A Fully Polynomial-Time Approximation Scheme for a Special Case of a Balanced 2-Clustering Problem

Alexander Kel'manov[1,2(✉)] and Anna Motkova[2(✉)]

[1] Sobolev Institute of Mathematics, 4 Koptyug Avenue, 630090 Novosibirsk, Russia
kelm@math.nsc.ru
[2] Novosibirsk State University, 2 Pirogova Street, 630090 Novosibirsk, Russia
anitamo@mail.ru

Abstract. We consider the strongly NP-hard problem of partitioning a set of Euclidean points into two clusters so as to minimize the sum (over both clusters) of the weighted sum of the squared intracluster distances from the elements of the clusters to their centers. The weights of sums are 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 determined as the geometric center (centroid), i.e. the average value over all points in the cluster. We analyze the variant of the problem with cardinality constraints. We present an approximation algorithm for the problem and prove that it is a fully polynomial-time approximation scheme when the space dimension is bounded by a constant.

Keywords: NP-hardness · Euclidian space · Fixed dimension · FPTAS

1 Introduction

The subject of this study is a strongly NP-hard quadratic Euclidean problem of partitioning a finite set of points into two clusters. We will show a fully polynomial-time approximation scheme (FPTAS) for a special case of the problem.

Our research is motivated by insufficient study of the problem from an algorithmic direction and its importance in some applications including geometry, cluster analysis, statistical problems of joint evaluation and hypotheses testing with heterogeneous samples, data interpretation problem, etc.

The paper has the following structure. Section 2 contains the problem formulation, some applications, and some closely related problems. Additionally, known and our new results are discussed. In Sect. 3 we formulate and prove some basic properties exploited by our algorithm. In Sect. 4, an approximation algorithm is presented. Finally, also in Sect. 4 we show that our algorithm is a fully polynomial-time approximation scheme when the space dimension is fixed.

Y. Kochetov et al. (Eds.): DOOR 2016, LNCS 9869, pp. 182–192, 2016.
DOI: 10.1007/978-3-319-44914-2_15

2 Problem Formulation, Its Origin, Related Problems, known and New Results

Everywhere below we use the standard notations, namely: $\mathbb{R}$ is the set of the 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 also [1,2]).

Problem 1 (*Balanced 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 $\mathcal{Y}$ into two non-empty clusters $\mathcal{C}$ and $\mathcal{Y} \setminus \mathcal{C}$ such that

$$F(\mathcal{C}) = |\mathcal{C}| \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + |\mathcal{Y} \setminus \mathcal{C}| \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \longrightarrow \min, \tag{1}$$

where $\overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum\limits_{y \in \mathcal{C}} y$ is the geometric center (centroid) of $\mathcal{C}$ and such that $|\mathcal{C}| = M$.

The problem has an obvious geometrical interpretation. It is a partition of a finite set of points in Euclidean space into two geometrical structures minimizing (1). In formula (1) the weights of the sums are the cardinalities of the desired clusters. So, Problem 1 can be interpreted as the problem of optimal weighted (by the cardinalities of the clusters) summing and also as a problem of balanced partitioning (or clustering).

In addition, the problem has applications in Data mining problem (see, for example, [3–5]). The essence of this multifaceted problem is the approximation of data by some mathematical model that allows to plausibly explain the origin of the data in terms of the model. In particular, the next statistical hypothesis can be used as such mathematical model: it is true that the input data $\mathcal{Y}$ is the inhomogeneous sample from two distributions, and that one of these distributions has zero mean while another mean is unknown and non-equal to zero. To test this hypothesis, first we need to find an optimal solution to Problem 1, and only then we will be able to use the classical results in the field of statistical hypothesis testing.

It is widely known that applied researchers, who study and analyze data, use algorithms as the basic mathematical tools for solving a variety of clustering problems in which clusters consist of similar or related by certain criteria objects. Creating such mathematical tools for solving data mining problems causes the development of new algorithms with guaranteed performance estimates of accuracy and time complexity.

The strong NP-hardness of Problem 1 was proved in [1,2]. This fact implies that, unless P=NP, there are neither exact polynomial-time nor exact pseudopolynomial-time algorithms for it [6]. In addition, in [1,2], the nonexistence of an FPTAS was shown (unless P=NP) for Problem 1. So, finding subclasses of this problem for which there exists an FPTAS is a question of topical interest.

Note that there is only one algorithmic result for Problem 1, i.e. an exact algorithm [7] for the case of integer components of the input points. The time complexity of this algorithm is $\mathcal{O}(qN(2MB+1)^q)$, where B is the maximum absolute value of the components of the input points. If the dimension q of the space is bounded by a constant, then the time complexity of the algorithm is $\mathcal{O}(N(MB)^q)$. So, in this case the algorithm is pseudopolynomial.

At the same time, there are a lot of results for problems closely related to Problem 1. Properties of algorithms for these problems can be found in the papers cited below.

The NP-hard *Balanced variance-based 2-clustering* problem is one of the most closely related to Problem 1. The objective function in this problem is different from (1) in that the center of cluster $\mathcal{Y} \setminus \mathcal{C}$ is not fixed:

$$|\mathcal{C}| \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + |\mathcal{Y} \setminus \mathcal{C}| \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y - \overline{y}(\mathcal{Y} \setminus \mathcal{C})\|^2 \longrightarrow \min . \tag{2}$$

In problem (2) the centroids $\overline{y}(\mathcal{C})$ and $\overline{y}(\mathcal{Y} \setminus \mathcal{C})$ of both clusters $\mathcal{C}$ and $\mathcal{Y} \setminus \mathcal{C}$ are the functions of $\mathcal{C}$. It is well-known that this problem is equivalent to *Min-sum all-pairs 2-clustering* problem in which it is required to find a partition such that

$$\sum_{x \in \mathcal{C}} \sum_{z \in \mathcal{C}} \|x - z\|^2 + \sum_{x \in \mathcal{Y} \setminus \mathcal{C}} \sum_{z \in \mathcal{Y} \setminus \mathcal{C}} \|x - z\|^2 \longrightarrow \min . \tag{3}$$

Algorithmic questions for problems (2) and (3) were studied, for example, in [1,2,8–13].

The well-known NP-hard [14] *Minimum sum-of-squares 2-clustering* problem is close to Problem 1. In this problem (related to classical work by Fisher [15] and also called 2-*Means* [16]), we need to find two clusters $\mathcal{C}$ and $\mathcal{Y} \setminus \mathcal{C}$ such that

$$\sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y - \overline{y}(\mathcal{Y} \setminus \mathcal{C})\|^2 \longrightarrow \min . \tag{4}$$

In problem (4) as well as in problem (2) the centroids of both clusters are the functions of $\mathcal{C}$, but in problem (4) the sums are not weighted by the cluster cardinalities. Thousands of publications are dedicated to problem (4) and its applications.

The strongly NP-hard problem *Minimum sum-of-squares 2-clustering with given center* has been actively studied in the last decade. In this problem we need to find a 2-partition such that

$$\sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \longrightarrow \min . \tag{5}$$

Problem (5) differs from Problem 1 in that the sums are not weighted by the cardinalities of the desired clusters. The algorithmic results for this problem can be found in [17–24].

In the considered Problem 1, the centroid $\overline{y}(\mathcal{C})$ of the cluster $\mathcal{C}$ is unknown and the center of the cluster $\mathcal{Y} \setminus \mathcal{C}$ is given at the origin as in the problem (5).

Since Problem 1 is neither equivalent nor a special case of the problems (2)–(5), the previous algorithmic results for these closely related problems do not apply to Problem 1. We need new explorations for this problem.

In this work we present an approximation algorithm for Problem 1. Given a relative error ε, the algorithm finds a $(1+\varepsilon)$-approximate solution in $\mathcal{O}\left(qN^2(\sqrt{\frac{2q}{\varepsilon}}+1)^q\right)$ time. In the case of a fixed space dimension q the running time of the algorithm is equal to $\mathcal{O}\left(N^2\left(\frac{1}{\varepsilon}\right)^{q/2}\right)$ and so, it implements a fully polynomial-time approximation scheme.

3 Foundations of the Algorithm

In this section, we provide some basic statements exploited by our algorithm.

The following two lemmas are well known. Their proofs are presented in many publications (see, for example, [25,26]).

Lemma 1. *For an arbitrary point $x \in \mathbb{R}^q$ and a finite set $\mathcal{Z} \subset \mathbb{R}^q$, it is true that*

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

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

Lemma 2. *Let the conditions of Lemma 1 hold. If a point $u \in \mathbb{R}^q$ is closer (in terms of distance) to the centroid $\overline{z}$ of $\mathcal{Z}$ than any point in $\mathcal{Z}$, then*

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

Lemma 3. *Let*

$$S(\mathcal{C},x) = |\mathcal{C}|\sum_{y\in\mathcal{C}}\|y-x\|^2 + |\mathcal{Y}\setminus\mathcal{C}|\sum_{y\in\mathcal{Y}\setminus\mathcal{C}}\|y\|^2,\ \mathcal{C}\subseteq\mathcal{Y},\ x\in\mathbb{R}^q, \tag{6}$$

where $\mathcal{Y}$ is the input set of Problem 1. Then it is true that

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

Proof. Applying Lemma 1 to the set $\mathcal{C}$ and its centroid, we have

$$\sum_{y\in\mathcal{C}}\|y-x\|^2 = \sum_{y\in\mathcal{C}}\|y-\overline{y}(\mathcal{C})\|^2 + |\mathcal{C}|\cdot\|x-\overline{y}(\mathcal{C})\|^2 . \tag{7}$$

After the substitution of (7) in the definition (6), we obtain

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

□

For any function $f(x, y)$, we denote by $f^x(y)$ the function when the argument x is fixed and by $f^y(x)$ the function when the argument y is fixed.

Lemma 4. *For the conditional minimums of the function (6) the next statements are true:*

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

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

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

where

$$G^x(\mathcal{C}) = \sum_{y \in \mathcal{C}} g^x(y) \,, \tag{8}$$

$$g^x(y) = (2M - N)\|y\|^2 - 2M \langle y, x \rangle \,, \quad y \in \mathcal{Y} \,, \tag{9}$$

and

$$\min_{\mathcal{C} \subseteq \mathcal{Y}} G^x(\mathcal{C}) = \sum_{y \in \mathcal{B}^x} g^x(y) \,, \tag{10}$$

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

Proof. The first statement follows from Lemma 3.

Since $|\mathcal{Y}| = N$ and $|\mathcal{C}| = M$, the second statement follows from the next chain of equalities:

$$\begin{aligned} S^x(\mathcal{C}) &= M \sum_{y \in \mathcal{C}} \|y - x\|^2 + (N - M) \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \\ &= M \sum_{y \in \mathcal{C}} \|y\|^2 + M^2 \|x\|^2 - 2M \sum_{y \in \mathcal{C}} \langle y, x \rangle + (N - M) \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \\ &= (N - M) \sum_{y \in \mathcal{Y}} \|y\|^2 + M^2 \|x\|^2 + (2M - N) \sum_{y \in \mathcal{C}} \|y\|^2 - 2M \sum_{y \in \mathcal{C}} \langle y, x \rangle \\ &= (N - M) \sum_{y \in \mathcal{Y}} \|y\|^2 + M^2 \|x\|^2 + \sum_{y \in \mathcal{C}} g^x(y) = (N - M) \sum_{y \in \mathcal{Y}} \|y\|^2 + M^2 \|x\|^2 + G^x(\mathcal{C}) \,. \end{aligned}$$

It remains to note that in the last two equalities the first two addends do not depend on $\mathcal{C}$. The formula (10) is obvious. □

Lemma 5. *Let the conditions of Lemma 4 hold and $\mathcal{C}^*$ be the optimal solution of Problem 1. Then, for a fixed point $x \in \mathbb{R}^q$, the following inequality is true*

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

Proof. The definitions (1) and (6), and Lemma 4 imply

$$F(\mathcal{B}^x) = S^{\overline{y}(\mathcal{B}^x)}(\mathcal{B}^x) \le S^x(\mathcal{B}^x) \le S^x(\mathcal{C}^*)\,. \tag{11}$$

Applying Lemma 3 to the right-hand side of (11), we obtain

$$S^x(\mathcal{C}^*) = F(\mathcal{C}^*) + M^2||x - \overline{y}(\mathcal{C}^*)||^2\,. \tag{12}$$

Combining (11) and (12) yields the statement of the lemma. □

Lemma 6. *Let the conditions of Lemma 5 hold and* $t = \arg\min\limits_{y\in\mathcal{C}^*} \|y - \overline{y}(\mathcal{C}^*)\|^2$ *be the point from the subset* $\mathcal{C}^*$ *closest to its centroid. Then the following inequality is true*

$$\|t - \overline{y}(\mathcal{C}^*)\|^2 \le \frac{1}{M^2} F(\mathcal{B}^t)\,, \tag{13}$$

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

Proof. By the definition of point t we have

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

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

$$M\|t - \overline{y}(\mathcal{C}^*)\|^2 \le \sum_{y\in\mathcal{C}^*} \|y - \overline{y}(\mathcal{C}^*)\|^2\,. \tag{14}$$

Since $\mathcal{C}^*$ is the optimal solution,

$$F(\mathcal{C}^*) \le F(\mathcal{B}^t)\,. \tag{15}$$

Then (14), (1) and (15) imply

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

□

Lemma 7. *Let the conditions of Lemma 6 hold. Let*

$$\|x - \overline{y}(\mathcal{C}^*)\|^2 \le \frac{\varepsilon}{2M^2} F(\mathcal{B}^t) \tag{16}$$

for some $\varepsilon > 0$ *and* $x \in \mathbb{R}^q$*. Then the subset* $\mathcal{B}^x$ *(defined in Lemma 4) is a* $(1+\varepsilon)$*-approximate solution of Problem 1.*

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

$$F(\mathcal{B}^t) = S^{\overline{y}(\mathcal{B}^t)}(\mathcal{B}^t) \le S^t(\mathcal{B}^t) \le S^t(\mathcal{C}^*)\,. \tag{17}$$

Applying Lemma 2 to the set $\mathcal{C}^*$ and the point t, we have

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

Therefore, definition (6) yields

$$S^t(\mathcal{C}^*) = M \sum_{y \in \mathcal{C}^*} \|y - t\|^2 + (N - M) \sum_{y \in \mathcal{Y} \setminus \mathcal{C}^*} \|y\|^2$$
$$\leq 2M \sum_{y \in \mathcal{C}^*} \|y - \overline{y}(\mathcal{C}^*)\|^2 + (N - M) \sum_{y \in \mathcal{Y} \setminus \mathcal{C}^*} \|y\|^2 \leq 2F(\mathcal{C}^*)\,. \quad (18)$$

Combining (16), (17), and (18) we obtain

$$\|x - \overline{y}(\mathcal{C}^*)\|^2 \leq \frac{\varepsilon}{2M^2} F(\mathcal{B}^t) \leq \frac{\varepsilon}{2M^2} S^t(\mathcal{C}^*) \leq \frac{\varepsilon}{M^2} F(\mathcal{C}^*)\,. \quad (19)$$

Finally, from Lemma 5 and (19) for the subset $\mathcal{B}^x$ we obtain the following estimate of the value of the objective function

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

This estimate means that the subset $\mathcal{B}^x$ is a $(1 + \varepsilon)$-approximate solution for Problem 1. □

4 Approximation Algorithm

In this section, we present our approximation algorithm for Problem 1. Its main idea is 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 ε of the solution, a lattice (a grid) is generated that discretizes the cube with a uniform step in all coordinates. For each lattice node, a subset of M points from the input set that have the smallest values of the function (9) is formed (the minimum of (8) is reached at that subset). The resulting set is declared as a solution candidate. The candidate that minimizes the objective function is chosen to be the final solution.

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

$$\mathcal{D}(x, h, H) = \{d \in \mathbb{R}^q \,|\, d = x + h \cdot (i_1, \ldots, i_q), i_k \in \mathbb{Z}, |hi_k| \leq H, k \in \{1, \ldots, q\}\} \quad (20)$$

which is a cubic lattice of size $2H$ centered at the point x with node spacing h.

For any point $x \in \mathbb{R}^q$ the number of nodes in this lattice is

$$|\mathcal{D}(x, h, H)| \leq \left(2\left\lfloor \frac{H}{h} \right\rfloor + 1\right)^q \leq \left(2\frac{H}{h} + 1\right)^q . \quad (21)$$

Remark 1. If some point z from $\mathbb{R}^q$ and some node x from the lattice $\mathcal{D}(x, h, H)$ satisfy the inequality $\|z - x\| \leq H$ then the distance from z to the nearest node of the lattice obviously does not exceed $\frac{h\sqrt{q}}{2}$.

For constructing an algotithmic solution we need to determine adaptively the size H of the lattice and its node spacing h for each point y of the input set $\mathcal{Y}$ so that the domain of the lattice contains the centroid of the desired subset. The node spacing is defined by the relative error ε. To this end we define the functions:

$$H(y) = \frac{1}{M}\sqrt{F(\mathcal{B}^y)}, \quad y \in \mathcal{Y}\,, \tag{22}$$

$$h(y,\varepsilon) = \frac{1}{M}\sqrt{\frac{2\varepsilon}{q}F(\mathcal{B}^y)}, \quad y \in \mathcal{Y},\ \varepsilon \in \mathbb{R}_+\,, \tag{23}$$

where $\mathcal{B}^y$ is a set determined in Lemma 4, if $x = y$.

Note that all calculations in the algorithm described below are based on constructing candidate (approximate) solutions of Problem 1 as a subset $\mathcal{B}^x$ (defined in Lemma 4) for any point x from the support set of points. In this way we use two support sets. The first of them is the input set $\mathcal{Y}$ and the second one is the set of nodes of the lattice $\mathcal{D}(y,h,H)$ centered at y. The lattice is adaptively calculated by formulae (22) and (23) for each input point $y \in \mathcal{Y}$. The approximation factor is finally bounded using the basic statements in Sect. 3.

Remark 2. For any point $y \in \mathcal{Y}$ the cardinality $|\mathcal{D}(y,h,H)|$ of the lattice does not exceed the value

$$L = \left(\sqrt{\frac{2q}{\varepsilon}} + 1\right)^q$$

due to (21), (22), and (23).

Below is the step-by-step description of the algorithm.

Algorithm $\mathcal{A}$.

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

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

Step 1. Compute the values $g^y(z)$, $z \in \mathcal{Y}$, using formula (9); find a subset $\mathcal{B}^y \subseteq \mathcal{Y}$ with M smallest values $g^y(z)$, compute $F(\mathcal{B}^y)$ using 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 using formulae (22) and (23).

Step 4. Construct the lattice $\mathcal{D}(y,h,H)$ using formula (20).

Step 5. For each node x of the lattice $\mathcal{D}(y,h,H)$ compute the values $g^x(y)$, $y \in \mathcal{Y}$, using formula (9) and find a subset $\mathcal{B}^x \subseteq \mathcal{Y}$ with M smallest values $g^x(y)$. Compute $F(\mathcal{B}^x)$ using formula (1), 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 $\{\mathcal{B}^x | \, x \in \mathcal{D}(y,h,H), y \in \mathcal{Y}\}$ of candidate sets that have been constructed in Steps 1–6, choose as a solution $\mathcal{C}_\mathcal{A}$ the set $\mathcal{B}^x$ for which $F(\mathcal{B}^x)$ is minimal.

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

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

Proof. Let us bound the approximation factor of the algorithm. If the equality $F(\mathcal{B}^y) = 0$ holds at Step 2 for some point $y \in \mathcal{Y}$, then the subset $\mathcal{B}^y \subseteq \mathcal{Y}$ is an optimal solution of Problem 1, since, for any set $\mathcal{C} \subseteq \mathcal{Y}$, it is true that $F(\mathcal{C}) \geq 0$. We get an optimal solution at Step 6 in the same way.

Consider the case when the condition $F(\mathcal{B}^y) = 0$ at Step 2 does not hold. Obviously, there exists a point $t \in \mathcal{Y}$ such that $t = \arg\min\limits_{y \in C^*} ||y - \overline{y}(C^*)||$ and the algorithm meets it at least once in the set $\mathcal{Y}$ while running. By Lemma 6, inequality (13) holds for this point. This inequality and (22) mean that $\|t - \overline{y}(\mathcal{C}^*)\| \leq H(t)$, so the centroid of the optimal subset lies within the lattice $D(t, h, H)$ of the size $H = H(t)$ and the node spacing $h = h(t, \varepsilon)$.

Let $x^* = \arg\min\limits_{x \in \mathcal{D}(t,h,H)} \|x - \overline{y}(\mathcal{C}^*)\|$ be a node of the grid $\mathcal{D}(t, h, H)$, the nearest to the centroid of the optimal subset. Since the squared distance from the optimal centroid $\overline{y}(\mathcal{C}^*)$ to the nearest node x^* of the lattice does not exceed $\frac{h^2 q}{4}$ (by remark 1), we have the estimate

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

Therefore, the point x^* satisfies the conditions of Lemma 7 and, hence, the set $\mathcal{B}^{x^*}$ is a $(1+\varepsilon)$-approximate solution of Problem 1.

It is clear, that any subset $\mathcal{B}^x$ in the family of candidate solutions on Step 7 constructed for node x such that $\|x - \overline{y}(C^*)\|^2 \leq \|x^* - \overline{y}(C^*)\|^2$ guarantees a $(1+\varepsilon)$-approximation also.

Let us evaluate the time complexity of the algorithm.

At Step 1 calculation of $g^y(z)$ requires at most $\mathcal{O}(qN)$-time. Finding the M smallest elements in the set of N elements is performed in $\mathcal{O}(N)$ operations (for example, using the algorithm of finding the n-th smallest value in an unordered array [27]). Computation of the value $F(\mathcal{B}^y)$ takes $\mathcal{O}(qN)$ time.

Steps 2, 3 and 6 are executed in $\mathcal{O}(1)$ operations. It requires $\mathcal{O}(qL)$ operations for generating the lattice at Step 4 (by remark 2).

At Step 5, computation of the elements of the set $\mathcal{B}^x$ for each node of the grid requires $\mathcal{O}(qN)$ time, and the same is true for the computation of $F(\mathcal{B}^x)$ (as computations at Step 1). Thus, at this step the computational time for all nodes of the grid is $\mathcal{O}(qNL)$.

Since Steps 1–6 are performed N times, the time complexity of these steps is $\mathcal{O}(qN^2L)$. The time complexity of Step 7 is bounded by $\mathcal{O}(NL)$, and the total time complexity of all Steps is $\mathcal{O}(qN^2L)$. Therefore, the time complexity of Algorithm $\mathcal{A}$ is $\mathcal{O}\left(qN^2\left(\sqrt{\frac{2q}{\varepsilon}}+1\right)^q\right)$. □

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

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

and it means that Algorithm $\mathcal{A}$ is an FPTAS.

Remark 4. It is clear that the constructed algorithm can be applied for solving a problem in which the cardinalities of the clusters are the optimized variables. For this purpose, it is sufficient 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 these solutions in the sense of minimizing the objective function. The time complexity of this algorithm obviously equals $\mathcal{O}\left(N^3 \left(\frac{1}{\varepsilon}\right)^{q/2}\right)$. But it is interesting to construct algorithms with less time complexity without searching for such candidate solutions.

5 Conclusion

In this paper we presented an approximation algorithm for one strongly NP-hard quadratic Euclidian problem of balanced partitioning a finite set of points into two clusters. It was proved that our algorithm is a fully polynomial-time approximation scheme if the space dimension is bounded by a constant.

In the algorithmical sense, the considered problem is poorly studied. Therefore, it seems important to continue studying the questions on algorithmical approximability of the problem.

Acknowledgments. This work was supported by the RFBR, projects 15-01-00462, 16-31-00186 and 16-07-00168.

References

1. Kel'manov, A.V., Pyatkin, A.V.: NP-hardness of some quadratic euclidean 2-clustering problems. Doklady Math. **92**(2), 634–637 (2015)
2. Kel'manov, A.V., Pyatkin, A.V.: On the complexity of some quadratic euclidean 2-clustering problems. Comput. Math. Math. Phys. **56**(3), 491–497 (2016)
3. Aggarwal, C.C.: Data Mining: The Textbook. Springer International Publishing, Switzerland (2015)
4. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer Science+ Business Media, LLC, New York (2006)
5. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, New York (2001)
6. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. Freeman, San Francisco (1979)
7. Kel'manov, A.V., Motkova, A.V.: An exact pseudopolynomial algorithm for a special case of a euclidean balanced variance-based 2-clustering problem. In: Abstracts of the VI International Conference "Optimization and Applications" (OPTIMA-2015), P. 98. Petrovac, Montenegro (2015)
8. Sahni, S., Gonzalez, T.: P-complete approximation problems. J. ACM **23**, 555–566 (1976)
9. Brucker, P.: On the complexity of clustering problems. Lect. Notes Econ. Math. Syst. **157**, 45–54 (1978)
10. Inaba, M., Katoh, N., Imai, H.: Applications of Weighted Voronoi Diagrams and Randomization toVariance-Based k-Clustering: (extended abstract). Stony Brook, NY, USA, pp. 332–339 (1994)

11. Hasegawa, S., Imai, H., Inaba, M., Katoh, N., Nakano, J.: Efficient algorithms for variance-based k-clustering. In: Proceedings of the 1st Pacific Conference on Computer Graphics andApplications (Pacific Graphics 1993, Seoul, Korea),World Scientific, River Edge, NJ. 1, pp. 75–89 (1993)
12. de la Vega, F., Kenyon, C.: A randomized approximation scheme for metric max-cut. J. Comput. Syst. Sci. **63**, 531–541 (2001)
13. de la Vega, F., Karpinski, M., Kenyon, C., Rabani, Y.: Polynomial Time Approximation Schemes for Metric Min-Sum Clustering. Electronic Colloquium on Computational Complexity (ECCC), 25 (2002)
14. Aloise, D., Deshpande, A., Hansen, P., Popat, P.: NP-hardness of euclidean sum-of-squares clustering. Mach. Learn. **75**(2), 245–248 (2009)
15. Fisher, R.A.: Statistical Methods and Scientific Inference. Hafner Press, New York (1956)
16. Rao, M.: Cluster analysis and mathematical programming. J. Amer. Statist. Assoc. **66**, 626–662 (1971)
17. Gimadi, E.K., Kel'manov, A.V., Kel'manova, M.A., Khamidullin, S.A.: Aposteriori finding of a quasiperiodic fragment with a given number of repetitions in a numerical sequence (in Russian). Sib. Zh. Ind. Mat. **9**(25), 55–74 (2006)
18. Gimadi, E.K., Kel'manov, A.V., Kel'manova, M.A., Khamidullin, S.A.: A posteriori detecting a quasiperiodic fragment in a numerical sequence. Pattern Recogn. Image Anal. **18**(1), 30–42 (2008)
19. Dolgushev, A.V., Kel'manov, A.V.: An approximation algorithm for solving a problem of cluster analysis. J. Appl. Indust. Math. **5**(4), 551–558 (2011)
20. Dolgushev, A.V., Kel'manov, A.V., Shenmaier, V.V.: A polynomial-time approximation scheme for a problem of partitioning a finite set into two clusters (in Russian). Trudy Inst. Mat. i Mekh. UrO. RAN. **21**(3), 100–109 (2015)
21. Kel'manov, A.V., Khandeev, V.I.: A 2-approximation polynomial algorithm for a clustering problem. J. Appl. Indust. Math. **7**(4), 515–521 (2013)
22. Kel'manov, A.V., Khandeev, V.I.: A randomized algorithm for two-cluster partition of a set of vectors. Comput. Math. Math. Phys. **55**(2), 330–339 (2015)
23. Kel'manov, A.V., Khandeev, V.I.: An exact pseudopolynomial algorithm for a problem of the two-cluster partitioning of a set of vectors. J. Appl. Indust. Math. **9**(4), 497–502 (2015)
24. Kel'manov, A.V., Khandeev, V.I.: 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)
25. Kel'manov, A.V., Romanchenko, S.M.: An approximation algorithm for solving a problem of search for a vector subset. J. Appl. Ind. Math. **6**(1), 90–96 (2012)
26. Kel'manov, A.V., Romanchenko, S.M.: An FPTAS for a vector subset search problem. J. Appl. Indust. Math. **8**(3), 329–336 (2014)
27. Wirth, N.: Algorithms + Data Structures = Programs. Prentice Hall, New Jersey (1976)