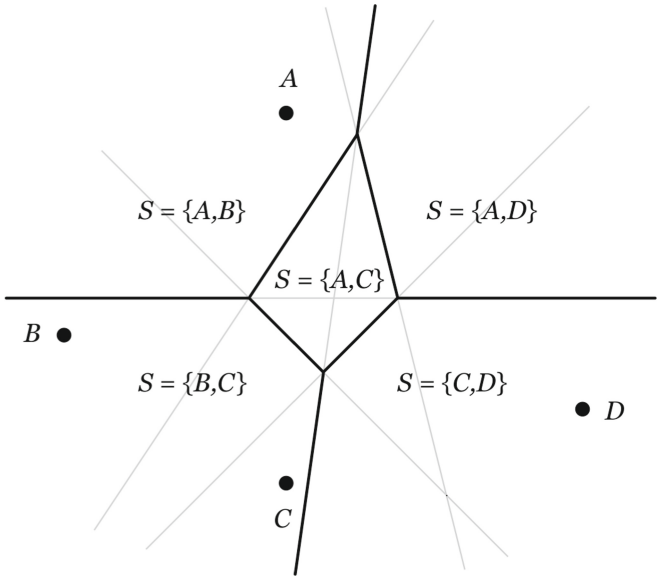

Fig. 1. The m-order Voronoi diagram of $X = \{A, B, C, D\}$ for $m = 2$

Every Voronoi cell $V(S, X)$ consists of the points of $\mathbb{R}^d$ for which the distances to the elements of S are less than those to the other elements of X. This means that $V(S, X)$ is the polytope formed by the intersection of all the open half-spaces $\left\{z \in \mathbb{R}^d \,\middle|\, \|z - x\| < \|z - y\|\right\}$, $x \in S$, $y \in X \setminus S$ (see Fig. 1).

It is easy to see that the closure $\overline{V(S, X)}$ of any non-empty cell $V(S, X)$ consists of the points $z \in \mathbb{R}^d$ satisfying the inequalities $\|z - x\| \leq \|z - y\|$ for all $x \in S$, $y \in X \setminus S$. It immediately implies the following observation.

Fact 1. *If $p \in \overline{V(S, X)}$, where $S \subset X$ and $|S| = m$, then the set S consists of m points of X closest to p.*

Given a point $p \in \mathbb{R}^d$, let $S_m(p)$ be a set of m points of X closest to p (in the case of ambiguity, we choose any of such sets). Obviously, if the distances from p to different points of X are not equal, then $S_m(p)$ is uniquely defined and $p \in V(S_m(p), X)$. It follows that the cells of $V_m(X)$ cover at least all the points of $\mathbb{R}^d$ lying outside the hyperplanes $\left\{z \in \mathbb{R}^d \,\middle|\, \|z - x\| = \|z - y\|\right\}$, $x, y \in X$, $x \neq y$. Hence, the closures of these cells cover the whole space $\mathbb{R}^d$.

Fact 2. [19] *For any n-element set $X \subset \mathbb{R}^d$, $d \geq 3$, and $m \in \{1, \ldots, n-1\}$, the diagram $V_m(X)$ consists of $s = \mathcal{O}\big(n^{\lfloor (d+1)/2 \rfloor} m^{\lceil (d+1)/2 \rceil}\big)$ cells and can be constructed in time $\mathcal{O}(s \log n + n^2 m^d)$.*

Fact 3. [18] *For any n-element set $X \subset \mathbb{R}^2$ and $m \in \{1, \ldots, n-1\}$, the diagram $V_m(X)$ consists of $\mathcal{O}(mn)$ cells and can be constructed in time $\mathcal{O}(m^2 n \log n)$.*

Based on the above facts, we suggest the following algorithm, which computes an approximate solution of Capacitated Geometric Median.

Algorithm $\mathcal{A}_1$.
Input: a set X of n points in $\mathbb{R}^d$; a parameter $\varepsilon \in (0,1)$.
Step 0. If $d = 1$, return a point $x \in X$ and the set $S_m(x)$ with the minimum value of $cost(S_m(x), x)$. If $m = n$, apply the algorithm from [10] to the whole set X and return the resulting $(1+\varepsilon)$-approximate geometric median of X.
Step 1. By using the algorithms from [18,19], construct the diagram $V_m(X)$.
Step 2. For each cell $C \in V_m(X)$, denote by $S(C)$ the subset of X labeling C, i.e., such that $C = V(S(C), X)$. By using the algorithm from [10], find a point $p(C)$ which is a $(1+\varepsilon)$-approximate geometric median of $S(C)$.
Step 3. Output a point $p(C)$ and the set $S(C)$, $C \in V_m(X)$, with the minimum value of $cost(S(C), p(C))$.

Theorem 1. *For any $\varepsilon \in (0,1)$, with constant probability, Algorithm $\mathcal{A}_1$ computes a $(1+\varepsilon)$-approximate solution of Capacitated Geometric Median in time $\mathcal{O}\big(dn^{\lfloor (d+1)/2 \rfloor} m^{\lceil (d+3)/2 \rceil} \log^3(m/\varepsilon)\big)$.*

Proof. If $d = 1$, the statement easily follows from the obvious fact that, for any set of points on the real line, one of the points of this set is its geometric median. If $m = n$, the statement is a direct corollary of the result of [10]. Next, let a point $c^* \in \mathbb{R}^d$ and a subset $S^* \subset X$ be an optimal solution of the problem in the case when $d \geq 2$ and $m \leq n-1$.

Since the closures of cells of $V_m(X)$ cover the whole space $\mathbb{R}^d$, there exists a cell $C \in V_m(X)$ whose closure contains c^*. Then, by Fact 1, the set $S(C)$ consists of m points of X closest to c^*. So

$$cost(S^*, c^*) \geq cost(S(C), c^*) \geq cost(S(C), \mu(S(C))),$$

where $\mu(S(C))$ is the geometric median of $S(C)$. Therefore, the point $\mu(S(C))$ and the set $S(C)$ are also an optimum solution of the problem. But the set $S(C)$ is computed at Step 2 of Algorithm $\mathcal{A}_1$. Hence, the objective function value on the output of this algorithm is at most

$$cost(S(C), p(C)) \leq (1+\varepsilon)\, cost(S(C), \mu(S(C))) = (1+\varepsilon)\, cost(S^*, c^*).$$

The time complexity of Algorithm $\mathcal{A}_1$ follows from Facts 2, 3, and the result of [10]. The probability of success is defined by that of the algorithm from [10]. The theorem is proved. □

2.2 Algorithm for High Dimensions

If the dimension of space is not fixed, a more productive idea for finding approximate solutions of Capacitated Geometric Median is based on using the framework from [24,25], which allows to compute a polynomial-cardinality set of points containing approximations of every point of space with respect to the distances to all n input points.

Definition 2. *Given a finite set $X \subset \mathbb{R}^d$ and $\varepsilon > 0$, a* $(1+\varepsilon)$-approximate centers collection *or, shortly, a* $(1+\varepsilon)$-collection *for X is a set $K \subseteq \mathbb{R}^d$ such that, for every point $p \in \mathbb{R}^d$, there is a point $p' \in K$ for which the distances from p' to all the elements of X are at most $1+\varepsilon$ of those from p.*

Fact 4. [25] *For any n-element set $X \subset \mathbb{R}^d$ and each fixed $\varepsilon \in (0,1]$, there exists a $(1+\varepsilon)$-collection for X which consists of $N(n,\varepsilon) = \mathcal{O}\big(n^{\lceil \log(2/\varepsilon)/\varepsilon \rceil}\big)$ elements and can be constructed in time $\mathcal{O}\big(dN(n,\varepsilon)\big)$.*

Note that the cardinality of the $(1+\varepsilon)$-collection mentioned in Fact 4 does not depend on d, which is useful when we consider the case of high dimensions. This result gives a universal approximation-preserving reduction of geometric center-based problems with continuity-type objective functions to their discrete versions, where the desired centers are selected from a polynomial-cardinality set of points (see [24,25] for details). In the case of Capacitated Geometric Median, this reduction leads to the following approximation algorithm.

Algorithm $\mathcal{A}_2$.

Input: a set X of n points in $\mathbb{R}^d$; a parameter $\varepsilon \in (0,1]$.

Step 1. By using the algorithm from [25], construct a $(1+\varepsilon)$-collection K for X.

Step 2. Output a point $c \in K$ and the set $S_m(c)$ with the minimum value of $cost(S_m(c), c)$.

Theorem 2. *For any fixed $\varepsilon \in (0,1]$, Algorithm $\mathcal{A}_2$ finds a $(1+\varepsilon)$-approximate solution of Capacitated Geometric Median in time $\mathcal{O}\big(dn^{\lceil \log(2/\varepsilon)/\varepsilon \rceil + 1}\big)$.*

Proof. Let a point $c^* \in \mathbb{R}^d$ and a subset $S^* \subseteq X$ be an optimal solution of the problem. By the definition of a $(1+\varepsilon)$-collection, the set K contains a point c such that $\|c - x\| \le (1+\varepsilon)\,\|c^* - x\|$ for all $x \in X$. It follows the inequality $cost(S^*, c) \le (1+\varepsilon)\, cost(S^*, c^*)$. On the other hand, by the construction of the set $S_m(c)$, we have $cost(S_m(c), c) \le cost(S^*, c)$. Hence, the objective function value on the output of Algorithm $\mathcal{A}_2$ is at most $(1+\varepsilon)\, cost(S^*, c^*)$.

It remains to estimate the time complexity of this algorithm. By Fact 4, the set K consists of $N(n,\varepsilon) = \mathcal{O}\big(n^{\lceil \log(2/\varepsilon)/\varepsilon \rceil}\big)$ elements and can be constructed in time $\mathcal{O}\big(dN(n,\varepsilon)\big)$. Each set $S_m(.)$ can be computed in linear time by using the known algorithm for finding an mth smallest value in an array [3]. It follows that Algorithm $\mathcal{A}_2$ runs in time $\mathcal{O}\big(dnN(n,\varepsilon)\big)$. The theorem is proved. □

Remark 1. The above technique gives also an approximation scheme PTAS for the more general problem in which we need to find a center $c \in \mathbb{R}^d$ and an m-element subset $S \subseteq X$ minimizing the value of

$$cost_{w,\alpha}(S, c) = \sum_{x \in S} w(x) \|x - c\|^{\alpha(x)},$$

where $w(.)$ are any non-negative weights and $\alpha(.)$ are any non-negative degrees bounded by arbitrary positive constant α. Indeed, it is easy to show that a $(1+\varepsilon)^\alpha$-approximate solution of this problem can be computed in time $\mathcal{O}\big(dn^{\lceil \log(2/\varepsilon)/\varepsilon \rceil + 1}\big)$ by the version of Algorithm $\mathcal{A}_2$ which outputs a point $c \in K$ and the set $S_m(c)$ with the minimum value of $cost_{w,\alpha}(S_m(c), c)$.

3 Complexity

In this section, we prove that the Capacitated Geometric Median problem is strongly NP-hard and does not admit an approximation scheme FPTAS unless P = NP. The proof is done by a reduction from the well-known APX-hard problem of finding a maximum bisection in a 3-regular graph [13].

Max-Bisection|3. Given a 3-regular n-vertex undirected graph $G = (V, E)$, where n is even. Find a partition of the set of its vertices into two equal-size subsets S and $V \setminus S$ to maximize the number $cut(S, V \setminus S)$ of the edges with one endpoint in S and the other in $V \setminus S$.

Let us define the instance of Capacitated Geometric Median corresponding to an instance of Max-Bisection|3. Suppose that G is any 3-regular undirected graph with a set of vertices V and a set of edges E, an input of Max-Bisection|3. Fix arbitrary orientation on the edges of this graph, i.e., for every edge $e \in E$, choose an endpoint of this edge which it is "outgoing from" and one which it is "incoming to". Next, we map each vertex $v \in V$ to the point $x_v \in \mathbb{R}^{E \cup V}$ with the following coordinates: $x_v(e) = 1$ for every edge $e \in E$ outgoing from v and $x_v(e) = -1$ for incoming ones; $x_v(v) = M$, where M is some large integer which will be specified later; all the other coordinates are zero (see Fig. 2). Define the instance of Capacitated Geometric Median corresponding to the graph G as the set $X = \{x_v | v \in V\}$ and the value $m = n/2$.

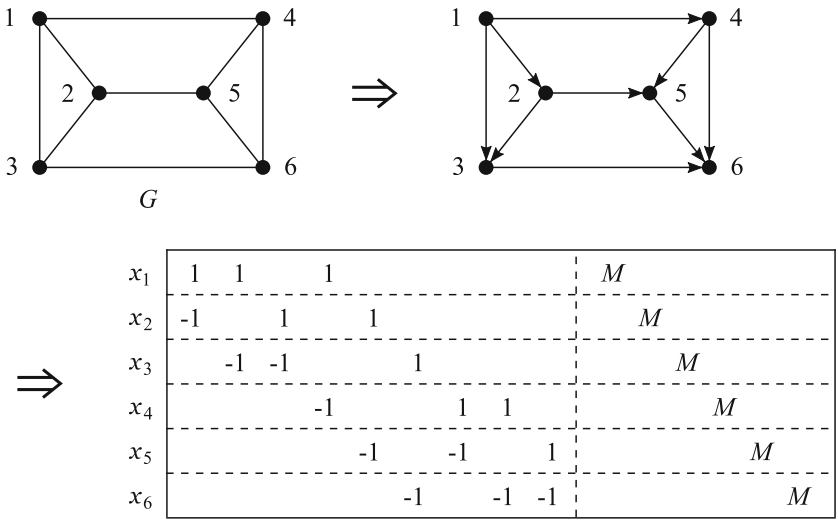

Fig. 2. Reduction scheme: constructing the vectors x_v, $v \in V$

Idea of the Reduction. Setting the coordinates $x_v(v)$, $v \in V$, to a large value M ensures that the geometric median of any subset $Y \subseteq X$ becomes very close

to its mean (Lemma 1) and the total distance from the mean to the elements of Y has an almost affine dependence on the sum of pairwise squared distances between these elements (see the proof of Lemma 2). At the same time, each pairwise squared distance $\|x_v - x_u\|^2$ equals $2M^2 + 4 + (1+1)^2 = 2M^2 + 8$ for adjacent vertices v, u and $2M^2 + 6$ for non-adjacent ones. Based on these observations, we will prove that, given an m-element subset of vertices $S \subseteq V$, the value of $cost(X_S, \mu(X_S))$, where $X_S = \{x_v | v \in S\}$ and $\mu(X_S)$ is the geometric median of X_S, monotonously depends on the number of inner edges in S and, therefore, on the value of $cut(S, V \setminus S)$ (Lemma 3). So, if the set X_S is an optimal solution of the Capacitated Geometric Median problem on the set X, then the set S is an optimal bisection in the graph G.

For any finite set $Y \subset \mathbb{R}^{E \cup V}$, denote by $c(Y)$ and $\mu(Y)$ the mean and the geometric median of this set, respectively:

$$c(Y) = \frac{1}{|Y|} \sum_{x \in Y} x \quad \text{and} \quad \mu(Y) = \arg\min_{c \in \mathbb{R}^{E \cup V}} cost(Y, c).$$

Consider any subset of vertices $S \subseteq V$ with arbitrary cardinality $m \geq 3$. Let ℓ_S be the number of inner edges in S. For every vertex $v \in S$, define the vector $z_v = x_v - c(X_S)$ and the value $\zeta_v = \sum_{e \in E} z_v^2(e)$. Then $\sum_{v \in S} z_v$ is the zero vector and the following estimates hold.

Property 1. *For every vertex $v \in S$, we have* (a) $\zeta_v < 3.38$;
(b) $\|z_v\| = \sqrt{A^2 + \zeta_v} < A + \dfrac{1.69}{A}$, *where* $A = M\sqrt{1 - \dfrac{1}{m}}$.

Proof. (a) Given a vector $x \in \mathbb{R}^{E \cup V}$, let $E(x) = \{e \in E \,|\, x(e) \neq 0\}$. Then it is easy to see that the set $E(c(X_S))$ consists exactly of the edges connecting S with $V \setminus S$, while the set $E(z_v)$ consists of all the elements of $E(c(X_S))$ and also the edges connecting v with the other vertices of S. Hence, by the 3-regularity of the graph G, we have $|E(c(X_S))| = 3m - 2\ell_S$ and $|E(z_v)| = 3m - 2\ell_S + \Delta_S^v$, where Δ_S^v is the degree of v in S. For Δ_S^v coordinates $e \in E(z_v)$, the values of $z_v(e)$ are ± 1; for $3 - \Delta_S^v$ coordinates, these values are $\pm(1 - 1/m)$; for the other coordinates from $E(z_v)$, these values are $\pm 1/m$. Then

$$\zeta_v = \Delta_S^v + (3 - \Delta_S^v)\Big(1 - \frac{1}{m}\Big)^2 + \frac{3(m-1) - 2\ell_S + \Delta_S^v}{m^2}$$

$$= 3\Big(1 - \frac{1}{m}\Big)^2 + \frac{2\Delta_S^v}{m} - \frac{\Delta_S^v}{m^2} + \frac{3}{m} - \frac{3}{m^2} - \frac{2\ell_S}{m^2} + \frac{\Delta_S^v}{m^2} = 3 - \frac{3}{m} - \frac{2\ell_S}{m^2} + \frac{2\Delta_S^v}{m}.$$

But $\ell_S \geq \Delta_S^v$ and $\Delta_S^v \leq 3$, so $\zeta_v \leq 3 - \dfrac{3}{m} - \dfrac{2\Delta_S^v}{m^2} + \dfrac{2\Delta_S^v}{m} \leq 3 + \dfrac{3}{m} - \dfrac{6}{m^2}$. The latter is maximized when $m = 4$, therefore, we have $\zeta_v \leq 27/8 < 3.38$.

(b) It is easy to see that $\|z_v\|^2 = M^2\Big(1 - \dfrac{1}{m}\Big)^2 + (m-1)\Big(\dfrac{M}{m}\Big)^2 + \zeta_v = A^2 + \zeta_v$, so $\|z_v\| = \sqrt{A^2 + \zeta_v} < A + \dfrac{\zeta_v}{2A}$. By (a), the latter is less than $A + \dfrac{1.69}{A}$. The property is proved. □

Next, we formulate the main geometric statement underlying the proposed reduction. It claims that the distance between $c(X_S)$ and $\mu(X_S)$ is close to zero for large M.

Lemma 1. *Suppose that* $A \ge 200m$. *Then* $\|\mu(X_S) - c(X_S)\| < \dfrac{40}{mA}$.

The proof of Lemma 1 is omitted in this preliminary version. Based on this lemma, we estimate the value of $cost(X_S, \mu(X_S))$.

Lemma 2. *Suppose that* $A \ge 200m$. *Then, for some* $\gamma \in [-1, 1]$, *we have*

$$cost(X_S, \mu(X_S)) = mA + \frac{3(m-1)}{2A} + \frac{\ell_S}{mA} + \gamma\frac{121m}{A^3}.$$

Proof. Let $y = \mu(X_S) - c(X_S)$ and $\delta = \|y\|$. Then, by the cosine theorem and Property 1, the value of $cost(X_S, \mu(X_S))$ equals

$$\sum_{v\in S} \sqrt{\|z_v\|^2 + \|y\|^2 - 2\langle y, z_v\rangle} = \sum_{v\in S} \sqrt{A^2 + \zeta_v + \delta^2 - 2\langle y, z_v\rangle},$$

where $\langle .\,, .\rangle$ is the dot product. Next, Property 1, Lemma 1, and the condition for A imply that

$$|\zeta_v + \delta^2 - 2\langle y, z_v\rangle| < 3.38 + \Big(\frac{40}{mA}\Big)^2 + 2\frac{40}{mA}\Big(A + \frac{1.69}{A}\Big) < 31$$

and $\Big|\dfrac{\zeta_v + \delta^2 - 2\langle y, z_v\rangle}{A^2}\Big| < \dfrac{31}{A^2} < 0.001$. On the other hand, by using Taylor's theorem (in the Lagrange remainder form), we obtain the equation

$$\sqrt{1+\varepsilon} = 1 + \frac{\varepsilon}{2} - \theta\frac{\varepsilon^2}{7.99} \text{ for some } \theta \in [0, 1] \text{ if } |\varepsilon| \le 0.001.$$

Therefore, we have

$$cost(X_S, \mu(X_S)) = \sum_{v\in S} A\Big(1 + \frac{\zeta_v + \delta^2 - 2\langle y, z_v\rangle}{2A^2} - \theta_v\frac{(\zeta_v + \delta^2 - 2\langle y, z_v\rangle)^2}{7.99A^4}\Big),$$

where $\theta_v \in [0, 1]$. But $\sum\limits_{v\in S} z_v$ is the zero vector, so the sum of terms $\langle y, z_v\rangle$ is zero. Taking into account the inequalities $\delta < \dfrac{40}{mA}$ and $|\zeta_v + \delta^2 - 2\langle y, z_v\rangle| < 31$, it follows that $cost(X_S, \mu(X_S)) =$

$$\sum_{v\in S}\Big(A + \frac{\zeta_v}{2A}\Big) + \theta_1\frac{40^2}{2mA^3} - \theta_2\frac{31^2 m}{7.99A^3} = mA + \sum_{v\in S}\frac{\zeta_v}{2A} + \gamma\frac{121m}{A^3},$$

where $\theta_1, \theta_2 \in [0, 1]$ and $\gamma \in [-1, 1]$.

Next, we estimate the value of $\sum_{v\in S}\zeta_v$. Given a vertex $v \in S$, define the vector $\tilde{x}_v \in \mathbb{R}^E$ such that $\tilde{x}_v(e) = x_v(e)$ for every $e \in E$. Then each term ζ_v equals the squared distance from the vector $\tilde{x}_v$ to the mean of these vectors. But it is well known (e.g., see [14,20]) that the sum of such squared distances equals the sum of all the pairwise squared distances divided by $2m$:

$$\sum_{v\in S}\zeta_v = \frac{1}{2m}\sum_{v\in S}\sum_{u\in S}\|\tilde{x}_v - \tilde{x}_u\|^2.$$

At the same time, by the 3-regularity of the graph G and by the construction of vectors $\tilde{x}_v$, each pairwise squared distance $\|\tilde{x}_v - \tilde{x}_u\|^2$ equals

$$\begin{cases} 8 \text{ if the vertices } v, u \text{ are adjacent,} \\ 6 \text{ otherwise.} \end{cases}$$

So we have $\sum_{v\in S}\zeta_v = \dfrac{6(m^2-m)+2\ell_S\cdot 2}{2m} = 3(m-1)+\dfrac{2\ell_S}{m}$. It follows the required equation. The lemma is proved. □

Lemma 3. *Let $A \geq 200m$. Then $cost(X_S, \mu(X_S)) = f(M, m, cut(S, V\setminus S), \gamma)$, where $f(M, m, t, \gamma) = mA + \dfrac{3m}{2A} - \dfrac{t}{2mA} + \gamma\dfrac{121m}{A^3}$ and $\gamma \in [-1, 1]$.*

Proof. By the 3-regularity of the graph G, we have $cut(S, V\setminus S) = 3m - 2\ell_S$. Then $\ell_S = (3m - cut(S, V\setminus S))/2$ and, by Lemma 2, we obtain the required equation. The lemma is proved. □

Theorem 3. *Capacitated Geometric Median is strongly NP-hard and does not admit an approximation scheme FPTAS unless $P = NP$.*

Proof. Suppose that the graph G consists of $n \geq 6$ vertices and set $M = 124n$. Then $m = n/2 \geq 3$ and $A \geq M\sqrt{2/3} > 200m$. By Lemma 3, it follows that $cost(X_S, \mu(X_S)) = f(M, m, cut(S, V\setminus S), \gamma)$, where $\gamma \in [-1, 1]$. On the other hand, since $A > 200m$, the absolute value of the term $\gamma\dfrac{121m}{A^3}$ in the expression for f is at most $\dfrac{2\cdot 121}{200^2} < 0.01$ times of the value $\dfrac{1}{2mA}$, the minimum possible non-zero change of the term $\dfrac{cut(S, V\setminus S)}{2mA}$. Therefore, if the minimum median cost over all the m-element subsets of X is attained on the set X_S, then S is a maximum bisection in the graph G.

Thus, Max-Bisection|3 is reduced to Capacitated Geometric Median. Taking into account that M is an integer bounded by a polynomial in the length of the input, it gives the strong NP-hardness of our problem.

Moreover, by the above, if $cut(S, V\setminus S) \leq cut(T, V\setminus T) - 1$, where S and T are any m-element subsets of vertices, then

$$cost(X_S, \mu(X_S)) - cost(X_T, \mu(X_T)) > \frac{1 - 2\cdot 0.01}{2mA} > \frac{0.98}{nM} > \frac{0.007}{n^2}.$$

At the same time, Lemma 3 and the inequality $A > 200m$ imply that, both cost values, for X_S and X_T, are less than $m\Big(A + \dfrac{3}{400m} + \dfrac{121}{(200m)^3}\Big) < mM = 62n^2$. It follows that Capacitated Geometric Median is NP-hard to approximate within a factor of $1 + \dfrac{0.007}{62n^4}$. But, for arbitrary polynomial $poly(n)$, any approximation scheme FPTAS allows to get a $\Big(1 + \dfrac{1}{poly(n)}\Big)$-approximation in polynomial time. Hence, the existence of such schemes is impossible unless P = NP. The theorem is proved. □

Remark 2. A slightly more complicated reduction to Capacitated Geometric Median can be constructed from the problem of finding an m-element independent set in a general graph (with arbitrary vertex degrees). Since the latter problem is W[1]-hard with respect to the parameter m, this reduction additionally gives the W[1]-hardness of our problem.

4 Conclusion

The question of the polynomial-time solvability and approximability of the Capacitated Geometric Median problem is studied. We give a simple "almost exact" polynomial-time algorithm for this problem in fixed dimensions and also an approximation scheme PTAS for the general case. On the other hand, we prove that the problem is strongly NP-hard and does not admit a scheme FPTAS unless P = NP. A possible direction for future work is constructing an efficient polynomial-time approximation scheme (EPTAS). Another interesting question is the complexity of the closely related Geometric 2-Median problem.

References

1. Agarwal, P.K., Har-Peled, S., Varadarajan, K.R.: Geometric approximation via coresets. Combinatorial and Computational Geometry, MSRI Publications **52**, 1–30 (2005). http://library.msri.org/books/Book52/files/01agar.pdf
2. Aggarwal, A., Imai, H., Katoh, N., Suri, S.: Finding k points with minimum diameter and related problems. J. Algorithms **12**(1), 38–56 (1991). https://doi.org/10.1016/0196-6774(91)90022-Q
3. Aho, A.V., Hopcroft, J.E., Ullman, J.D.: The Design and Analysis of Computer Algorithms. Addison-Wesley, Boston (1974). https://doi.org/10.1002/zamm.19790590233
4. Bajaj, C.: The algebraic degree of geometric optimization problems. Discrete Comput. Geom. **3**(2), 177–191 (1988). https://doi.org/10.1007/BF02187906
5. Bădoiu, M., Har-Peled, S., Indyk, P.: Approximate clustering via core-sets. In: Proceedings of the 34th ACM Symposium on Theory of Computing (STOC 2002), pp. 250–257 (2002). https://doi.org/10.1145/509907.509947
6. Bhattacharya, A., Jaiswal, R., Kumar, A.: Faster algorithms for the constrained k-means problem. Theory Comput. Syst. **62**(1), 93–115 (2018). https://doi.org/10.1007/s00224-017-9820-7

7. Charikar, M., Khuller, S., Mount, D.M., Narasimhan, G.: Algorithms for facility location problems with outliers. In: Proceedings of the 12th ACM-SIAM Symposium on Discrete Algorithms (SODA 2001), pp. 642–651 (2001). https://dl.acm.org/doi/10.5555/365411.365555
8. Chen, K.: A constant factor approximation algorithm for k-median clustering with outliers. In: Proceedings of the 19th ACM-SIAM Symposium on Discrete Algorithms (SODA 2008), pp. 826–835 (2008). https://dl.acm.org/doi/10.5555/1347082.1347173
9. 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). https://doi.org/10.1137/070699007
10. Cohen, M.B., Lee, Y.T., Miller, G., Pachocki, J., Sidford, A.: Geometric median in nearly linear time. arXiv:1606.05225 [cs.DS] (2016). https://arxiv.org/abs/1606.05225
11. Cohen-Addad, V., Feldmann, A.E., Saulpic, D.: Near-linear time approximations schemes for clustering in doubling metrics. In: Proceedings of the 60th Symposium on Foundations of Computer Science (FOCS 2019), pp. 540–559 (2019). https://doi.org/10.1109/FOCS.2019.00041
12. ElGindy, H., Keil, J.M.: Efficient algorithms for the capacitated 1-median problem. ORSA J. Comput. **4**(4), 418–425 (1992). https://doi.org/10.1287/ijoc.4.4.418
13. Feige, U., Karpinski, M., Langberg, M.: A note on approximating max-bisection on regular graphs. Inf. Proc. Letters **79**(4), 181–188 (2001). https://doi.org/10.1016/S0020-0190(00)00189-7
14. Inaba, M., Katoh, N., Imai, H.: Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering. In: Proceedings of the 10th ACM Symposium on Computational Geometry, pp. 332–339 (1994). https://doi.org/10.1145/177424.178042
15. Kel'manov, A.V., Pyatkin, A.V.: NP-completeness of some problems of choosing a vector subset. J. Appl. Industr. Math. **5**(3), 352–357 (2011). https://doi.org/10.1134/S1990478911030069
16. Krishnaswamy, R., Li, S., Sandeep, S.: Constant approximation for k-median and k-means with outliers via iterative rounding. In: Proceedings of the 50th ACM Symposium on Theory of Computing (STOC 2018), pp. 646–659 (2018). https://doi.org/10.1145/3188745.3188882
17. Kumar, A., Sabharwal, Y., Sen, S.: Linear-time approximation schemes for clustering problems in any dimensions. J. ACM. **57**(2), 1–32 (2010). https://doi.org/10.1145/1667053.1667054
18. Lee, D.T.: On k-nearest neighbor Voronoi diagrams in the plane. IEEE Trans. Comput. **31**(6), 478–487 (1982). https://doi.org/10.1109/TC.1982.1676031
19. Mulmuley, K.: Output sensitive and dynamic constructions of higher order Voronoi diagrams and levels in arrangements. J. Comp. Syst. Sci. **47**(3), 437–458 (1993). https://doi.org/10.1016/0022-0000(93)90041-T
20. Shenmaier, V.V.: An approximation scheme for a problem of search for a vector subset. J. Appl. Industr. Math. **6**(3), 381–386 (2012). https://doi.org/10.1134/S1990478912030131
21. Shenmaier, V.V.: The problem of a minimal ball enclosing k points. J. Appl. Industr. Math. **7**(3), 444–448 (2013). https://doi.org/10.1134/S1990478913030186
22. Shenmaier, V.V.: Complexity and approximation of the smallest k-enclosing ball problem. European J. Comb. **48**, 81–87 (2015). https://doi.org/10.1016/j.ejc.2015.02.011

23. Shenmaier, V.V.: Solving some vector subset problems by Voronoi diagrams. J. Appl. Industr. Math. **10**(4), 560–566 (2016). https://doi.org/10.1134/S199047891604013X
24. Shenmaier, V.V.: A structural theorem for center-based clustering in high-dimensional Euclidean space. In: Nicosia, G., Pardalos, P., Umeton, R., Giuffrida, G., Sciacca, V. (eds.) LOD 2019. LNCS, vol. 11943, pp. 284–295. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-37599-7_24
25. Shenmaier, V.V.: Some estimates on the discretization of geometric center-based problems in high dimensions. In: Kochetov, Y., Bykadorov, I., Gruzdeva, T. (eds.) MOTOR 2020. CCIS, vol. 1275, pp. 88–101. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58657-7_10