On Complexity of Searching a Subset of Vectors with Shortest Average under a Cardinality Restriction

Anton V. Eremeev[1], Alexander V. Kelmanov[2], and Artem V. Pyatkin[2]

[1] Omsk Branch of Sobolev Institute of Mathematics, Siberian Branch of Russian Academy of Sciences, Omsk State University n.a. F.M. Dostoevsky, Omsk, Russia, eremeev@ofim.oscsbras.ru,

[2] Sobolev Institute of Mathematics, Siberian Branch of Russian Academy of Sciences, Novosibirsk State University, Novosibirsk, Russia, {kelm,artem}@math.nsc.ru

Abstract. In this paper, we study the computational complexity of the following subset search problem in a set of vectors. Given a set of N Euclidean q-dimensional vectors and an integer M, choose a subset of at least M vectors minimizing the Euclidean norm of the arithmetic mean of chosen vectors. This problem is induced, in particular, by a problem of clustering a set of points into two clusters where one of the clusters consists of points with a mean close to a given point. Without loss of generality the given point may be assumed to be the origin.
We show that the considered problem is NP-hard in the strong sense and it does not admit any approximation algorithm with guaranteed performance, unless P=NP. An exact algorithm with pseudo-polynomial time complexity is proposed for the special case of the problem, where the dimension q of the space is bounded from above by a constant and the input data are integer.

Keywords: vectors sum, subset selection, Euclidean norm, NP-hardness, pseudo-polymonial time.

1 Introduction

In this paper, we study a discrete extremal problem of searching a subset of vectors with shortest average under a cardinality restriction. The goal of the study is finding out the computational complexity of this problem and its approximability. The research is motivated by significance of the problem in many applications (see below).

The Subset with the Shortest Average under Cardinality Restriction (SSA) problem is formulated as follows.

Given: a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points (vectors) from $\mathbb{R}^q$ and a positive integer M.

Find: a subset $\mathcal{C} \subseteq \mathcal{Y}$ such that $|\mathcal{C}| \geq M$ and

$$\frac{1}{|\mathcal{C}|} \| \sum_{y \in \mathcal{C}} y \| \to \min, \tag{1}$$

where $\|\cdot\|$ denotes the Euclidean norm.

Note that the above formulation involves a norm of a sum of elements of the desired subset $\mathcal{C}$. Therefore this problem may be viewed as an optimal summation problem and has an obvious geometrical interpretation. At the same time, this problem may be considered as a problem of clustering a set of points into two clusters ($\mathcal{C}$ and $\mathcal{Y} \setminus \mathcal{C}$) when one of the clusters consists of points with a mean close to the origin. Obviously, given any other vector instead of the origin, the problem can be easily reduced to the mentioned above. This type of 2-clustering problems can be used for censoring the input data, if the expectation of an observed variable is known in advance.

Also SSA problem has applications in the diverse and multidisciplinary area of Data Mining (see e.g. [1,2],[13]). One of the central problems in this area consists in approximation of data by some mathematical model which allows to interpret the data adequately and explain their emergence. In particular, such a model may be expressed as a statistical hypothesis that the input data $\mathcal{Y}$ are sampled from a mixture of several distributions and at least M observations correspond to a distribution with zero mean. First one can solve SSA problem with the given data, after that the classical methods of statistical hypothesis testing may be applied to the obtained SSA solution and finally the data interpretation may be done on the basis of hypothesis testing results.

Another area where the SSA problem emerges is the trading hubs construction for electricity markets under locational marginal pricing [4].

It can be seen from the form of the optimization criterion (1) that the problem under consideration may be easily interpreted as a version of important classical problems in physics that ask for a balanced subset of forces (vectors). Besides that, if the given points of the Euclidean space correspond to people so that the coordinates of points are equal to some characteristics of these people (w.r.t. some matters), then the formulated problem may be treated as a problem of finding a balanced group (a subset) of people.

The formulation of SSA problem is resembling the formulation of optimal summation problems with a *maximization* criterion which first arose in studying the problem of noise-proof off-line search for an unknown repeating fragment in a discrete signal [17]. The maximization criterion in [17] has a different scaling compared to (1):

$$\frac{1}{|\mathcal{C}|}\Big\|\sum_{y\in\mathcal{C}} y\Big\|^2 \to \max\,. \tag{2}$$

The strong NP-hardness of the maximization problems with criterion (1) was proved in [3],[9], [10],[22,23] under different restrictions on the cardinality of the desired set. These problems, their generalizations and special cases were also studied in [5,6],[8], [10,11,12], [14,15,16], [18,19,20,21], [24]. In particular, it was proved in [11],[20] that in the case of the fixed dimension q of the space, the problems with criterion (2) are polynomially solvable in time $\mathcal{O}(N^{2q})$.

The complexity and approximability status of SSA problem was not completely known up to now. An equivalent single hub selection problem was studied in [4] where it was shown to be NP-hard in the 2-dimensional Euclidean space.

A modification of the single hub selection problem problem, where the size of the sought subset $\mathcal{C}$ is given in the input, was shown to be strongly NP-hard in [25]. SSA problem may be transformed to $\mathcal{O}(N)$ instances of the problem from [25] but this does not help to identify the complexity status of SSA in the general case. In the next section, we provide a detailed study of computational complexity of the SSA problem and its approximability.

2 Analysis of Computational Complexity and Approximability

Note that in the general formulation of the SSA problem given above, the dimension q of the space is a part of the input data. The following theorem states the complexity status of this problem.

Theorem 1. *SSA problem is NP-hard in the strong sense.*

Proof. Let us prove the strong NP-completeness of the equivalent decision problem, which implies the strong NP-hardness (see e.g. [7]). Let us formulate SSA problem in the form of decision problem.

Instance: A set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$, a positive integer K and a positive integer M.

Question: Is there a nonempty subset $\mathcal{C} \subseteq \mathcal{Y}$ of size at least M, such that the value of objective function (1) is at most K?

SSA decision problem obviously belongs to class NP. In what follows we will consider a special case of this problem, where $K = 0$, denoting it by SSA0. Let us reduce a classical NP-complete problem [7] Exact Cover by 3-Sets to SSA0.

Exact Cover by 3-Sets.

Instance: A finite set $\mathcal{Z}$ such that $|\mathcal{Z}| = 3n$ and a collection $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_k\}$ of 3-element subsets of the set $\mathcal{Z}$.

Question: Does $\mathcal{X}$ contain an exact cover for set $\mathcal{Z}$, i.e. a collection $\{\mathcal{X}_{i_1}, \mathcal{X}_{i_2}, \ldots, \mathcal{X}_{i_n}\} \subseteq \mathcal{X}$ such that $\cup_{j=1}^{n} \mathcal{X}_{i_j} = \mathcal{Z}$?

Given an instance of Exact Cover by 3-Sets, let us construct an equivalent instance of SSA0 problem. Put $q = 3n$ and $M = n + 1$. For each subset $\mathcal{X}_i$, $i = 1, \ldots, k$, a $3n$-dimensional point y_i is assigned, whose j-th coordinate ($j = 1, 2, \ldots, 3n$) is defined as $y_i^{(j)} = 1$, if $j \in \mathcal{X}_i$, and $y_i^{(j)} = 0$ otherwise. Let $y_{k+1} = (-1, \ldots, -1)$, $N = k + 1$ and $\mathcal{Y} = \{y_1, y_2, \ldots, y_k, y_{k+1}\}$.

Note that the objective function of SSA problem equals zero iff $z := \sum_{y \in \mathcal{C}} y = 0$.

If the instance of Exact Cover by 3-Sets has the answer "Yes", then obviously the subset $\mathcal{C} = \{y_{i_1}, \ldots, y_{i_n}, y_N\}$ turns the objective function (1) into 0.

Now let the optimal value of the objective function in SSA problem be equal to 0. Then subset $\mathcal{C}$ must contain the point $y_N = (-1, \ldots, -1)$, because otherwise all coordinates of point z are non-negative and at least one of them is positive.

In this case, the rest of the points in the subset $\mathcal{C}$ altogether should contain exactly one 1 in each coordinate, so there should be exactly n such points and the subsets corresponding to them form an exact cover. Note that the equality $|\mathcal{C}| = n + 1 = M$ holds.

Finally, the strong NP-hardness of SSA problem follows from the fact that an NP-complete in the strong sense problem Exact Cover by 3-Sets is reduced to a special case of SSA problem with binary input and the objective function values are bounded by a polynomial in n. □

Now let us consider complexity and approximability of SSA problem when the dimension q is fixed.

Let $\rho > 1$. A polynomial-time algorithm that finds a feasible solution to a minimization problem, such that the value of objective function in this solution is at most ρ times the optimal value (if the problem is solvable) is called a *ρ-approximation algorithm*. The corresponding feasible solution is called a *ρ-approximate solution*.

Below we denote by $\mathbb{N}$ the set of positive integers.

Theorem 2. *For any function $r : \mathbb{N} \to (1, \infty)$, the problem of searching an $r(N)$-approximate solution to SSA problem is NP-hard even in the special case of $q = 2$.*

Proof. Let us reduce the following modification of the NP-complete Partition problem (see e.g. [7]), which we call Bounded Partition, to the decision problem SSA0.

Bounded Partition

Instance: An even number n of positive integers $\alpha_j,\ j = 1, 2, \ldots, n$.

Question: Is there a subset $\mathcal{I} \subset \{1, 2, \ldots, n\}$ such that $|\mathcal{I}| = n/2$ and $\sum_{i \in \mathcal{I}} \alpha_i = \frac{1}{2} \sum_{i=1}^{n} \alpha_i$?

Given a set of integers $\alpha_1, \alpha_2, \ldots, \alpha_n$ we construct an instance of SSA0 with $q = 2$, $N = n + 1$ and $M = n/2 + 1$. Let $L = \sum_{i=1}^{n} \alpha_i$. Put $y_i = (L, \alpha_i)$ for $i = 1, 2, \ldots, n$, $y_{n+1} = (-Ln/2, -L/2)$ and for each subset $\mathcal{I} \subseteq \{1, 2, \ldots, N\}$ denote $S(\mathcal{I}) = \sum_{i \in \mathcal{I}} y_i$.

If the set $\mathcal{I}$ required in Bounded Partition problem exists, then it is easy to see that $S(\mathcal{I} \cup \{n + 1\}) = 0$, and therefore the objective function (1) turns into zero.

Suppose there exists a set $\mathcal{C}^*$ of cardinality at least M such that the value of the objective function on this set is zero. Let z denote the sum of elements of $\mathcal{C}^*$. Now since the first coordinate of z equals 0, we have $|\mathcal{C}^*| = n/2 + 1$ and $y_{n+1} \in \mathcal{C}^*$. Then, due to zero value in the second coordinate of z we have $\sum_{i \in \mathcal{I}} \alpha_i = L/2 = \frac{1}{2} \sum_{i=1}^{n} \alpha_i$, where $\mathcal{I} = \{i \mid y_i \in \mathcal{C}^*\} \setminus \{n + 1\}$.

The observed properties of the reduction imply the NP-completeness of SSA0 problem for $q = 2$. Under this reduction, the objective function value of an optimal solution to the SSA problem instance equals zero iff the Bounded Partition problem instance has the answer "Yes", and the same applies to any $r(N)$-approximate solution to SSA problem. Finally, since the objective

function of SSA problem is efficiently computable, the problem of searching an $r(N)$-approximate solution is NP-hard. □

Theorem 2 implies that unless P=NP, SSA problem does not admit approximation algorithms with any non-trivial guaranteed approximation ratio, and, in particular it does not admit a fully polynomial time approximation scheme (FPTAS).

SSA problem with a fixed $q \geq 2$ can not be solved by a polynomial-time algorithm, unless P=NP. Nevertheless, as shown below, it is solvable in a pseudo-polynomial time, provided that all points of set $\mathcal{Y}$ have integer coordinates and the dimension q of the space is fixed.

For any two sets $\mathcal{P}, \mathcal{Q} \subset \mathbb{R}^q$ we introduce the following rule of summation:

$$\mathcal{P} + \mathcal{Q} = \{x \in \mathbb{R}^q \mid x = y + y',\ y \in \mathcal{P},\ y' \in \mathcal{Q}\}. \tag{3}$$

For any positive integer r we denote by $\mathcal{B}(r)$ the set of integer points in $\mathbb{R}^q$ with absolute values of all coordinates at most r. Then $|\mathcal{B}(r)| \leq (2r+1)^q$.

Let us denote the maximal absolute value of coordinates of the input points $y_1, y_2, \ldots, y_N$ by b. The proposed algorithm for solving SSA problem consists in consequent computing of subsets $\mathcal{S}_k \subseteq \mathcal{B}(bk)$, $k = 0, 1, \ldots, M$, that can be obtained by summing at most k different elements of the set of points $y_1, y_2, \ldots, y_k$. First we assume $\mathcal{S}_0 = \{0\}$. After that we compute $\mathcal{S}_k = \mathcal{S}_{k-1} + \{0, y_k\}$ for all $k = 1, 2, \ldots, N$ using formula (3). For each element $z \in \mathcal{S}_k$ we store an integer parameter n_z, equal to the maximum number of addends that can be used to produce z and the n_z-element set of these addends $\mathcal{C}_z \subseteq \mathcal{Y}$.

Finally, find in the subset $\mathcal{S}_N$ an element $z \in \mathcal{S}_N$ with $n_z \geq M$ and the minimum value of $\|z\|/n_z$ and output the subset $\mathcal{C}_z$ corresponding to such z.

Computation of $\mathcal{S}_k$ takes $\mathcal{O}(q \cdot |\mathcal{S}_{k-1}|)$ operations. Therefore the following theorem holds

Theorem 3. *If the coordinates of the points of input set $\mathcal{Y}$ are integer and b is the maximum absolute value of these coordinates then SSA problem is solvable in $\mathcal{O}(qN(2bN+1)^q)$ time.*

In the case of fixed dimension q, i.e. $q = \mathcal{O}(1)$, the complexity of the algorithm presented above is $\mathcal{O}(N(bN)^q)$ and SSA problem is solvable in a pseudo-polynomial time in this special case.

Conclusion

The obtained results imply that there exist no exact polynomial or pseudo-polynomial algorithms for SSA problem, unless P=NP.

In the case when the dimension of the space is not a part of the input (i.e. the dimension is fixed), SSA problem is NP-hard even on the plane and no approximation algorithms with non-trivial guaranteed approximation ratio exist for this problem, unless P=NP. SSA problem is solvable, however, within a

pseudo-polynomial time if the coordinates of the input points are all integer and the dimension is fixed.

The obtained results indicate that in spite of simplicity of formulation of the considered problem, efficient algorithms finding an exact or even an approximate solution to it are unlikely to exist. An exception is the special case where the space dimension is bounded by a constant and coordinates of the input points are bounded by a polynomial in N. We expect that obtaining "positive" results for SSA would require analysis of the special cases, which reflect the specifics of applications area.

Acknowledgements. This research is supported by RFBR, projects 15-01-00462, 16-01-00740 and 15-01-00976.

References

1. Aggarwal C. C.: Data Mining: The Textbook. Springer International Publishing (2015)
2. Bishop M. C.: Pattern Recognition and Machine Learning. New York: Springer Science+Business Media, LLC (2006)
3. Baburin A. E., Gimadi E.Kh., Glebov N. I. and Pyatkin A. V.: The problem of finding a subset of vectors with the maximum total weight. Journal of Applied and Industrial Mathematics. 2 (1), 32–38 (2008)
4. Borisovsky P.A., Eremeev A.V., Grinkevich E.B., Klokov S.A. and Vinnikov A.V.: Trading Hubs Construction for Electricity Markets. In: Kallrath, J., Pardalos, P.M., Rebennack, S., Scheidt, M. (eds.) Optimization in the Energy Industry. pp. 29–58. Springer, Berlin, Heidelberg (2009)
5. 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)
6. Dolgushev A.V., Kel'manov A.V., Shenmaier V.V.: Polynomial-time approximation scheme for a problem of partitioning a finite set into two clusters (in Russian). Trudy Instituta Matematiki i Mekhaniki UrO RAN. 21 (3), 100–109 (2015)
7. Garey, M.R. and Johnson, D.S.: Computers and intractability. A guide to the theory of NP-completeness. W.H. Freeman and Company, San Francisco (1979)
8. Gimadi E.Kh., Glazkov Yu.V., Rykov I.A.: On two problems of choosing some subset of vectors with integer coordinates that has maximum norm of the sum of elements in euclidean space. J. Appl. Indust. Math. 3 (3), 343–352 (2009)
9. Gimadi E.Kh., Kel'manov A.V., Kel'manova M.A., Khamidullin S.A.: Aposteriori finding a quasiperiodic fragment with given number of repetitions in a number sequence (in Russian). Sibirskii Zhurnal Industrial'noi Matematiki. 9 (25), 55–74 (2006)
10. Gimadi E.Kh., Kel'manov A.V., Kel'manova M.A., Khamidullin S.A.: A posteriori detecting a quasiperiodic fragment in a numerical sequence. Pattern Recognition and Image Analysis. 18 (1), 30–42 (2008)
11. Gimadi E.Kh., Pyatkin A.V., Rykov I.A.: On polynomial solvability of some problems of a vector subset choice in a Euclidean space of fixed dimension. J. Appl. Indust. Math. 4 (4), 48–53 (2010)

12. Gimadi E.Kh., Rykov I.A.: A randomized algorithm for finding a subset of vectors. J. Appl. Indust. Math. 9 (3), 351–357 (2015)
13. Hastie T., Tibshirani R., Friedman J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, New York (2001)
14. Kel'manov A.V.: Off-line detection of a quasi-periodically recurring fragment in a numerical sequence. Proceedings of the Steklov Institute of Mathematics. 263 (S2), 84–92 (2008)
15. Kel'manov A.V.: On the complexity of some data analysis problems. Comput. Math. and Math. Phys. 50 (11), 1941–1947 (2010)
16. Kel'manov A.V.: On the complexity of some cluster analysis problems. Comput. Math. and Math. Phys. 51 (11), 1983–1988 (2011)
17. Kel'manov A.V., Khamidullin S.A., Kel'manova M.A.: Joint finding and evaluation of a repeating fragment in noised number sequence with given number of quasiperiodic repetitions (in Russian). In: Book of Abstracts of the Russian Conference "Discret Analysis and Operations Reserch" (DAOR-2004), p. 185. Sobolev Institute of Mathematics SB RAN, Novosibirsk (2004)
18. 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)
19. Kel'manov A.V., Khandeev V.I.: A randomized algorithm for two-cluster partition of a set of vectors. Comput. Math. and Math. Phys. 55 (2), 330–339 (2015)
20. 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)
21. 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. and Math. Phys. 56 (2), 334–341 (2016)
22. Kel'manov A.V., Pyatkin A.V.: On the complexity of a search for a subset of "similar" vectors. Doklady Mathematics. 78 (1), 574–575 (2008)
23. Kel'manov A.V., Pyatkin A.V.: On a version of the problem of choosing a vector subset. J. Appl. Indust. Math. 3 (4), 447–455 (2009)
24. Kel'manov A.V., Pyatkin A.V.: Complexity of certain problems of searching for subsets of vectors and cluster analysis. Comput. Math. and Math. Phys. 49 (11), 1966–1971 (2009)
25. Tarasenko, E.: On complexity of single-hub selection problem (in Russian). In: Proc. of 24-th Regional Conference of Students "Molodezh tretjego tysacheletija". pp. 45–48. Omsk State University, Omsk (2010)