On searching for a vector subset with the minimum normalized squared sum length

Anton Eremeev1,2, Alexander Kel'manov2,3, Artem Pyatkin2,3

1. Omsk State University n.a. F.M. Dostoevsky,
2. Sobolev Institute of Mathematics,
3. Novosibirsk State University,

eremeev@ofim.oscsbras.ru,{kelm,artem}@math.nsc.ru

Supported by Russian Science Foundation (projects 16-11-10041 and 15-11-10009).

«11th International Conference on Intelligent Data Processing: Theory and Applications (IDP-2016)»
Barcelona, Spain.· October 10 - 14, 2016

Problem formulation

- The subject of study is one of the subset choice problems in a finite set of Euclidean vectors.

- The aim of the study is analysis of its complexity status and investigation of algorithmic approaches for it.

Problem formulation

- Problem 1 (Subset with the Minimum Normalized Length of Vectors Sum).
- Given: a set $Y = \{y1, \ldots, yN\}$ of vectors (points) from Rq.
- Find: a nonempty subset $C \subseteq Y$ such that

$$\frac{1}{|C|}\left\|\sum_{y \in C} y\right\|^2 \rightarrow min$$

Problem formulation

- Problem 2 (Subset with the Maximum Normalized Length of Vectors Sum).
- Given: a set $Y = \{y1, \ldots, yN\}$ of vectors from Rq.
- Find: a nonempty subset $C \subseteq Y$ such that

$$\frac{1}{|C|} \left\| \sum_{y \in C} y \right\|^2 \to max$$

Motivation and known results

- Problem 1 can be interpreted as searching for a balanced subset of forces from a given set.
- Another interpretation: choosing a group of experts for a talk-show where q issues would be discussed so that the mean opinion of the group is neutral.

Motivation and known results

- Problem 2 arose in studying the problem of noise-proof off-line search for an unknown repeating fragment in a discrete signal.
- Problem 2 is NP-hard.
- In case of fixed dimension q of the space Problem 2 is solvable in time $O(N^{2q})$. There is also an FPTAS of complexity $O\left(N^2\left(\frac{q-1}{\varepsilon}\right)^{q-1}\right)$.

Our results

- Problem 1 is NP-hard in a strong sense if q is a part of input
- Problem 1 is NP-hard in an ordinary sense even for $q=1$
- No approximation algorithm with a guaranteed approximation ratio is possible
- Pseudopolinomial algorithm for the case of fixed dimension and integer coordinates is presented

NP-hardness

- Theorem 1. Problem 1 is NP-hard in the strong sense.

- Theorem 2. Problem 1 is NP-hard in the ordinary sense even for $q=1$.

NP-hardness

- Decision version of Problem 1:
- Given: a set $Y = \{y1, \ldots, yN\}$ of vectors from Rq and positive K.
- Question: is there a nonempty subset $C \subseteq Y$ such that

$$\frac{1}{|C|} \left\| \sum_{y \in C} y \right\|^2 \leq K$$

NP-hardness

- Problem Exact Cover by 3-sets:
- Given: a set $Z=\{1,\ldots,3n\}$ and a family $X=\{X1,\ldots,Xk\}$ of its 3-element subsets.
- Question: is there an exact cover of Z in X, i.e. the subsets $Xi1,\ldots,Xin$ such that

$$\bigcup_{j=1}^{n} X_{i^j} = Z$$

NP-hardness

- Reduce an instance of Exact Cover by 3-sets to an instance of Problem 1.
- Put $q=3n$ and $K=0$. Let $N=k+1$, $Y=\{y1,\ldots,yN\}$ where $yk+1=(-1,\ldots,-1)$ and i-th coordinate of a vector yj for $j=1,\ldots,k$ is defined as follows:

$$y_j(i) = \begin{cases} 1, \text{if } i \in X_j \\ 0, \text{otherwise} \end{cases}$$

NP-hardness

- Denote by $z(C)$ the sum of all elements of a subset $C \subseteq Y$. Then the objective function in Problem 1 is zero if and only if $z(C)=0$
- If there is an exact cover $X_{i1}, \ldots, X_{in}$ then for $C=\{y_{i1}, \ldots, y_{in}, y_N\}$, clearly, $z(C)=0$
- On the other hand, if $z(C)=0$ then C must contain y_N, and n other vectors that sum up to $(1,\ldots,1)$, i.e. the corresponding subsets form an exact 3-cover

NP-hardness

- Theorem 2 follows from the fact that for $q=1$ Problem 1 contains as a partial case the following known NP-complete Subset Sum Problem:
- Given: a finite set of integers A.
- Question: is there a non-empty subset $B \subseteq A$ such that the sum of its elements is 0?

Pseudopolynomial algorithm

- Theorem 3. If the coordinates of the input vectors from Y are integer and each of them is at most b by the absolute value then Problem 1 is solvable in $O(qN(2bN+1)q)$ time.

Pseudopolynomial algorithm

- For arbitrary sets $P,Q \subset Rq$ define their sum as

$$P+Q=\{x \in Rq \mid x=y+z,\ y \in P,\ z \in Q\}$$

- For every positive integer r denote by $B(r)$ the set of all vectors in Rq whose coordinates are integer and at most r by absolute value. Then $|B(r)|=(2r+1)q$

- Denote by Sk the set of all vectors that can be

Pseudopolynomial algorithm

- First put $S1=\{0,y1\}$. Then for all $k=2,\ldots,N$ calculate $Sk=Sk-1+\{0,yk\}$.

- For each element $z\in Sk$ we store an integer parameter nz and a subset $Cz\subseteq Y$ of nz non-zero elements whose sum is y, and the value nz is maximum possible.

Pseudopolynomial algorithm

- Finally, find in the subset SN an element $z^* \in SN$ with $nz^* \neq 0$ such that the value $\|z^*\|2/nz^*$ is minimum, and output the corresponding subset Cz^*.

- Clearly, computation of Sk takes $O(q|Sk-1|) \leq O(q(2bN+1)q)$ operations and we need to count N of them ($S1, \dots, SN$).

Linear program

- Let Problem 1(M) be a version of Problem 1 with an additional restriction that $|C|=M$.

- For each $M=1,\ldots,N$ we solve Problem 1(M) and choose the best one as a solution of Problem 1

Linear program

- Let a Boolean variable xi be 1 if $yi \in C$ and 0 otherwise
- Define the auxiliary variables $zkl = xkxl$ for all $k = 2,\ldots,N;\ l = 1,\ldots,k-1$
- This is equivalent to
- $zkl \leq xk;\ zkl \leq xl;\ zkl \geq xk + xl - 1;\ zkl \geq 0;\ k = 2,\ldots,N;\ l = 1,\ldots,k-1$

Linear program

- Note that

$$\frac{1}{|C|}\left\|\sum_{i=1}^{N} x_i y_i\right\|^2 = \frac{1}{|C|}\sum_{i=1}^{N}\|y_i\|^2 x_i + \frac{2}{|C|}\sum_{k=2}^{N}\sum_{l=1}^{k-1}\langle y_k, y_l\rangle x_k x_l$$

- Put $C_i = \dfrac{\|y_i\|^2}{M}$ and $B_{kl} = \dfrac{2\langle y_k, y_l\rangle}{M}$

Linear program

- Then we have the following LP:

$$\sum_{i=1}^{N} C_i x_i + \sum_{k=2}^{N} \sum_{l=1}^{k-1} B_{kl} z_{kl} \to min$$

$$\sum_{i=1}^{N} x_i = M$$

$zkl \le xk$; $zkl \le xl$; $zkl \ge xk + xl - 1$; $k=2,\ldots,N$; $l=1,\ldots,k-1$

$xi \in \{0,1\}$, $zkl \ge 0$; $i=1,\ldots,N$; $k=2,\ldots,N$; $l=1,\ldots,k-1$

- Note that the size of LP does not depend on q

Thanks for your attention!