On Some Finite Set Clustering Problems in Euclidean Space

Alexander Kel'manov
Sobolev Institute of Mathematics
Acad. Koptyug avenue, 4,
630090 Novosibirsk, Russia
Novosibirsk State University
Pirogova str. 1,
630090 Novosibirsk, Russia.
kelm@math.nsc.ru

Artem Pyatkin
Sobolev Institute of Mathematics
Acad. Koptyug avenue, 4,
630090 Novosibirsk, Russia
Novosibirsk State University
Pirogova str. 1,
630090 Novosibirsk, Russia.
artem@math.nsc.ru

Abstract

Problems of partitioning a finite set of Euclidean points (vectors) into clusters are considered. The criterion is minimizing the sum over all clusters of: (1) normalized by the cardinality squared norms of the sum of cluster elements, (2) squared norms of the sum of cluster elements, (3) norms of the sum of cluster elements. It is proved that all these problems are strongly NP-hard if the number of clusters is a part of the input, and NP-hard in the ordinary sense if the number of clusters is not a part of the input (is fixed). Moreover, the problems are NP-hard even in the case of dimension 1 (on a line).

1 Introduction

The subject of this study includes several related discrete optimization problems of partitioning a finite set of Euclidean points into a family of clusters. Our goal is to analyze the computational complexity of these problems. This study was motivated by the lack of published results concerning the complexity status of these problems and by their importance, specifically, for combinatorial geometry, statistics, physics, mathematical problems of clustering, and interdisciplinary problems regarding data mining and big data. This paper supplements and develops the results obtained in [Kel'manov & Pyatkin, 2016], [Eremeev et al., 2016], [Kel'manov & Pyatkin, 2009].

2 Problems Formulation

Throughout this paper, $\mathbb{R}$ denotes the set of real numbers and $\|\cdot\|$ is the Euclidean norm. The following problems are considered.

Problem 1 (Minimum sum of normalized squares of norms clustering).

Given a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$ and a positive integer $J > 1$.

In: Yu. G. Evtushenko, M. Yu. Khachay, O. V. Khamisov, Yu. A. Kochetov, V.U. Malkova, M.A. Posypkin (eds.): Proceedings of the OPTIMA-2017 Conference, Petrovac, Montenegro, 02-Oct-2017, published at http://ceur-ws.org

Find a partition $\mathcal{C}_1, \ldots, \mathcal{C}_J$ of $\mathcal{Y}$ into nonempty clusters (subsets) such that

$$\sum_{j=1}^{J} \frac{1}{|\mathcal{C}_j|} \Big\| \sum_{y \in \mathcal{C}_j} y \Big\|^2 \longrightarrow \min .$$

Problem 2 (Minimum sum of squared norms clustering).
Given a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$ and a positive integer $J > 1$.
Find a partition $\mathcal{C}_1, \ldots, \mathcal{C}_J$ of $\mathcal{Y}$ into nonempty clusters such that

$$\sum_{j=1}^{J} \Big\| \sum_{y \in \mathcal{C}_j} y \Big\|^2 \longrightarrow \min .$$

Problem 3 (Minimum sum-of-norms clustering).
Given a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$ and a positive integer $J > 1$.
Find a partition $\mathcal{C}_1, \ldots, \mathcal{C}_J$ of $\mathcal{Y}$ into nonempty clusters such that

$$\sum_{j=1}^{J} \Big\| \sum_{y \in \mathcal{C}_j} y \Big\| \longrightarrow \min .$$

These partitioning problems can be interpreted as optimal summing ones. On the other hand, they have an obvious geometric interpretation, namely, in a finite set of Euclidean points, we search for a family of disjoint geometric structures (clusters) with an extremal property described by the corresponding objective function.

The example of an input set of 1000 points from two clusters on a plane is presented at Fig. 1. The points of both clusters are scattered around the origin (the points of different clusters are colored differently). One needs to find a partition of this set into two clusters in accordance with the objective function of one of above formulated problems.
3[mm]

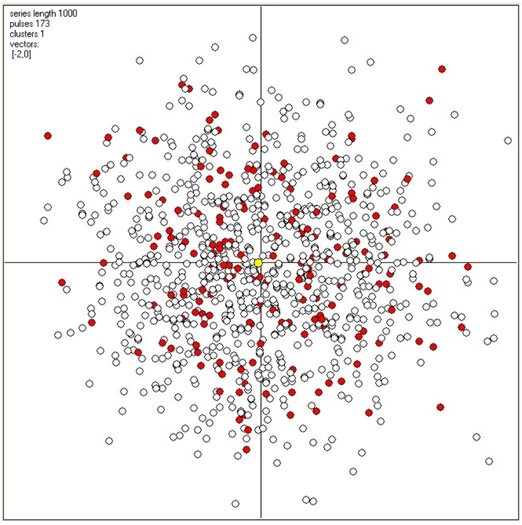

Figure 1: Example of the input set with two clusters

Another example of an input set of 500 points on a plane is presented at Fig. 2. In this example, the points from three clusters are scattered around the origin.

In [Kel'manov & Pyatkin, 2009] the similar maximization problems were considered. In particular, it was proved that Problem 1 on maximum is strongly NP-hard. Since changing the optimization direction can vary the complexity quite unpredictably, finding out the complexity status of Problems 1–3 looks interesting.

On the other hand, Problems 1–3 have some applications in Data mining, statistics and natural sciences. But first let us remind some relative problems that were studied in [Eremeev et al., 2016].

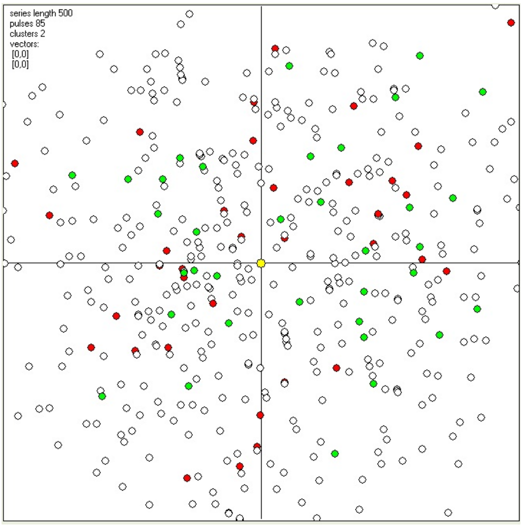

Figure 2: Example of the input set with three clusters

Problem 4 (Subset with the minimum normalized length of vectors sum).
Given a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$.
Find a nonempty subset $\mathcal{C} \subseteq \mathcal{Y}$ such that

$$\frac{1}{|\mathcal{C}|}\Big\|\sum_{y\in\mathcal{C}} y\Big\|^2 \longrightarrow \min .$$

Problem 5 (Subset with shortest vectors sum, arbitrary cardinality).
Given a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$.
Find a nonempty subset $\mathcal{C} \subseteq \mathcal{Y}$ such that

$$\Big\|\sum_{y\in\mathcal{C}} y\Big\| \longrightarrow \min .$$

Problem 6 (Subset with shortest vectors sum, given cardinality).
Given a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$ and a positive integer M.
Find a subset $\mathcal{C} \subseteq \mathcal{Y}$ of cardinality M such that

$$\Big\|\sum_{y\in\mathcal{C}} y\Big\| \longrightarrow \min .$$

It was proved in [Eremeev et al., 2016] that if the dimension of the space is a part of the input then Problems 4–6 are NP-hard in the strong sense, and if the dimension of the space is fixed then they are NP-hard in the ordinary sense even on the plane ($q = 2$). It was also shown there that there are no approximation algorithms with guaranteed performance for them unless P=NP. However, if the coordinates of the input points are integer and the dimension of the space is bounded by a constant then Problems 4–6 can be solved in a pseudopolynomial time.

From the natural sciences point of view, the objective function of Problems 4–6 can be treated as searching for a subset of compensated (balanced) forces. Then Problems 1–3 can be interpreted as physical problems of partitioning a set of forces into subsets of collectively balanced forces. If the elements of the subset are treated as the sequence of coordinates of the jth Brownian particle that moves in space starting at the origin, then is the length of the total displacement of this particle from the starting position under the action of chaotic jolts. Therefore, Problems 13 can be treated as ones of partitioning a set of points into Brownian clusters with the minimum sum of mean squared displacements (Problem 1), with the minimum sum of squared displacements (Problem 2), or with the minimum sum of displacements (Problem 3).

On the other hand, Problems 1–3 have roots in interdisciplinary problems related to data mining and big data (see, e.g., [Aggarwal, 2015, Bishop, 2006, Hastie et al., 2001]). In these problems, a major task is to partition a set into clusters consisting of objects that are similar or close according to some criterion, that can be of various types. The partitioning criteria in Problems 1–3 have not been investigated thus far.

Finally, Problems 1–3 have applications in statistics. Indeed, the input set can be treated as an inhomogeneous sample of several distributions where the correspondence of sample elements to distributions is not given. Is it true that these unknown distributions have different finite variances and identical (zero) expectations? Finding the optimal solution of Problems 1–3 could give an answer to this question. Indeed, if the hypothesis is true then by Central limit theorem for each of the optimal samples (i. e. clusters $\mathcal{C}_1^*, \ldots, \mathcal{C}_J^*$) the following convergence in distribution holds:

$$\frac{1}{\sigma_j |\mathcal{C}_j^*|} \sum_{y \in \mathcal{C}_j^*} y \xrightarrow[|\mathcal{C}_j^*| \to \infty]{} \Phi_{0,1},$$

where σ_j is the dispersion of jth distribution and $\Phi_{0,1}$ is Gaussian distribution with parameters $(\mathbf{0}, \mathbf{1})$. Therefore, the standard statistical hypothesis testing methods can be applied for each of the optimal subsets.

Thus, the search for an optimal solution of Problems 1–3 in polynomial time is an important task from both theoretical and application points of view. In this paper, we show that Problems 1–3 are strongly NP-hard if the number of clusters is a part of the input and are NP-hard in the ordinary sense if the number of clusters is not a part of the input. Moreover, all these problems are NP-hard even for $q = 1$, i.e., on a line. This result means that all the applied problems mentioned above are hard to solve.

3 Main Results

The first result of this paper consists in the following

Theorem 1. Problems 1–3 are strongly NP-hard even for the dimension $q = 1$.

Proof. First let us present the Problems 1–3 for $q = 1$ in the properties verification form.

Problems 1–3.

Given the set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of reals ("vectors" of dimension 1), positive integer $J > 1$ and positive K.

Question: Is there a prstition of the set $\mathcal{Y}$ into nonempty subsets (clusters) $\mathcal{C}_1, \ldots, \mathcal{C}_J$ such that the corresponding objective function is at most K?

To prove the theorem we show that the following well-known strongly NP-hard problem [Garey & Johnson, 1979] is polynomially reduced to Problems 1–3.

3-Partition problem.

Given a positive integer B and a set A of $3n$ positive integers from the interval $(B/4, B/2)$ such that their sum is equal to nB.

Question: can it be partitioned into n subsets such that the sum of the numbers in each of them is equal to B?

Then we use the following reduction. Given an arbitrary instance of 3-*Partition* construct the following joint instance of the Problems 1–3.

Put $K = 0, J = n$ and $\mathcal{Y} = \mathcal{A} \cup \mathcal{B}$, where

$$\mathcal{B} = \{nB + i\alpha, -(nB + B + i\alpha) \mid i = 1, 2, \ldots, n\}$$

and

$$\alpha = 1/(n + 1).$$

Assume that in the instance of 3-*Partition* the set $\mathcal{A}$ can be partitioned into n subsets such that the sum of the elements in each subset is B. Then adding to the ith of these subsets $(i = 1, 2, \ldots, n)$ the numbers $nB + i\alpha$ and $-(nB + B + i\alpha)$ from the set $\mathcal{B}$ results in the partition of the set $\mathcal{Y}$ in Problems 1–3 into $J = n$ clusters with all three objective functions equal to 0.

Now suppose that in Problems 1—3 the set $\mathcal{Y}$ can be partitioned into $J = n$ clusters such that the corresponding objective function is 0. Then, clearly, the sum of the elements inside each cluster is 0. Hence, each cluster contains at least one negative number. Since there are n negative numbers in $\mathcal{Y}$, we can assume ith cluster sontains the number $-(nB + B + i\alpha)$.

Next, since

$$(nB + \alpha) + (nB + 2\alpha) = 2nB + 3\alpha > (n + 1)B + n\alpha,$$

each cluster contains at most one positive number from the set $\mathcal{B}$. Since the addend $-i\alpha$ cannot be turned to zero by the addends from the set $\mathcal{A}$, the ith cluster must contain exactly one positive number from $\mathcal{B}$, namely,

$(nB + i\alpha)$. But then the sum of other elements in the cluster (note that all of them are from$\mathcal{A}$) is equal to B, i. e. they induce a partition of the set $\mathcal{A}$ into n subsets satisfying the requirements of 3-*Partition*.

Theorem 1 is proved.

Now denote by Problems 1(J)–3(J) the variants of the Problems 1–3 where the number of clusters J is not a part of input. Here is their statement in the properties verification form.

Problems 1(J)–3(J).

Given: Set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of reals and positive K.

Question: Is there a partition of the set $\mathcal{Y}$ into nonempty clusters $\mathcal{C}_1, \ldots, \mathcal{C}_J$ such that the corresponding objective function is at most K?

The second, main result of our paper is the following

Theorem 2. Problems 1(J)–3(J) are NP-hard even for the dimension $q = 1$.

Proof. We use the reduction of the classic [Garey & Johnson, 1979] NP-hard

Partition problem.

Given a set A of positive integers whose sum is equal to $2W$.

Question: Can it be partitioned into two subsets with the sum of the numbers in each equal to W?

Given an arbitrary instance of *Partition* problem construct the following general instance of Problems 1(J)–3(J). Put $K = 0$ and $\mathcal{Y} = \mathcal{A} \cup \mathcal{B}$, where the set

$$\mathcal{B} = \{-W, -3W, 2W\} \cup \{-iW, iW \mid i = 4, 5 \ldots, J+1\}.$$

Assume the *Partition* instance contains two nonintersecting subsets such that the sum of elements in each of them is equal to W. Consider the following partition of the set $\mathcal{Y}$ in Problems 1(J)–3(J). As first two clusters consider the optimal subsets from *Partition*, supplemented by the number $-W$ and a couple of numbers $2W$ and $-3W$, respectively. The remaining elements of the set $\mathcal{Y}$ are partitioned into $J - 2$ clusters of type $\{-iW, iW\}$, where $i = 4, 5 \ldots, J+1$. Clearly, the constructed J clusters is a solution of Problems 1(J)–3(J) providing value 0 of the objective function.

Now let the Problems 1(J)–3(J) allow a solution with the corresponding objective function equal to 0. Note that each of the objective functions of the Problems 1(J)–3(J) is 0 if and only if the sum of the elements in each cluster is 0. Since $\mathcal{Y}$ contains J negative numbers, each of J clusters has exactly one negative number. Then it is easy to show by induction on j from $J + 1$ down to 4 that the number jW must be in the same cluster as $-jW$. Then clearly the number $2W$ shares a cluster with the number $-3W$, and the elements of $\mathcal{A}$ lying in the last two clusters sums to W, i. e. they induce the desired partition of the set $\mathcal{A}$.

Theorem 2 is proved.

Note that the simpler proofs of Theorems 1 and 2 for the multisets versions of Problems 1–3 were earlier presented in [Kel'manov & Pyatkin, 2016].

4 Conclusion

The obtained "negative" results show that in spite of the simplicity of the statement of the considered problems, there are no exact or even approximation polynomial algorithms unless $P = NP$. However, obtaining "positive" algorithmic results looks promising for special cases of these problems that contain additional restrictions excluding the zero value of the objective function.

Acknowledgements

This work was supported by Russian Science Foundation, Project 16-11-10041.

References

[Aggarwal, 2015] Aggarwal C. C. (2015). *Data Mining: The Textbook.* (2015). Springer International Publishing

[Bishop, 2006] Bishop M. C. (2006). *Pattern Recognition and Machine Learning.* New York: Springer Science+Business Media, LLC.

[Garey & Johnson, 1979] Garey M. R., Johnson D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness.* San Francisco: Freeman.

[Hastie et al., 2001] Hastie T., Tibshirani R., Friedman J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer-Verlag.

[Eremeev et al., 2016] Eremeev A.V., Kel'manov A.V., Pyatkin A.V. (2016). On the Complexity of Some Euclidean Optimal Summing Problems. *Doklady Mathematics, 93(3)*, 286-288. doi: 0.1134/S1064562416030157

[Kel'manov & Pyatkin, 2009] Kel'manov A. V., Pyatkin A. V. (2009). Complexity of Certain Problems of Searching for Subsets of Vectors and Cluster Analysis *Comput. Math. Math. Phys. 49(11)*, 1966-1971. doi: 10.1134/S0965542509110128

[Kel'manov & Pyatkin, 2016] Kel'manov A.V., Pyatkin A.V. (2016). On the Complexity of Some Euclidean Problems of Partitioning a Finite Set of Points. *Doklady Mathematics, 94(3)*, 635-638. doi: 10.1134/S1064562416060089