Approximation Scheme for a Quadratic Euclidean Weighted 2-Clustering Problem

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

[a] *Sobolev Institute of Mathematics, Siberian Branch, Russian Academy of Sciences, pr. Akad. Koptyuga 4, Novosibirsk, Russia*
[b] *Novosibirsk State University, ul. Pirogova 2, Novosibirsk, Russia*
**e-mail: kelm@math.nsc.ru*
***e-mail: anitamo@mail.ru*

Abstract—We consider the strongly NP-hard problem of partitioning a finite set of Euclidean points into two clusters so as to minimize the sum (over both clusters) of the weighted sums of the squared intra-cluster distances from the elements of the cluster to its center. The weights of the sums are equal to the cardinalities of the clusters. The center of one of the clusters is given as input, while the center of the other cluster is unknown and is determined as the mean value over all points in this cluster, i.e., as the geometric center (centroid). The version of the problem with constrained cardinalities of the clusters is analyzed. We construct an approximation algorithm for the problem and show that it is a fully polynomial-time approximation scheme (FPTAS) if the space dimension is bounded by a constant.

Keywords: data analysis, weighted 2-clustering, Euclidean space, NP-hardness, fixed space dimension, FPTAS

DOI: 10.1134/S105466181801008X

INTRODUCTION

The subject of this study is the strongly NP-hard quadratic problem of partitioning a finite set of points in Euclidean space into two clusters with a fixed center of one of the clusters. Our goal is to substantiate an approximation algorithm that solves this problem.

The research is motivated by the fact that the problem has been poorly studied from an algorithmic standpoint and is important for applications including, in particular, problems in cluster analysis, approximation theory, and data interpretation; statistical problems of joint evaluation and hypothesis testing with heterogeneous samples; geometric problems, etc. (see, for example, [1–12], the references therein, and the next section).

The paper[1] develops the results presented in [1–3] and is structured as follows. Section 1 contains the formulation of the problem and examples of its applications. Additionally, known results are given and the result obtained in this paper is announced. In Section 2, geometric statements are formulated and proved that provide the establishment of performance estimates (accuracy and time complexity) for the proposed algorithm. Finally, in Section 3, an approximation algorithm is substantiated and it is shown that the proposed algorithm for the space of fixed dimension is a fully polynomial-time approximation scheme (FPTAS).

1. FORMULATION OF THE PROBLEM, ITS INTERPRETATION, AND APPLICATIONS. KNOWN AND OBTAINED RESULTS

In what follows, $\mathbb{R}$ is the set of real numbers, $\mathbb{R}_+$ is the set of positive real numbers, $\mathbb{Z}$ is the set of integers, $\|\cdot\|$ is the Euclidean norm, and $\langle\cdot,\cdot\rangle$ is the scalar product.

The problem under consideration is formulated as follows (see [1–3], [23]).

Problem 1 *(Cardinality-weighted variance-based 2-clustering with given center). Given:* a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$ and a positive integer M. *Find:* a partition of the set $\mathcal{Y}$ into two nonempty clusters $\mathcal{C}$ and $\mathcal{Y}\setminus\mathcal{C}$ such that

$$F(\mathcal{C}) = |\mathcal{C}| \sum_{y \in C} \|y - \bar{y}(\mathcal{C})\|^2 + |\mathcal{Y}\setminus\mathcal{C}| \sum_{y \in \mathcal{Y}\setminus\mathcal{C}} \|y\|^2 \to \min, \quad (1)$$

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

Problem 1 has a simple (easily verifiable) geometrical interpretation. It represents partitioning a finite set of points in Euclidean space into two subsets by the optimal (according to (1)) second-order separating surface. As is known, the construction and application

[1] A short preliminary version of the paper has been published in the proceedings of the conference [23].

Received August 5, 2017

of optimal nonlinear separating surfaces (decision functions), in particular of second-order surfaces, is one of the traditional directions of research in data analysis and pattern recognition (see, for example, [5–12] and the reference therein).

In formula (1), the cardinalities of the desired clusters are the weight factors of the intra-cluster sums. Therefore, Problem 1 can be treated as a problem of optimal summing weighted by the cardinalities of the clusters.

In addition, Problem 1 has applications in an interdisciplinary data-mining problem (see, e.g., [4], [8], [10], [11]). The essence of this multifaceted problem is the approximation of the data with some mathematical model that allows us to adequately interpret these data and plausibly explain their origin in terms of the approximating model. In particular, the following statistical hypothesis may be such a model: whether it is true that the input data $\mathcal{Y}$ is an inhomogeneous sample from two probability distributions, where one of these distributions has a zero mean, while the mean of the second is unknown and not equal to zero. It is assumed that the correspondence of the sample elements to the distribution is not known and the sample data from the cluster $\mathcal{Y}\setminus\mathcal{C}$ are taken from the distribution with zero mean. To test this hypothesis, we first need to find the optimal solution of Problem 1 (i.e., the partition into two clusters—homogeneous samples), and only then we will be able to use the classical results in the field of statistical hypothesis testing for homogeneous samples.

As is known, the basic mathematical tools for applied researchers who study and analyze data are algorithms for solving a variety of clustering problems in which the clusters consist of similar objects or objects close by a certain criterion. The design of new mathematical tools for solving data-mining problems causes the development of effective algorithms with guaranteed performance estimates of accuracy and time complexity.

Recall that Problem 1 under consideration (with one unknown centroid) is close in its statement to the intractable problem of weighted clustering (with two unknown centroids) [13, 14], in which, instead of sum (1), it is required to minimize the sum $|\mathcal{C}|\sum_{y\in\mathcal{C}}\|y-\bar{y}(\mathcal{C})\|^2 + |\mathcal{Y}\setminus\mathcal{C}|\sum_{y\in\mathcal{Y}\setminus\mathcal{C}}\|y-\bar{y}(\mathcal{Y}\setminus\mathcal{C})\|^2$. For this problem, a number of algorithmical results have been obtained, particularly in [14–18]. These results cannot be transferred directly to Problem 1 because the latter is not equivalent to the above-mentioned known problem and is not its particular case. Problem 1 requires individual algorithmic studies.

It is shown in [2, 3] that Problem 1 is strongly NP-hard. Therefore, according to [19], there exist neither exact polynomial-time nor exact pseudopolynomial-time algorithms for this problem if the hypothesis P≠NP is true. In addition, it is proved in [2, 3] that there is no FPTAS for Problem 1 with numerical inputs unless P = NP. For this reason, it is of interest to find subclasses of this problem for which such schemes exist.

At present, there is only one algorithmic result obtained for Problem 1; namely, for the case of this problem with integer input points, an exact algorithm has been constructed [1]. This algorithm finds a solution in $\mathbb{O}(qN(2MB+1)^q)$ time, where B is the maximum absolute value of the coordinates in the input set of points. If the dimension q of the space is bounded by a constant, then the runtime of the algorithm is estimated as $\mathbb{O}(N(MB)^q)$ and its time complexity is pseudopolynomial.

In the present paper, we construct an approximation algorithm for Problem 1. Given a relative error ε, this algorithm allows us to find an $(1+\varepsilon)$-approximate solution of the problem in $\mathbb{O}\left(qN^2\left(\sqrt{\frac{2q}{\varepsilon}}+2\right)^q\right)$ time. In the case of the space dimension q bounded by a constant, the runtime of the algorithm is estimated as $\mathbb{O}\left(N^2\left(\frac{1}{\varepsilon}\right)^{q/2}\right)$ and it implements an FPTAS.

2. GEOMETRIC FOUNDATIONS OF THE ALGORITHM

In order to substantiate the algorithm we need several basic statements.

The following two lemmas are well-known. Their proofs are presented in many publications (see, for example, [20, 21]).

Lemma 1. *For an arbitrary point $x\in\mathbb{R}^q$ and a finite set $\mathcal{Z}\subset\mathbb{R}^q$, we have the equality*

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

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

Lemma 2. *Let the conditions of Lemma 1 hold. If some point $u\in\mathbb{R}^q$ lies closer (in the distance sense) to the centroid $\bar{z}$ of the set $\mathcal{Z}$ than any point of this set, then*

$$\sum_{z\in\mathcal{Z}}\|z-u\|^2\le 2\sum_{z\in\mathcal{Z}}\|z-\bar{z}\|^2.$$

Hereinafter, we use $f^x(y)$ to denote the function $f(x,y)$ provided that the argument x of this function is fixed; the notation $f^y(x)$ has the analoguous sense. The following two lemmas were proved in [1].

Lemma 3. *Let*

$$S(\mathcal{C},x)=|\mathcal{C}|\sum_{y\in\mathcal{C}}\|y-x\|^2+|\mathcal{Y}\setminus\mathcal{C}|\sum_{y\in\mathcal{Y}\setminus\mathcal{C}}\|y\|^2,\quad \mathcal{C}\subseteq\mathcal{Y},\quad x\in\mathbb{R}^q, \tag{2}$$

where $\mathcal{Y}$ is the input set of points in Problem 1. *Then the following equality holds:*

$$S(\mathcal{C}, x) = F(\mathcal{C}) + |\mathcal{C}|^2 \|x - \overline{y}(\mathcal{C})\|^2.$$

Lemma 4. *Let the conditions of Lemma* 3 *hold. Then the following assertions are valid for the conditioinal minima of function* (2):

(1) *for each nonempty fixed subset $\mathcal{C} \subseteq \mathcal{Y}$, the minimum of the function $S^{\mathcal{C}}(x)$ over $x \in \mathbb{R}^q$ is attained at the point $x = \overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|}\sum_{y\in\mathcal{C}} y$ and is equal to $F(\mathcal{C})$;*

(2) *if $|\mathcal{C}| = M = const$, then, for every fixed point $x \in \mathbb{R}^q$, the minimum of the function $S^x(\mathcal{C})$ over $\mathcal{C} \subseteq \mathcal{Y}$ satisfied the equality*

$$\arg\min_{\mathcal{C}\subseteq\mathcal{Y}} S^x(\mathcal{C}) = \arg\min_{\mathcal{C}\subseteq\mathcal{Y}} G^x(\mathcal{C}),$$

where

$$G^x(\mathcal{C}) = \sum_{y\in\mathcal{C}} g^x(y), \tag{3}$$

$$g^x(y) = (2M - N)\|y\|^2 - 2M\langle y, x\rangle, \quad y \in \mathcal{Y}; \tag{4}$$

moreover,

$$\min_{\mathcal{C}\subseteq\mathcal{Y}} G^x(\mathcal{C}) = \sum_{y\in\mathcal{B}^x} g^x(y),$$

and the set $\mathcal{B}^x$ consists of those M points of the set $\mathcal{Y}$ at which the function $g^x(y)$ takes the smallest values.

Lemma 5. *Let the conditions of Lemma* 4 *hold and let $\mathcal{C}^*$ be the optimal solution of Problem* 1. *Then, for a fixed point $x \in \mathbb{R}^q$, the value $F(\mathcal{C})$ on the set $\mathcal{C} = \mathcal{B}^x$ satisfies the bound*

$$F(\mathcal{B}^x) \le F(\mathcal{C}^*) + M^2\|x - \overline{y}(\mathcal{C}^*)\|^2.$$

Proof. From definitions (1) and (2) and Lemma 1, we obtain

$$F(\mathcal{B}^x) = S^{\overline{y}(\mathcal{B}^x)}(\mathcal{B}^x) \le S^x(\mathcal{B}^x) \le S^x(\mathcal{C}^*). \tag{5}$$

In addition, Lemma 3 for the right-hand side of (5) implies the equality

$$S^x(\mathcal{C}^*) = F(\mathcal{C}^*) + M^2\|x - \overline{y}(\mathcal{C}^*)\|^2. \tag{6}$$

Combining (5) and (6) yields the statement of the lemma. The lemma is proved.

Lemma 6. *Suppose that the conditions of Lemma 5 hold and $t = \arg\min_{y\in\mathcal{C}^*}\|y - \overline{y}(\mathcal{C}^*)\|^2$ is the point of the set $\mathcal{C}^*$ nearest to its centroid. Then the squared distance from the point t to the centroid of the subset $\mathcal{C}^*$ satisfies the bound*

$$\|t - \overline{y}(\mathcal{C}^*)\|^2 \le \frac{1}{M^2}F(\mathcal{B}^t), \tag{7}$$

where $\mathcal{B}^t$ is the set defined in Lemma 4 *for $x = t$.*

Proof. By the definition of the point t, we have the inequality

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

for each point $y \in \mathcal{C}^*$. Summing up this inequality over all $y \in \mathcal{C}^*$, we obtain

$$M\|t - \overline{y}(\mathcal{C}^*)\|^2 \le \sum_{y\in\mathcal{C}^*}\|y - \overline{y}(\mathcal{C}^*)\|^2. \tag{8}$$

Furthermore, since the subset $\mathcal{C}^*$ is optimal, we have the bound

$$F(\mathcal{C}^*) \le F(\mathcal{B}^t), \tag{9}$$

and definition (1) implies the inequality

$$\sum_{y\in\mathcal{C}^*}\|y - \overline{y}(\mathcal{C}^*)\|^2 \le \frac{1}{M}F(\mathcal{C}^*). \tag{10}$$

Combining (8), (10), and (9), we obtain

$$M\|t - \overline{y}(\mathcal{C}^*)\|^2 \le \frac{1}{M}F(\mathcal{B}^t).$$

The lemma is proved.

Lemma 7. *Let the conditions of Lemma* 6 *hold. If, for arbitrary $\varepsilon > 0$ and a point $x \in \mathbb{R}^q$, the inequality*

$$\|x - \overline{y}(\mathcal{C}^*)\|^2 \le \frac{\varepsilon}{2M^2}F(\mathcal{B}^t), \tag{11}$$

is valid, then the subset $\mathcal{B}^x$ defined in Lemma 4 *is an $(1+\varepsilon)$-approximate solution of Problem* 1.

Proof. From (1), Lemma 4, and the definition of the point t, we have

$$F(\mathcal{B}^t) = S^{\overline{y}(\mathcal{B}^t)}(\mathcal{B}^t) \le S^t(\mathcal{B}^t) \le S^t(\mathcal{C}^*). \tag{12}$$

Further, applying Lemma 2 to the set $\mathcal{L} = \mathcal{C}^*$ and the point $u = t$, we obtain

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

Using this inequality and definition (2), we find the estimate for the right-hand side of (12):

$$\begin{aligned} S^t(\mathcal{C}^*) &= M\sum_{y\in\mathcal{C}^*}\|y - t\|^2 + (N - M)\sum_{y\in\mathcal{Y}\setminus\mathcal{C}^*}\|y\|^2 \\ &\le 2M\sum_{y\in\mathcal{C}^*}\|y - \overline{y}(\mathcal{C}^*)\|^2 + (N - M)\sum_{y\in\mathcal{Y}\setminus\mathcal{C}^*}\|y\|^2 \le 2F(\mathcal{C}^*). \end{aligned} \tag{13}$$

Combining (11), (12), and (13), we obtain

$$\begin{aligned} \|x - \overline{y}(\mathcal{C}^*)\|^2 &\le \frac{\varepsilon}{2M^2}F(\mathcal{B}^t) \\ &\le \frac{\varepsilon}{2M^2}S^t(\mathcal{C}^*) \le \frac{\varepsilon}{M^2}F(\mathcal{C}^*). \end{aligned} \tag{14}$$

Finally, from Lemma 5 and (14), we find the following estimate for the value of the objective function (1) on the subset $\mathcal{B}^x$:

$$F(\mathcal{B}^x) \le F(\mathcal{C}^*) + M^2 \left\|x - \overline{y}(\mathcal{C}^*)\right\|^2 \le (1+\varepsilon)F(\mathcal{C}^*).$$

The resulting estimate means that the subset $\mathcal{B}^x$ is a $(1+\varepsilon)$-approximate solution of Problem 1. The lemma is proved.

3. APPROXIMATION ALGORITHM

The main idea of the proposed approach to the search for an approximate solution of the problem is as follows. For each point of the input set, we construct a domain (of cubic form) so that at least one of these domains would necessarily contain the centroid of the desired subset. Given a required relative error of the solution in the input, a lattice is constructed that discretizes this domain with a uniform step in all the coordinates. The lattice size and step are calculated adaptively (see below) for each of the input points. For each node of the lattice, we form a subset of M points of the input set at which function (4) has the smallest values and the minimum of the auxiliary objective function (3) is attained. The formed set is declared as a candidate for the solution. The subset of the constructed family that minimizes the objective function of Problem 1 is taken to be the final solution.

This inherently adaptive grid approach was previously used in [21] and [22] to solve related strongly NP-hard clustering problems. In these reference papers, the auxiliary objective functions differ from (3) since the structure of the optimal solutions of the above related problems is different from the structure of the optimal solution of Problem 1 stated in Lemma 4. It is these differences that determine the features of the adaptive grid computations in the algorithm proposed below. The present paper demonstrates the effectiveness of the adaptive grid approach to solving Problem 1 under consideration.

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

$$\begin{aligned}\mathcal{D}(x,h,H) = \{d \in \mathbb{R}^q \,|\, d = x + h\cdot(i_1,\ldots,i_q),\\ i_k \in \mathcal{I},\ |hi_k| \le H,\ k \in \{1,\ldots,q\}\},\end{aligned} \tag{15}$$

which is a multidimensional cubic lattice of size $2H$ centered at point x with coordinate-wise spacing h between the nodes.

Remark 1. For arbitrary points x and z from $\mathbb{R}^q$ such that $\|z - x\| \le H$, the distance from the point z to the nearest node of the lattice $|\mathcal{D}(x,h,H+h/2)|$ obviously does not exceed $\dfrac{h\sqrt{q}}{2}$.

The cardinality of the lattice satisfies the estimate

$$|\mathcal{D}(x,h,H+h/2)| \le \left(2\left\lfloor\frac{H+h/2}{h}\right\rfloor + 1\right)^q \le \left(2\frac{H}{h} + 2\right)^q \tag{16}$$

for each point $x \in \mathbb{R}^q$.

To construct an algorithmic solution, we should adaptively determine the half-size H of the lattice and its step h for each point y of the input set $\mathcal{Y}$ so that the domain of the lattice would necessarily contain the centroid of the desired subset, while the step of the lattice is defined by the given relative error ε. For this reason, we define the functions

$$H(y) = \frac{1}{M}\sqrt{F(\mathcal{B}^y)}, \quad y \in \mathcal{Y}, \tag{17}$$

$$h(y,\varepsilon) = \frac{1}{M}\sqrt{\frac{2\varepsilon}{q}F(\mathcal{B}^y)}, \quad y \in \mathcal{Y}, \quad \varepsilon \in \mathbb{R}_+, \tag{18}$$

where the set $\mathcal{B}^y$ is defined in Lemma 4 for $x = y$.

Note that all calculations in the algorithm described below are based on constructing a feasible (approximating) solution of Problem 1 in the form of a subset $\mathcal{B}^x$ (defined in Lemma 4) for a certain point x of the support set of points. As a support set, we consider the input set $\mathcal{Y}$ and the set of nodes of the lattice $\mathcal{D}(y,h,H+h/2)$ centered at point y, which is adaptively calculated by (17) and (18) for each input point $y \in \mathcal{Y}$. The accuracy of the found solution is estimated using the basic statements in Section 2.

Remark 2. For an arbitrary point $y \in \mathcal{Y}$, the cardinality $|\mathcal{D}(y,h,H+h/2)|$ of the lattice does not exceed the quantity

$$L = \left(\sqrt{\frac{2q}{\varepsilon}} + 2\right)^q$$

by (16), (17), and (18).

Let us give the step-by-step description of the algorithm.

Algorithm $\mathcal{A}$.

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

For each point $y \in \mathcal{Y}$, perform Steps 1–6.

Step 1. Compute the values $g^y(z)$, $z \in \mathcal{Y}$, by formula (4); find the subset $\mathcal{B}^y \subseteq \mathcal{Y}$ with the M smallest values $g^y(z)$, and compute the value $F(\mathcal{B}^y)$ by formula (1).

Step 2. If $F(\mathcal{B}^y) = 0$, then put $\mathcal{C}_{\mathcal{A}} = \mathcal{B}^y$; Exit.

Step 3. Compute H and h by formulae (17) and (18), respectively.

Step 4. Construct the lattice $\mathcal{D}(y,h,H+h/2)$ by formula (15).

Step 5. For each node x of the lattice $\mathcal{D}(y, h, H + h/2)$, compute the values $g^x(y)$, $y \in \mathcal{Y}$, by formula (4) and find the subset $\mathcal{B}^x \subseteq \mathcal{Y}$ with the M smallest values $g^x(y)$. Compute the value $F(\mathcal{B}^x)$ by formula (1) and remember this value and the set $\mathcal{B}^x$.

Step 6. If $F(\mathcal{B}^x) = 0$, then put $\mathcal{C}_{\mathcal{A}} = \mathcal{B}^x$; Exit.

Step 7. In the family $\{\mathcal{B}^x | x \in \mathcal{D}(y, h, H + h/2), y \in \mathcal{Y}\}$ of feasible sets constructed at Steps 1–6, choose as the solution $\mathcal{C}_{\mathcal{A}}$ the set $\mathcal{B}^x$ for which $F(\mathcal{B}^x)$ is minimal.

Output of the algorithm: the set $\mathcal{C}_{\mathcal{A}}$.

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

Proof. Let us estimate the accuracy of the algorithm. Consider the case when the condition $F(\mathcal{B}^y) = 0$ holds at Step 2 for some input point $y \in \mathcal{Y}$. In this case, the subset $\mathcal{B}^y \subseteq \mathcal{Y}$ is an optimal solution of Problem 1, since, for any set $\mathcal{C} \subseteq \mathcal{Y}$, we have $F(\mathcal{C}) \geq 0$. We obtain a similar optimal result at Step 6.

Consider the case when $F(\mathcal{B}^y) = 0$ is not fulfilled at Step 2. It is obvious that, while running, the algorithm finds a point among all points of the set $\mathcal{Y}$ that is nearest to the centroid of the optimal set, i.e., a point $t \in \mathcal{Y}$ such that $t = \arg\min\limits_{y \in \mathcal{C}^*} \|y - \bar{y}(\mathcal{C}^*)\|$. By Lemma 6, inequality (7) holds for this point. This means, according to (17), that $\|t - \bar{y}(\mathcal{C}^*)\| \leq H(t)$; i.e., the centroid of the optimal subset lies within the lattice with the edge $2H(t)$ and the center at the point t.

Let $x^* = \arg\min\limits_{x \in \mathcal{D}(t, h, H + h/2)} \|x - \bar{y}(\mathcal{C}^*)\|$ be a node of the lattice $\mathcal{D}(t, h, H + h/2)$, nearest to the centroid of the optimal subset. According to Remark 1, the squared distance from the optimal centroid $\bar{y}(\mathcal{C}^*)$ to the nearest node x^* of the lattice does not exceed $\frac{h^2 q}{4}$. Therefore, by the definition of the point x^*, we obtain

$$\|x^* - \bar{y}(\mathcal{C}^*)\|^2 \leq \frac{h^2 q}{4} = \frac{\varepsilon}{2M^2} F(\mathcal{B}^t).$$

This estimate means that the point x^* meets the conditions of Lemma 7 and the corresponding subset $\mathcal{B}^{x^*}$ is an $(1 + \varepsilon)$-approximate solution of Problem 1.

It is clear that any subset $\mathcal{B}^x$ in the family of feasible solutions constructed by the algorithm for a node x such that $\|x - \bar{y}(\mathcal{C}^*)\|^2 \leq \|x^* - \bar{y}(\mathcal{C}^*)\|^2$ also guarantees an approximation with relative error not exceeding ε.

Let us evaluate the time complexity of the algorithm.

At Step 1, it takes at most $\mathbb{O}(qN)$ operations to compute the values $g^y(z)$. Finding the M smallest elements in the set of N elements requires $\mathbb{O}(N)$ operations (for example, by using the algorithm for finding the Mth smallest value in a unordered array (heap) [24]). The computation time of the value $F(\mathcal{B}^y)$ is $\mathbb{O}(qN)$.

Steps 2, 3, and 6 are executed in a constant time $O(1)$. The construction of the lattice at Step 4 requires $\mathbb{O}(qL)$ operations (by Remark 2).

At Step 5, the computation of the elements of the set $\mathcal{B}^x$ for each node x of the lattice is done in $\mathbb{O}(qN)$ time, and the same is true for the computation of the value $F(\mathcal{B}^x)$ (by analogy with the computations at Step 1). Therefore, the total computational time for all nodes of the lattice at this step equals $\mathbb{O}(qNL)$.

Since Steps 1–6 are executed N times, the time complexity of these steps is $\mathbb{O}(qN^2L)$. The time complexity of Step 7 is estimated by $\mathbb{O}(NL)$, while the total costs of all steps are equal to $\mathbb{O}(qN^2L)$. Therefore, the time complexity of the algorithm $\mathcal{A}$ equals $\mathbb{O}\left(qN^2\left(\sqrt{\frac{2q}{\varepsilon}} + 2\right)^q\right)$. The theorem is proved.

Remark 3. In the case when the dimension q of the space is bounded by a constant and $\varepsilon < 2q$, we have the estimate

$$qN^2\left(2 + \sqrt{\frac{2q}{\varepsilon}}\right)^q \leq qN^2 3^q \left(\frac{2q}{\varepsilon}\right)^{q/2} = \mathbb{O}\left(N^2\left(\frac{1}{\varepsilon}\right)^{q/2}\right),$$

which means that Algorithm $\mathcal{A}$ implements an FPTAS.

Remark 4. It is clear that the constructed algorithm can be applied to solve a problem in which the cardinalities of the desired clusters are unknown. For this purpose, it suffices to solve Problem 1 N times with the help of Algorithm $\mathcal{A}$ for each $M = 1, \ldots, N$ and then choose the best of the found solutions according to the minimum of the objective function. The time complexity of this algorithmic solution is obviously equal to $\mathbb{O}\left(N^3\left(\frac{1}{\varepsilon}\right)^{q/2}\right)$. However, it is interesting to construct less time-consuming algorithms without searching N admissible values of the cardinality $|\mathcal{C}|$ of the desired subset.

CONCLUSIONS

In the paper, we have constructed an approximation algorithm for the strongly NP-hard quadratic Euclidean problem of partitioning a finite set of points into two clusters. It has been shown that the algorithm is a fully polynomial-time approximation scheme if the dimension of the space is bounded by a constant.

The considered problem is among the discrete optimization problems poorly studied in the algorithmical sense. Therefore, it seems relevant to continue the study of the algorithmic approximability of the problem. An important direction is the construction of fast randomized algorithms that provide a solution with probabilistic guarantees in linear and sublinear time.

ACKNOWLEDGMENTS

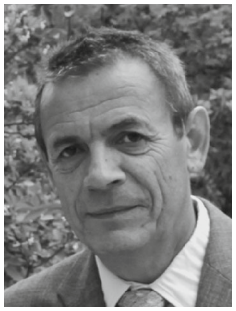
The authors are grateful to V.V. Shenmaier for valuable comments and suggestions on the strengthening and generalization of the obtained result.

This work was supported by the Russian Foundation for Basic Research, project nos. 15-01-00462, 16-07-00168, 16-31-00186-mol-a, and 18-31-00398-mol-a.

REFERENCES

1. A. V. Kel'manov and A. V. Motkova, "Exact pseudo-polynomial algorithms for a balanced 2-clustering problem," J. Appl. Indust. Math. **10** (3), 349–355 (2016).
2. A. V. Kel'manov and A. V. Pyatkin, "NP-hardness of some quadratic Euclidean 2-clustering problems," Dokl. Math. **92** (2), 634–637 (2015).
3. A. V. Kel'manov and A. V. Pyatkin, "On the complexity of some quadratic Euclidean 2-clustering problems," Comput. Math. Math. Phys. **56** (3), 491–497 (2015).
4. C. C. Aggarwal, *Data Mining: The Textbook* (Springer International Publishing, 2015).
5. C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer Science + Business Media, New York, 2006).
6. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (John Wiley & Sons, New York, 1973; Mir, Moscow, 1976).
7. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. (Academic Press, New York, 1990).
8. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer-Verlag, New York, 2009).
9. A. K. Jain, "Data clustering: 50 years beyond k-means," Pattern Recogn. Lett. **31** (8), 651–666 (2010).
10. G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Application in R* (Springer Science + Business Media, New York, 2013).
11. T.-C. Fu, "A review on time series data mining," Eng. Appl. Artif. Intell. **24** (1), 164–181 (2011).
12. J. T. Tou and R. C. Gonzalez, *Pattern Recognition Principles* (Addison-Wesley, Reading, MA, 1974; Mir, Moscow, 1978).
13. P. Brucker, "On the complexity of clustering problems," in *Optimization and Operations Research: Proc. of the Workshop Held at University Bonn (Bonn, Germany, October 2–8, 1977),* Lecture Notes Econom. Math. Syst. **157**, 45–54 (1978).
14. S. Sahni and T. Gonzalez, "P-complete approximation problems," J. ACM. **23** 555–566 (1976).
15. S. Hasegawa, H. Imai, M. Inaba, N. Katoh, and J. Nakano, "Efficient algorithms for variance-based k-Clustering," in *Proc. 1st Pacific Conf. on Computer Graphics and Applications (Seoul, Korea, August 30–September 2, 1993)*, Vol. 1 (World Scientific, River Edge, NJ, 1993), pp. 75–89.
16. M. Inaba, N. Katoh, and H. Imai, "Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering: (extended abstract)," in *Proc. 10th ACM Symposium on Computational Geometry (Stony Brook, New York, USA, June 6–8, 1994)*, (ACM, New York, 1994), pp. 332–339.
17. F. de la Vega, M. Karpinski, C. Kenyon, and Y. Rabani, "Polynomial time approximation schemes for metric min-sum clustering," *Electronic Colloquium on Computational Complexity* (ECCC), Report No. 25 (2002).
18. F. de la Vega and C. Kenyon, "A randomized approximation scheme for metric max-cut," J. Comput. Syst. Sci. **63**, 531–541 (2001).
19. M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness* (W. H. Freeman and Co., San Francisco, 1979; Mir, Moscow, 1982).
20. A. V. Kel'manov and S. M. Romanchenko, "An approximation algorithm for solving a problem of search for a vector subset," J. Appl. Indust. Math. **6** (1), 90–96 (2012).
21. A. V. Kel'manov and S. M. Romanchenko, "An FPTAS for a vector subset search problem," J. Appl. Indust. Math. **8** (3), 329–336 (2014).
22. A. V. Kel'manov and V. I. Khandeev, "Fully polynomial-time approximation scheme for a special case of a quadratic Euclidean 2-clustering problem," Comput. Math. Math. Phys. **56** (2), 334–341 (2016).
23. A. V. Kel'manov and A. V. Motkova, "A fully polynomial-time approximation scheme for a special case of a balanced 2-clustering problem," in *Discrete Optimization and Operations Research: Proc. 9th Intern. Conf. DOOR 2016 (Vladivostok, Russia, September 19–23, 2016),* Lecture Notes Comp. Sci. **9869**, 182–192 (2016).
24. N. Wirth, *Algorithms + Data Structures = Programs* (Prentice Hall, New Jersey, 1976; Mir, Moscow, 1985).

Translated by I. Tselishcheva



Alexander Vasilyevich Kel'manov. Born 1952. Graduated from Izhevsk State Technical University in 1974 with specialty in Applied Mathematics. Received Candidate's Degree in Engineering Cybernetics and Information Theory in 1980 and Doctor of Sciences degree in Physics and Mathematics in 1994. Currently with Sobolev Institute of Mathematics, Siberian Branch of the Russian Academy of Sciences, head of Data Analysis Laboratory. Scientific interests: data analysis, data mining, pattern recognition, clusterization, discrete optimization, NP-hard problems, efficient algorithms with performance guarantees. Author of more than 200 publications.

Anna Vladimirovna Motkova. Born 1993. Graduated from Novosibirsk State University in 2017 with specialty in Mathematics. Currently with Sobolev Institute of Mathematics, Siberian Branch of the Russian Academy of Sciences, Ph.D. student in Data Analysis Laboratory. Scientific interests: data analysis, pattern recognition, clustering, discrete optimization, NP-hard problems, efficient algorithms with performance guarantees. Author of 8 publications.