On some clustering problems: Complexity and efficient algorithms with performance garantees

Alexander Kel'manov

Sobolev Institute of Mathematics
Siberian Branch of the Russian Academy of Sciences,
Novosibirsk State University,
Novosibirsk, Russia

XIth International Conference
Intelligent Data Processing (IDP-11)
Barcelona, Spain
October 9-14, 2016

Introduction

Outline

1. Introduction (subject, goal and motivation of investigation)

2. Some quadratic Euclidean clustering problems:

   1. complexity and algorithms
   2. successful techniques and practices for these problems
   3. open questions

3. Conclusion

Introduction

The subject of investigation is

some quadratic Euclidean clustering problems.

The goal is

short review of some new results on the complexity of these problems, and on efficient algorithms with performance guarantees for their solutions.

The research is motivated by

poorly studied of the problems and its relevance to many applications.

Introduction

Applications and origins

1. Geometric problems, approximation problems, problems of joint statistical hypotheses testing and estimating while the sample is heterogeneous.

2. Data clustering, Data mining, Machine learning, Big data.

3. Applied problems in technical and medical diagnostics, remote monitoring, electronic intelligence, biometrics, bioinformatics, econometrics, criminology, processing of experimental data, processing and recognition of signals, etc.

Introduction

List of considered quadratic Euclidean clustering problems

1. Searching a family of disjoint subsets
(Поиск семейства непересекающихся подмножеств)

2. Balanced Variance-based 2-clustering with one given center
(Сбалансированное разбиение на два кластера при заданном центре одного из кластеров)

3. Finding a subsequence
(Поиск подпоследовательности)

List of of considered problems

4. Partitioning a sequence into clusters with restrictions on their cardinalities
(Разбиение последовательности на кластеры при ограничениях на их мощность)

5. Partitioning a sequence into clusters
(Разбиение последовательности на кластеры)

6. Minimum sum of normalized squares of norms clustering (Brownian clustering)
(разбиение на броуновские кластеры)

1. Searching a family of disjoint subsets

One of the well-known (Fisher, 1958) data analysis problems is the MSSC (Minimum Sum-of-Squares Clustering) problem which is strongly NP-hard (Aloise D., Deshpande A., Hansen P., Popat P., 2009) and has the following formulation.

Problem MSSC (Minimum Sum-of-Squares Clustering)

Given a set $\mathcal{Y} = \{y_1, \dots, y_N\}$ of points from $\mathbb{R}^q$ and positive integers $J > 1$.
Find a partition of $\mathcal{Y}$ into non-empty clusters $\{\mathcal{C}_1, \dots, \mathcal{C}_J\}$ such that

$$\sum_{j=1}^{J} \sum_{y \in \mathcal{C}_j} \|y - \overline{y}(\mathcal{C}_j)\|^2 \to \min,$$

where $\overline{y}(\mathcal{C}_j) = \frac{1}{|\mathcal{C}_j|} \sum\limits_{y \in \mathcal{C}_j} y$ is the centroid (geometrical center) of $\mathcal{C}_j$.

1. Searching a family of disjoint subsets

In MSSC problem the cardinalities of the required clusters are unknown and we have to find a partition of the set. But in the considered problem we have to find a family of subsets which union could not cover input set and the cardinalities of the required clusters are given.

Problem 1. Searching a family of disjoint subsets

Given a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$ and some positive integers $M_1, \ldots, M_J$.
Find a family $\{\mathcal{C}_1, \ldots, \mathcal{C}_J\}$ of disjoint subsets of $\mathcal{Y}$ such that

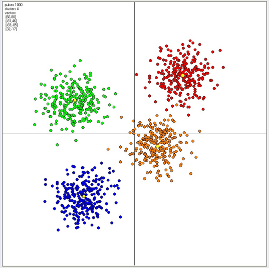
$$\sum_{j=1}^{J} \sum_{y \in \mathcal{C}_j} \|y - \overline{y}(\mathcal{C}_j)\|^2 \to \min,$$

where $\overline{y}(\mathcal{C}_j) = \frac{1}{|\mathcal{C}_j|} \sum\limits_{y \in \mathcal{C}_j} y$ is the centroid (geometrical center) of the subset $\mathcal{C}_j$, under constraints $|\mathcal{C}_j| = M_j, j = 1, \ldots, J$, on the cardinalities of the required subsets.

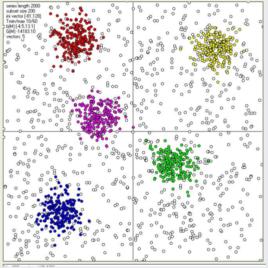
1. Searching a family of disjoint subsets

Two-dimensional examples

Minimum Sum-of-Squares Clustering



Searching a family of disjoint subsets



1. Searching a family of disjoint subsets

Known results

1. The strong NP-hardness of the problem is implied from the results of Kel'manov and Pyatkin (it was proved that the special case of the problem when $J = 1$ is strongly NP-hard, 2011).

2. A 2-approximation algorithm, $\mathcal{O}(N^2(N^{J+1} + q))$, Galashov, Kel'manov, 2014;

Some results were obtained for the **case** of the problem when $J = 1$.

3. An exact algorithm, $\mathcal{O}(qN^{q+1})$, Aggarwal, Imai, Katoh, Suri (1991).

4. A 2-approximation polynomial-time algorithm, $\mathcal{O}(qN^2)$, Kel'manov, Romanchenko (2012).

1. Searching a family of disjoint subsets

Known results

5. An exact pseudopolynomial algorithm, for the integer-valued variant of the data input and for the case of fixed space dimension; $\mathcal{O}(N(MB)^q)$ running time, where B is the maximum absolute value of the coordinates of the input points, Kel'manov, Romanchenko (2012).

6. FPTAS for the case of fixed space dimension with $\mathcal{O}(N^2(M/\varepsilon)^q)$-time complexity, where ε is a guaranteed relative error, Kel'manov, Romanchenko (2014).

7. PTAS of complexity $\mathcal{O}(qN^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon})$, where ε is a guaranteed relative error, Shenmaier (2012).

1. Searching a family of disjoint subsets

New result (Galashov, Kel'manov, 2016)

An **exact algorithm** for the case of integer components of the input points with

$$\mathcal{O}(N(N^2 + qJ)(2MB + 1)^{qJ} + (J - 1)\lg N)$$

running time, where B is the maximum absolute value of the coordinates of the input points and M is the least common multiple for the numbers $M_1, \ldots, M_J$.

In the case of the fixed dimension q of the space and of the fixed number J of required subsets, the proposed algorithm is pseudopolynomial and its time complexity is bounded by

$$\mathcal{O}(N^3(MB)^{qJ})$$

.

1. Searching a family of disjoint subsets

The idea of algorithm which implements a grid approach

1. Find the least common multiple M for the numbers $M_1, \dots, M_J$ and the maximum absolute value B of the coordinates of the input points. Construct the multi-dimensional grid $\mathcal{D}$ using formula

$$\mathcal{D} = \{z \in \mathbb{R}^q |\ (z)^k = \frac{1}{M}(v)^k, (v)^k \in \mathbb{Z}, |(v)^k| \le B, k = 1, \dots, q\},$$

2. For every tuple $d = (d_1, \dots, d_J) \in \mathcal{D}^J$ of J points from the grid find and save the exact solution $\{\mathcal{B}_1(d), \dots, \mathcal{B}_J(d)\}$ of the auxiliary problem

$$G^d(\mathcal{B}_1, \dots, \mathcal{B}_J) = \sum_{j=1}^{J} \sum_{y \in \mathcal{B}_j} \|y - d_j\|^2 \to \min .$$

Save the value of G^d.

3. Take as a solution the family $\{\mathcal{C}_1^A, \dots, \mathcal{C}_J^A\}$ of the subsets with minimum value of the function G^d, $d \in \mathcal{D}^J$.

2. Balanced Variance-based 2-clustering with given center

Problem 2. Balanced Variance-based 2-clustering with given center

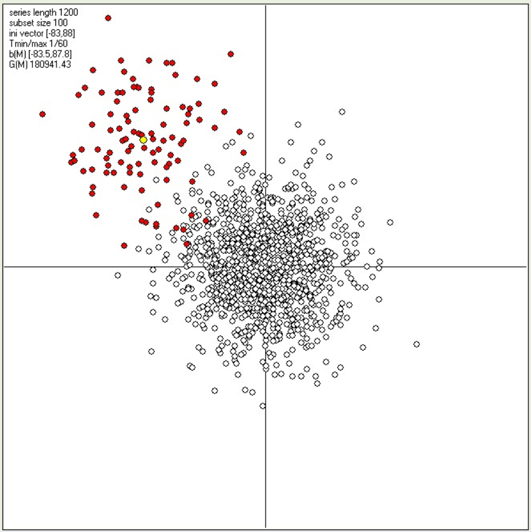
Given a set $\mathcal{Y} = \{y_1, \dots, y_N\}$ of points from $\mathbb{R}^q$ and a positive integer M.
Find a partition of $\mathcal{Y}$ into two non-empty clusters $\mathcal{C}$ and $\mathcal{Y} \setminus \mathcal{C}$ such that

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

where $\overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum\limits_{y \in \mathcal{C}} y$ is the geometric center (centroid) of $\mathcal{C}$, subject to constraint $|\mathcal{C}| = M$.

2. Balanced Variance-based 2-clustering with given center

2-dimesional example



series length 1200
subset size 100
ini vector [-83,88]
Tmin/max 1/60
b(M) [-83.5,87.8]
G(M) 180941.43

2. Balanced Variance-based 2-clustering with given center

Known results

1. The strong NP-hardness of the problem was proved in 2015, Kel'manov, Pyatkin.

2. An exact algorithm for the case of integer components of the input points was presented. If the dimension q of the space is bounded by a constant, then this algorithm has a pseudopolynomial $\mathcal{O}(N(MB)^q)$-time complexity, where B is the maximum absolute value of the coordinates of the input points; Kel'manov, Motkova (2015).

New result (Kel'manov, Motkova, 2016)

Here we present an approximation algorithm that allows to find a $(1+\varepsilon)$-approximate solution in $\mathcal{O}(qN^2(\sqrt{\frac{2q}{\varepsilon}}+1)^q)$ time for a given relative error ε. If the space dimension q is bounded by a constant this algorithm implements a fully polynomial-time approximation scheme with $\mathcal{O}\left(N^2\left(\frac{1}{\varepsilon}\right)^{q/2}\right)$-time complexity.

2. Balanced Variance-based 2-clustering with given center

The main idea of algorithm which implements an adaptive-grid-approach

For each point $y \in \mathcal{Y}$ steps 1-2 are executed:
1. Construct the cubic grid centered at the point y with node spacing h and edge cube size $2H$:

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

where

$$H = H(y) = \frac{1}{M}\sqrt{F(\mathcal{B}^y)}, \;\; h = h(y, \varepsilon) = \frac{1}{M}\sqrt{\frac{2\varepsilon}{q}F(\mathcal{B}^y)},$$

where $\mathcal{B}^y$ is a subset of $\mathcal{Y}$ with M smallest values of the function

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

2. Balanced Variance-based 2-clustering with given center

The main idea of algorithm which implements an adaptive-grid-approach

2. For each node x of the grid $\mathcal{D}(y, h, H)$ find a subset $\mathcal{B}^x \subseteq \mathcal{Y}$ with M smallest values of

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

Compute $F(\mathcal{B}^x)$ and remember this value and the subset $\mathcal{B}^x$.

3. In the family $\{\mathcal{B}^x | \, x \in \mathcal{D}(y, h, H), y \in \mathcal{Y}\}$ choose as a solution $\mathcal{C}_{\mathcal{A}}$ the set $\mathcal{B}^x$ for which the value of $F(\mathcal{B}^x)$ is the smallest.

3. Finding a subsequence

Problem 3. Finding a subsequence

Given a sequence $\mathcal{Y} = (y_1, \ldots, y_N)$ of points from $\mathbb{R}^q$, and some positive integer numbers $T_{\min}$, $T_{\max}$ and M.
Find a subset $\mathcal{M} = \{n_1, \ldots, n_M\}$ of $\mathcal{N} = \{1, \ldots, N\}$ such that

$$F(\mathcal{M}) = \sum_{j \in \mathcal{M}} \|y_j - \overline{y}(\mathcal{M})\|^2 \to \min,$$

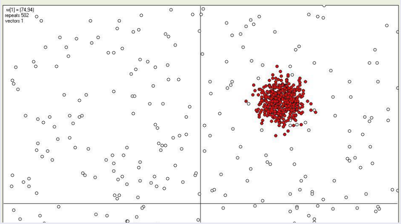
where $\overline{y}(\mathcal{M})$ is the centroid of $\{y_j | \, j \in \mathcal{M}\}$, under constraints

$$T_{\min} \leq n_m - n_{m-1} \leq T_{\max} \leq N, \;\; m = 2, \ldots, M, \tag{1}$$

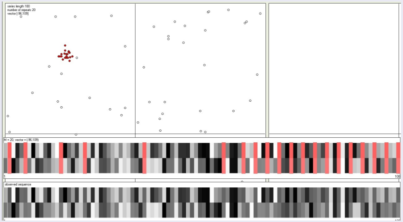
on the elements of $(n_1, \ldots, n_M)$.

3. Finding a subsequence: some examples

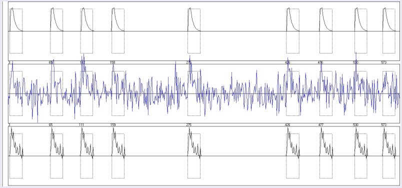
Finding a subset, $q = 2$


w[1] = (74,94)
repeats 502
vectors 1

Finding a subsequence, $q = 2$


series length 100
number of repeats 20
vector [-96,109]

Finding a subsequence of repeating fragments (fragment size $q = 20$) in a one-dimensional sequence



3. Finding a subsequence

Known results

1. The strong NP-hardness of the problem is implied from the results of Kel'manov, Pyatkin (2010).

2. A 2-approximation polynomial algorithm having $\mathcal{O}(N^2(N+q))$ running time was proposed in 2012 (Kel'manov, Romanchenko, Khamidullin).

3. For the case of fixed space dimension and integer input points coordinates, an exact pseudopolynomial algorithm with $\mathcal{O}(N^3(MD)^q)$-time complexity where D is the maximum absolute coordinate value of the points was presented in 2013 (Kel'manov, Romanchenko, Khamidullin).

Currently, there are no other algorithmic results for the problem.

New result (Kel'manov, Romanchenko, Khamidullin, 2016)

Here we present a FPTAS for the case of fixed space dimension with $\mathcal{O}(MN^3(1/\varepsilon)^{q/2})$-time complexity for an arbitrary relative error ε.

3. Finding a subsequence

The main idea of algorithm which implements an adaptive-grid-approach

For each point $y \in \mathcal{Y}$ steps 1-2 are executed:
1. Construct the cubic grid centered at the point y with node spacing h and edge cube size $2H$

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

where

$$H = H(y) = \sqrt{\frac{1}{M} F(\mathcal{M}^y)}, \;\; h = h(y, \varepsilon) = \sqrt{\frac{2\varepsilon}{qM} F(\mathcal{M}^y)},$$

where $\mathcal{M}^y$ is the optimal solution (tuple of indices) of the auxiliary problem

$$\sum_{i \in \mathcal{M}} \|y_i - y\|^2 \to \min_{\mathcal{M} \subseteq \mathcal{N}} .$$

3. Finding a subsequence

The main idea of algorithm which implements an adaptive-grid-approach

2. For each node d of the grid $\mathcal{D}(y, h, H)$ find the optimal solution (tuple of indices) $\mathcal{M}^d$ of the auxiliary problem

$$\sum_{i \in \mathcal{M}} \|y_i - d\|^2 \to \min_{\mathcal{M} \subseteq \mathcal{N}} .$$

Compute $F(\mathcal{M}^d)$ and remember this value and the subset $\mathcal{M}^d$.

3. In the family $\{\mathcal{M}^d | \, d \in \mathcal{D}(y, h, H), y \in \mathcal{Y}\}$ choose as a solution $\mathcal{M}_{\mathcal{A}}$ the set $\mathcal{M}^d$ for which the value of $F(\mathcal{M}^d)$ is the smallest.

Problem 4. Partitioning a sequence into clusters with restrictions on their cardinalities

Given a sequence $\mathcal{Y} = (y_1, \ldots, y_N)$ of points from $\mathbb{R}^q$ and some positive integers $T_{\min}$, $T_{\max}$, L, and M.
Find nonempty disjoint subsets $\mathcal{M}_1, \ldots, \mathcal{M}_L$ of $\mathcal{N} = \{1, \ldots, N\}$ such that

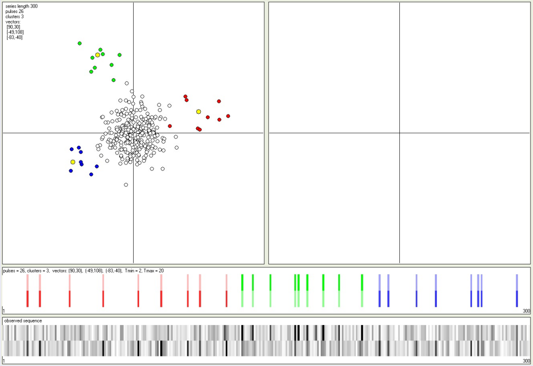
$$\sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} \|y_j - \overline{y}(\mathcal{M}_l)\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}} \|y_i\|^2 \to \min,$$

where $\mathcal{M} = \bigcup_{l=1}^{L} \mathcal{M}_l$, and $\overline{y}(\mathcal{M}_l)$ is the centroid of subset $\{y_j | j \in \mathcal{M}_l\}$, under the following constraints:
(i) the cardinality of $\mathcal{M}$ is equal to M,
(ii) concatenation of elements of subsets $\mathcal{M}_1, \ldots, \mathcal{M}_L$ is an increasing sequence, provided that the elements of each subset are in ascending order,
(iii) the following inequalities for the elements of $\mathcal{M} = \{n_1, \ldots, n_M\}$ are satisfied:

$$T_{\min} \le n_m - n_{m-1} \le T_{\max} \le N, \;\; m = 2, \ldots, M.$$

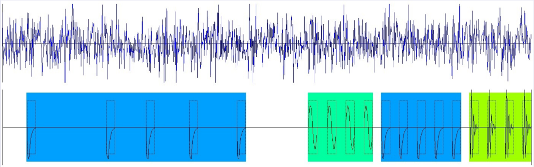
4. Partitioning a sequence into clusters with restrictions on their cardinalities

2-dimensional example



4. Partitioning a sequence into clusters with restrictions on their cardinalities

Example. Partitioning a one-dimensional sequence into clusters containing a repeated fragment (fragment size $q = 20$)



4. Partitioning a sequence into clusters with restrictions on their cardinalities

Known results

1. The strong NP-hardness of the problem is implied from the results of Kel'manov, Pyatkin (2011, 2013).

At present, for Problem 4, except for its **particular case** when $L = 1$, there are no efficient algorithms with guaranteed accuracy. For the mentioned case of Problem 4 the following results were obtained.

2. A 2-approximation polynomial-time algorithm having $\mathcal{O}(N^2(MN + q))$ running time time was proposed in 2014 (Kel'manov, Khamidullin).

3. For the case of fixed space dimension and integer input points coordinates, an exact pseudopolynomial algorithm with $\mathcal{O}(MN^2(MD)^q)$-time complexity where D is the maximum absolute coordinate value of the points was presented in 2015 (Kel'manov, Khamidullin, Khandeev).

4. Partitioning a sequence into clusters with restrictions on their cardinalities

Known results

3. For the case of fixed space dimension and integer input points coordinates, an exact pseudopolynomial algorithm with $\mathcal{O}(MN^2(MD)^q)$-time complexity where D is the maximum absolute coordinate value of the points was presented in 2015 (Kel'manov, Khamidullin, Khandeev).

Currently, there are no other algorithmic results for Problem 4.

New result (Kel'manov, Mikhailova, Khamidullin, Khandeev, 2016)

Here we present an algorithm that allows to find a 2-approximate solution of Problem 4 in $\mathcal{O}(LN^{L+1}(MN+q))$ time, which is polynomial if L is fixed (bounded by some constant).

4. Partitioning a sequence into clusters with restrictions on their cardinalities

The main idea of algorithm

For each point $y \in \mathcal{Y}$ steps 1-2 are executed:

1. For every tuple $x = (x_1, \ldots, x_L) \in \mathcal{Y}^L$ of elements of the sequence $\mathcal{Y}$, using **special dynamic programming scheme**, find the optimal solution $\{\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x\}$ of the auxiliary problem

$$G^x(\mathcal{M}_1, \ldots, \mathcal{M}_L) = \sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} (2\langle y_j, x_l\rangle - \|x_l\|^2) \to \max .$$

2. Find a tuple $x(A) = \arg\max_{x \in \mathcal{Y}^L} G^x(\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x)$ and a family $\{\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A\} = \{\mathcal{M}_1^{x(A)}, \ldots, \mathcal{M}_L^{x(A)}\}$.

If the optimum is taken by several tuples, we choose any of them.

Problem 5. Partitioning a sequence into clusters

Given a sequence $\mathcal{Y} = (y_1, \ldots, y_N)$ of points from $\mathbb{R}^q$ and some positive integers $T_{\min}$, $T_{\max}$, and L.
Find nonempty disjoint subsets $\mathcal{M}_1, \ldots, \mathcal{M}_L$ of $\mathcal{N} = \{1, \ldots, N\}$ such that

$$\sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} \|y_j - \overline{y}(\mathcal{M}_l)\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}} \|y_i\|^2 \to \min,$$

where $\mathcal{M} = \bigcup_{l=1}^{L} \mathcal{M}_l$, and $\overline{y}(\mathcal{M}_l)$ is the centroid of subset $\{y_j | \, j \in \mathcal{M}_l\}$, under the following constraints:
(i) concatenation of elements of subsets $\mathcal{M}_1, \ldots, \mathcal{M}_L$ is an increasing sequence, provided that the elements of each subset are in ascending order,
(ii) the following inequalities for the elements of $\mathcal{M} = \{n_1, \ldots, n_M\}$ are satisfied:

$$T_{\min} \le n_m - n_{m-1} \le T_{\max} \le N, \;\; m = 2, \ldots, M.$$

(the cardinality of $\mathcal{M}$ assumed to be unknown)

5. Partitioning a sequence into clusters

Known results

1. The strong NP-hardness of the problem is implied from the results of Kel'manov, Pyatkin (2011, 2013).

At present, for Problem 5, except for its **particular case when** $L = 1$, there are no efficient algorithms with guaranteed accuracy. For the mentioned case of Problem 4 the following results were obtained.

2. A 2-approximation polynomial-time algorithm having $\mathcal{O}(N^2(N+q))$ running time time was proposed in 2015 (Kel'manov, Khamidullin).

Currently, there are no other algorithmic results for Problem 5.

New result (Kel'manov, Mikhailova, Khamidullin, Khandeev, 2016)

Here we present an algorithm that allows to find a 2-approximate solution of Problem in $\mathcal{O}(LN^{L+1}(N+q))$ time, which is polynomial if the number L of clusters is fixed.

5. Partitioning a sequence into clusters

The main idea of algorithm

For each point $y \in \mathcal{Y}$ steps 1-2 are executed:

1. For every tuple $x = (x_1, \ldots, x_L) \in \mathcal{Y}^L$ of elements of the sequence $\mathcal{Y}$, using **special dynamic programming scheme**, find the optimal solution $\{\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x\}$ of the auxiliary problem

$$G^x(\mathcal{M}_1, \ldots, \mathcal{M}_L) = \sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} (2\langle y_j, x_l \rangle - \|x_l\|^2) \to \max .$$

2. Find a tuple $x(A) = \arg\max_{x \in \mathcal{Y}^L} G^x(\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x)$ and a family $\{\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A\} = \{\mathcal{M}_1^{x(A)}, \ldots, \mathcal{M}_L^{x(A)}\}$.

If the optimum is taken by several tuples, we choose any of them.

6. Brownian clustering

New unexplored problem

Problem 6. Minimum sum of normalized squares of norms clustering (Brownian clustering)

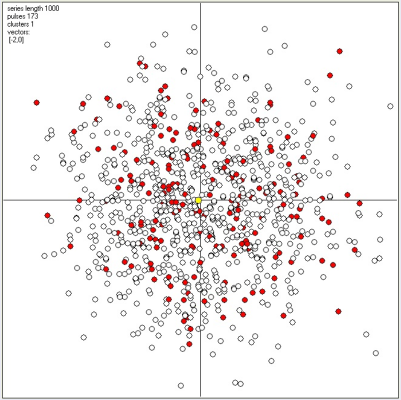
Given a set $\mathcal{Y} = \{y_1, \dots, y_N\}$ of points from $\mathbb{R}^q$, and positive integer $J > 1$.

Find a partition of $\mathcal{Y}$ into nonempty subsets $\mathcal{C}_1, \dots, \mathcal{C}_J$ such that

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

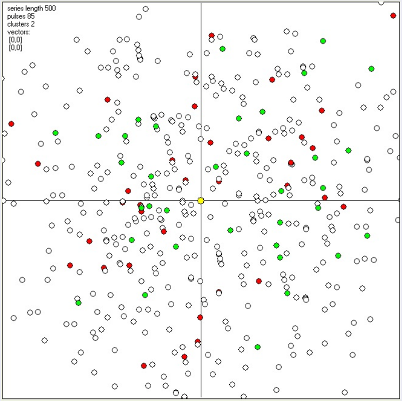
6. Brownian clustering

Example. Finding a 2-partition on a plane



```
series length 1000
pulses 173
clusters 1
vectors:
 [-2,0]
```

Example. Finding a 3-partition on a plane



```
series length 500
pulses 85
clusters 2
vectors:
 [0,0]
 [0,0]
```

6. Brownian clustering. Related problem

Related hard problem (Eremeev, Kel'manov, Pyatkin, 2015)

Subset with a minimum normalized length of vectors sum

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

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

Interpretation

1. **Searching a subset of balanced forces** from a given set.
2. **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

Problem is NP-hard in a **strong sense** if q is a part of input.
Problem is NP-hard in an **ordinary sense** even for $q = 1$.
No approximation algorithm with a guaranteed approximation ratio is possible (unless P=NP). An exact pseudopolinomial algorithm for the case of fixed q and integer coordinates was presented.

6. Brownian clustering

New result (Kel'manov, Pyatkin, 2016)

(1) Problem 6 is **strongly NP-hard** if the number of clusters is a part of the input.

(2) Problem 6 is **NP-hard** if the ordinary sense if the number of clusters is not a part of the input (is fixed).

(3) Problem 6 is **NP-hard even** in the case of dimension 1 (**on a line**).

(4) No approximation algorithm with a guaranteed approximation ratio is possible (unless P=NP).

The main of idea of the proof

As proof, we show that the problem is NP-complete, which implies that it is NP-hard.
For the case, when the number of clusters is a part of the input, we reduce strongly NP-complete **3-partition** problem to Problem 6.
For the case, when the number of clusters is not a part of the input, we reduce NP-complete **Partition** problem to Problem 6.

6. Brownian clustering

3-partition problem (strongly NP-complete)

Given positive integer B and a set $\mathcal{A} = \{a_1, \ldots, a_{3n}\}$ of positive integers from $(\frac{B}{4}, \frac{B}{2})$ such that the sum of the numbers equal to nB.
Question: can $\mathcal{A}$ be partitioned into n subsets of size 3, such that the sum of the numbers in each subset equal to B?

Brownian clustering

Given a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of real numbers, positive integer $J > 1$, and nonnegative real number K.
Question: does there exist a partitioning of $\mathcal{Y}$ into nonempty clusters $\mathcal{C}_1, \ldots, \mathcal{C}_J$, such that

$$\sum_{j=1}^{J} \frac{1}{|\mathcal{C}_j|} \Big\| \sum_{y \in \mathcal{C}_j} y \Big\|^2 \le K\,?$$

Exampe of Brownian clustering problem

$J = n$, $K = 0$, $\mathcal{Y} = \mathcal{A} \cup \mathcal{B}$, where multiset $\mathcal{B} = \{-B, \ldots, -B\}$, $|\mathcal{B}| = n$

6. Brownian clustering

Partition problem (NP-complete)

Given a set of positive integer, such that the sum of the numbers equal to $2W$.
Question: can the set be partitioned into 2 subsets, such that the sum of the numbers in each subset equal to W?

Brownian clustering (J)

Given a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of real numbers and nonnegative real number K.
Question: does there exist a partitioning of $\mathcal{Y}$ into nonempty clusters $\mathcal{C}_1, \ldots, \mathcal{C}_J$, such that $\sum_{j=1}^{J} \frac{1}{|\mathcal{C}_j|} \| \sum_{y \in \mathcal{C}_j} y \|^2 \leq K$?

Exampe of Brownian clustering (J)

$K = 0$, $\mathcal{Y} = \mathcal{A} \cup \mathcal{B}$, where multiset $\mathcal{B}$ contains: 2 copies of number $-W$, $J-2$ copies of number $-2W-1$, and $J-2$ copies of number $2W+1$.

Conclusion

All considered problems are still poorly studied in the algorithmical sense. Therefore, it seems important to continue studying the questions on algorithmical approximability of these problems.

Thank you for your attention!

Спасибо за внимание!