ISSN 0965-5425, Computational Mathematics and Mathematical Physics, 2016, Vol. 56, No. 2, pp. 334–341. © Pleiades Publishing, Ltd., 2016.
Original Russian Text © A.V. Kel'manov, V.I. Khandeev, 2016, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2016, Vol. 56, No. 2, pp. 332–340.

Fully Polynomial-Time Approximation Scheme for a Special Case of a Quadratic Euclidean 2-Clustering Problem

A. V. Kel'manov[a, b] **and V. I. Khandeev**[a]

[a] *Sobolev Institute of Mathematics, Siberian Branch, Russian Academy of Sciences, pr. Akademika Koptyuga 4, Novosibirsk, 630090 Russia*

[b] *Novosibirsk State University, ul. Pirogova 2, Novosibirsk, 630090 Russia*

e-mail: kelm@math.nsc.ru, khandeev@math.nsc.ru

Received March 2, 2015

Abstract—The strongly NP-hard problem of partitioning a finite set of points of Euclidean space into two clusters of given sizes (cardinalities) minimizing the sum (over both clusters) of the intracluster sums of squared distances from the elements of the clusters to their centers is considered. It is assumed that the center of one of the sought clusters is specified at the desired (arbitrary) point of space (without loss of generality, at the origin), while the center of the other one is unknown and determined as the mean value over all elements of this cluster. It is shown that unless P = NP, there is no fully polynomial-time approximation scheme for this problem, and such a scheme is substantiated in the case of a fixed space dimension.

Keywords: cluster analysis, partition, Euclidean space, minimum of the sum of squares of distances, NP-hardness, fixed space dimension, FPTAS.

DOI: 10.1134/S0965542516020111

INTRODUCTION

The subject of this study is a strongly NP-hard optimization problem. Our goal is to investigate the approximability of this problem and to substantiate a fully polynomial-time approximation scheme (FPTAS) for a special case of the problem.

In terms of theory, the considered problem of partitioning a finite set of points of Euclidean space into two subsets (clusters) is important, for example, in computer geometry [1], statistical analysis of data [2], and pattern recognition [3]. Among its natural-science and engineering applications, we can note the classification and interpretation of observations, image processing, etc. (see, e.g., [4–7] and the references therein).

The problem under consideration is to partition a finite set of points of Euclidean space into two clusters minimizing the sum (over both clusters) of the sums of squared distances from the centers of the clusters to their elements. The center of one of the sought clusters is specified at the desired (arbitrary) point of space (without loss of generality, at the origin). The center of the other cluster is unknown and determined as the mean value over all elements of this cluster. In fact, this center is the centroid or geometric center of the cluster points.

This problem is induced, for example, by the problem of testing the statistical hypothesis of two means (one being zero, and the other unknown) in an inhomogeneous sample of two multidimensional Gaussian distributions with identical given covariance matrices containing identical diagonal elements assuming that the correspondence of the sample units to the distributions is not known. Problems in noise-resistant data analysis that also induce the considered problem can be found, for example, in [8–13].

The formulation of the problem under consideration is close, but not equivalent to the NP-hard Minimum Sum-of-Squares Clustering (MSSC) problem [14]. The latter, which is also referred to as k-means, is one of the best known (for more than 50 years) problems in cluster data analysis (see [2, 3, 7, 15, 16]). In the simplest, but nevertheless NP-hard baseline two-cluster case, MSSC is to partition a finite set of spatial points into two subsets (clusters) minimizing the sum (over both clusters) of the intracluster sums of squared distances from the elements of the clusters to their desired centroids. In contrast to the baseline two-cluster case of MSSC, the desired center of one of the sought clusters in the considered problem is given as input at the origin. The optimal centroid of this cluster may not coincide with the given center,

which can be seen directly from the objective function of the problem (see the next section). Similarly, in statistical hypothesis testing, the sample mean (centroid) may not coincide with the Gaussian expectation. As is well known, the detection of this noncoincidence is a typical problem in statistics.

Recall that new clustering problems in which the desired centers for some of the clusters are given as input at some points of space were formulated and studied in [8–13, 17–21]. It was proved there that these problems, which are close to MSSC, are NP-hard even in the 2-cluster case, when only one of the desired centers is given. Note that the problems in [8–13, 17–21] were equipped with brief names reflecting their formulation. Specifically, in the last publications, MSSC was referred to as 1-MSSC-F when the cluster sizes (cardinalities) were given as input [13, 22] or as 1-MSSC-NF when the cluster sizes were unknown (optimized variables) [23]. Below, both these cases are referred to as Minimum Sum-of-Squares 2-Clustering with a Given Center. In our view, this name more precisely reflects the essence of the problem and its difference from and similarity to MSSC. Below, attention is focused primarily on the case of given cluster sizes (i.e., the sizes of the desired clusters are given as input).

The mathematical formulation of the problem is given in Section 1. Additionally, the characteristics of existing algorithms are presented. Note that the strong NP-hardness of the problem follows (see below) from the results of [8–10, 20, 24]. The strong NP-hardness of the problem implies [25] that, unless P = NP, there is neither an exact polynomial-time nor an exact pseudopolynomial-time algorithm for it. Accordingly, a task of interest is to analyze the approximability of the problem. Specifically, since the problem has numerical input, the existence of an FPTAS for it is a question of topical interest.

In this paper, we show that unless P = NP, there is no FPTAS for Minimum Sum-of-Squares 2-Clustering with a Given Center and given cluster sizes, and such a scheme is constructed in the case of a fixed space dimension.

1. FORMULATION OF THE PROBLEM AND PREVIOUS AND PRESENT RESULTS

The problem under consideration is formulated as follows (see also [13, 17, 22]).

Problem 1 (Minimum Sum-of-Squares 2-Clustering with Given Center and cluster sizes). Given a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points of $\mathbb{R}^q$ and a positive integer M, find a partition of $\mathcal{Y}$ into two clusters $\mathcal{C}$ and $\mathcal{Y}\setminus\mathcal{C}$ minimizing the objective function

$$S(\mathcal{C}) = \sum_{y\in\mathcal{C}} \|y - \bar{y}(\mathcal{C})\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{C}} \|y\|^2, \tag{1.1}$$

where $\bar{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|}\sum_{y\in\mathcal{C}} y$ is the centroid of $\mathcal{C}$ under the constraint $|\mathcal{C}| = M$.

In [12] an algorithm was proposed that finds a 2-approximate solution of this problem in $\mathbb{O}(qN^2)$ time. A PTAS having $\mathbb{O}(qN^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon})$ time complexity, where ε is the guaranteed relative error, was substantiated in [26]. A randomized algorithm was proposed in [13]. It was shown that, for a prescribed relative error $\varepsilon > 0$, a given failure probability $\gamma \in (0, 1)$, and a certain value of parameter k, the algorithm finds a $(1 + \varepsilon)$-approximate solution of the problem in $\mathbb{O}(2^k q(k + N))$ time. Additionally, the conditions were established under which this algorithm is asymptotically exact and has $\mathbb{O}(qN^2)$ time complexity. It was shown in [22] that the problem is solvable in $\mathbb{O}(q^2N^{2q})$ time, which is polynomial when the space dimension q is fixed. Additionally, an exact pseudopolynomial-time algorithm was constructed in [22] in the case of integer-valued coordinates of the input points. The time complexity of the algorithm is $\mathbb{O}(qN(2MD + 1)^q)$, where D is the maximum absolute coordinate value in the input set. For a fixed space dimension, the time complexity of this algorithm is $\mathbb{O}(N(MD)^q)$.

The main result of this paper is an FPTAS that, given a relative error ε, finds a $(1 + \varepsilon)$-approximate solution of Problem 1 in time $\mathbb{O}(N^2(1/\varepsilon)^{q/2})$, which is a polynomial in the input size of the problem and in $1/\varepsilon$ in the case of a fixed space dimension.

2. NONEXISTENCE OF FPTAS

First, we show that the Minimum Sum-of-Squares 2-Clustering with a Given Center is a strongly NP-hard problem. Indeed, it is easy to see that the objective function (1.1) can be represented in the form

$$S(\mathcal{C}) = \sum_{y\in\mathcal{Y}} \|y\|^2 - \frac{1}{|\mathcal{C}|}\left\|\sum_{y\in\mathcal{C}} y\right\|^2.$$

This equality shows that, for a given cardinality of the desired subset $\mathcal{C}$, the minimization of $S(\mathcal{C})$ in Problem 1 is polynomial-time equivalent to the strongly NP-hard problem of maximizing the norm of the sum (see [9, 10, 20]) on the right-hand side of the equality, since the minuend sum is a constant. Therefore, Problem 1 is also strongly NP-hard. In the case of an unknown cardinality of $\mathcal{C}$, Problem 1 is also strongly NP-hard, since the maximization of the second term on the right-hand side of this equality is a strongly NP-hard problem (see [8, 24]).

Now we consider the approximability of Problem 1, which is an important issue.

Theorem 1. *There is no* FPTAS *for Problem* 1 *unless* P = NP.

Proof. For any nonempty finite set $\mathcal{Z}$ of points of $\mathfrak{R}^q$ and its centroid $\overline{z}(\mathcal{Z}) = \frac{1}{|\mathcal{Z}|}\sum_{z\in\mathcal{Z}} z$, we have the easy-to-check identity

$$\sum_{x\in\mathcal{Z}}\sum_{z\in\mathcal{Z}} \|x - z\|^2 = 2|\mathcal{Z}|\sum_{z\in\mathcal{Z}} \|z - \overline{z}(\mathcal{Z})\|^2.$$

Applying this identity to the subset $\mathcal{C}$ and its centroid in the first term of equality (1.1), we obtain

$$2MS(\mathcal{C}) = \sum_{x\in\mathcal{C}}\sum_{y\in\mathcal{C}} \|x - y\|^2 + 2M \sum_{y\in\mathcal{Y}\setminus\mathcal{C}} \|y\|^2, \tag{2.1}$$

since $|\mathcal{C}| = M$.

First, we note that, for integer input data, the right-hand side of (2.1) is obviously an integer bounded by a polynomial in the input size of the problem (since $M \leqslant N$) and in the maximum (in absolute value) of the coordinates of the points in the input set. Second, the existence of an FPTAS for the minimization of the right-hand side of (2.1) for a given M implies the existence of an FPTAS for Problem 1 of minimizing $S(\mathcal{C})$ by virtue of the polynomial equivalence following from (2.1). Then, by Theorem 8.5 from [27], it follows that there is no FPTAS for the strongly NP-hard Problem 1 with numerical input unless P = NP. Theorem 1 is proved.

3. GEOMETRIC FOUNDATIONS OF THE ALGORITHM

To construct an algorithm, we need several auxiliary statements.

Lemma 1 (see [28]). *For an arbitrary point $x \in \mathfrak{R}^q$ and a finite set $\mathcal{Z} \in \mathfrak{R}^q$, it is true that*

$$\sum_{z\in\mathcal{Z}} \|z - x\|^2 = \sum_{z\in\mathcal{Z}} \|z - \overline{z}\|^2 + |\mathcal{Z}|\,\|x - \overline{z}\|^2,$$

where $\overline{z}$ is the centroid of $\mathcal{Z}$.

Lemma 2 (see [12]). *Let $\mathcal{Z}$ be a nonempty finite set of points of $\mathbb{R}^q$ and $\overline{z}$* be *the centroid of $\mathcal{Z}$. If a point $x \in \mathbb{R}^q$ satisfies the conditions*

$$\|x - \overline{z}\| \leq \|z - \overline{z}\| \quad \forall z \in \mathcal{Z},$$

then

$$\sum_{z\in\mathcal{Z}} \|z - x\|^2 \leq 2\sum_{z\in\mathcal{Z}} \|z - \overline{z}\|^2.$$

For a finite set $\mathcal{Z}$ of points of $\mathfrak{R}^q$, a positive integer $M \leqslant |\mathcal{Z}|$, and an arbitrary point $x \in \mathfrak{R}^q$, we define the set consisting of M elements of $\mathcal{Z}$ having the largest projections onto the direction specified by x:

$$\mathcal{B}_M(x, \mathcal{Z}) = \{z_i | \langle z_i, x\rangle \geq \langle z_j, x\rangle;\ z_i, z_j \in \mathcal{Z},\ i \leq M,\ j > M\},$$

where $\langle\cdot,\cdot\rangle$ is the scalar product.

Lemma 3 (see [13]). *Let $\mathcal{Z}$ be a nonempty finite set of points from $\mathbb{R}^q$ and*

$$G(\mathcal{B}, x) = \sum_{z\in\mathcal{B}} \|z - x\|^2 + \sum_{z\in\mathcal{Z}\setminus\mathcal{B}} \|z\|^2, \quad \mathcal{B} \subseteq \mathcal{Z}, \quad x \in \mathbb{R}^q. \tag{3.1}$$

Then the following assertions hold:

(1) *For any fixed subset $\mathcal{B} \subseteq \mathcal{Z}$ the minimum of functional* (3.1) *over $x \in \mathbb{R}^q$ is reached at the point $x = \overline{z}(\mathcal{B}) = \frac{1}{|\mathcal{B}|}\sum_{z\in\mathcal{B}} z$ and is equal to $S(\mathcal{B})$.*

(2) *If $|\mathcal{B}| = M$, then, for any fixed point $x \in \mathbb{R}^q$, the minimum of functional* (3.1) *over $\mathcal{B} \subseteq \mathcal{Z}$ is reached on the set $\mathcal{B}_M(x, \mathcal{Z})$.*

Lemma 4. *Let $\mathcal{C}^*$ be an optimal solution of Problem* 1. *Then, for an arbitrary point $x \in \mathbb{R}^q$,*

$$S(\mathcal{B}_M(x, \mathcal{Y})) \le S(\mathcal{C}^*) + M\|x - \overline{y}(\mathcal{C}^*)\|^2, \tag{3.2}$$

where $\overline{y}(\mathcal{C}^*) = \frac{1}{M}\sum_{y\in\mathcal{C}^*} y$ is the centroid of $\mathcal{C}^*$.

Proof. The first assertion in Lemma 3 implies the estimate

$$S(\mathcal{B}_M(x, \mathcal{Y})) = G(\mathcal{B}_M(x, \mathcal{Y}), \overline{z}(\mathcal{B}_M(x, \mathcal{Y}))) \le G(\mathcal{B}_M(x, \mathcal{Y}), x), \tag{3.3}$$

while the second assertion yields the inequality

$$G(\mathcal{B}_M(x, \mathcal{Y}), x) \le G(\mathcal{C}^*, x). \tag{3.4}$$

Applying Lemma 1 to the point x and the set $\mathcal{C}^*$, we obtain

$$\sum_{y\in\mathcal{C}^*} \|y - x\|^2 = \sum_{y\in\mathcal{C}^*} \|y - \overline{y}(\mathcal{C}^*)\|^2 + |\mathcal{C}^*|\,\|x - \overline{y}(\mathcal{C}^*)\|^2. \tag{3.5}$$

Finally, combining (3.3)–(3.5) gives the estimate

$$\begin{aligned} S(\mathcal{B}_M(x, \mathcal{Y})) \le G(\mathcal{B}_M(x, \mathcal{Y}), x) \le G(\mathcal{C}^*, x) &= \sum_{y\in\mathcal{C}^*} \|y - x\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{C}^*} \|y\|^2 \\ = \sum_{y\in\mathcal{C}^*} \|y - \overline{y}(\mathcal{C}^*)\|^2 + |\mathcal{C}^*|\,\|x - \overline{y}(\mathcal{C}^*)\|^2 &+ \sum_{y\in\mathcal{Y}\setminus\mathcal{C}^*} \|y\|^2 = S(\mathcal{C}^*) + M\|x - \overline{y}(\mathcal{C}^*)\|^2. \end{aligned}$$

Lemma 4 is proved.

Lemma 5. *Let the conditions of Lemma 4 hold and $t = \arg\min_{y\in\mathcal{C}^*} \|y - \overline{y}(\mathcal{C}^*)\|$ be the point of the set $\mathcal{C}^*$ nearest to its centroid. Then, given a fixed $\varepsilon > 0$, for the set $\mathcal{B}_M(x, \mathcal{Y})$ to be a $(1 + \varepsilon)$-approximate solution of Problem* 1, *it is sufficient that the point x satisfies the inequality*

$$\|x - \overline{y}(\mathcal{C}^*)\|^2 \le \frac{\varepsilon}{2M} S(\mathcal{B}_M(t, \mathcal{Y})). \tag{3.6}$$

Proof. The first assertion of Lemma 3 implies the estimate

$$S(\mathcal{B}_M(t, \mathcal{Y})) = G(\mathcal{B}_M(t, \mathcal{Y}), \overline{z}(\mathcal{B}_M(t, \mathcal{Y}))) \le G(\mathcal{B}_M(t, \mathcal{Y}), t), \tag{3.7}$$

while the second assertion yields the inequality

$$G(\mathcal{B}_M(t, \mathcal{Y}), t) \le G(\mathcal{C}^*, t). \tag{3.8}$$

Consider the set $\mathcal{C}^*$ and the point t. Since they satisfy the conditions of Lemma 2, we have

$$\sum_{y\in\mathcal{C}^*} \|y - t\|^2 \le 2\sum_{y\in\mathcal{C}^*} \|y - \overline{y}(\mathcal{C}^*)\|^2$$

and, hence,

$$\begin{aligned} G(\mathcal{C}^*, t) = \sum_{y\in\mathcal{C}^*} \|y - t\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{C}^*} \|y\|^2 &\le 2\sum_{y\in\mathcal{C}^*} \|y - \overline{y}(\mathcal{C}^*)\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{C}^*} \|y\|^2 \\ \le 2\sum_{y\in\mathcal{C}^*} \|y - \overline{y}(\mathcal{C}^*)\|^2 + 2\sum_{y\in\mathcal{Y}\setminus\mathcal{C}^*} \|y\|^2 &= 2S(\mathcal{C}^*). \end{aligned} \tag{3.9}$$

Combining (3.6)–(3.9) yields

$$\|x - \bar{y}(\mathcal{C}^*)\|^2 \le \frac{\varepsilon}{2M} S(\mathcal{B}_M(t, \mathcal{Y})) \le \frac{\varepsilon}{2M} G(\mathcal{B}_M(t, \mathcal{Y}), t) \le \frac{\varepsilon}{2M} G(\mathcal{C}^*, t) = \frac{\varepsilon}{M} S(\mathcal{C}^*). \tag{3.10}$$

Finally, applying (3.10) to the right-hand side of (3.2), we obtain the estimate

$$S(\mathcal{B}_M(x, \mathcal{Y})) \le (1 + \varepsilon) S(\mathcal{C}^*),$$

which proves Lemma 5.

Lemma 6. *Let $\mathcal{C}^*$ be an optimal solution of Problem* 1. *Then the point $t = \arg\min\limits_{y \in \mathcal{C}^*} \|y - \bar{y}(\mathcal{C}^*)\|$ satisfies the estimate*

$$\|t - \bar{y}(\mathcal{C}^*)\|^2 \le \frac{1}{M} S(\mathcal{B}_M(t, \mathcal{Y})).$$

Proof. The definition of t implies that, for any $y \in \mathcal{C}^*$,

$$\|t - \bar{y}(\mathcal{C}^*)\|^2 \le \|y - \bar{y}(\mathcal{C}^*)\|^2.$$

Summing up both sides of this inequality over $y \in \mathcal{C}^*$, we obtain

$$M\|t - \bar{y}(\mathcal{C}^*)\|^2 \le \sum_{y \in \mathcal{C}^*} \|y - \bar{y}(\mathcal{C}^*)\|^2. \tag{3.12}$$

Since $\mathcal{B}_M(t, \mathcal{Y})$ is a feasible solution of Problem 1 and $\mathcal{C}^*$ is its optimal solution, we have the inequality

$$S(\mathcal{C}^*) \le S(\mathcal{B}_M(t, \mathcal{Y})). \tag{3.13}$$

Combining (3.12) and (3.13) yields the estimate

$$M\|t - \bar{y}(\mathcal{C}^*)\|^2 \le \sum_{y \in \mathcal{C}^*} \|y - \bar{y}(\mathcal{C}^*)\|^2 \le S(\mathcal{C}^*) \le S(\mathcal{B}_M(t, \mathcal{Y})),$$

which proves inequality (3.11). Lemma 6 is proved.

4. FPTAS ALGORITHM

The idea of the algorithmic solution proposed (FPTAS algorithm) can be described as follows. For each point of the input set, a domain (cube) is constructed so that the center of the desired subset necessarily belongs to one of these domains. Given (as input) the prescribed relative error of the solution, a grid (lattice) is generated that discretizes the cube with a uniform step in all coordinates. For each lattice node, a set of M points of the original set that have the largest projections onto the direction specified by this node is formed. The resulting set is declared a contender for the solution. The contender subset that minimizes the objective function is chosen to be the final solution.

For an arbitrary point $z \in \mathbb{R}^q$ and positive numbers h and H, we define the set of points

$$\mathcal{D}(z, h, H) = \{d | d = z + h(j_1, \ldots, j_q),\ j_i \in \mathbb{Z},\ |hj_i| \le H,\ i = 1, \ldots, q\},$$

which is a cubic lattice of size $2H$ centered at the point z with mesh spacing h. The number of nodes in this grid is

$$|\mathcal{D}(z, h, H)| \le \left(2\left\lfloor \frac{H}{h} \right\rfloor + 1\right)^q \le \left(2\frac{H}{h} + 1\right)^q.$$

Moreover, for any $x \in \mathbb{R}^q$ such that $\|z - x\| \leqslant H$, the distance to the nearest node of $\mathcal{D}(z, h, H)$ obviously does not exceed $(h\sqrt{q})/2$.

Note that, in fact, Lemma 6 (the right-hand side of (3.11)) determines the size of the lattice that necessarily contains the centroid of the optimal solution to the problem only if the point t of the input set $\mathcal{Y}$ is the nearest to this centroid. Therefore, for the size of the lattice, we set

$$H(y) = \sqrt{\frac{1}{M} S(\mathcal{B}_M(y, \mathcal{Y}))}, \quad y \in \mathcal{Y}. \tag{4.1}$$

Moreover, Lemma 5 establishes the condition on the lattice mesh spacing under which, among the nodes, there is one close (in the sense of the guaranteed error ε) to the centroid of the optimal solution. Therefore, for the mesh spacing, we set

$$h(y,\varepsilon) = \sqrt{\frac{2\varepsilon}{qM} S(\mathcal{B}_M(y,\mathcal{Y}))}, \quad y \in \mathcal{Y}, \quad \varepsilon > 0. \tag{4.2}$$

Let us formulate the following algorithm for solving the problem.

Algorithm $\mathcal{A}$

Input: a set $\mathcal{Y}$ and numbers M and ε.

For each point $y \in \mathcal{Y}$ Steps 1–5 are executed.

Step 1. Construct the set $\mathcal{B}_M(y, \mathcal{Y})$.

Step 2. Compute $S(\mathcal{B}_M(y, \mathcal{Y}))$, h, and H using formulas (1.1), (4.2), and (4.1).

Step 3. If $S(\mathcal{B}_M(y, \mathcal{Y})) = 0$, then return the set $\mathcal{B}_M(y, \mathcal{Y})$ as the result produced by the algorithm and exit; otherwise, go to the next step.

Step 4. Construct the lattice $\mathcal{D}(y, h, H)$.

Step 5. For each point d of the lattice $\mathcal{D}(y, h, H)$, construct the set $\mathcal{B}_M(d, \mathcal{Y})$ and compute $S(\mathcal{B}_M(d, \mathcal{Y}))$.

Step 6. In the family $\{\mathcal{B}_M(d, \mathcal{Y}) | d \in \mathcal{D}(y, h, H), y \in \mathcal{Y}\}$ of sets, choose, as a solution, the set $\mathcal{B}_M(d, \mathcal{Y})$ for which $\mathcal{B}_M(d, \mathcal{Y})$ is minimal.

Exit.

Theorem 2. *For any fixed $\varepsilon > 0$, algorithm $\mathcal{A}$ finds a $(1 + \varepsilon)$-approximate solution of Problem 1 in* $\mathbb{O}\left(qN^2\left(\sqrt{\frac{2q}{\varepsilon}} + 1\right)^q\right)$ *time.*

Proof. Let $t = \arg\min_{y \in \mathcal{C}^*} \|y - \bar{y}(\mathcal{C}^*)\|$ be the point of the set $\mathcal{C}^*$ nearest to its centroid. If the equality $S(\mathcal{B}_M(t, \mathcal{Y})) = 0$ holds for this point at Step 3, then the set $\mathcal{B}_M(t, \mathcal{Y})$ is an optimal solution of Problem 1, since, for any set $\mathcal{C} \subseteq \mathcal{Y}$, it is true that $S(\mathcal{C}) \geq 0$.

Consider the case $S(\mathcal{B}_M(t, \mathcal{Y})) > 0$. By Lemma 6, inequality (3.11) holds for the point t. This inequality and (4.1) imply that $\|t - \bar{y}(\mathcal{C}^*)\| \leqslant H$. In other words, the centroid $\bar{y}(\mathcal{C}^*)$ of the optimal set lies within the grid $\mathcal{D}(t, h, H)$.

Let $d^* = \arg\min_{d \in \mathcal{D}(t,h,H)} \|d - \bar{y}(\mathcal{C}^*)\|$. Since the distance from $\bar{y}(\mathcal{C}^*)$ to the nearest node d^* of $\mathcal{D}(t, h, H)$ does not exceed $(h\sqrt{q})/2$, we have the estimate

$$\|d^* - \bar{y}(\mathcal{C}^*)\|^2 \leq \frac{h^2 q}{4} = \frac{\varepsilon}{2M} S(\mathcal{B}_M(t, \mathcal{Y})).$$

Therefore, the point d^* satisfies the conditions of Lemma 5 and, hence, the set $\mathcal{B}_M(d^*, \mathcal{Y})$ is a $(1 + \varepsilon)$-approximate solution of Problem 1.

Let us estimate the time complexity of the algorithm. At Step 1, (1) N projections of points of the set $\mathcal{Y}$ onto the direction specified by the point y are computed in $\mathbb{O}(qN)$ operations, and (2) in the resulting collection, M largest projections and the corresponding points of the set $\mathcal{Y}$ are chosen in $\mathbb{O}(N)$ operations without sorting (see, e.g., [29]). Step 2 requires at most $\mathbb{O}(qN)$ operations, while Step 3 is executed in $\mathbb{O}(1)$ operations.

The complexity of generating $\mathcal{D}(y, h, H)$ at Step 4 is $\mathbb{O}(q|\mathcal{D}(y, h, H)|)$.

At Step 5, each of $|\mathcal{D}(y, h, H)|$ sets $\mathcal{B}_M(d, \mathcal{Y})$ is constructed in $\mathbb{O}(qN)$ operations, and the same is true for the computation of $S(\mathcal{B}_M(d, \mathcal{Y}))$.

For each of N points $y \in \mathcal{Y}$ Steps 1–5 are executed in $\mathbb{O}(qN|\mathcal{D}(y, h, H)|)$ operations. Step 6 (choosing the least element) requires at most $\mathbb{O}\left(\sum_{y \in \mathcal{Y}} |\mathcal{D}(y, h, H)|\right)$ operations.

It remains to be noted that the cardinality of $\mathcal{D}(y, h, H)$ satisfies the estimate

$$|\mathcal{D}(y,h,H)| \le \left(2\frac{H}{h}+1\right)^q \le \left(\sqrt{\frac{2q}{\varepsilon}}+1\right)^q.$$

Therefore, the time complexity of the algorithm is $\mathbb{O}\left(qN^2\left(\sqrt{\frac{2q}{\varepsilon}}+1\right)^q\right)$. Theorem 2 is proved.

Let us show that algorithm $\mathcal{A}$ is an FPTAS if the space dimension q is fixed. Indeed, if $\varepsilon \in (0, 2q]$, then

$$\left(\sqrt{\frac{2q}{\varepsilon}}+1\right)^q \le 2^q\left(\sqrt{\frac{2q}{\varepsilon}}\right)^q = 2^{3q/2}q^{q/2}(1/\varepsilon)^{q/2} = \mathbb{O}((1/\varepsilon)^{q/2}).$$

Therefore, under the indicated conditions, the running time of the algorithm is $\mathbb{O}(N^2(1/\varepsilon)^{q/2})$, which is bounded by a polynomial in the input size of the problem and in $1/\varepsilon$. Thus, the algorithm implements an FPTAS.

Remark. The FPTAS constructed can be used to solve Minimum Sum-of-Squares 2-Clustering with a Given Center, in which the cardinalities of the clusters are optimization variables. For this purpose, it is sufficient to find $\mathbb{O}(N)$ solutions of the problem with the help of the FPTAS for each admissible size of the desired subset and to choose the best of these solutions. Obviously, the time complexity of this algorithm is $\mathbb{O}\left(qN^3\left(\sqrt{\frac{2q}{\varepsilon}}+1\right)^q\right)$. For a fixed space dimension, it is $\mathbb{O}(N^3(1/\varepsilon)^{q/2})$.

CONCLUSIONS

The existence of an FPTAS was analyzed for a strongly NP-hard problem with numerical input, namely, for the problem of partitioning a finite set of points of Euclidean space into two clusters of given cardinalities minimizing the sum (over both clusters) of the intracluster sums of squared distances from the elements of the clusters to their centers. It is assumed that the center of one of the desired clusters is given at the origin, while the center of the other is unknown and determined as the mean value over all elements of this cluster.

It was proved that unless P = NP there is no FPTAS for the problem, and such a scheme was constructed in the case of a fixed space dimension.

It was shown that the algorithm substantiated can be used to construct an FPTAS for the problem in which the cardinalities of the clusters are optimization variables. However, a task of interest is to design another (less expensive) FPTAS without exhaustive search of all admissible sizes of the desired clusters. The substantiation of such a scheme is an important issue to be addressed in the nearest future.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research, project nos. 13-07-00070 and 15-01-00462.

REFERENCES

1. M. Bern and D. Eppstein, "Approximation algorithms for geometric problems," in *Approximation Algorithms for NP-Hard Problems*, Ed. by D. S. Hochbaum (PWS, Boston, 1997), pp. 296–345.
2. G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning* (Springer Science, New York, 2013).
3. M. C. Bishop, *Pattern Recognition and Machine Learning* (Springer Science, New York, 2006).
4. P. Flach, *Machine Learning: The Art and Science of Algorithms That Make Sense of Data* (Cambridge University Press, New York, 2012).
5. C. Steger, M. Ulrich, and C. Wiedemann, *Machine Vision Algorithms and Applications* (Cambridge Univ. Press, New York, 2010).
6. S. Kaiser, *Biclustering: Methods, Software, and Application* (Faculty of Math. Comput. Sci. Stat., Munich, 2011).
7. A. K. Jain, "Data clustering: 50 years beyond k-means," Pattern Recogn. Lett. **31** (8), 651–666 (2010).

8. A. V. Kel'manov and A. V. Pyatkin, "On the complexity of a search for a subset of "similar" vectors," Dokl. Math **78** (1), 574–575 (2008).
9. E. Kh. Gimadi, A. V. Kel'manov, M. A. Kel'manova, and S. A. Khamidullin, "A posteriori detection of a quasiperiodic fragment with a given number of repetitions in a numerical sequence," Sib. Zh. Ind. Mat. **9** (1), 55–74 (2006).
10. E. Kh. Gimadi, A. V. Kel'manov, M. A. Kel'manova, and S. A. Khamidullin, "A posteriori detecting a quasiperiodic fragment in a numerical sequence," Pattern Recogn. Image Anal. **18** (1), 30–42 (2008).
11. A. V. Kel'manov, "Off-line detection of a quasi-periodically recurring fragment in a numerical sequence," Proc. Steklov Inst. Math. **263**, Suppl. 2, 84–92 (2008).
12. A. V. Dolgushev and A. V. Kel'manov, "An approximation algorithm for solving a problem of cluster analysis," J. Appl. Ind. Math. **5** (4), 551–558 (2011).
13. A. V. Kel'manov and V. I. Khandeev, "Randomized algorithm for two-cluster partition of a set of vectors," Comput. Math. Math. Phys. **55** (2), 330–339 (2015).
14. D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "NP-hardness of Euclidean sum-of-squares clustering," Machine Learning **75** (2), 245–248 (2009).
15. J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of the 5th Berkeley Symposium of Mathematical Statistics and Probability* (Univ. of California Press, Berkeley, 1967), Vol. 1, pp. 281–297.
16. M. Rao, "Cluster analysis and mathematical programming," J. Am. Stat. Assoc. **66**, 622–626 (1971).
17. A. V. Kel'manov and A. V. Pyatkin, "Complexity of certain problems of searching for subsets of vectors and cluster analysis," Comput. Math. Math. Phys. **49** (11), 1966–1971 (2009).
18. A. V. Kel'manov, "On the complexity of some data analysis problems," Comput. Math. Math. Phys. **50** (11), 1941–1947 (2010).
19. A. V. Kel'manov, "On the complexity of some cluster analysis problems," Comput. Math. Math. Phys. **51** (11), 1983–1988 (2011).
20. A. E. Baburin, E. Kh. Gimadi, N. I. Glebov, and A. V. Pyatkin, "The problem of finding a subset of vectors with the maximum total weight," J. Appl. Ind. Math. **2** (1), 32–38 (2008).
21. A. V. Kel'manov and A. V. Pyatkin, "NP-completeness of some problems of choosing a vector subset," J. Appl. Ind. Math. **5** (3), 352–357 (2011).
22. A. V. Kel'manov and V. I. Khandeev, "An exact pseudopolynomial algorithm for a problem of the two-cluster partitioning of a set of vectors," J. Appl. Ind. Math. **9** (4), 497–502 (2015).
23. A. V. Kel'manov and V. I. Khandeev, "A 2-approximation polynomial algorithm for a clustering problem," J. Appl. Ind. Math. **7** (4), 515–521 (2013).
24. A. V. Kel'manov and A. V. Pyatkin, "On a version of the problem of choosing a vector subset," J. Appl. Ind. Math. **3** (4), 447–455 (2009).
25. M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness* (Freeman, San Francisco, 1979).
26. A. V. Dolgushev, A. V. Kel'manov, and V. V. Shenmaier, "A PTAS for a problem of cluster analysis," *Proceedings of the 9th International Conference on Intelligent Information Processing, Budva, Montenegro* (Torus, Moscow, 2012), pp. 242–244.
27. V. V. Vazirani, *Approximation Algorithms* (Springer, Berlin, 2001).
28. A. V. Kel'manov and S. M. Romanchenko, "An FPTAS for a vector subset search problem," J. Appl. Ind. Math. **8** (3), 329–336 (2014).
29. I. Wirth, *Algorithms + Data Structures = Programs* (Prentice Hall, New Jersey, 1976).

Translated by I. Ruzanova