ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
УЧРЕЖДЕНИЕ НАУКИ
ИНСТИТУТ МАТЕМАТИКИ им. С.Л. СОБОЛЕВА
СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК

На правах рукописи

Хандеев Владимир Ильич

# АЛГОРИТМЫ С ОЦЕНКАМИ КАЧЕСТВА ДЛЯ КВАДРАТИЧНЫХ ЗАДАЧ КЛАСТЕРИЗАЦИИ С ФИКСИРОВАННЫМ ЦЕНТРОМ ОДНОГО ИЗ КЛАСТЕРОВ

Специальность 01.01.09 – дискретная математика и математическая кибернетика

Диссертация на соискание ученой степени
кандидата физико-математических наук

Научный руководитель:
д.ф-м.н. Кельманов Александр Васильевич

Новосибирск – 2017

# Оглавление

# Введение

## Общая характеристика работы

**Актуальность работы**[1]. Объект исследования работы — проблемы оптимизации. Предмет исследования — труднорешаемые экстремальные задачи разбиения множества и последовательности точек евклидова пространства. Цель исследования — изучение вопросов алгоритмической аппроксимируемости этих задач.

Рассматриваемые задачи в постановочном плане близки к классической NP-трудной в сильном смысле задаче MSSC (Minimum Sum-of-Squares Clustering). В этой задаче требуется разбить конечное множество точек евклидова пространства на несколько кластеров по критерию минимума суммы по всем кластерам внутрикластерных сумм квадратов расстояний от элементов кластеров до их геометрических центров (центроидов). Другое название задачи — $k$-means ($k$-средних). Основное отличие рассматриваемых NP-трудных в сильном смысле задач состоит в том, что для одного из кластеров внутрикластерная сумма равна сумме квадратов расстояний от элементов кластера до фиксированной точки — центра (без ограничения общности этой точкой может служить начало координат). В задачах кластеризации последовательности имеется дополнительное отличие — ограничения на элементы, входящие в кластеры.

[1]Работа поддержана грантами РФФИ 12-01-33028 мол-а-вед, 12-01-00090-а, 13-07-00070-а, 15-01-00462-а, 16-31-00186 мол-а, 16-07-00168-а, РНФ 16-11-10041.

Исследование мотивировано слабой изученностью задач и их актуальностью для анализа данных, распознавания образов, аппроксимации, компьютерной геометрии, статистики, а также для естественнонаучных и технических приложений, в которых требуется обработка и классификация данных численных экспериментов или результатов наблюдения за состояниями каких-либо объектов.

**Цель работы** — построение эффективных алгоритмов с гарантированными оценками качества (точности, трудоёмкости, вероятности несрабатывания) для модифицированной задачи MSSC, в которой центр одного из кластеров фиксирован, и для вариантов модифицированной задачи, ориентированных на разбиение последовательности при дополнительных ограничениях на элементы, входящие в кластеры.

**Основные результаты.**

1. Для квадратичной задачи разбиения конечного множества точек евклидова пространства на два кластера при фиксированном центре одного из кластеров:

(а) построен 2-приближённый полиномиальный алгоритм для случая задачи без ограничений на мощности кластеров;

(б) предложен рандомизированный алгоритм (ориентированный на случай задачи с ограничениями на мощности кластеров), который при заданных относительной ошибке и вероятности несрабатывания для установленных значений параметров находит приближённое решение за полиномиальное время; найдены условия, при которых этот алгоритм асимптотически точен.

2. Для квадратичных евклидовых задач 2-кластеризации конечных множества и последовательности (с ограничениями на выбор элементов, входящих в кластеры) при фиксированном центре одного из кластеров и дополнительном ограничении на мощности кластеров:

(а) построены точные алгоритмы для случая целочисленных входов задачи; при фиксированной размерности пространства алгоритмы псевдополиномиальны;

(б) показано, что для общих случаев задач не существует полностью полиномиальных приближённых схем (FPTAS), если $\mathrm{P} \neq \mathrm{NP}$; такие схемы построены для случаев задач, в которых размерность пространства фиксирована.

3. Для квадратичной евклидовой задачи многокластерного разбиения конечной последовательности точек с ограничениями на выбор внутрикластерных элементов при фиксированном центре одного из кластеров построены 2-приближённые алгоритмы, ориентированные как на случай задачи без ограничений на мощности кластеров, так и на случай с ограничениями; алгоритмы полиномиальны при фиксированном числе кластеров.

**Научная новизна**. Все результаты диссертации снабжены строгими доказательствами, а их новизну раскрывают следующие аргументы (по каждому основному результату).

1. Алгоритм из п. 1 (а) имеет меньшую трудоёмкость по сравнению с лучшим из известных алгоритмов при той же, как и у известного алгоритма, точности. Алгоритм из п. 1 (б) является первым алгоритмом рандомизированного типа, предложенным для задачи из этого пункта.

2. Алгоритм кластеризации множества из п. 2 (а) является новым решением задачи, послужившим важным промежуточным результатом, на котором основана идея построения оригинальных аппроксимационных схем из п. 2 (б). Алгоритм разбиения последовательности из п. 2 (а) построен впервые. Факт несуществования схемы FPTAS для общих случаев задач из п. 2 также установлен впервые; результаты по построению приближённых схем для указанных в п. 2 (б) случаев задач приоритетны.

3. На настоящее время результаты п. 3 являются единственными алгоритма-

ми с гарантированными оценками точности, предложенными для задач из этого пункта.

**Методы исследований.** В работе используются методы дискретной оптимизации, геометрии, теории вероятностей.

**Теоретическая значимость и практическая ценность.** Работа носит теоретический характер. Предложенные алгоритмы имеют теоретическую значимость для математических проблем анализа данных и распознавания образов, дискретной оптимизации, аппроксимации, компьютерной геометрии. Практическая ценность результатов (алгоритмов) состоит в том, что они могут быть использованы в естественно-научных и технических приложениях для создания эффективных компьютерных технологий с теоретическими гарантиями качества, ориентированных, в частности, на обработку сигналов, изображений, результатов численных экспериментов, дистанционный мониторинг и др.

**На защиту выносится** совокупность эффективных алгоритмов с гарантированными оценками качества для решения NP-трудных в сильном смысле квадратичных задач кластеризации конечных множества и последовательности точек евклидова пространства с фиксированным центром одного из кластеров.

**Личный вклад автора.** Постановки задач предложены научным руководителем. Подходы к построению алгоритмов найдены совместно. Доказательства утверждений получены соискателем лично. Конфликт интересов с соавторами отсутствует.

**Апробация работы.** Результаты работы докладывались на семинарах ИМ СО РАН (часть из них отмечены в качестве важнейших), кафедры Теоретической кибернетики НГУ (часть их них отмечены Премией им. А.А. Ляпунова) и на следующих всероссийских и международных конференциях: «Проблемы оптимизации и экономические приложения» (Омск, 2012, 2015); «Intelligent Data Processing: Theory and Applications» (Черногория, 2012, Греция, 2014, Испа-

ния, 2016); «Discrete Optimization and Operations Research» (Новосибирск, 2013, Владивосток, 2016); «Математические методы распознавания образов» (Казань, 2013, Светлогорск, 2015); «Optimization and applications» (Черногория, 2013, 2014, 2015, 2016); «Methods of Optimization and Their Applications» (Иркутск, 2014); «European Chapter on Combinatorial Optimization» (Италия, 2015); «Математическое программирование и приложения» (Екатеринбург, 2015).

**Публикации.** По теме диссертации опубликовано 29 работ, из них 20 — тезисы докладов, 9 работ — в изданиях, входящих в список ВАК, в том числе 5 — в журналах, индексируемых системой цитирования Web of Science, 9 — Scopus, 9 — RSCI (ядро РИНЦ).

**Структура и объем диссертации.** Работа состоит из введения, трёх глав, заключения и списка литературы. Объём диссертации — 111 страниц. Список литературы содержит 88 источников.

## Содержание работы

Во **введении** обоснована актуальность работы, приведено краткое изложение содержания работы, сформулированы основные результаты и раскрыта их новизна.

В **первой главе** представлены истоки задач, их трактовки и приложения, а также известные результаты для близкой в постановочном плане задачи MSSC. Обзор результатов для рассматриваемых задач приведён после их формулировок в соответствующих главах. Отмечено, что все рассматриваемые в работе задачи относятся к числу слабоизученных задач дискретной оптимизации. Этот факт определил направление исследований и характер полученных результатов.

Во **второй главе** представлены результаты исследования квадратичной задачи 2-кластеризации конечного множества точек евклидова пространства при

фиксированном (в начале координат) центре одного из кластеров. Рассматриваются два варианта задачи: 1) с оптимизируемыми мощностями кластеров; 2) с ограничениями на мощности кластеров.

Для задачи без ограничений на мощности кластеров построен 2-приближённый полиномиальный алгоритм. В этой же главе предложен точный алгоритм для случая целочисленных входов задачи с ограничениями на мощности кластеров. Показано, что при фиксированной размерности пространства алгоритм псевдополиномиален. Далее в этой главе доказано, что для задачи не существует схемы FPTAS, если P $\neq$ NP, и такая схема предложена для случая фиксированной размерности пространства. Наконец, в этой же главе предложен рандомизированный алгоритм, который при фиксированных относительной ошибке и вероятности несрабатывания позволяет находить приближённое решение задачи за время, линейное как от размерности пространства, так и от числа точек входного множества. Найдены условия, при которых алгоритм асимптотически точен и имеет трудоёмкость, линейную относительно размерности пространства и квадратичную относительно мощности входного множества.

**Третья глава** содержит результаты исследований квадратичных задач кластеризации конечной последовательности точек евклидова пространства при фиксированном центре одного из кластеров. Сначала представлены алгоритмические результаты для задачи 2-кластеризации с ограничениями на мощности искомых кластеров, а затем — результаты для двух вариантов задачи разбиения на произвольное заданное число кластеров: 1) с ограничениями на их мощности; 2) без ограничений. По сути, в этой главе развиты результаты главы 2. Развитие результатов опирается на технику (схемы) динамического программирования.

Для случая целочисленных входов задачи 2-кластеризации последовательности предложен точный алгоритм. Показано, что если размерность пространства ограничена константой, то алгоритм псевдополиномиален. В этой же главе

установлено, что если P $\neq$ NP, то для общего случая задачи не существует схемы FPTAS. Далее в этой главе построен приближённый алгоритм и показано, что в случае, когда размерность пространства фиксирована, этот алгоритм реализует схему FPTAS. Кроме того, в этой главе обоснован рандомизированный алгоритм. Все алгоритмы, предложенные для задачи 2-кластеризации последовательности, имеют те же вероятностные или аппроксимационные характеристики, что и алгоритмы для задачи 2-кластеризации множества (предложенные в главе 2), но увеличенные в полиномиальное число раз оценки трудоёмкости. Увеличение трудоёмкости обусловлено применением схем динамического программирования.

Наконец, в главе 3 для квадратичной евклидовой задачи многокластерного разбиения конечной последовательности точек с ограничениями на выбор внутрикластерных элементов построены 2-приближённые алгоритмы, ориентированные как на случай задачи без ограничений на мощности кластеров, так и на случай с ограничениями; алгоритмы полиномиальны при фиксированном числе кластеров.

В **заключении** сформулированы основные результаты работы и обозначены перспективы дальнейших исследований.

# Глава 1

# Истоки задач, их трактовка и приложения

Одной из известных [1–5] задач разбиения конечного множества точек евклидова пространства является задача

**Задача MSSC** (*Minimum Sum-of-Squares Clustering*)**.** *Дано:* $N$-элементное множество $\mathcal{Y} \subset \mathbb{R}^d$, натуральное число $J > 1$. *Найти:* разбиение множества $\mathcal{Y}$ на непустые подмножества (кластеры) $\mathcal{C}_1, \ldots, \mathcal{C}_J$ такое, что

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

где $\overline{y}(\mathcal{C}_j) = \frac{1}{|\mathcal{C}_j|} \sum\limits_{y \in \mathcal{C}_j} y$, $j = 1, \ldots, J$, — геометрический центр (центроид) $j$-го кластера.

Своими корнями задача уходит к классическим работам Фишера [6]. Во многих публикациях задача MSSC имеет краткое название $k$-means, которое соответствует названию одного из первых алгоритмов для её решения [4]. Вопросы построения алгоритмических решений этой задачи (с доказуемыми гарантиями и без них) рассматривались в тысячах публикаций (см., например, [5,7–24] и цитированные там работы). Ниже приведены наиболее значимые из полученных результатов.

В течение длительного времени задача MSSC считалась NP-трудной. Однако, окончательное доказательство её труднорешаемости было получено относительно недавно [7]. В [8] было показано, что одномерный случай этой задачи (когда $d = 1$) разрешим за полиномиальное время с помощью метода динамического программирования. Точный алгоритм для общего случая задачи, основанный на технике построения диаграмм Вороного, имеющий трудоёмкость $\mathcal{O}(dN^{dJ+1})$, предложен в [9].

Наиболее значимые приближённые алгоритмы представлены в работах [10–17]. В [10] для случая фиксированных $J$ и $d$ была предложена полиномиальная приближённая схема (PTAS), имеющая трудоёмкость $\mathcal{O}(N \log^J N \varepsilon^{-2J^2 d})$, где $\varepsilon$ — относительная погрешность алгоритма. Алгоритм, позволяющий находить $(1+\varepsilon)$-приближённое решение задачи за время $\mathcal{O}(N + J^{J+2}\varepsilon^{-(2d+1)J} \log^{J+1} N \log^J \frac{1}{\varepsilon})$, предложен в [11]. В [12] был обоснован алгоритм, позволяющий находить $(1+\varepsilon)$-приближённое решение задачи за время $\mathcal{O}(2^{(J/\varepsilon)^{\mathcal{O}(1)}}\mathrm{poly}(d)N \log^J N)$, где $\mathrm{poly}(d)$ — полином от $d$. В случае, когда число $J$ кластеров фиксировано, этот алгоритм реализует схему PTAS. В [13] предложен алгоритм, позволяющий с фиксированной вероятностью находить $(1 + \varepsilon)$-приближённое решение задачи за время $\mathcal{O}(2^{(J^3/\varepsilon^8)(\ln(J/\varepsilon))\ln J} dN \log^J N)$. Впоследствии в [14, 15] был построен улучшенный (по времени) алгоритм аналогичного плана, имеющий трудоёмкость $\mathcal{O}(2^{(J/\varepsilon)^{\mathcal{O}(1)}} dN)$.

Кроме того, в [16] обоснован алгоритм, который позволяет для любого $\varepsilon \in (0, 1)$ находить $(9 + \varepsilon)$-приближённое решение задачи за время $\mathcal{O}(N \log N + N\frac{1}{\varepsilon^d}\log(1/\varepsilon) + N^2 J^3 \log N)$, полиномиальное при фиксированном $d$. Алгоритм с гарантированной оценкой 50 точности и трудоёмкостью $\mathcal{O}(J^3 N^2 \log Nd)$ предложен в [17]. По-видимому, это один из первых приближённых алгоритмов, трудоёмкость которых полиномиально зависит от $N$, $d$ и $J$. В этой же работе были предложены два других алгоритма с гарантированными оценками точности,

один из которых имеет трудоёмкость, полиномиальную от $N$ и $d$, а другой позволяет с фиксированной вероятностью находить решение задачи за время, не зависящее от $N$ и полиномиально зависящее от $J$ и $d$.

Долгое время оставался открытым вопрос аппроксимируемости общего случая задачи MSSC (когда размерность $d$ пространства и число $J$ кластеров являются частью входа). Этот вопрос был разрешён совсем недавно (в 2015 г.) в работе [18], в которой было доказано, что задача MSSC является APX-трудной, т.е. для её общего случая не существует схемы PTAS, если P≠NP. Кроме того, в [19] показано, что если P≠NP, то для задачи MSSC не существует полиномиального алгоритма, гарантирующего отыскание решения с относительной погрешностью меньше, чем 0.0013.

Одна из возможных содержательных трактовок, которая индуцирует задачу MSSC, состоит в следующем. Имеется таблица, содержащая многократные результаты измерения набора числовых информационно значимых характеристик нескольких материальных объектов. В каждом результате измерения содержится ошибка, а соответствие результатов измерений объектам неизвестно. Требуется разбить исходные наборы на подмножества, соответствующие каждому из объектов, и оценить наборы характеристик этих объектов. Для примера на рисунке 1.1 изображено множество точек на плоскости, соответствующих результатам измерений двумерного набора характеристик четырёх объектов.

На практике зачастую объекты могут находиться в одном из двух состояний: активном и пассивном, причём в пассивном состоянии все характеристики объекта равны нулю, а в активном — значение хотя бы одной характеристики не равно нулю. В этом случае, как и в задаче MSSC, требуется найти подмножества, соответствующие активному состоянию каждого из объектов, оценить наборы характеристик этих объектов, и, кроме того, выделить кластер, соответствующий пассивному состоянию всех объектов (с учётом того, что данные содержат

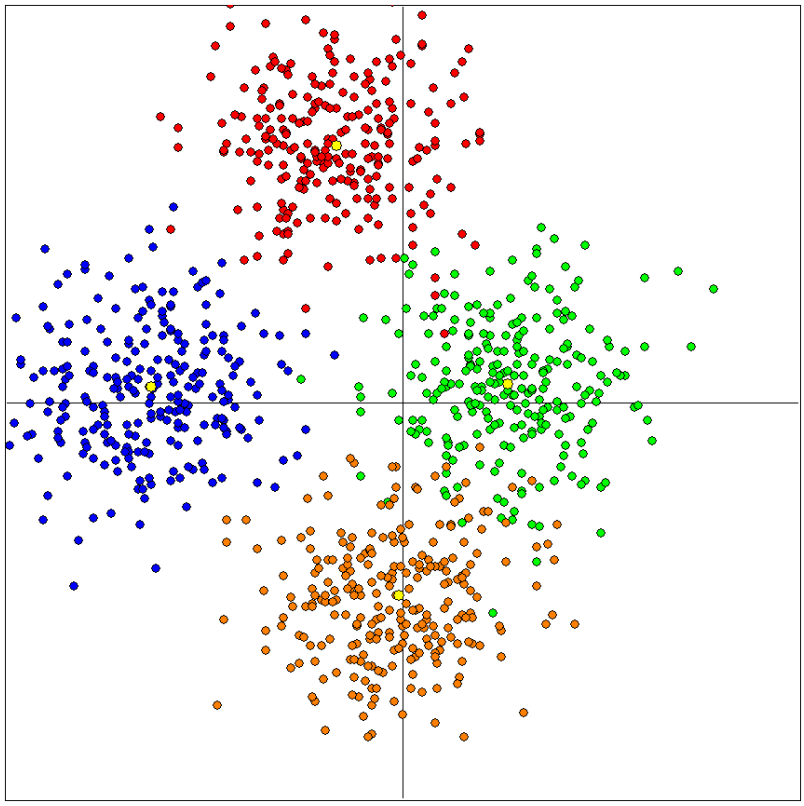

Рис. 1.1.

ошибку измерения). На рисунке 1.2 изображено множество точек на плоскости, соответствующих результатам измерений двумерного набора характеристик четырёх объектов, каждый из которых может находиться в одном из двух состояний: активном и пассивном. Точки, сгруппированные около начала координат, соответствуют пассивному состоянию, а остальные точки — одному из четырёх активных состояний. Проблемы подобного плана возникают при дистанционном мониторинге (в частности, радиолокационном, геофизическом, гидроакустическом, космическом) объектов. Экстремальные задачи, моделирующие эти проблемы, по-видимому, впервые были сформулированы Кельмановым А.В. в 90-х годах прошлого века в связи с решением некоторых задач оборонной тематики.

На протяжении почти десятилетия вопрос о сложностном статусе индуцированной экстремальной задачи даже в случае, когда в содержательной проблеме имеется один объект, который может находиться в двух состояниях — актив-

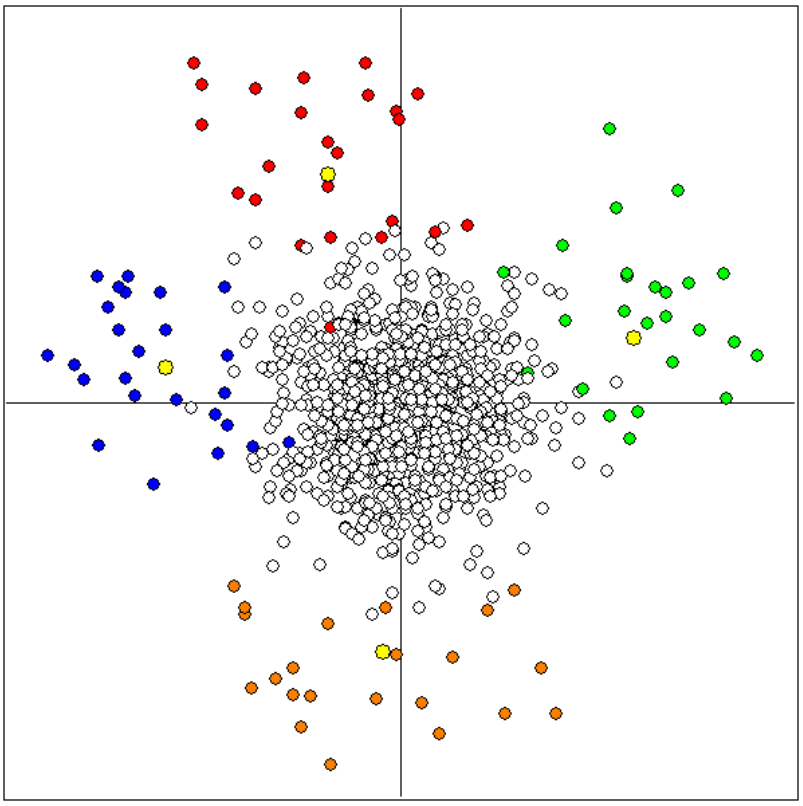

Рис. 1.2.

ном и пассивном, оставался открытым. Труднорешаемость этого случая была доказана в [25–29].

На рисунке 1.3 изображено множество точек, соответствующих результатам измерений двумерного набора характеристик одного объекта, находившегося в одном из двух состояний. Точки, изображённые тёмными кружками, соответствуют активному состоянию объекта, а точки, изображённые светлыми кружками — пассивному.

Следуя [25, 26], приведём формализацию содержательной задачи разбиения множества наборов на два подмножества, соответствующих активному и пассивному состояниям объекта. Пусть (ненаблюдаемая) последовательность $x_n \in \mathbb{R}^d$,

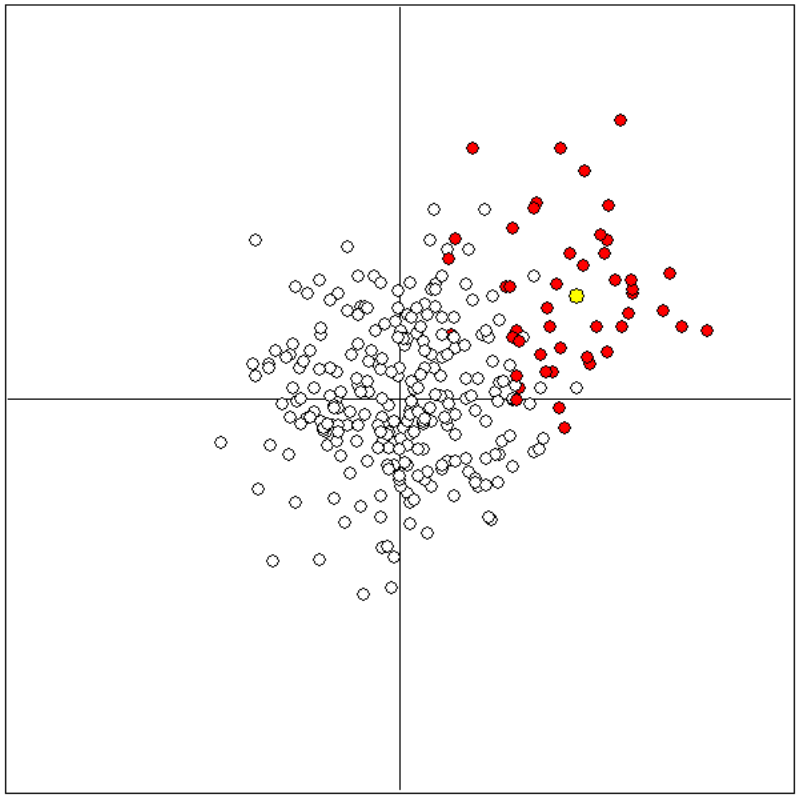

Рис. 1.3.

$n \in \mathcal{N}$, где $\mathcal{N} = \{1, 2, \ldots, N\}$, обладает свойством

$$x_n = \begin{cases} w, & n \in \mathcal{M}, \\ 0, & n \in \mathcal{N} \setminus \mathcal{M}, \end{cases} \tag{1.1}$$

где $\mathcal{M} \subseteq \mathcal{N}$.

Предположим, что доступная для обработки (наблюдаемая) последовательность имеет вид

$$y_n = x_n + e_n, \; n \in \mathcal{N}, \tag{1.2}$$

где $e_n \in \mathbb{R}^d$ — ошибка, независимая от $x_n$.

Учитывая структуру последовательности (1.1), определим квадратичный

функционал

$$S(\mathcal{M}, w) = \sum_{n \in \mathcal{N}} \|y_n - x_n\|^2, \quad \mathcal{M} \subseteq \mathcal{N},\ \omega \in \mathbb{R}^d, \tag{1.3}$$

и сформулируем следующую оптимизационную задачу.

**Задача.** *Дано:* последовательность $y_n \in \mathbb{R}^d$, $n \in \mathcal{N}$. *Найти:* подмножество $\mathcal{M} \subseteq \mathcal{N}$ и точку $\omega \in \mathbb{R}^d$, минимизирующие функционал (1.3), при условии, что структура входной последовательности задаётся формулами (1.1) и (1.2).

По сути, это задача аппроксимации последовательности $y_n$, $n \in \mathcal{N}$, последовательностью $x_n$, $n \in \mathcal{N}$, по критерию минимума суммы квадратов уклонений (расстояний). К этой же задаче аппроксимации приводит [25, 26] статистический подход к решению упомянутой содержательной проблемы с использованием критерия максимума правдоподобия в случае, когда $e_n$ является выборкой единичного объёма из $d$-мерного гауссовского распределения с параметрами $(0, \sigma^2 I)$, где $I$ — единичная матрица, $\sigma$ — фиксированный параметр.

Учитывая предположение (1.1) о структуре последовательности $x_n$, $n \in \mathcal{N}$, легко убедиться (например, с помощью дифференцирования), что минимум по $w \in \mathbb{R}^d$ функционала (1.3) доставляется вектором $w = \frac{1}{|\mathcal{M}|} \sum\limits_{n \in \mathcal{M}} y_n$. Этот факт индуцирует задачу 2-разбиения конечного множества точек евклидова пространства, которая рассматривается в главе 2. В настоящей работе исследуются два варианта этой задачи: с оптимизируемыми мощностями подмножеств и с ограничениями на мощности подмножеств. Формулировки этих вариантов задачи, а также характеристики существующих алгоритмов приведены в главе 2. Здесь лишь приведём отличия этой задачи от двухкластерного варианта задачи MSSC, который имеет следующую формулировку.

**Задача 2-MSSC.** *Дано:* $N$-элементное множество $\mathcal{Y} \subset \mathbb{R}^d$. *Найти:* разбиение множества $\mathcal{Y}$ на два кластера $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$ такое, что

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

где $\overline{y}(\mathcal{C}) = \dfrac{1}{|\mathcal{C}|} \sum\limits_{y \in \mathcal{C}} y$ и $\overline{y}(\mathcal{Y} \setminus \mathcal{C}) = \dfrac{1}{|\mathcal{Y} \setminus \mathcal{C}|} \sum\limits_{y \in \mathcal{Y} \setminus \mathcal{C}} y$ — геометрические центры (центроиды) кластеров.

Как и в задаче 2-MSSC, в задачах, рассматриваемых в главе 2, требуется разбить конечное множество точек евклидова пространства на два кластера по критерию минимума суммы по обоим кластерам суммарных внутрикластерных разбросов, причём одна из внутрикластерных сумм — такая же, как в задаче 2-MSSC, т.е. это сумма квадратов расстояний от элементов кластера до неизвестного центроида $\overline{y}(\mathcal{C})$, а другая — сумма квадратов расстояний от элементов кластера до фиксированного в произвольной точке евклидова пространства центра; без ограничения общности фиксированным центром может служить начало координат. Формально, в рассматриваемых задачах требуется минимизировать сумму $\sum\limits_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + \sum\limits_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2$ двух внутрикластерных разбросов. В этой сумме, в отличие от суммы, фигурирующей в целевой функции задачи 2-MSSC, вместо центроида $\overline{y}(\mathcal{Y} \setminus \mathcal{C})$ фигурирует фиксированный центр, совпадающий с началом координат.

Как и задача MSSC, задачи, в которых фиксирован центр одного из кластеров, характерны для проблем интерпретации данных (Data Mining), машинного обучения (Machine Learning), распознавания образов (Pattern Recognition), очистки данных (Data Cleaning), проблем редактирования данных (Data Reduction) [30–36].

Во многих естественнонаучных и технических приложениях требуется клас-

сификация упорядоченных по времени данных численных экспериментов или результатов наблюдения за состояниями каких-либо материальных объектов [37–39]. Рассмотрим одну из содержательных постановок, характерных для таких приложений.

Имеется последовательность, содержащая упорядоченные по времени результаты измерения набора характеристик некоторого объекта, который может находиться в одном из двух состояний: активном и пассивном. В пассивном состоянии все характеристики объекта равны нулю, а в активном — значение хотя бы одной характеристики не равно нулю. В каждом результате измерения содержится ошибка, а соответствие результатов измерений состояниям объекта неизвестно. Однако, известно, что временной интервал между двумя последовательными активными состояниями объекта ограничен сверху и снизу некоторыми константами $T_{\min}$ и $T_{\max}$. Требуется разбить последовательность на два кластера (подпоследовательности), соответствующих активному и пассивному состояниям объекта, и оценить набор характеристик объекта в активном состоянии.

Формализация приведённой содержательной проблемы, в частности, в виде задачи аппроксимации (суть которой описана ниже) по критерию минимума суммы квадратов уклонений индуцирует [25, 26, 40–42] рассматриваемую в главе 3 дискретную экстремальную задачу 2-разбиения последовательности $\mathcal{Y} = (y_1, \ldots, y_N)$ точек из $\mathbb{R}^d$ по критерию минимума функционала $\sum\limits_{j \in \mathcal{M}} \|y_j - \overline{y}(\mathcal{M})\|^2 + \sum\limits_{i \in \mathcal{N} \setminus \mathcal{M}} \|y_i\|^2$, где $\overline{y}(\mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum\limits_{j \in \mathcal{M}} y_j$, по $\mathcal{M} = (n_1, \ldots, n_M)$, где $n_m \in \mathcal{N} = \{1, \ldots, N\}$, $m = 1, \ldots, M$, с дополнительными ограничениями на разность между номерами элементов, входящих в первый кластер: $T_{\min} \leq n_m - n_{m-1} \leq T_{\max} \leq N$, $m = 2, \ldots, M$.

Задача аппроксимации, которая приводит к рассматриваемой экстремальной

задаче, состоит в поиске приближения входной последовательности $\mathcal{Y}$ искомой последовательностью вида

$$\ldots 0x0 \ldots 0x0 \ldots 0x0 \ldots, \tag{1.4}$$

в которой $x \in \mathbb{R}^d$ — неизвестная ненулевая точка (соответствующая активному состоянию), 0 — начало координат (соответствующее пассивному состоянию), а число нулей между ненулевыми точками неизвестно и определяется ограничениями на разность между номерами элементов, соответствующих активному состоянию. При этом найденная оптимальная аппроксимирующая последовательность имеет вид

$$\ldots 0\overline{y}(\mathcal{M})0 \ldots 0\overline{y}(\mathcal{M})0 \ldots 0\overline{y}(\mathcal{M})0 \ldots$$

В этой последовательности центроид $\overline{y}(\mathcal{M})$ кластера $\{y_j \,|\, j \in \mathcal{M}\}$ и номера из наборов $\mathcal{M}$ и $\mathcal{N} \setminus \mathcal{M}$ определяются в результате решения оптимизационной задачи. Кластеры $\{y_j \,|\, j \in \mathcal{M}\}$ и $\{y_i \,|\, i \in \mathcal{N} \setminus \mathcal{M}\}$ соответствуют активному и пассивному состояниям объекта, а центроид $\overline{y}(\mathcal{M})$ первого кластера является оценкой для точки $x$.

Одна из статистических постановок проблемы помехоустойчивого анализа временных рядов также индуцирует рассматриваемую дискретную экстремальную задачу разбиения последовательности (см., например, [25, 26, 40]).

От задач разбиения множества, рассматриваемых в главе 2, задачу 2-разбиения последовательности, рассматриваемую в главе 3, отличают несколько особенностей. Во-первых, входом задачи является не множество, а последовательность. Во-вторых, имеются ограничения на номера элементов подпоследовательностей, включаемых в кластеры. При этом, как и в задачах из главы 2,

предполагается, что центр одного из кластеров фиксирован в начале координат.

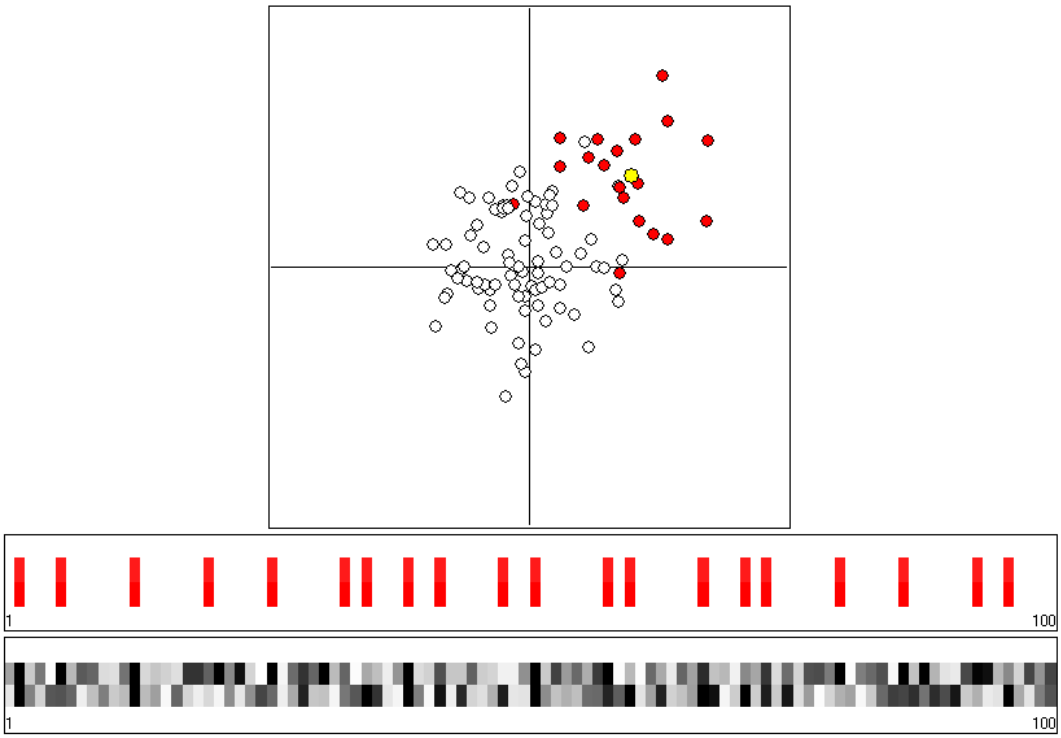


Рис. 1.4.

На рисунке 1.4 приведён пример, иллюстрирующий отличия задач кластеризации множества и последовательности. В верхней части рисунка, как и на рисунке 1.3, приведено множество $\{y_1, \ldots, y_N\}$ точек, соответствующих результатам измерений двумерного набора характеристик некоторого объекта, находившегося в одном из двух состояний — активном и пассивном. В нижней части рисунка эти же точки представлены в виде двух последовательностей вертикальных полос. Верхняя (ненаблюдаемая) последовательность соответствует (1.4). Интервалы между полосками соответствуют временным интервалам между двумя последовательными активными состояниями объекта. Нижняя последовательность является входом $\mathcal{Y}$ задачи. Эту последовательность требуется разбить на два кластера так, чтобы один из кластеров соответствовал активному состоянию (совокупности полосок верхней последовательности) объекта, а второй кластер — пассивному.

На практике нередко возникают сходные содержательные проблемы, в которых один и тот же объект может находиться в нескольких (более одного) активных состояниях и пассивном. Одна из таких проблем формулируется следующим образом.

Дана последовательность $\mathcal{Y}$, содержащая $N$ упорядоченных по времени результатов $y_1, \ldots, y_N$ измерения набора $y$ из $d$ числовых характеристик некоторого объекта, который может находиться в $L+1$ состояниях. Среди этих состояний $L$ активных и одно пассивное. В пассивном состоянии все элементы набора равны нулю, а в каждом из активных — хотя бы одна из компонент набора не равна нулю. Измерения сопровождаются инструментальной ошибкой. Известно, что объект некоторое время находится в одном из активных состояний, а затем переключается в другое активное состояние. При этом все активные состояния объекта сопровождаются переключениями в пассивное состояние на некоторое ограниченное сверху и снизу неизвестное время. Кроме того, известны (заданы) натуральные числа $T_{\min}$ и $T_{\max}$, которые соответствуют минимальному и максимальному интервалам времени между любыми двумя последовательными активными состояниями объекта. Соответствие элемента последовательности какому-либо состоянию объекта неизвестно. Требуется найти в последовательности все элементы, соответствующие активным состояниям объекта, и оценить характеристики объекта в каждом из активных состояний.

Легко проверить, что формализация этой содержательной проблемы с использованием критерия минимума суммы квадратов уклонений индуцирует следующую задачу аппроксимации. Дана последовательность $y_n \in \mathbb{R}^d$, $n \in \mathcal{N}$, натуральные числа $T_{\min}$, $T_{\max}$ и $L$. Требуется найти аппроксимирующую после-

довательность $z_n \in \mathbb{R}^d$, $n \in \mathcal{N}$, вида

$$z_n = \begin{cases} x_1, & n \in \mathcal{M}_1, \\ \dots & \\ x_L, & n \in \mathcal{M}_L, \\ 0, & n \in \mathcal{N} \backslash \mathcal{M}, \end{cases} \tag{1.5}$$

где $\mathcal{M}_1, \dots, \mathcal{M}_L$ — непустые непересекающиеся наборы номеров элементов последовательности $\mathcal{Y}$, $\mathcal{M} = \bigcup\limits_{l=1}^{L} \mathcal{M}_l$, а $x_1, \dots, x_L$ — неизвестные ненулевые точки из $\mathbb{R}^d$, такую, что

$$\sum_{i \in \mathcal{N}} \|y_i - z_i\|^2 \longrightarrow \min, \tag{1.6}$$

при ограничениях: (i) в последовательности, образованной конкатенацией наборов $\mathcal{M}_1, \dots, \mathcal{M}_L$, номера упорядочены по возрастанию при условии, что элементы каждого набора образуют возрастающую последовательность, и (ii) номера из объединённого набора $\mathcal{M} = (n_1, \dots, n_M)$ связаны неравенствами $T_{\min} \leq n_m - n_{m-1} \leq T_{\max} \leq N$, $m = 2, \dots, M$.

Схематически, участок последовательности $z_n$, $n \in \mathcal{N}$, можно представить в виде

$$\dots 0 x_{l-1} 0 \dots 0 x_{l-1} 0 \dots \dots 0 x_l 0 \dots 0 x_l 0 \dots \tag{1.7}$$

Здесь $x_{l-1}, x_l \in \mathbb{R}^d$ — неизвестные ненулевые точки (соответствующие $(l-1)$-му и $l$-му активным состояниям объекта), 0 — начало координат (соответствующее пассивному состоянию), а число нулей между ненулевыми точками неизвестно и лежит в допустимом интервале от $T_{\min} - 1$ до $T_{\max} - 1$ в соответствии с ограничениями (ii).

Раскрыв сумму (1.6) с учетом (1.5) и сгруппировав члены, нетрудно про-

верить с помощью дифференцирования, что оптимальными в смысле (1.6) являются значения $x_l = \overline{y}(\mathcal{M}_l) = \frac{1}{|\mathcal{M}_l|} \sum\limits_{j \in \mathcal{M}_l} y_j$, $l = 1, \ldots, L$, а сформулированная задача аппроксимации индуцирует задачу минимизации целевой функции $\sum\limits_{l=1}^{L} \sum\limits_{j \in \mathcal{M}_l} \|y_j - \overline{y}(\mathcal{M}_l)\|^2 + \sum\limits_{i \in \mathcal{N} \setminus \mathcal{M}} \|y_i\|^2$. Эта задача рассматривается в главе 3.

Легко заметить, что в найденной оптимальной аппроксимирующей последовательности участок, соответствующий (1.7), имеет вид:

$$\ldots 0\overline{y}(\mathcal{M}_{l-1})0 \ldots 0\overline{y}(\mathcal{M}_{l-1})0 \ldots\ldots 0\overline{y}(\mathcal{M}_l)0 \ldots 0\overline{y}(\mathcal{M}_l)0 \ldots$$

В этой последовательности для всех $l = 1, \ldots, L$ номера из набора $\mathcal{M}_l$, кластер $\{y_j \,|\, j \in \mathcal{M}_l\}$ и его центроид $\overline{y}(\mathcal{M}_l)$ определяются в результате решения задачи минимизации указанной выше целевой функции. Центроид $\overline{y}(\mathcal{M}_l)$ является оценкой для точки $x_l$.

Из приведённой выше схематичной строковой записи последовательностей видно, что их можно интерпретировать как последовательности, содержащие участки с серийными квазипериодическими (в силу ограничений (ii)) повторами. Если условиться о границах серий, например, по первому (или по последнему) повтору, то упомянутые выше задачи можно трактовать как задачи разбиения последовательности на серийные участки с квазипериодическими повторами неизвестных точек совместно с оцениванием точек и отысканием их положения в последовательности.

На рисунке 1.5 приведён пример, демонстрирующий отличительные особенности задачи разбиения последовательности на несколько серийных кластеров. В верхней части рисунка приведено множество точек, соответствующих результатам измерений двумерного набора характеристик некоторого объекта, находившегося в одном из пяти состояний, среди которых четыре активных и одно пассивное. В нижней части рисунка, как и на рисунке 1.4, эти же точки пред-

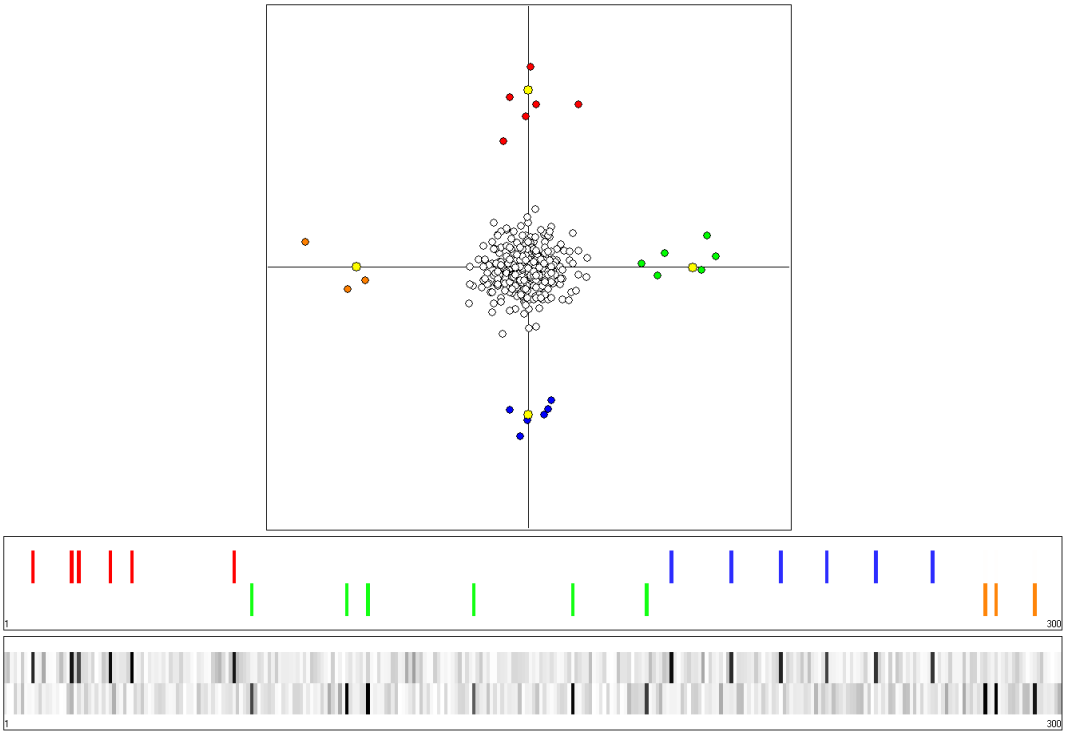


Рис. 1.5.

ставлены в виде двух последовательностей вертикальных полос. Верхняя последовательность (ненаблюдаемая) соответствует (1.5); как и на рисунке 1.4, интервалы между полосками соответствуют временным интервалам между двумя последовательными активными состояниями объекта. Нижняя последовательность является входом задачи. Эту последовательность требуется разбить на пять кластеров так, чтобы четыре из них соответствовали активным состояниям объекта, а один — пассивному.

Рисунок 1.6 иллюстрирует содержательную задачу совместного помехоустойчивого разбиения одномерного сигнала (последовательности), представленного на нижней части рисунка, на серийные участки, содержащие квазипериодически повторяющиеся идентичные импульсы, и обнаружения этих импульсов. Легко показать, что эта задача также индуцирует рассматриваемую в главе 3 задачу многокластерного разбиения последовательности.

В целях удобства восприятия материала обзор результатов по рассматри-

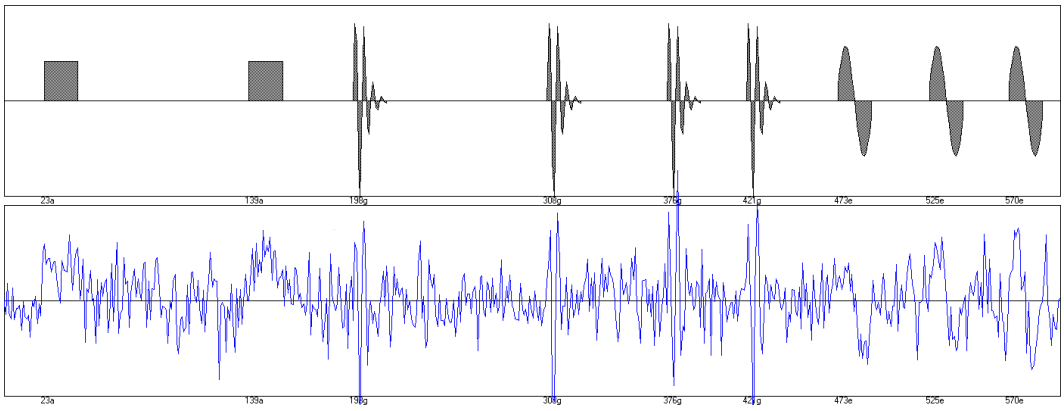


Рис. 1.6.

ваемым задачам приведен после их формулировок в соответствующих главах. Заметим, что все рассматриваемые задачи относятся к числу слабоизученных в алгоритмическом плане задач дискретной оптимизации. По-видимому, это связано с тем, что постановки этих задач появились в публикациях лишь в начале текущего века, а факты труднорешаемости задач установлены около десяти лет назад. Здесь следует напомнить, что (пока верна гипотеза тысячелетия о несовпадении классов P и NP) единого метода построения полиномиальных приближённых алгоритмов с гарантированными оценками качества для NP-трудных задач не существует. Каждая из таких задач требует индивидуального алгоритмического подхода. Перечисленные факты определили направления исследований и характер ряда полученных алгоритмических результатов, представленных ниже.

# Глава 2

# Кластеризация конечного множества точек евклидова пространства

В этой главе рассматриваются задачи 2-кластеризации конечного множества точек евклидова пространства при фиксированном (в начале координат) центре одного из кластеров.

## 2.1 Задача 2-кластеризации с оптимизируемыми мощностями кластеров

### 2.1.1 Формулировка задачи и известные результаты

**Задача 1.** *Дано:* множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ точек из $\mathbb{R}^d$. *Найти:* разбиение множества $\mathcal{Y}$ на два кластера $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$ такое, что

$$S(\mathcal{C}) = \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 \to \min, \tag{2.1}$$

где $\overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum\limits_{y \in \mathcal{C}} y$ — центроид множества $\mathcal{C}$.

Заметим, что для произвольного множества $\mathcal{Y}$ точек из $\mathbb{R}^d$ и для произвольного подмножества $\mathcal{C} \subseteq \mathcal{Y}$ справедливо следующее легко проверяемое равенство:

$$\sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y\|^2 = \sum_{y \in \mathcal{Y}} \|y\|^2 - \frac{1}{|\mathcal{C}|} \left\| \sum_{y \in \mathcal{C}} y \right\|^2 . \tag{2.2}$$

Поэтому задаче 1 полиномиально эквивалентна следующая задача поиска подмножества с максимальным нормированным квадратом длины суммы:

**Задача MNLVS** (subset with the Maximum Normalized Length of Vector Sum)**.** *Дано:* множество $\mathcal{Y} = \{y_1, \ldots, y_N\} \subseteq \mathbb{R}^d$. *Найти:* подмножество $\mathcal{C} \subseteq \mathcal{Y}$ такое, что

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

В [28, 29] была показана NP-трудность задачи MNLVS в сильном смысле путём сведения к ней классической труднорешаемой в сильном смысле задачи 3-SAT. Из этого результата следует NP-трудность в сильном смысле задачи 1.

Далее заметим, что из формулы (2.2), связывающей целевые функции задач на минимум и на максимум, следует, что точные алгоритмы для задачи MNLVS можно применять для получения оптимальных решений задачи 1. К таким алгоритмам относится предложенный в [43] точный алгоритм для задачи MNLVS, который имеет трудоёмкость $\mathcal{O}(d^2 N^{2d})$. Кроме того, в работе [44] был обоснован точный алгоритм для задачи MNLVS, имеющий временную сложность $\mathcal{O}(dN^{d+1})$. Из этих результатов следует, что обе задачи являются полиномиально разрешимыми при фиксированной размерности пространства.

В [29] для задачи MNLVS предложен алгоритм с гарантированной оценкой относительной погрешности $\varepsilon = (d-1)/(4l^2)$, где $l$ — целочисленный параметр алгоритма. Временная сложность алгоритма есть величина $\mathcal{O}\big(Nd(d+\log N)(2l+1)^{d-1}\big)$. Фактически, в [29] обоснована вполне полиномиальная аппроксимацион-

ная схема (FPTAS), устанавливающая полиномиальную относительно $N$ и $1/\varepsilon$ оценку временной сложности алгоритма, для случая, когда размерность $d$ пространства фиксирована. Однако, следует отметить, что этот алгоритм не обеспечивает гарантированной оценки точности для полиномиально эквивалентной задачи 1.

В работе [45] приведён 2-приближённый алгоритм с оценкой трудоёмкости $\mathcal{O}(dN^2)$ для варианта задачи 1, в котором мощность искомого кластера фиксирована. Очевидно, с помощью этого алгоритма можно получить приближённое решение задачи 1, перебирая возможные значения мощности кластера $\mathcal{C}$, за время $\mathcal{O}(dN^3)$. В данной работе предложен менее трудоёмкий эффективный алгоритм, имеющий временную сложность $\mathcal{O}(dN^2)$ при той же, что и в [45], оценке точности.

### 2.1.2 Основы алгоритма

Следующие две леммы относятся к числу хорошо известных, а их доказательства представлены во многих публикациях (см., например, [45, 46]).

**Лемма 2.1.** *Для произвольной точки $x \in \mathbb{R}^d$ и конечного множества $\mathcal{Z} \subset \mathbb{R}^d$ имеет место равенство*

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

*где $\overline{z}$ — центроид множества $\mathcal{Z}$.*

**Лемма 2.2.** *Пусть $\mathcal{Z}$ — непустое конечное множество точек из $\mathbb{R}^d$, а $\overline{z}$ — центроид множества $\mathcal{Z}$. Тогда если точка $x \in \mathbb{R}^d$ удовлетворяет условиям*

$$\|x - \overline{z}\| \leq \|z - \overline{z}\|, \qquad \forall z \in \mathcal{Z},$$

*то имеет место неравенство*

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

Заметим, что леммы 2.1 и 2.2 остаются справедливыми в случае, когда $\mathcal{Z}$ является мультимножеством.

Рассмотрим следующую вспомогательную задачу.

**Задача 1′.** *Дано:* множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ точек из $\mathbb{R}^d$. *Найти:* подмножество $\mathcal{B} \subseteq \mathcal{Y}$ и точку $x \in \mathcal{Y}$ такие, что

$$Q(\mathcal{B}, x) = \sum_{y\in\mathcal{B}} \|y-x\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{B}} \|y\|^2 \to \min\,. \tag{2.3}$$

Для целевой функции (2.3) этой задачи справедлива следующая

**Лемма 2.3.** *При любом фиксированном подмножестве $\mathcal{B} \subseteq \mathcal{Y}$ минимум функционала* (2.3) *достигается в точке $\overline{y}(\mathcal{B}) = \frac{1}{|\mathcal{B}|}\sum\limits_{y\in\mathcal{B}} y$ и равен $S(\mathcal{B})$.*

*Доказательство.* Справедливость этой леммы следует из леммы 2.1 и следующей цепочки оценок:

$$\begin{aligned} Q(\mathcal{B}, x) = \sum_{y\in\mathcal{B}} \|y-x\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{B}} \|y\|^2 = \\ \sum_{y\in\mathcal{B}} \|y-\overline{y}(\mathcal{B})\|^2 + |\mathcal{B}|\cdot\|x-\overline{y}(\mathcal{B})\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{B}} \|y\|^2 = \\ S(\mathcal{B}) + |\mathcal{B}|\cdot\|x-\overline{y}(\mathcal{B})\|^2 \ge S(\mathcal{B}). \end{aligned}$$

Лемма доказана.

Для фиксированной точки $x \in \mathcal{Y}$ положим $\mathcal{B}^*(x) = \arg\min\limits_{\mathcal{B}\subseteq\mathcal{Y}} Q(\mathcal{B}, x)$.

**Лемма 2.4.** *Пусть $x \in \mathcal{Y}$. Тогда для любых точек $u \in \mathcal{B}^*(x)$ и $v \in \mathcal{Y} \setminus \mathcal{B}^*(x)$ имеют место неравенства:* 1) $\|u - x\| \le \|u\|$; 2) $\|v - x\| \ge \|v\|$.

*Доказательство.* Докажем первое неравенство. Предположим противное: существует точка $b \in \mathcal{B}^*(x)$, для которой $\|b - x\| > \|b\|$. Тогда

$$\begin{aligned} Q\big(\mathcal{B}^*(x), x\big) &= \sum_{y \in \mathcal{B}^*(x)} \|y - x\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{B}^*(x)} \|y\|^2 \\ &= \sum_{y \in \mathcal{B}^*(x) \setminus \{b\}} \|y - x\|^2 + \|b - x\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{B}^*(x)} \|y\|^2 \\ &> \sum_{y \in \mathcal{B}^*(x) \setminus \{b\}} \|y - x\|^2 + \|b\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{B}^*(x)} \|y\|^2 \\ &= \sum_{y \in \mathcal{B}^*(x) \setminus \{b\}} \|y - x\|^2 + \sum_{y \in \mathcal{Y} \setminus (\mathcal{B}^*(x) \setminus \{b\})} \|y\|^2 = Q\big(\mathcal{B}^*(x) \setminus \{b\}, x\big), \end{aligned}$$

что противоречит определению множества $\mathcal{B}^*(x)$.

Аналогичным образом, для доказательства второго неравенства допустим противное: существует точка $b \in \mathcal{Y} \setminus \mathcal{B}^*(x)$, для которой $\|b - x\| < \|b\|$. Тогда

$$\begin{aligned} Q\big(\mathcal{B}^*(x), x\big) &= \sum_{y \in \mathcal{B}^*(x)} \|y - x\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{B}^*(x)} \|y\|^2 \\ &= \sum_{y \in \mathcal{B}^*(x)} \|y - x\|^2 + \|b\|^2 + \sum_{y \in \mathcal{Y} \setminus (\mathcal{B}^*(x) \cup \{b\})} \|y\|^2 \\ &> \sum_{y \in \mathcal{B}^*(x)} \|y - x\|^2 + \|b - x\|^2 + \sum_{y \in \mathcal{Y} \setminus (\mathcal{B}^*(x) \cup \{b\})} \|y\|^2 \\ &= \sum_{y \in \mathcal{B}^*(x) \cup \{b\}} \|y - x\|^2 + \sum_{y \in \mathcal{Y} \setminus (\mathcal{B}^*(x) \cup \{b\})} \|y\|^2 = Q\big(\mathcal{B}^*(x) \cup \{b\}, x\big), \end{aligned}$$

что противоречит определению множества $\mathcal{B}^*(x)$. Лемма доказана.

Лемма 2.4 имеет наглядный геометрический смысл. Действительно, равенство $\|y - x\| = \|y\|$ равносильно равенству $2\langle y, x\rangle = \|x\|^2$, которое при фиксиро-

ванной точке $x$ определяет гиперплоскость, перпендикулярную отрезку, соединяющему начало координат с точкой $x$, и проходящую через его середину. При этом оптимальное множество $\mathcal{B}^*(x)$ лежит в полупространстве, не включающем начало координат.

Кроме того, для целевой функции (2.3) задачи $1'$ справедливо следующее очевидное

**Свойство.** *Пусть* $x \in \mathcal{Y}$, $\mathcal{B} \subseteq \mathcal{Y}$. *Тогда если* $b \in \mathcal{B}$ *и* $\|b - x\| = \|b\|$, *то* $Q(\mathcal{B}, x) = Q\big(\mathcal{B} \setminus \{b\}, x\big)$.

В соответствии с леммой 2.4 определим множества

$$\mathcal{B}(x) = \Big\{ y \in \mathcal{Y} \;\Big|\; 2\langle y, x\rangle > \|x\|^2 \Big\}; \tag{2.4}$$

$$\mathcal{B}'(x) = \Big\{ y \in \mathcal{Y} \;\Big|\; 2\langle y, x\rangle \geq \|x\|^2 \Big\}.$$

Из леммы 2.4 и сформулированного свойства функции $Q(\mathcal{B}, x)$ вытекает

**Следствие 2.1.** *Для любой точки* $x \in \mathcal{Y}$ *множество* $\mathcal{B}^*$ *доставляет минимум функции* $Q(\mathcal{B}, x)$ *тогда и только тогда, когда* $\mathcal{B}(x) \subseteq \mathcal{B}^* \subseteq \mathcal{B}'(x)$.

Это утверждение позволяет сформулировать следующий алгоритм решения задачи $1'$.

**Алгоритм $\mathcal{A}_1'$.**

*Вход алгоритма:* множество $\mathcal{Y}$.

Для каждой точки $x \in \mathcal{Y}$ выполним шаги 1, 2.

`Шаг 1`. Формируем множество $\mathcal{B}(x)$ по формуле (2.4).

`Шаг 2`. Вычислим значение $Q\big(\mathcal{B}(x), x\big)$ целевой функции (2.3).

`Шаг 3`. В качестве решения задачи выберем такую точку $x_A \in \mathcal{Y}$ и соответствующее ей множество $\mathcal{B}_A = \mathcal{B}(x_A)$, что значение функции $Q(\mathcal{B}_A, x_A)$ минимально. Если оптимальных значений несколько, то выберем любое из них.

*Выход алгоритма:* множество $\mathcal{B}_A$ и точка $x_A$.

**Лемма 2.5.** *Алгоритм $\mathcal{A}_1'$ находит оптимальное решение задачи $1'$ за время $\mathcal{O}(dN^2)$.*

*Доказательство.* Оптимальность вытекает из следствия 2.1 и цепочки равенств

$$\min_{\mathcal{B}\subseteq\mathcal{Y},\, x\in\mathcal{Y}} Q(\mathcal{B},x) = \min_{x\in\mathcal{Y}} \min_{\mathcal{B}\subseteq\mathcal{Y}} Q(\mathcal{B},x) = \min_{x\in\mathcal{Y}} Q\big(\mathcal{B}(x),x\big).$$

Оценим временную сложность алгоритма.

Для фиксированной точки $x \in \mathcal{Y}$ формирование множества $\mathcal{B}(x)$ требует $\mathcal{O}(dN)$ операций, как и вычисление функции $Q\big(\mathcal{B}(x),x\big)$. Поэтому для каждой из $N$ точек $x \in \mathcal{Y}$ трудоёмкость шагов 1 и 2 составляет $\mathcal{O}(dN)$. Шаг 3 — поиск наименьшего элемента — требует не более $\mathcal{O}(N)$ операций. Таким образом, итоговая временная сложность алгоритма есть величина $\mathcal{O}(dN^2)$. Лемма доказана.

**Замечание 2.1.** *В алгоритме $\mathcal{A}_1'$ можно вместо множеств $\mathcal{B}(x)$ использовать множества $\mathcal{B}'(x)$. Ввиду следствия* 2.1 *полученное таким образом решение тоже будет оптимальным.*

### 2.1.3 2-приближённый полиномиальный алгоритм

Сформулируем алгоритм решения задачи 1.

**Алгоритм $\mathcal{A}_1$.**

*Вход алгоритма:* множество $\mathcal{Y}$.

`Шаг 1`. По заданному множеству $\mathcal{Y}$ находим оптимальное решение $\mathcal{B}_A$, $x_A$ вспомогательной задачи $1'$ с помощью алгоритма $\mathcal{A}_1'$.

`Шаг 2`. Подмножество $\mathcal{B}_A$ объявляем решением задачи 1.

*Выход алгоритма:* множество $\mathcal{B}_A$.

**Теорема 2.1.** *Алгоритм $\mathcal{A}_1$ находит 2-приближённое решение задачи 1 за время $\mathcal{O}(dN^2)$. Оценка 2 точности алгоритма достижима.*

*Доказательство.* Из леммы 2.3 следует оценка

$$S(\mathcal{B}_A) \le Q(\mathcal{B}_A, x_A). \tag{2.5}$$

Пусть $\mathcal{C}^*$ — оптимальное решение задачи 1. Рассмотрим множество $\mathcal{C}^*$ и точку $t = \arg\min\limits_{y \in \mathcal{C}^*} \|y - y^*\|$, где $y^* = \overline{y}(\mathcal{C}^*) = \frac{1}{|\mathcal{C}^*|} \sum\limits_{y \in \mathcal{C}^*} y$ (точка $t$ — ближайшая к $y^*$ точка в оптимальном множестве $\mathcal{C}^*$). Поскольку они удовлетворяют условиям леммы 2.2, имеет место неравенство

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

и, следовательно,

$$\begin{aligned} Q(\mathcal{C}^*, t) = \sum_{y \in \mathcal{C}^*} \|y - t\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}^*} \|y\|^2 &\le 2 \sum_{y \in \mathcal{C}^*} \|y - y^*\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}^*} \|y\|^2 \\ &\le 2 \sum_{y \in \mathcal{C}^*} \|y - y^*\|^2 + 2 \sum_{y \in \mathcal{Y} \setminus \mathcal{C}^*} \|y\|^2 = 2S(\mathcal{C}^*). \end{aligned} \tag{2.6}$$

Кроме того, заметим, что $\mathcal{C}^*$, $t$ — допустимое решение вспомогательной задачи $1'$, а $\mathcal{B}_A$, $x_A$ — её оптимальное решение. Поэтому

$$Q(\mathcal{B}_A, x_A) \le Q(\mathcal{C}^*, t). \tag{2.7}$$

Объединяя (2.5)–(2.7), получим оценку

$$S(\mathcal{B}_A) \le Q(\mathcal{B}_A, x_A) \le Q(\mathcal{C}^*, t) \le 2S(\mathcal{C}^*).$$

Достижимость оценки точности алгоритма следует из следующего примера.

Пусть $d = 2,\ N = 2,\ y_1 = (0, \alpha),\ y_2 = (1, \alpha)$. Тогда если $0 < \alpha < 1$, то $\mathcal{B}_A = \{y_2\}$, $\mathcal{C}^* = \{y_1, y_2\}$, $S(\mathcal{B}_A) = \alpha^2$, $S(\mathcal{C}^*) = 1/2$. Таким образом, отношение $S(\mathcal{B}_A)/S(\mathcal{C}^*) = 2\alpha^2$ может быть сколь угодно близко к 2 при $\alpha \to 1$.

Если же $\alpha = 1$, то имеем два оптимальных решения: либо $\mathcal{B}_A = \{y_1, y_2\}$, либо $\mathcal{B}_A = \{y_2\}$. При этом для второго решения $S(\mathcal{B}_A) = 1$; $\mathcal{C}^* = \{y_1, y_2\}$, $S(\mathcal{C}^*) = 1/2$ и $S(\mathcal{B}_A)/S(\mathcal{C}^*) = 2$, т.е. оценка точности алгоритма достижима. Теорема доказана.

## 2.2 Задача 2-кластеризации с ограничениями на мощности кластеров

### 2.2.1 Формулировка задачи и известные результаты

**Задача 2.** *Дано:* множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ точек из $\mathbb{R}^d$ и натуральное число $M$. *Найти:* разбиение множества $\mathcal{Y}$ на два кластера $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$ такое, что целевая функция (2.1) минимальна, при ограничении $|\mathcal{C}| = M$.

Из равенства (2.2) следует, что задаче 2 эквивалентна задача поиска подмножества с максимальной длиной суммы:

**Задача LVS** (subset with the Longest Vector Sum)**.** *Дано:* множество $\mathcal{Y} = \{y_1, \ldots, y_N\} \subseteq \mathbb{R}^d$ и натуральное число $M$. *Найти:* подмножество $\mathcal{C} \subseteq \mathcal{Y}$ такое, что

$$\left\| \sum_{y \in \mathcal{C}} y \right\| \to \max,$$

при ограничении $|\mathcal{C}| = M$.

В [25–27] было доказано, что задача LVS является NP-трудной в сильном смысле. Из этого факта и равенства (2.2) следует, что задача 2 также NP-трудна в сильном смысле.

В работе [43] был предложен алгоритм, позволяющий находить точное решение задачи LVS за время $\mathcal{O}(d^2N^{2d})$. Позже в работе [44] был обоснован точный алгоритм, имеющий временную сложность $\mathcal{O}(dN^{d+1})$. Таким образом, и задача LVS, и задача 2 являются полиномиально разрешимыми при фиксированной размерности пространства.

Для случая задачи LVS, когда входные данные целочисленны, в [27] и [47] построены точные алгоритмы с трудоёмкостью $\mathcal{O}(Nd^{d+1}(MD)^{d-1})$ и $\mathcal{O}(dMN(2MD)^{d-1})$, соответственно, где $D$ — максимальное абсолютное значение компонент входных точек. При фиксированной размерности пространства оба алгоритма являются псевдополиномиальными.

В [27] для задачи LVS был обоснован алгоритм с гарантированной оценкой относительной погрешности $\frac{1}{8}(d-1)/l^2$ и трудоёмкостью $\mathcal{O}\big(d^2N(2l+1)^{d-1}\big)$, где $l$ — параметр алгоритма.

В [48] был предложен рандомизированный алгоритм, позволяющий за время $\mathcal{O}(dNl)$ находить $(1+\varepsilon)$-приближённое решение задачи LVS с вероятностью $1-\delta$, где $\varepsilon \le \varphi_0^2/2$, $\delta \le \exp\left(-\frac{\left(7/4\sin\frac{\varphi_0}{2}\right)^{d-1}}{\pi\sqrt{d}}l\right)$, а $l$ и $\varphi_0$ — параметры алгоритма; эти оценки впоследствии были улучшены в [49]. Кроме того, в [48] были найдены параметры алгоритма, для которых при фиксированной размерности пространства алгоритм является асимптотически точным и полиномиальным.

Отметим, что приближённые алгоритмы для задачи LVS не обеспечивают гарантированной оценки точности для полиномиально эквивалентной задачи 2. К числу найденных к настоящему времени эффективных алгоритмических решений с гарантированными оценками точности для задачи 2 относятся следующие.

В [45] предложен 2-приближённый полиномиальный алгоритм, трудоёмкость которого есть величина $\mathcal{O}(dN^2)$.

Полиномиальная приближённая схема (PTAS) обоснована в [50,51]. Эта схема позволяет решать задачу 2 с произвольной относительной погрешностью $\varepsilon$ за время $\mathcal{O}(dN^{2/\varepsilon+1}(9/\varepsilon)^{3/\varepsilon})$.

В настоящей работе для задачи 2 предложены точный псевдополиномиальный алгоритм для специального случая, аппроксимационная схема и рандомизированный алгоритм. Характеристики предложенных алгоритмов приведены в соответствующих параграфах.

### 2.2.2 Основы алгоритмов

В условиях задачи 2 для произвольной точки $x \in \mathbb{R}^d$ определим множество, состоящее из $M$ элементов множества $\mathcal{Y}$, имеющих наибольшие проекции на направление, задаваемое этой точкой:

$$\mathcal{B}_M(x) = \big\{ y_i \;\big|\; \langle y_i, x\rangle \geq \langle y_j, x\rangle;\ y_i, y_j \in \mathcal{Y},\ i \leq M,\ j > M \big\},$$

где $\langle \cdot, \cdot \rangle$ — скалярное произведение.

Справедлива следующая

**Лемма 2.6.** *Для любой фиксированной точки $x \in \mathbb{R}^d$ минимум функционала* (2.3) *по всем подмножествам $\mathcal{B} \subseteq \mathcal{Y}$ мощности $M$ достигается на множестве $\mathcal{B}_M(x)$.*

*Доказательство.* Справедливость этой леммы вытекает из следующей цепочки

равенств:

$$Q(\mathcal{B}, x) = \sum_{y \in \mathcal{B}} \|y - x\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{B}} \|y\|^2$$
$$= \sum_{y \in \mathcal{B}} (\|y - x\|^2 - \|y\|^2) + \sum_{y \in \mathcal{Y}} \|y\|^2 = \sum_{y \in \mathcal{Y}} \|y\|^2 + M \cdot \|x\|^2 - 2 \sum_{y \in \mathcal{B}} \langle y, x \rangle.$$

Остаётся заметить, что первые два слагаемых в правой части полученного выражения являются константами. Лемма доказана.

### 2.2.3 Точный псевдополиномиальный алгоритм для специального случая задачи

Суть предлагаемого алгоритмического решения состоит в следующем. В области пространства, определяемой максимальным абсолютным значением координат входных точек, строится многомерная равномерная по каждой координате решётка (сетка) кубической формы с рациональным шагом. Шаг решётки выбирается так, чтобы один из её узлов совпал с геометрическим центром одного из оптимизируемых кластеров. Для каждого узла построенной решётки решается задача минимизации вспомогательной целевой функции. В результате решения находится набор точек, доставляющий минимум этой функции. Найденный набор объявляется претендентом на решение. В качестве окончательного решения выбирается тот набор, для которого значение вспомогательной целевой функции минимально.

Допустим, что точки из множества $\mathcal{Y}$ имеют целочисленные координаты. Положим

$$D = \max_{y \in \mathcal{Y}} \max_{j \in \{1,\ldots,d\}} |(y)^j|, \tag{2.8}$$

где $(y)^j$ — $j$-я координата точки $y$.

Определим множество

$$\mathcal{G} = \left\{ x \;\middle|\; x \in \mathbb{R}^d, (x)^j = \frac{1}{M}(v)^j, (v)^j \in \mathbb{Z}, |(v)^j| \leq MD, j = 1, \ldots, d \right\}$$

— многомерную решётку кубической формы размера $2D$ с расстоянием $\frac{1}{M}$ между узлами и центром в начале координат. Заметим, что

$$|\mathcal{G}| = (2MD + 1)^d.$$

Сформулируем следующий алгоритм решения задачи 2.

**Алгоритм $\mathcal{A}_2$.**

*Вход алгоритма:* множество $\mathcal{Y}$, натуральное число $M$.

Шаг 1. Для каждой точки $x \in \mathcal{G}$ построим множество $\mathcal{B}_M(x)$. Вычислим значение $Q\big(\mathcal{B}_M(x), x\big)$.

Шаг 2. Найдём точку $x_A = \arg\min\limits_{x \in \mathcal{G}} Q\big(\mathcal{B}_M(x), x\big)$ (если минимальному значению соответствует несколько точек, то выберем любую из них) и соответствующее ей множество $\mathcal{B}(x_A)$. В качестве решения задачи возьмём множество $\mathcal{C}_A = \mathcal{B}(x_A)$.

*Выход алгоритма:* множество $\mathcal{C}_A$.

**Теорема 2.2.** *Пусть в условиях задачи* 2 *точки из множества $\mathcal{Y}$ имеют целочисленные координаты из интервала $[-D, D]$. Тогда алгоритм $\mathcal{A}_2$ находит оптимальное решение задачи* 2 *за время $\mathcal{O}\big(dN(2MD + 1)^d\big)$.*

*Доказательство.* Пусть $\mathcal{C}^*$ — оптимальное решение задачи 2, а $\mathcal{C}_A$ — множество, полученное в результате работы алгоритма $\mathcal{A}_2$.

Из определения (2.8) следует, что центр $\overline{y}(\mathcal{C}) = \frac{1}{M} \sum\limits_{y \in \mathcal{C}} y$ любого подмножества $\mathcal{C} \subseteq \mathcal{Y}$ мощности $M$ лежит во множестве $\mathcal{G}$. Следовательно, и центр $y^* = \overline{y}(\mathcal{C}^*)$ оптимального подмножества $\mathcal{C}^*$ лежит в этом же множестве.

По определению шага 2 имеем

$$Q(\mathcal{C}_A, x_A) \leq Q\big(\mathcal{B}_M(y^*), y^*\big). \tag{2.9}$$

Из леммы 2.3 следует неравенство

$$S(\mathcal{C}_A) \leq Q(\mathcal{C}_A, x_A), \tag{2.10}$$

а из леммы 2.6 — оценка

$$Q\big(\mathcal{B}_M(y^*), y^*\big) \leq Q(\mathcal{C}^*, y^*) = S(\mathcal{C}^*). \tag{2.11}$$

Объединяя (2.9)–(2.11), получаем оценку $S(\mathcal{C}_A) \leq S(\mathcal{C}^*)$.

С другой стороны, так как множество $\mathcal{C}_A$ является допустимым решением задачи 2, то справедливо неравенство $S(\mathcal{C}^*) \leq S(\mathcal{C}_A)$, что устанавливает равенство значений $S(\mathcal{C}^*)$ и $S(\mathcal{C}_A)$.

Оценим временную сложность алгоритма. Шаг 1 выполняется $|\mathcal{G}|$ раз. При этом для каждой точки $x \in \mathcal{G}$ вычисление проекций на направление, задаваемое этой точкой, требует $\mathcal{O}(dN)$ операций, а выбор $M$ элементов, имеющих наибольшие проекции, можно осуществить за $\mathcal{O}(N)$ операций без сортировки (см., например, [52]). Затраты на вычисление значения функции $Q\big(\mathcal{B}_M(x), x\big)$ составляют $\mathcal{O}(dN)$ операций. Поскольку $|\mathcal{G}| = (2MD+1)^d$, трудоёмкость шага 1 оценивается величиной $\mathcal{O}\big(dN(2MD+1)^d\big)$. Шаг 2 — поиск наименьшего элемента — требует $\mathcal{O}\big((2MD+1)^d\big)$ операций. Таким образом, итоговая временная сложность алгоритма есть величина $\mathcal{O}\big(dN(2MD+1)^d\big)$. Теорема доказана.

Покажем, что в случае фиксированной размерности пространства алго-

ритм $\mathcal{A}_2$ псевдополиномиален. Действительно, поскольку $MD \geq \frac{1}{2}$, то

$$(2MD+1)^d = 2^d\left(MD+\frac{1}{2}\right)^d \leq 4^d(MD)^d.$$

Отсюда следует, что при указанных условиях время работы алгоритма оценивается величиной $\mathcal{O}\big(N(MD)^d\big)$.

Время работы известного [43] алгоритма, гарантирующего оптимальное решение общего случая задачи, при фиксированной размерности пространства есть величина $\mathcal{O}(N^{2d})$. Поэтому при $MD < N^{2-\frac{1}{d}}$ предложенный для частного случая псевдополиномиальный алгоритм $\mathcal{A}_2$ более эффективен по сравнению с точным алгоритмом, ориентированным на общий случай.

Однако, для этого же частного случая задачи LVS в [27] и [47] обоснованы точные алгоритмы, при фиксированной размерности пространства имеющие трудоёмкость $\mathcal{O}\big(N(MD)^{d-1}\big)$ и $\mathcal{O}\big(NM(MD)^{d-1}\big)$, соответственно. Поскольку задачи 2 и LVS полиномиально эквивалентны, эти алгоритмы гарантируют отыскание точного решения задачи 2 в $MD$ и в $D$ раз быстрее, чем предложенный алгоритм $\mathcal{A}_2$. Тем не менее, изложенный подход к построению алгоритма является полезным как ещё один эффективный инструмент решения сходных в постановочном плане задач. В частности, этот — по своей сути сеточный — подход более привлекателен в плане распараллеливания алгоритма. Кроме того, этот подход послужил важным промежуточным результатом, на котором основана идея построения точного алгоритма для рассматриваемой в следующей главе задачи 2-разбиения последовательности, а также идея построения оригинальных аппроксимационных схем для задачи 2 и задачи разбиения последовательности.

### 2.2.4 Аппроксимационная схема

Рассмотрим теперь один из важных вопросов об аппроксимируемости задачи 2. Справедлива следующая

**Теорема 2.3.** *Если* P≠NP*, то для задачи* 2 *не существует схемы* FPTAS*.*

*Доказательство.* Для любого непустого конечного множества $\mathcal{Z}$ точек из $\mathbb{R}^d$ и его центроида $\overline{z}(\mathcal{Z}) = \frac{1}{|\mathcal{Z}|}\sum\limits_{z\in\mathcal{Z}} z$ справедливо легко проверяемое тождество

$$\sum_{x\in\mathcal{Z}}\sum_{z\in\mathcal{Z}} \|x-z\|^2 = 2|\mathcal{Z}|\sum_{z\in\mathcal{Z}} \|z-\overline{z}(\mathcal{Z})\|^2.$$

Применив это тождество к подмножеству $\mathcal{C}$ и его центроиду в первом члене равенства (2.1), получим

$$2M\cdot S(\mathcal{C}) = \sum_{x\in\mathcal{C}}\sum_{y\in\mathcal{C}} \|x-y\|^2 + 2M\sum_{y\in\mathcal{Y}\setminus\mathcal{C}} \|y\|^2, \tag{2.12}$$

так как $|\mathcal{C}| = M$.

Далее заметим, во-первых, что при целочисленных входных данных значение правой части (2.12), очевидно, целочисленно и ограничено полиномом от размера входа задачи (так как $M \le N$) и максимального по модулю значения координат точек входного множества. Во-вторых, несуществование схемы FPTAS для задачи минимизации правой части (2.12) при заданном $M$ влечет несуществование схемы FPTAS для задачи 2 минимизации функции $S(\mathcal{C})$ в силу полиномиальной эквивалентности, которая следует из (2.12). Отсюда, в соответствии с [55] (теорема 8.5), следует несуществование FPTAS для NP-трудной в сильном смысле задачи 2 с числовыми входами, в предположении, что P≠NP. Теорема доказана.

Суть предлагаемого алгоритмического решения — аппроксимационной схемы — состоит в следующем. Для каждой точки входного множества строится область (многомерный куб) с адаптивными шагом и размером так, что одна из этих областей гарантировано включает центр искомого подмножества. По заданной на входе желаемой относительной погрешности решения строится решётка (сетка), дискретизирующая куб с равномерным по всем координатам шагом. Для каждого узла решётки формируется набор из $M$ элементов исходного множества, имеющих наибольшие проекции на направление, задаваемое этим узлом. Сформированный набор объявляется претендентом на решение. В качестве окончательного решения выбирается то подмножество-претендент, которое доставляет наименьшее значение целевой функции.

Для построения алгоритма нам потребуется несколько вспомогательных утверждений.

**Лемма 2.7.** *Пусть $\mathcal{C}^*$ — оптимальное решение задачи 2. Тогда для произвольной точки $x \in \mathbb{R}^d$ справедлива оценка*

$$S\big(\mathcal{B}_M(x)\big) \le S(\mathcal{C}^*) + M\|x - \overline{y}(\mathcal{C}^*)\|^2, \tag{2.13}$$

*где $\overline{y}(\mathcal{C}^*) = \frac{1}{M}\sum\limits_{y\in\mathcal{C}^*} y$ — центроид множества $\mathcal{C}^*$.*

*Доказательство.* Из леммы 2.3 следует оценка

$$S\big(\mathcal{B}_M(x)\big) = Q\big(\mathcal{B}_M(x), \overline{y}\big(\mathcal{B}_M(x)\big)\big) \le Q\big(\mathcal{B}_M(x), x\big), \tag{2.14}$$

а из леммы 2.6 — неравенство

$$Q\big(\mathcal{B}_M(x), x\big) \le Q(\mathcal{C}^*, x). \tag{2.15}$$

Далее, применяя лемму 2.1 к точке $x$ и множеству $\mathcal{C}^*$, получим

$$\sum_{y\in\mathcal{C}^*}\|y-x\|^2=\sum_{y\in\mathcal{C}^*}\|y-\overline{y}(\mathcal{C}^*)\|^2+|\mathcal{C}^*|\cdot\|x-\overline{y}(\mathcal{C}^*)\|^2. \tag{2.16}$$

Наконец, объединяя (2.14)–(2.16), получим оценку

$$S\big(\mathcal{B}_M(x)\big)\le Q\big(\mathcal{B}_M(x),x\big)\le Q(\mathcal{C}^*,x)=\sum_{y\in\mathcal{C}^*}\|y-x\|^2+\sum_{y\in\mathcal{Y}\setminus\mathcal{C}^*}\|y\|^2$$
$$=\sum_{y\in\mathcal{C}^*}\|y-\overline{y}(\mathcal{C}^*)\|^2+|\mathcal{C}^*|\cdot\|x-\overline{y}(\mathcal{C}^*)\|^2+\sum_{y\in\mathcal{Y}\setminus\mathcal{C}^*}\|y\|^2=S(\mathcal{C}^*)+M\|x-\overline{y}(\mathcal{C}^*)\|^2.$$

Лемма доказана.

**Лемма 2.8.** *Пусть выполнены условия леммы* 2.7, *и* $t=\arg\min\limits_{y\in\mathcal{C}^*}\|y-\overline{y}(\mathcal{C}^*)\|$ *— точка из множества* $\mathcal{C}^*$, *ближайшая к центроиду этого множества. Тогда для того чтобы при фиксированном* $\varepsilon>0$ *множество* $\mathcal{B}_M(x)$ *было* $(1+\varepsilon)$*-приближённым решением задачи* 2, *достаточно, чтобы точка* $x$ *удовлетворяла неравенству*

$$\|x-\overline{y}(\mathcal{C}^*)\|^2\le\frac{\varepsilon}{2M}S\big(\mathcal{B}_M(t)\big). \tag{2.17}$$

*Доказательство.* Из леммы 2.3 следует оценка

$$S\big(\mathcal{B}_M(t)\big)=Q\big(\mathcal{B}_M(t),\overline{y}\big(\mathcal{B}_M(t)\big)\big)\le Q\big(\mathcal{B}_M(t),t\big), \tag{2.18}$$

а из леммы 2.6 — неравенство

$$Q\big(\mathcal{B}_M(t),t\big)\le Q(\mathcal{C}^*,t). \tag{2.19}$$

Рассмотрим множество $\mathcal{C}^*$ и точку $t$. Поскольку они удовлетворяют условиям

леммы 2.2, имеет место неравенство

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

и, следовательно,

$$\begin{aligned} Q(\mathcal{C}^*,t) = \sum_{y\in\mathcal{C}^*}\|y-t\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{C}^*}\|y\|^2 &\le 2\sum_{y\in\mathcal{C}^*}\|y-\overline{y}(\mathcal{C}^*)\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{C}^*}\|y\|^2 \\ &\le 2\sum_{y\in\mathcal{C}^*}\|y-\overline{y}(\mathcal{C}^*)\|^2 + 2\sum_{y\in\mathcal{Y}\setminus\mathcal{C}^*}\|y\|^2 = 2S(\mathcal{C}^*). \quad (2.20) \end{aligned}$$

Далее, объединив (2.17)–(2.20), получим

$$\|x-\overline{y}(\mathcal{C}^*)\|^2 \le \frac{\varepsilon}{2M}S\big(\mathcal{B}_M(t)\big) \le \frac{\varepsilon}{2M}Q\big(\mathcal{B}_M(t),t\big) \le \frac{\varepsilon}{2M}Q(\mathcal{C}^*,t) \le \frac{\varepsilon}{M}S(\mathcal{C}^*). \quad (2.21)$$

Наконец, применив (2.21) к правой части неравенства (2.13), получим оценку

$$S\big(\mathcal{B}_M(x)\big) \le (1+\varepsilon)S(\mathcal{C}^*),$$

из которой следует справедливость утверждения леммы. Лемма доказана.

**Лемма 2.9.** *Пусть $\mathcal{C}^*$ — оптимальное решение задачи 2. Тогда для точки $t = \arg\min\limits_{y\in\mathcal{C}^*}\|y-\overline{y}(\mathcal{C}^*)\|$ справедлива оценка*

$$\|t-\overline{y}(\mathcal{C}^*)\|^2 \le \frac{1}{M}S\big(\mathcal{B}_M(t)\big). \quad (2.22)$$

*Доказательство.* Из определения точки $t$ следует, что для любого $y\in\mathcal{C}^*$ справедливо неравенство

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

Просуммировав обе части этого неравенства по $y \in \mathcal{C}^*$, получим

$$M\|t - \overline{y}(\mathcal{C}^*)\|^2 \le \sum_{y \in \mathcal{C}^*} \|y - \overline{y}(\mathcal{C}^*)\|^2. \tag{2.23}$$

Поскольку $\mathcal{B}_M(t)$ — допустимое решение задачи 2, а $\mathcal{C}^*$ — её оптимальное решение, имеем неравенство

$$S(\mathcal{C}^*) \le S\big(\mathcal{B}_M(t)\big). \tag{2.24}$$

Объединяя (2.23) и (2.24), получим оценку

$$M\|t - \overline{y}(\mathcal{C}^*)\|^2 \le \sum_{y \in \mathcal{C}^*} \|y - \overline{y}(\mathcal{C}^*)\|^2 \le S(\mathcal{C}^*) \le S\big(\mathcal{B}_M(t)\big),$$

которая устанавливает справедливость неравенства (2.22). Лемма доказана.

Для произвольной точки $z \in \mathbb{R}^d$ и положительных чисел $h$, $H$ определим множество точек

$$\mathcal{G}(z, h, H) = \big\{x \mid x = z + h(j_1, \ldots, j_d),\ j_i \in \mathbb{Z},\ |h \cdot j_i| \le H,\ i = 1, \ldots, d\big\}$$

— многомерную решётку кубической формы размера $2H$ с расстоянием $h$ между узлами с центром в точке $z$. При этом для любого $x \in \mathbb{R}^d$ такого, что $\|z - x\| \le H$, расстояние до ближайшего узла решётки $\mathcal{G}(z, h, H + h/2)$, очевидно, не превосходит $h\sqrt{d}/2$.

Число узлов решётки $\mathcal{G}(z, h, H + h/2)$ равно

$$|\mathcal{G}(z, h, H + h/2)| \le \left(2\left\lfloor \frac{H + h/2}{h} \right\rfloor + 1\right)^d \le \left(2\frac{H}{h} + 2\right)^d.$$

Заметим, что лемма 2.9 (правая часть неравенства (2.22)), фактически, определяет размер решётки, которая гарантированно содержит центроид оптимального решения задачи, если только точка $t$ из входного множества $\mathcal{Y}$ — ближайшая к этому центроиду точка. Поэтому положим

$$H(y) = \sqrt{\frac{1}{M} S\big(\mathcal{B}_M(y)\big)}, \quad y \in \mathcal{Y}. \tag{2.25}$$

Кроме того, лемма 2.8 устанавливает условие на размер шага решётки, при котором среди её узлов найдется элемент, близкий (в смысле гарантированной погрешности $\varepsilon$) к центроиду оптимального решения. Поэтому для шага решётки положим

$$h(y, \varepsilon) = \sqrt{\frac{2\varepsilon}{dM} S\big(\mathcal{B}_M(y)\big)}, \quad y \in \mathcal{Y},\ \varepsilon > 0. \tag{2.26}$$

Сформулируем следующий алгоритм решения задачи 2.

**Алгоритм $\mathcal{A}_3$.**

*Вход алгоритма:* множество $\mathcal{Y}$, натуральное число $M$ и число $\varepsilon > 0$.

Для каждой точки $y \in \mathcal{Y}$ выполним шаги 1–5.

`Шаг 1.` Построим множество $\mathcal{B}_M(y)$.

`Шаг 2.` Вычислим $S\big(\mathcal{B}_M(y)\big)$, $h$ и $H$ по формулам (2.1), (2.26) и (2.25).

`Шаг 3.` Если $S\big(\mathcal{B}_M(y)\big) = 0$, то множество $\mathcal{B}_M(y)$ объявим результатом $\mathcal{C}_A$ работы алгоритма; выход. Иначе переходим к следующему шагу.

`Шаг 4.` Построим решётку $\mathcal{G}(y, h, H + h/2)$.

`Шаг 5.` Для каждой точки $x$ решётки $\mathcal{G}(y, h, H + h/2)$ построим множество $\mathcal{B}_M(x)$ и вычислим значение $S\big(\mathcal{B}_M(x)\big)$.

`Шаг 6.` В семействе $\big\{\mathcal{B}_M(x) \bigm| x \in \mathcal{G}(y, h, H + h/2),\ y \in \mathcal{Y}\big\}$ множеств в качестве решения $\mathcal{C}_A$ выберем то множество $\mathcal{B}_M(x)$, для которого значение $S\big(\mathcal{B}_M(x)\big)$ минимально. Если оптимуму соответствует несколько множеств, то выберем лю-

бое из них.

*Выход алгоритма:* множество $\mathcal{C}_A$.

**Теорема 2.4.** *Для любого фиксированного $\varepsilon > 0$ алгоритм $\mathcal{A}_3$ находит $(1+\varepsilon)$-приближённое решение задачи 2 за время*[1] $\mathcal{O}\left(dN^2\left(\sqrt{2d/\varepsilon}+2\right)^d\right)$.

*Доказательство.* Пусть $t = \arg\min\limits_{y\in\mathcal{C}^*} \|y - \overline{y}(\mathcal{C}^*)\|$ — точка из множества $\mathcal{C}^*$, ближайшая к центроиду этого множества. Если для этой точки на шаге 3 алгоритма выполнено равенство $S\big(\mathcal{B}_M(t)\big) = 0$, то в этом случае множество $\mathcal{B}_M(t)$ является оптимальным решением задачи 2, так как для любого множества $\mathcal{C} \subseteq \mathcal{Y}$ справедливо неравенство $S(\mathcal{C}) \geq 0$.

Рассмотрим случай, когда $S\big(\mathcal{B}_M(t)\big) > 0$. По лемме 2.9 для точки $t$ выполнено неравенство (2.22). Из этого неравенства и (2.25) следует, что $\|t - \overline{y}(\mathcal{C}^*)\| \leq H$.

Положим $x^* = \arg\min\limits_{x\in\mathcal{G}(t,h,H+h/2)} \|x - \overline{y}(\mathcal{C}^*)\|$. Поскольку расстояние от $\overline{y}(\mathcal{C}^*)$ до ближайшего узла $x^*$ решётки $\mathcal{G}(t,h,H+h/2)$ не превосходит $\frac{h\sqrt{d}}{2}$, имеем оценку

$$\|x^* - \overline{y}(\mathcal{C}^*)\|^2 \leq \frac{h^2 d}{4} = \frac{\varepsilon}{2M} S\big(\mathcal{B}_M(t)\big).$$

Поэтому точка $x^*$ удовлетворяет условиям леммы 2.8, и, следовательно, множество $\mathcal{B}_M(x^*)$ является $(1+\varepsilon)$-приближённым решением задачи 2.

Оценим временную сложность алгоритма. На шаге 1: 1) вычисляется совокупность из $N$ проекций элементов множества $\mathcal{Y}$ на направление, задаваемое точкой $y$, что требует $\mathcal{O}(dN)$ операций, и 2) в найденной совокупности выбираются $M$ наибольших проекций и соответствующих им точек множества $\mathcal{Y}$, что можно осуществить за $\mathcal{O}(N)$ операций без сортировки (см., например, [52]). Шаг 2 требует не более $\mathcal{O}(dN)$ операций, а шаг 3 выполняется за $\mathcal{O}(1)$ операций.

[1] При аккуратной организации вычислений формула трудоёмкости алгоритма именно такая, что слегка отличается от формулы $\mathcal{O}(dN^2(\sqrt{2d/\varepsilon}+1)^d)$, указанной в статье [63].

Трудоёмкость построения решётки $\mathcal{G}(y, h, H + h/2)$ на шаге 4 равна $\mathcal{O}\big(d|\mathcal{G}(y, h, H + h/2)|\big)$.

Построение каждого из $|\mathcal{G}(y, h, H + h/2)|$ множеств $\mathcal{B}_M(x)$ на шаге 5 выполняется за $\mathcal{O}(dN)$ операций, как и вычисление значений $S\big(\mathcal{B}_M(x)\big)$.

Для каждой из $N$ точек $y \in \mathcal{Y}$ шаги 1–5 выполняются за $\mathcal{O}\big(dN|\mathcal{G}(y, h, H + h/2)|\big)$ операций. Шаг 6 — выбор наименьшего элемента — требует не более $\mathcal{O}\left(\sum\limits_{y\in\mathcal{Y}} |\mathcal{G}(y, h, H + h/2)|\right)$ операций.

Остается заметить, что для мощности решётки $\mathcal{G}(y, h, H + h/2)$ справедлива оценка

$$|\mathcal{G}(y, h, H + h/2)| \le (2H/h + 2)^d \le \left(\sqrt{2d/\varepsilon} + 2\right)^d.$$

Поэтому временная сложность алгоритма есть величина $\mathcal{O}\left(dN^2\left(\sqrt{2d/\varepsilon} + 2\right)^d\right)$. Теорема доказана.

Покажем, что в случае фиксированной размерности $d$ пространства алгоритм $\mathcal{A}_3$ реализует схему FPTAS. Действительно, если $\varepsilon \in (0,\, d/2]$, то

$$\left(\sqrt{2d/\varepsilon} + 2\right)^d \le 2^d\left(\sqrt{2d/\varepsilon}\right)^d = 2^{3d/2}d^{d/2}(1/\varepsilon)^{d/2} = \mathcal{O}\Big((1/\varepsilon)^{d/2}\Big).$$

Поэтому при указанных условиях время работы алгоритма есть величина $\mathcal{O}\Big(N^2(1/\varepsilon)^{d/2}\Big)$, которая ограничена полиномом как от размера входа задачи, так и от $1/\varepsilon$. Таким образом, предложенный алгоритм реализует схему FPTAS.

### 2.2.5 Рандомизированный алгоритм

Суть предлагаемого алгоритмического решения состоит в следующем. Из множества $\mathcal{Y}$ путем случайного независимого выбора (с возвращением) его элементов формируется конечное мультимножество. Для каждого из непустых подмножеств этого мультимножества вычисляется геометрический центр и фор-

мируется набор из $M$ элементов исходного множества, имеющих наибольшие проекции на направление, задаваемое этим центром. Сформированный набор объявляется претендентом на решение. В качестве окончательного решения выбирается то подмножество-претендент, для которого значение целевой функции минимально.

Всюду далее под пересечением мультимножества $\mathcal{A}$ и множества $\mathcal{B}$ будем подразумевать мультимножество, т.е. будем считать, что если $x \in \mathcal{B}$ кратно входит в $\mathcal{A}$, то в пересечение $\mathcal{A} \cap \mathcal{B}$ он входит с той же кратностью.

**Лемма 2.10.** *Пусть $\mathcal{Z}$ — произвольное множество точек из $\mathbb{R}^d$ мощности $N$; $\mathcal{C} \subseteq \mathcal{Z}$, $|\mathcal{C}| = M$; $\mathcal{T}$ — мультимножество, полученное $k$ случайными независимыми одноэлементными выборками с возвращением из множества $\mathcal{Z}$. Пусть, кроме того, $\overline{z}(\mathcal{C}) = \frac{1}{M}\sum\limits_{z\in\mathcal{C}} z$ и $\overline{z}(\mathcal{T}\cap\mathcal{C}) = \frac{1}{|\mathcal{T}\cap\mathcal{C}|}\sum\limits_{z\in\mathcal{T}\cap\mathcal{C}} z$ — центры множества $\mathcal{C}$ и мультимножества $\mathcal{T}\cap\mathcal{C}$, соответственно. Тогда для любого натурального $t \le k$ справедливо:*

$$\mathbb{E}\big(\overline{z}(\mathcal{T}\cap\mathcal{C}) \;\big|\; |\mathcal{T}\cap\mathcal{C}| \ge t\big) = \overline{z}(\mathcal{C}), \tag{2.27}$$

$$\mathbb{E}\Big(\|\overline{z}(\mathcal{T}\cap\mathcal{C}) - \overline{z}(\mathcal{C})\|^2 \;\Big|\; |\mathcal{T}\cap\mathcal{C}| \ge t\Big) \le \frac{1}{tM}\sum_{z\in\mathcal{C}} \|z - \overline{z}(\mathcal{C})\|^2, \tag{2.28}$$

*где $\mathbb{E}$ — символ математического ожидания.*

*Доказательство.* Пусть $\Omega$ — множество всех $k$-элементных мультиподмножеств множества $\mathcal{Z}$. Положим $\Omega_t = \big\{\mathcal{T} \;\big|\; \mathcal{T}\in\Omega \text{ и } |\mathcal{T}\cap\mathcal{C}| \ge t\big\}$. Заметим, что

$$|\Omega_t| = \sum_{i=t}^{k} \binom{k}{i} M^i (N-M)^{k-i},$$

так как $\Omega_t$ есть множество тех мультиподмножеств множества $\mathcal{Z}$, которые содержат не менее $t$ элементов из множества $\mathcal{C}$ (с учётом кратности).

Покажем справедливость равенства (2.27). По определению математического

ожидания имеем

$$\mathbb{E}\big(\overline{z}(\mathcal{T}\cap\mathcal{C}) \;\big|\; |\mathcal{T}\cap\mathcal{C}|\ge t\big) = \frac{1}{|\Omega_t|}\sum_{\mathcal{T}\in\Omega_t}\frac{1}{|\mathcal{T}\cap\mathcal{C}|}\sum_{z\in\mathcal{T}\cap\mathcal{C}} z$$
$$= \frac{1}{|\Omega_t|}\sum_{z\in\mathcal{C}}\left(\sum_{i=t}^{k}\frac{1}{i}k\binom{k-1}{i-1}M^{i-1}(N-M)^{k-i}\right)z$$
$$= \frac{1}{|\Omega_t|}\frac{1}{M}\sum_{z\in\mathcal{C}}\left(\sum_{i=t}^{k}\binom{k}{i}M^{i}(N-M)^{k-i}\right)z = \frac{1}{M}\sum_{z\in\mathcal{C}} z = \overline{z}(\mathcal{C}).$$

Второе равенство в этой цепочке следует из того, что в двойной сумме $\sum\limits_{\mathcal{T}\in\Omega_t}\sum\limits_{z\in\mathcal{T}\cap\mathcal{C}} z$ каждая точка $z\in\mathcal{C}$ входит в множества $\mathcal{T}\in\Omega_t$, для которых $|\mathcal{T}\cap\mathcal{C}| = i$ (при $i\in\{t,\dots,k\}$), ровно $k\binom{k-1}{i-1}M^{i-1}(N-M)^{k-i}$ раз, где $k$ — число позиций в наборе $\mathcal{T}$, на которых может стоять точка $z$, $\binom{k-1}{i-1}$ — количество вариантов расстановки позиций для остальных $i-1$ точек из $\mathcal{T}\cap\mathcal{C}$, $M^{i-1}$ — количество вариантов выбора этих точек, а $(N-M)^{k-i}$ — количество вариантов выбора $k-i$ точек из $\mathcal{T}\cap(\mathcal{Z}\setminus\mathcal{C})$.

Покажем справедливость неравенства (2.28). Заметим сначала, что

$$\sum_{z\in\mathcal{C}}\|z-\overline{z}(\mathcal{C})\|^2 = \sum_{z\in\mathcal{C}}\|z\|^2 - \frac{1}{M}\left\langle\sum_{z\in\mathcal{C}} z, \sum_{z\in\mathcal{C}} z\right\rangle$$
$$= \left(1-\frac{1}{M}\right)\sum_{z\in\mathcal{C}}\|z\|^2 - \frac{1}{M}\sum_{x,z\in\mathcal{C};\; x\ne z}\langle x,z\rangle$$
$$= \frac{M-1}{M}\sum_{z\in\mathcal{C}}\|z\|^2 - \frac{1}{M}\sum_{x,z\in\mathcal{C};\; x\ne z}\langle x,z\rangle. \quad (2.29)$$

Далее, используя известное свойство дисперсии случайной величины, найдем

$$\begin{aligned}
&\mathbb{E}\Big(\|\overline{z}(\mathcal{T}\cap\mathcal{C})-\overline{z}(\mathcal{C})\|^2 \;\Big|\; |\mathcal{T}\cap\mathcal{C}|\geq t\Big)\\
&\qquad=\mathbb{E}\Big(\|\overline{z}(\mathcal{T}\cap\mathcal{C})\|^2 \;\Big|\; |\mathcal{T}\cap\mathcal{C}|\geq t\Big)-\big(\mathbb{E}\big(\overline{z}(\mathcal{T}\cap\mathcal{C}) \;\big|\; |\mathcal{T}\cap\mathcal{C}|\geq t\big)\big)^2\\
&\qquad=\frac{1}{|\Omega_t|}\left(\sum_{\mathcal{T}\in\Omega_t}\frac{1}{|\mathcal{T}\cap\mathcal{C}|^2}\left\langle\sum_{z\in\mathcal{T}\cap\mathcal{C}}z,\sum_{z\in\mathcal{T}\cap\mathcal{C}}z\right\rangle\right)-\frac{1}{M^2}\left\langle\sum_{z\in\mathcal{C}}z,\sum_{z\in\mathcal{C}}z\right\rangle\\
&=\frac{1}{|\Omega_t|}\left(S_1\sum_{z\in\mathcal{C}}\|z\|^2+S_2\sum_{x,z\in\mathcal{C};\,x\neq z}\langle x,z\rangle\right)-\frac{1}{M^2}\left(\sum_{z\in\mathcal{C}}\|z\|^2+\sum_{x,z\in\mathcal{C};\,x\neq z}\langle x,z\rangle\right),
\end{aligned}\tag{2.30}$$

где

$$\begin{aligned}
S_1=\sum_{i=t}^{k}\frac{1}{i^2}k\binom{k-1}{i-1}M^{i-1}(N-M)^{k-i}\\
+\sum_{i=\max\{2,t\}}^{k}\frac{1}{i^2}k(k-1)\binom{k-2}{i-2}M^{i-2}(N-M)^{k-i},
\end{aligned}$$

$$S_2=\sum_{i=\max\{2,t\}}^{k}\frac{1}{i^2}k(k-1)\binom{k-2}{i-2}M^{i-2}(N-M)^{k-i}.$$

Выполнив несложные преобразования этих равенств, получим

$$\begin{aligned}
S_1&=\frac{1}{M}\sum_{i=t}^{k}\frac{1}{i}\binom{k}{i}M^i(N-M)^{k-i}+\frac{1}{M^2}\sum_{i=t}^{k}\frac{i-1}{i}\binom{k}{i}M^i(N-M)^{k-i}\\
&=\frac{M-1}{M^2}\sum_{i=t}^{k}\frac{1}{i}\binom{k}{i}M^i(N-M)^{k-i}+\frac{1}{M^2}\sum_{i=t}^{k}\binom{k}{i}M^i(N-M)^{k-i}\\
&=\frac{M-1}{M^2}\sum_{i=t}^{k}\frac{1}{i}\binom{k}{i}M^i(N-M)^{k-i}+\frac{|\Omega_t|}{M^2},
\end{aligned}$$

$$S_2 = \frac{1}{M^2}\sum_{i=t}^{k}\frac{i-1}{i}\binom{k}{i}M^i(N-M)^{k-i}$$
$$= \frac{1}{M^2}\sum_{i=t}^{k}\binom{k}{i}M^i(N-M)^{k-i} - \frac{1}{M^2}\sum_{i=t}^{k}\frac{1}{i}\binom{k}{i}M^i(N-M)^{k-i}$$
$$= \frac{|\Omega_t|}{M^2} - \frac{1}{M^2}\sum_{i=t}^{k}\frac{1}{i}\binom{k}{i}M^i(N-M)^{k-i}.$$

Наконец, подставив полученные выражения для $S_1$ и $S_2$ в равенство (2.30), используя (2.29), найдем оценку

$$\mathbb{E}\Big(\|\overline{z}(\mathcal{T}\cap\mathcal{C}) - \overline{z}(\mathcal{C})\|^2 \mid |\mathcal{T}\cap\mathcal{C}| \geq t\Big)$$
$$= \frac{M-1}{|\Omega_t|M^2}\left(\sum_{i=t}^{k}\frac{1}{i}\binom{k}{i}M^i(N-M)^{k-i}\right)\sum_{z\in\mathcal{C}}\|z\|^2$$
$$- \frac{1}{|\Omega_t|M^2}\left(\sum_{i=t}^{k}\frac{1}{i}\binom{k}{i}M^i(N-M)^{k-i}\right)\sum_{x,z\in\mathcal{C};\, x\neq z}\langle x,z\rangle$$
$$= \frac{1}{|\Omega_t|}\left(\sum_{i=t}^{k}\frac{1}{i}\binom{k}{i}M^i(N-M)^{k-i}\right)\left(\frac{M-1}{M^2}\sum_{z\in\mathcal{C}}\|z\|^2 - \frac{1}{M^2}\sum_{x,z\in\mathcal{C};\, x\neq z}\langle x,z\rangle\right)$$
$$\leq \frac{1}{tM}\sum_{z\in\mathcal{C}}\|z-\overline{z}(\mathcal{C})\|^2.$$

Лемма доказана.

Далее нам потребуется известное [53]

**Неравенство Маркова.** *Пусть $\xi$ — положительная случайная величина, имеющая конечное математическое ожидание $\mathbb{E}(\xi)$. Тогда для произвольного $a>0$ имеет место неравенство*

$$\Pr(\xi \geq a) \leq \frac{\mathbb{E}(\xi)}{a}.$$

**Лемма 2.11.** *Пусть выполнены условия леммы* 2.10. *Тогда для произвольного* $\delta \in (0,1)$ *справедлива оценка*

$$\Pr\left(\sum_{z\in\mathcal{C}}\|z-\overline{z}(\mathcal{T}\cap\mathcal{C})\|^2 \geq \left(1+\frac{1}{\delta t}\right)\sum_{z\in\mathcal{C}}\|z-\overline{z}(\mathcal{C})\|^2 \;\Big|\; |\mathcal{T}\cap\mathcal{C}|\geq t\right) \leq \delta. \quad (2.31)$$

*Доказательство.* Применив неравенство Маркова к $\xi = \|\overline{z}(\mathcal{T}\cap\mathcal{C}) - \overline{z}(\mathcal{C})\|^2$ и $a = \frac{1}{\delta t M}\sum\limits_{z\in\mathcal{C}}\|z-\overline{z}(\mathcal{C})\|^2$, используя неравенство (2.28), найдем следующую оценку для условной вероятности:

$$\Pr\left(\|\overline{z}(\mathcal{T}\cap\mathcal{C}) - \overline{z}(\mathcal{C})\|^2 \geq \frac{1}{\delta t M}\sum_{z\in\mathcal{C}}\|z-\overline{z}(\mathcal{C})\|^2 \;\Big|\; |\mathcal{T}\cap\mathcal{C}|\geq t\right)$$
$$\leq \delta t M \frac{\mathbb{E}\big(\|\overline{z}(\mathcal{T}\cap\mathcal{C}) - \overline{z}(\mathcal{C})\|^2 \;\big|\; |\mathcal{T}\cap\mathcal{C}|\geq t\big)}{\sum\limits_{z\in\mathcal{C}}\|z-\overline{z}(\mathcal{C})\|^2} \leq \delta. \quad (2.32)$$

Заметим, что по лемме 2.1

$$\sum_{z\in\mathcal{C}}\|z-\overline{z}(\mathcal{T}\cap\mathcal{C})\|^2 = \sum_{z\in\mathcal{C}}\|z-\overline{z}(\mathcal{C})\|^2 + M\cdot\|\overline{z}(\mathcal{T}\cap\mathcal{C}) - \overline{z}(\mathcal{C})\|^2. \quad (2.33)$$

Преобразуем неравенство

$$\|\overline{z}(\mathcal{T}\cap\mathcal{C}) - \overline{z}(\mathcal{C})\|^2 \geq \frac{1}{\delta t M}\sum_{z\in\mathcal{C}}\|z-\overline{z}(\mathcal{C})\|^2,$$

условная вероятность которого фигурирует в оценке (2.32). Умножая обе части этого неравенства на $M$, прибавляя к каждой из них слагаемое $\sum\limits_{z\in\mathcal{C}}\|z-\overline{z}(\mathcal{C})\|^2$ и используя (2.33), получим эквивалентное неравенство

$$\sum_{z\in\mathcal{C}}\|z-\overline{z}(\mathcal{T}\cap\mathcal{C})\|^2 \geq \left(1+\frac{1}{\delta t}\right)\sum_{z\in\mathcal{C}}\|z-\overline{z}(\mathcal{C})\|^2.$$

Отсюда следует, что оценка (2.32) равносильна неравенству (2.31). Лемма доказана.

Формула (2.31) даёт оценку условной вероятности. Для того, чтобы оценить безусловную вероятность, нам потребуется оценить вероятность события $|\mathcal{T} \cap \mathcal{C}| \geq t$, фигурирующего в формуле (2.31). Для этого воспользуемся равенством $\Pr\bigl(|\mathcal{T} \cap \mathcal{C}| \geq t\bigr) = 1 - \Pr\bigl(|\mathcal{T} \cap \mathcal{C}| < t\bigr)$, связывающим вероятности противоположных событий, заметив, что

$$\Pr\bigl(|\mathcal{T} \cap \mathcal{C}| < t\bigr) = \sum_{i=0}^{t-1} \binom{k}{i} \left(\frac{M}{N}\right)^i \left(1 - \frac{M}{N}\right)^{k-i}. \tag{2.34}$$

Для оценки сверху этой вероятности нам потребуется известная (см., например, [54])

**Теорема** (Чернова)**.** *Пусть* $\xi_1, \ldots, \xi_k \in \{0, 1\}$ — *независимые случайные величины,* $\xi = \sum\limits_{i=1}^{k} \xi_i$ *и* $\mu = \mathbb{E}(\xi)$*. Тогда для произвольного* $\nu \in (0, 1)$ *справедлива оценка*

$$\Pr\bigl(\xi \leq (1 - \nu)\mu\bigr) \leq e^{-\frac{\nu^2 \mu}{2}}.$$

**Лемма 2.12.** *Пусть выполнены условия леммы* 2.10*. Тогда для произвольного* $\nu \in (0, 1)$ *справедлива оценка*

$$\Pr\left(|\mathcal{T} \cap \mathcal{C}| \leq (1 - \nu)\frac{M}{N}k\right) \leq e^{-\frac{\nu^2 Mk}{2N}}. \tag{2.35}$$

*Доказательство.* Пусть в условиях леммы 2.10 при каждом $i \in \{1, \ldots, k\}$ случайная величина $\xi_i = 1$, если $i$-ый элемент, выбранный из множества $\mathcal{Z}$, принадлежит множеству $\mathcal{C}$, и $\xi_i = 0$, иначе. Тогда $\mathbb{E}(\xi_i) = \frac{|\mathcal{C}|}{|\mathcal{Z}|} = \frac{M}{N}$. Кроме того, $\sum\limits_{i=1}^{k} \xi_i = |\mathcal{T} \cap \mathcal{C}|$ и $\mathbb{E}\left(\sum\limits_{i=1}^{k} \xi_i\right) = \frac{M}{N}k$. Следовательно, по теореме Чернова для произвольного $\nu \in (0, 1)$ выполнено неравенство (2.35). Лемма доказана.

Сформулируем рандомизированный алгоритм решения задачи 2.

**Алгоритм $\mathcal{A}_4$.**

*Вход алгоритма:* множество $\mathcal{Y}$, натуральное число $M$, натуральный параметр $k$.

Шаг 1. Сформируем мультимножество $\mathcal{T}$ точек с помощью $k$ независимых случайных выборок (с возвращением) по одному элементу из множества $\mathcal{Y}$.

Шаг 2. Для каждого непустого $\mathcal{H} \subseteq \mathcal{T}$ вычислим центроид $\overline{y}(\mathcal{H})$ и сформируем множество $\mathcal{C} = \mathcal{B}_M\big(\overline{y}(\mathcal{H})\big)$. Вычислим значение $S(\mathcal{C})$.

Шаг 3. В семействе множеств, найденных на шаге 2, выберем то множество $\mathcal{C}_A = \mathcal{C}$, для которого $S(\mathcal{C})$ минимально. Если оптимальных значений несколько, то выберем любое из них.

*Выход алгоритма:* множество $\mathcal{C}_A$.

**Теорема 2.5.** *Для произвольных вещественного $\delta \in (0,1)$ и натуральных $t \leq k$ алгоритм $\mathcal{A}_4$ находит $\big(1+\frac{1}{\delta t}\big)$-приближённое решение задачи 2 за время $\mathcal{O}\big(2^k d(k+N)\big)$ с вероятностью не менее $1-(\delta+\alpha)$, где $\alpha = \sum\limits_{i=0}^{t-1} \binom{k}{i} \Big(\frac{M}{N}\Big)^i \Big(1-\frac{M}{N}\Big)^{k-i}$.*

*Доказательство.* Пусть $\mathcal{C}^*$ — оптимальное решение задачи 2, а $\mathcal{C}_A$ — множество, полученное в результате работы алгоритма $\mathcal{A}_4$.

Предположим, что на шаге 1 работы алгоритма мультимножество $\mathcal{T}$ выбрано так, что выполнено событие $|\mathcal{T} \cap \mathcal{C}^*| \geq t$. Заметим, что в этом случае $\mathcal{T} \cap \mathcal{C}^*$ является непустым подмножеством мультимножества $\mathcal{T}$. Следовательно, мультимножество $\mathcal{H} = \mathcal{T} \cap \mathcal{C}^*$ было рассмотрено на шаге 2 работы алгоритма и по этому мультимножеству было построено некоторое множество $\mathcal{C} = \mathcal{C}'$. По определению шага 3

$$S(\mathcal{C}_A) \leq S(\mathcal{C}'). \tag{2.36}$$

Из леммы 2.3 следует неравенство

$$S(\mathcal{C}') \le Q\big(\mathcal{C}', \overline{y}(\mathcal{H})\big). \tag{2.37}$$

Поскольку множество $\mathcal{C}'$ состоит из $M$ элементов множества $\mathcal{Y}$, имеющих наибольшие проекции на направление, задаваемое точкой $\overline{y}(\mathcal{H})$, для правой части (2.37) из леммы 2.6 имеем оценку

$$Q\big(\mathcal{C}', \overline{y}(\mathcal{H})\big) \le Q\big(\mathcal{C}^*, \overline{y}(\mathcal{H})\big). \tag{2.38}$$

Применяя лемму 2.11 для $\mathcal{Z} = \mathcal{Y}$ и $\mathcal{C} = \mathcal{C}^*$, получим, что с вероятностью более $1 - \delta$ выполнено неравенство

$$\sum_{y \in \mathcal{C}^*} \|y - \overline{y}(\mathcal{H})\|^2 < \left(1 + \frac{1}{\delta t}\right) \sum_{y \in \mathcal{C}^*} \|y - \overline{y}(\mathcal{C}^*)\|^2. \tag{2.39}$$

Наконец, объединяя (2.36)–(2.39), получим, что при $|\mathcal{H}| \ge t$ с вероятностью более $1 - \delta$ справедлива цепочка неравенств

$$\begin{aligned} S(\mathcal{C}_A) \le S(\mathcal{C}') &\le Q\big(\mathcal{C}', \overline{y}(\mathcal{H})\big) \le Q\big(\mathcal{C}^*, \overline{y}(\mathcal{H})\big) \\ &< \left(1 + \frac{1}{\delta t}\right) \sum_{y \in \mathcal{C}^*} \|y - \overline{y}(\mathcal{C}^*)\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}^*} \|y\|^2 \le \left(1 + \frac{1}{\delta t}\right) S(\mathcal{C}^*). \end{aligned}$$

Таким образом, для условной вероятности события $S(\mathcal{C}_A) < \big(1 + \frac{1}{\delta t}\big) S(\mathcal{C}^*)$ справедлива оценка

$$\Pr\left( S(\mathcal{C}_A) < \left(1 + \frac{1}{\delta t}\right) S(\mathcal{C}^*) \;\middle|\; |\mathcal{H}| \ge t \right) > 1 - \delta.$$

Переходя от условной вероятности к безусловной, получим

$$\Pr\left(S(\mathcal{C}_A) < \left(1+\frac{1}{\delta t}\right)S(\mathcal{C}^*)\right) > 1-\big(\delta+\Pr\big(|\mathcal{H}|<t\big)\big)$$
$$= 1-\big(\delta+\Pr\big(|\mathcal{T}\cap\mathcal{C}^*|<t\big)\big),$$

что, с учётом равенства (2.34) при $C = \mathcal{C}^*$, устанавливает оценку точности алгоритма.

Оценим временную сложность алгоритма. Шаг 1 — формирование множества $\mathcal{T}$ — требует $\mathcal{O}(k)$ операций. Для каждого из его подмножеств вычисление центра требует $\mathcal{O}(dk)$ операций, вычисление проекций на направление, задаваемое этим центром — $\mathcal{O}(dN)$ операций, а выбор $M$ элементов, имеющих наибольшие проекции — $\mathcal{O}(N)$ операций (см., например, [52]). Поскольку число всех подмножеств мультимножества $\mathcal{T}$ равно $2^k$, суммарная трудоёмкость шага 2 есть величина $\mathcal{O}\big(2^k d(k+N)\big)$. Шаг 3 — поиск наименьшего элемента — требует $\mathcal{O}(2^k)$ операций. Суммируя затраты на всех шагах, устанавливаем оценку временной сложности алгоритма. Теорема доказана.

**Замечание 2.2.** *В соответствии с формулировкой теоремы* 2.5 *величину* $\frac{1}{\delta t}$ *можно трактовать как относительную погрешность алгоритма, величину* $1-(\delta+\alpha)$ *— как вероятность успешного алгоритмического события, т.е. как вероятность срабатывания алгоритма, а величину* $\delta+\alpha$ *— как вероятность несрабатывания алгоритма.*

В следующем утверждении указывается значение параметра $k$, определяемое через некоторые вещественные $\beta$, $\gamma$, $\varepsilon$, при котором вероятность несрабатывания алгоритма оценивается величиной $\gamma$, а относительная погрешность — величиной $\varepsilon$.

**Следствие 2.2.** *Пусть* $M \geq \beta N$, *где* $\beta \in (0,1)$ — *некоторая константа. Тогда для заданных* $\varepsilon > 0$ *и* $\gamma \in (0,1)$ *при фиксированном параметре* $k = \max\Big(\Big\lceil \frac{2}{\beta}\Big\lceil \frac{2}{\gamma\varepsilon}\Big\rceil\Big\rceil, \Big\lceil \frac{8}{\beta}\ln\frac{2}{\gamma}\Big\rceil\Big)$ *алгоритм* $\mathcal{A}_4$ *находит* $(1+\varepsilon)$*-приближённое решение задачи* 2 *с вероятностью не менее* $1-\gamma$ *за время* $\mathcal{O}(dN)$.

*Доказательство.* Положим $\delta = \frac{\gamma}{2}$, $t = \big\lceil \frac{1}{\delta\varepsilon}\big\rceil = \Big\lceil \frac{2}{\gamma\varepsilon}\Big\rceil$. Заметим, что в этом случае при указанном в условии значении параметра $k$ выполнены неравенства $k \geq \frac{2t}{\beta}$ и $k \geq \frac{8}{\beta}\ln\frac{2}{\gamma}$. Кроме того, по лемме 2.12 при $\nu = \frac{1}{2}$ и $\mathcal{C} = \mathcal{C}^*$ выполнено неравенство

$$\Pr\left(|\mathcal{T}\cap\mathcal{C}^*| \leq \frac{kM}{2N}\right) \leq e^{-\frac{kM}{8N}}.$$

Учитывая эти замечания, из условий следствия для вероятности $\alpha$ имеем

$$\alpha = \Pr\big(|\mathcal{T}\cap\mathcal{C}^*| < t\big) \leq \Pr\left(|\mathcal{T}\cap\mathcal{C}^*| < \frac{\beta k}{2}\right) \leq \Pr\left(|\mathcal{T}\cap\mathcal{C}^*| \leq \frac{kM}{2N}\right)$$
$$\leq e^{-\frac{kM}{8N}} \leq e^{-\frac{M}{\beta N}\ln\frac{2}{\gamma}} \leq e^{-\ln\frac{2}{\gamma}} = \frac{\gamma}{2}.$$

Тогда по теореме 2.5 при указанном в условии значении параметра $k$ алгоритм $\mathcal{A}_4$ находит решение с относительной погрешностью $\frac{1}{\delta t} = \Big(\frac{\gamma}{2}\Big\lceil\frac{2}{\gamma\varepsilon}\Big\rceil\Big)^{-1} \leq \varepsilon$ за время $\mathcal{O}\big(2^k d(k+N)\big)$ при вероятности несрабатывания $\delta+\alpha \leq \frac{\gamma}{2}+\frac{\gamma}{2} = \gamma$. Поскольку параметр $k$ фиксирован, при указанных условиях время работы алгоритма есть величина $\mathcal{O}(dN)$. Следствие доказано.

Для установления условий асимптотической точности алгоритма $\mathcal{A}_4$ мы связываем (см. формулировку теоремы 2.6) относительную погрешность и вероятность несрабатывания алгоритма с мощностью $N$ входного множества.

**Теорема 2.6.** *Пусть в условиях теоремы* 2.5 *имеют место равенства* $k = \lceil\log_2 N\rceil$, $\delta = (\log_2 N)^{-1/2}$, $t = \Big\lceil\frac{kM}{2N}\Big\rceil$. *Тогда если* $M \geq \beta N$, *где* $\beta \in (0,1)$ — *некоторая константа, то алгоритм* $\mathcal{A}_4$ *находит* $(1+\varepsilon_N)$*-приближённое*

*решение задачи* 2 *с вероятностью* $1-\gamma_N$ *за время* $\mathcal{O}(dN^2)$, *где*

$$\varepsilon_N \le \frac{2}{\beta}(\log_2 N)^{-1/2} \xrightarrow[N\to\infty]{} 0,$$

$$\gamma_N \le (\log_2 N)^{-1/2} + N^{-\frac{\beta}{8\ln 2}} \xrightarrow[N\to\infty]{} 0.$$

*Доказательство.* При $t = \left\lceil \frac{kM}{2N} \right\rceil$ для вероятности $\alpha$ имеем цепочку неравенств

$$\alpha = \Pr\left( |\mathcal{T} \cap \mathcal{C}^*| < \left\lceil \frac{kM}{2N} \right\rceil \right) \le \Pr\left( |\mathcal{T} \cap \mathcal{C}^*| \le \frac{kM}{2N} \right) \le e^{-\frac{kM}{8N}}. \tag{2.40}$$

Последнее неравенство в цепочке (2.40) следует из леммы 2.12 при $\nu = \frac{1}{2}$ и $\mathcal{C} = \mathcal{C}^*$. Так как по условию $k = \lceil \log_2 N \rceil$ и $M \ge \beta N$, то для правой части формулы (2.40) имеем

$$e^{-\frac{kM}{8N}} \le e^{-\frac{\beta\lceil \log_2 N \rceil}{8}} \le e^{-\frac{\beta \log_2 N}{8}} \le N^{-\frac{\beta}{8\ln 2}}.$$

Следовательно, при $\delta = (\log_2 N)^{-1/2}$ для вероятности несрабатывания алгоритма справедлива оценка

$$\gamma_N \le \delta + \alpha \le (\log_2 N)^{-1/2} + N^{-\frac{\beta}{8\ln 2}}.$$

При этом для относительной погрешности алгоритма справедлива оценка

$$\varepsilon_N = \frac{1}{\delta t} = \frac{1}{(\log_2 N)^{-1/2}\left\lceil \frac{kM}{2N} \right\rceil} \le \frac{1}{(\log_2 N)^{-1/2}\frac{kM}{2N}} = \frac{2N(\log_2 N)^{1/2}}{M\lceil \log_2 N \rceil}$$
$$\le \frac{2N}{M(\log_2 N)^{1/2}} \le \frac{2}{\beta}(\log_2 N)^{-1/2}.$$

Оценка временной сложности алгоритма следует из того, что при $k =$

$\lceil \log_2 N \rceil$ справедливо

$$2^k d(k+N) = \mathcal{O}\Big(2^{\log_2 N} d(\log_2 N + N)\Big) = \mathcal{O}(dN^2).$$

Теорема доказана.

# Глава 3

# Кластеризация конечной последовательности точек евклидова пространства

В этой главе рассматриваются задачи кластеризации конечной последовательности точек евклидова пространства при фиксированном центре одного из кластеров и ограничениях на элементы, входящие в кластеры. По сути, результаты из этой главы развивают результаты главы 2. Учет ограничений в задачах кластеризации последовательности потребовал применения техники (схем) динамического программирования.

## 3.1 Задача 2-кластеризации

### 3.1.1 Формулировка задачи и известные результаты

**Задача 3.** *Дано:* последовательность $\mathcal{Y} = (y_1, \ldots, y_N)$ точек из $\mathbb{R}^d$, натуральные числа $T_{\min}$, $T_{\max}$ и $M > 1$. *Найти:* набор $\mathcal{M} = (n_1, \ldots, n_M)$, где $n_m \in \mathcal{N} = \{1, \ldots, N\}$, $m = 1, \ldots, M$, номеров элементов последовательности $\mathcal{Y}$ такой, что минимальна целевая функция

$$F(\mathcal{M}) = \sum_{j \in \mathcal{M}} \|y_j - \overline{y}(\mathcal{M})\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}} \|y_i\|^2,$$

где $\overline{y}(\mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum\limits_{j \in \mathcal{M}} y_j$ — центроид $\{y_j \mid j \in \mathcal{M}\}$, при ограничениях

$$T_{\min} \leq n_m - n_{m-1} \leq T_{\max} \leq N, \;\; m = 2, \ldots, M, \tag{3.1}$$

на элементы набора $\mathcal{M}$.

Частный случай этой задачи, в котором $T_{\min} = 1$ и $T_{\max} = N$, эквивалентен NP-трудной в сильном смысле задаче 2. Поэтому в случае, когда $T_{\min}$ и $T_{\max}$ являются частью входа, задача 3 также NP-трудна в сильном смысле.

В [42] анализировался случай задачи 3, в котором $T_{\min}$ и $T_{\max}$ являются параметрами, и было установлено, что задача 3 NP-трудна в сильном смысле для любых $T_{\min} < T_{\max}$. В тривиальном случае, когда $T_{\min} = T_{\max}$, задача разрешима за полиномиальное время.

К настоящему времени для задачи 3 был получен лишь один алгоритмический результат, а именно: в [56] предложен 2-приближённый полиномиальный алгоритм, временная сложность которого есть величина $\mathcal{O}(N^2(MN + d))$.

Ниже для задачи 3 представлены точный псевдополиномиальный алгоритм для специального случая задачи, аппроксимационная схема и рандомизированный алгоритм. Доказательства оценок качества этих алгоритмов опираются на результаты главы 2. Изложение доказательств оценок качества алгоритмов кластеризации последовательности аналогично изложению доказательств оценок качества алгоритмов кластеризации множества и приводится ниже ради полноты изложения. Следует лишь заметить, что отличие в изложении обусловлено учётом ограничений на элементы входной последовательности, включаемые в искомые кластеры.

### 3.1.2 Основы алгоритмов

Всюду далее будем использовать обозначение $f^y(x)$ для функции $f(x, y)$ при фиксированном аргументе $y$.

Положим

$$W(\mathcal{M}, x) = \sum_{n \in \mathcal{M}} \|y_n - x\|^2 + \sum_{n \in \mathcal{N} \setminus \mathcal{M}} \|y_n\|^2,\ \mathcal{M} \subseteq \mathcal{N},\ x \in \mathbb{R}^d, \tag{3.2}$$

$$G(\mathcal{M}, x) = 2 \sum_{n \in \mathcal{M}} \langle y_n, x \rangle - M \cdot \|x\|^2,\ \mathcal{M} \subseteq \mathcal{N},\ x \in \mathbb{R}^d, \tag{3.3}$$

где $y_n$ — элементы последовательности $\mathcal{Y}$, а элементы набора $\mathcal{M}$ удовлетворяют ограничениям (3.1).

**Лемма 3.1.** *При любом фиксированном $\mathcal{M} \subseteq \mathcal{N}$ минимум по $x$ функции* (3.2) *доставляется точкой $x = \overline{y}(\mathcal{M})$ и равен $F(\mathcal{M})$. При любой фиксированной точке $x \in \mathbb{R}^d$ минимум по $\mathcal{M}$ функции $W^x(\mathcal{M})$ достигается на наборе, доставляющем максимум функции $G^x(\mathcal{M})$.*

*Доказательство.* Справедливость первого утверждения вытекает из леммы 2.1 при $\mathcal{Z} = \{y_j \mid j \in \mathcal{M}\}$. Справедливость второго утверждения следует из следующей цепочки равенств:

$$\begin{aligned} W(\mathcal{M}, x) &= \sum_{n \in \mathcal{M}} \|y_n - x\|^2 + \sum_{n \in \mathcal{N} \setminus \mathcal{M}} \|y_n\|^2 \\ &= \sum_{n \in \mathcal{N}} \|y_n\|^2 + \sum_{n \in \mathcal{M}} (\|y_n - x\|^2 - \|y_n\|^2) = \sum_{n \in \mathcal{N}} \|y_n\|^2 - G^x(\mathcal{M}). \end{aligned} \tag{3.4}$$

Остаётся заметить, что первое слагаемое в правой части полученного выражения не зависит от $\mathcal{M}$. Лемма доказана.

Для произвольной фиксированной точки $x \in \mathbb{R}^d$ положим

$$g^x(n) = 2\langle y_n, x\rangle - \|x\|^2,\ n \in \mathcal{N}, \tag{3.5}$$

где $y_n$ — $n$-й элемент входной последовательности $\mathcal{Y}$. Тогда согласно (3.3) имеем

$$G^x(\mathcal{M}) = \sum_{n\in\mathcal{M}} g^x(n),\ \mathcal{M} \subseteq \mathcal{N}, \tag{3.6}$$

где элементы набора $\mathcal{M} = (n_1, \ldots, n_M)$ удовлетворяют ограничениям (3.1), и, кроме того, в соответствии со вторым утверждением леммы 3.1 имеет место равенство

$$\mathcal{M}^x = \arg\min_{\mathcal{M}} W^x(\mathcal{M}) = \arg\max_{\mathcal{M}} G^x(\mathcal{M}).$$

Рассмотрим следующую вспомогательную задачу.

**Задача 2′.** *Дано:* последовательность $\mathcal{Y} = (y_1, \ldots, y_N)$ точек из $\mathbb{R}^d$, точка $x \in \mathbb{R}^d$, натуральные числа $T_{\min}$, $T_{\max}$ и $M > 1$. *Найти:* набор $\mathcal{M} = (n_1, \ldots, n_M)$, где $n_m \in \mathcal{N}$, $m = 1, \ldots, M$, номеров элементов последовательности $\mathcal{Y}$, доставляющий максимум целевой функции (3.6), при ограничениях (3.1) на элементы набора $\mathcal{M}$.

В следующей лемме и следствии к ней приведена схема динамического программирования, гарантирующая отыскание оптимального решения $\mathcal{M}^x$ задачи 2′. Схема опирается на результаты из [56, 57] и приводится здесь ради полноты изложения.

**Лемма 3.2.** *Для любого натурального $M > 1$ такого, что $(M-1)T_{\min} \le N-1$, и для произвольной точки $x \in \mathbb{R}^d$ оптимальное значение $G^x_{\max} = \max\limits_{\mathcal{M}} G^x(\mathcal{M})$*

*целевой функции задачи* $2'$ *находится по формуле*

$$G^x_{\max} = \max_{n\in\omega_M} G^x_M(n), \tag{3.7}$$

*а значения функции* $G^x_M(n)$, $n \in \omega_M$, *вычисляются по следующим рекуррентным формулам:*

$$G^x_m(n) = \begin{cases} g^x(n), & \text{если } n \in \omega_1,\ m = 1; \\ g^x(n) + \max\limits_{j\in\gamma^-_{m-1}(n)} G^x_{m-1}(j), & \text{если } n \in \omega_m,\ m = 2,\dots,M, \end{cases} \tag{3.8}$$

*где множества* $\omega_m$ *и* $\gamma^-_{m-1}(n)$ *задаются следующими формулами:*

$$\omega_m = \big\{n \;\big|\; 1 + (m-1)T_{\min} \le n \le N - (M-m)T_{\min}\big\},\ m = 1,\dots,M,$$

$$\gamma^-_{m-1}(n) = \big\{j \;\big|\; \max\big\{1 + (m-2)T_{\min}, n - T_{\max}\big\} \le j \le n - T_{\min}\big\},$$
$$n \in \omega_m,\ m = 2,\dots,M.$$

**Следствие 3.1.** *Элементы* $n^x_1,\dots,n^x_M$ *оптимального набора* $\mathcal{M}^x$ *находятся по следующим рекуррентным формулам:*

$$n^x_M = \arg\max_{n\in\omega_M} G^x_M(n), \tag{3.9}$$

$$n^x_{m-1} = \arg\max_{n\in\gamma^-_{m-1}(n^x_m)} G^x_{m-1}(n),\ \ m = M, M-1,\dots,2. \tag{3.10}$$

Запишем алгоритм, реализующий приведённую схему, в пошаговом виде.

**Алгоритм** $\mathcal{A}'_2$**.**

*Вход алгоритма*: последовательность $\mathcal{Y}$, точка $x$, числа $T_{\min}$, $T_{\max}$ и $M$.

Шаг 1. Вычислим значения $g^x(n)$, $n \in \mathcal{N}$, по формуле (3.5).

Шаг 2. Используя рекуррентные формулы (3.8), вычислим значения $G_m^x(n)$ для каждого $n \in \omega_m$ и $m = 1, \ldots, M$.

Шаг 3. Найдём значение $G_{\max}^x$ максимума целевой функции $G^x$ по формуле (3.7) и оптимальный набор $\mathcal{M}^x = (n_1^x, \ldots, n_M^x)$ по формулам (3.9), (3.10).

*Выход:* набор $\mathcal{M}^x$.

В [56] установлено, что алгоритм $\mathcal{A}_2'$ находит оптимальное решение задачи $2'$ за время $\mathcal{O}\big(N\big(M(T_{\max} - T_{\min} + 1) + d\big)\big)$. В этом выражении значение $T_{\max} - T_{\min} + 1$ не превосходит $N$. Поэтому время работы алгоритма оценивается величиной $\mathcal{O}\big(N(MN + d)\big)$.

### 3.1.3 Точный псевдополиномиальный алгоритм для специального случая задачи

Изложим алгоритм решения задачи 3.

**Алгоритм $\mathcal{A}_5$.**

*Вход алгоритма:* последовательность $\mathcal{Y}$, натуральные числа $T_{\min}$, $T_{\max}$, $M$.

Шаг 1. Для каждой точки $x \in \mathcal{G}$, где $\mathcal{G}$ — многомерная решётка кубической формы размера $2D$ с расстоянием $\frac{1}{M}$ между узлами и центром в начале координат, а $D$ — максимальное абсолютное значение координат точек последовательности $\mathcal{Y}$, найдем оптимальное решение $\mathcal{M}^x$ и значение $G_{\max}^x$ целевой функции задачи $2'$ с помощью алгоритма $\mathcal{A}_2'$.

Шаг 2. Найдем точку $x_A = \arg\max\limits_{x \in \mathcal{G}} G_{\max}^x$ (если максимальному значению соответствует несколько точек, то выберем любую из них) и соответствующий ей набор $\mathcal{M}_A = \mathcal{M}^{x_A}$.

*Выход алгоритма:* набор $\mathcal{M}_A$.

**Теорема 3.1.** *Пусть в условиях задачи* 3 *точки последовательности* $\mathcal{Y}$ *имеют целочисленные координаты из интервала* $[-D, D]$. *Тогда алгоритм* $\mathcal{A}_5$ *находит оптимальное решение задачи* 3 *за время* $\mathcal{O}\big(N\big(M(T_{\max} - T_{\min} + 1) + d\big)(2MD + 1)^d\big)$.

*Доказательство.* Пусть $\mathcal{M}^*$ — оптимальное решение задачи 3, а $\mathcal{M}_A$ — набор, полученный в результате работы алгоритма $\mathcal{A}_5$.

Из определения решётки $\mathcal{G}$ следует, что центр $\overline{y}(\mathcal{M}) = \frac{1}{M}\sum\limits_{j\in\mathcal{M}} y_j$ любой подпоследовательности $\big\{y_j \;\big|\; j \in \mathcal{M}\big\} \subseteq \mathcal{Y}$, содержащей $M$ элементов, лежит во множестве $\mathcal{G}$. Следовательно, и центр $y^* = \overline{y}(\mathcal{M}^*)$ оптимальной подпоследовательности $\big\{y_j \;\big|\; j \in \mathcal{M}^*\big\}$ лежит в этом же множестве.

Из равенства (3.4) и определения шага 2 имеем

$$W(\mathcal{M}_A, x_A) = \sum_{n\in\mathcal{N}} \|y_n\|^2 - G^{x_A}(\mathcal{M}_A) \le \sum_{n\in\mathcal{N}} \|y_n\|^2 - G^{y^*}(\mathcal{M}^{y^*}). \tag{3.11}$$

Далее, так как алгоритм $\mathcal{A}_2'$ находит оптимальное решение задачи $2'$, то

$$G^{y^*}(\mathcal{M}^{y^*}) \ge G^{y^*}(\mathcal{M}^*).$$

Следовательно,

$$\sum_{n\in\mathcal{N}} \|y_n\|^2 - G^{y^*}(\mathcal{M}^{y^*}) \le \sum_{n\in\mathcal{N}} \|y_n\|^2 - G^{y^*}(\mathcal{M}^*) = W(\mathcal{M}^*, y^*) = F(\mathcal{M}^*). \tag{3.12}$$

Кроме того, из первого утверждения леммы 3.1 следует неравенство

$$F(\mathcal{M}_A) \le W(\mathcal{M}_A, x_A). \tag{3.13}$$

Объединяя (3.11)–(3.13), получаем оценку $F(\mathcal{M}_A) \le F(\mathcal{M}^*)$.

С другой стороны, так как набор $\mathcal{M}_A$ является допустимым решением задачи 3, то справедливо неравенство $F(\mathcal{M}^*) \leq F(\mathcal{M}_A)$, что устанавливает равенство значений $F(\mathcal{M}^*)$ и $F(\mathcal{M}_A)$.

Оценим временную сложность алгоритма. Шаг 1 выполняется $|\mathcal{G}| = (2MD+1)^d$ раз. При этом для каждой точки $x \in \mathcal{G}$ нахождение оптимального решения $\mathcal{M}$ вспомогательной задачи с помощью алгоритма $\mathcal{A}_2'$ требует $\mathcal{O}\big(N\big(M(T_{\max} - T_{\min} + 1) + d\big)\big)$ операций. Таким образом, итоговая временная сложность алгоритма есть величина $\mathcal{O}\big(N\big(M(T_{\max} - T_{\min} + 1) + d\big)(2MD+1)^d\big)$. Теорема доказана.

Покажем, что в случае, когда размерность пространства ограничена константой, алгоритм $\mathcal{A}_5$ псевдополиномиален. Действительно, поскольку $MD \geq \frac{1}{2}$, то

$$(2MD+1)^d = 2^d\left(MD + \frac{1}{2}\right)^d \leq 4^d(MD)^d.$$

Отсюда следует, что при указанных условиях время работы алгоритма оценивается величиной $\mathcal{O}\big(N\big(M(T_{\max} - T_{\min} + 1) + d\big)(MD)^d\big)$. Так как значение $(T_{\max} - T_{\min} + 1)$ не превосходит $N$, трудоёмкость алгоритма есть величина $\mathcal{O}\big(MN^2(MD)^d\big)$. Поскольку $D$ — числовое значение на входе задачи, алгоритм псевдополиномиален.

### 3.1.4 Аппроксимационная схема

Заметим сначала, что поскольку задача 3 — обобщение задачи 2, для неё, как и для задачи 2, не существует схемы FPTAS, если P $\neq$ NP. Для построения такой схемы для случая фиксированной размерности пространства нам потребуется несколько базовых утверждений.

**Лемма 3.3.** *Пусть $\mathcal{M}^*$ — оптимальное решение задачи 3, $x \in \mathbb{R}^d$ — произвольная фиксированная точка, $\mathcal{M}^x$ — набор, доставляющий минимум функ-*

*ции* $W^x(\mathcal{M})$, $\mathcal{M} \subseteq \mathcal{N}$, *при ограничениях* (3.1) *на элементы набора* $\mathcal{M}$, *а* $t = \arg \min\limits_{y \in \{y_i \,|\, i \in \mathcal{M}^*\}} \|y - \overline{y}(\mathcal{M}^*)\|$ *— точка из мультимножества* $\left\{y_i \;\middle|\; i \in \mathcal{M}^*\right\}$, *ближайшая к центроиду этого мультимножества. Тогда для того чтобы при фиксированном* $\varepsilon > 0$ *набор* $\mathcal{M}^x$ *был* $(1+\varepsilon)$*-приближённым решением задачи* 3, *достаточно, чтобы точка* $x$ *удовлетворяла неравенству*

$$\|x - \overline{y}(\mathcal{M}^*)\|^2 \leq \frac{\varepsilon}{2M} F(\mathcal{M}^t),$$

*где* $\mathcal{M}^t$ *— набор, доставляющий минимум функции* $W^t(\mathcal{M})$, $\mathcal{M} \subseteq \mathcal{N}$, *при ограничениях* (3.1) *на элементы набора* $\mathcal{M}$.

**Лемма 3.4.** *Пусть выполнены условия леммы* 3.3. *Тогда для точки* $t = \arg \min\limits_{y \in \{y_i \,|\, i \in \mathcal{M}^*\}} \|y - \overline{y}(\mathcal{M}^*)\|$ *справедлива оценка*

$$\|t - \overline{y}(\mathcal{M}^*)\|^2 \leq \frac{1}{M} F(\mathcal{M}^t). \tag{3.14}$$

Доказательства лемм 3.3 и 3.4 аналогичны доказательствам лемм 2.8 и 2.9.

Сформулируем следующий алгоритм решения задачи 3.

**Алгоритм $\mathcal{A}_6$.**

*Вход алгоритма*: последовательность $\mathcal{Y}$, натуральные числа $T_{\min}$, $T_{\max}$, $M$ и число $\varepsilon > 0$.

Для каждой точки $y \in \mathcal{Y}$ выполним шаги 1–5.

Шаг 1. С помощью алгоритма $\mathcal{A}_2'$ найдём оптимальное решение $\mathcal{M}^y$ задачи $2'$ при $x = y$.

Шаг 2. Вычислим $F(\mathcal{M}^y)$,

$$h = \sqrt{\frac{2\varepsilon}{dM} F(\mathcal{M}^y)}$$

и

$$H = \sqrt{\frac{1}{M} F(\mathcal{M}^y)}.$$

Шаг 3. Если $F(\mathcal{M}^y) = 0$, то набор $\mathcal{M}^y$ объявим результатом работы алгоритма; выход. Иначе переходим к следующему шагу.

Шаг 4. Построим решётку $\mathcal{G}(y, h, H + h/2)$.

Шаг 5. Для каждой точки $x$ решётки $\mathcal{G}(y, h, H + h/2)$ с помощью алгоритма $\mathcal{A}_2'$ построим оптимальное решение $\mathcal{M}^x$ задачи $2'$ и вычислим значение $F(\mathcal{M}^x)$.

Шаг 6. В семействе $\left\{\mathcal{M}^x \;\middle|\; x \in \mathcal{G}(y, h, H + h/2),\ y \in \mathcal{Y}\right\}$ наборов в качестве решения выберем тот набор $\mathcal{M}^x$, для которого значение $F(\mathcal{M}^x)$ минимально. Если минимальному значению соответствует несколько наборов, то выберем любой из них.

*Выход алгоритма:* набор $\mathcal{M}_A$.

**Теорема 3.2.** *Для любого фиксированного $\varepsilon > 0$ алгоритм $\mathcal{A}_6$ находит $(1 + \varepsilon)$-приближённое решение задачи 3 за время*[1] $\mathcal{O}\Big(N^2\big(M(T_{\max} - T_{\min} + 1) + d\big)\big(\sqrt{2d/\varepsilon} + 2\big)^d\Big)$.

*Доказательство.* Пусть $t = \arg \min\limits_{y \in \{y_i \mid i \in \mathcal{M}^*\}} \|y - \overline{y}(\mathcal{M}^*)\|$ — точка из мультимножества $\left\{y_i \;\middle|\; i \in \mathcal{M}^*\right\}$, ближайшая к центроиду этого мультимножества. В случае, когда для этой точки на шаге 3 алгоритма выполнено равенство $F(\mathcal{M}^t) = 0$, набор $\mathcal{M}^t$ является оптимальным решением задачи 3, так как для любого набора $\mathcal{M} \subseteq \mathcal{N}$ справедливо неравенство $F(\mathcal{M}) \geq 0$.

Рассмотрим случай, когда $F(\mathcal{M}^t) > 0$. В соответствии с леммой 3.4 для

[1]При аккуратной организации вычислений формула трудоёмкости алгоритма именно такая, что слегка отличается от формулы $\mathcal{O}(N^2(M(T_{\max} - T_{\min} + 1) + d)(\sqrt{2d/\varepsilon} + 1)^d)$, указанной в статье [65].

точки $t$ выполнено неравенство (3.14), т.е.

$$\|t - \overline{y}(\mathcal{M}^*)\| \leq \sqrt{\frac{1}{M} F(\mathcal{M}^y)} = H. \tag{3.15}$$

Положим $x^* = \arg \min\limits_{x \in \mathcal{G}(t,h,H+h/2)} \|x - \overline{y}(\mathcal{M}^*)\|$. Из (3.15) следует, что расстояние от $\overline{y}(\mathcal{M}^*)$ до $x^*$ — ближайшего узла решётки $\mathcal{G}(t, h, H + h/2)$ — не превосходит $h\sqrt{d}/2$. Поэтому справедлива оценка

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

Таким образом, точка $x^*$ удовлетворяет условиям леммы 3.3, а набор $\mathcal{M}^{x^*}$ является $(1 + \varepsilon)$-приближённым решением задачи 3.

Оценим временную сложность алгоритма. Шаг 1 — решение вспомогательной задачи — требует $\mathcal{O}\big(N\big(M(T_{\max} - T_{\min} + 1) + d\big)\big)$ операций [56]. На шаге 2 требуется $\mathcal{O}(dN)$ операций, а шаг 3 выполняется за $\mathcal{O}(1)$ операций.

Для построения решётки $\mathcal{G}(y, h, H + h/2)$ на шаге 4 потребуется $\mathcal{O}\big(d|\mathcal{G}(y, h, H + h/2)|\big)$ операций. Построение каждого из $|\mathcal{G}(y, h, H + h/2)|$ наборов $\mathcal{M}^x$ на шаге 5 и вычисление значений $F(\mathcal{M}^x)$ выполняется, как и на шаге 1, за $\mathcal{O}\big(N\big(M(T_{\max} - T_{\min} + 1) + d\big)\big)$ операций.

В итоге, для каждой из $N$ точек $y \in \mathcal{Y}$ выполнение шагов 1–5 потребует $\mathcal{O}\big(N\big(M(T_{\max} - T_{\min} + 1) + d\big)|\mathcal{G}(y, h, H + h/2)|\big)$ операций. Наконец, на шаге 6 для выбора наименьшего элемента требуется $\mathcal{O}\left(\sum\limits_{y \in \mathcal{Y}} |\mathcal{G}(y, h, H + h/2)|\right)$ операций.

Остаётся заметить, что для мощности решётки $\mathcal{G}(y, h, H + h/2)$ справедлива оценка

$$|\mathcal{G}(y, h, H + h/2)| \leq \big(2H/h + 2\big)^d \leq \left(\sqrt{2d/\varepsilon} + 2\right)^d.$$

Поэтому временная сложность алгоритма есть величина $\mathcal{O}\Big(N^2\big(M(T_{\max}-T_{\min}+1)+d\big)\big(\sqrt{2d/\varepsilon}+2\big)^d\Big)$. Теорема доказана.

Покажем, что в случае фиксированной размерности $d$ пространства алгоритм $\mathcal{A}_6$ реализует схему FPTAS. Действительно, если $\varepsilon \in (0,\, d/2]$, то

$$\Big(\sqrt{2d/\varepsilon}+2\Big)^d \le 2^d\Big(\sqrt{2d/\varepsilon}\Big)^d = 2^{3d/2}d^{d/2}(1/\varepsilon)^{d/2} = \mathcal{O}\Big((1/\varepsilon)^{d/2}\Big).$$

Следовательно, так как величина $T_{\max}-T_{\min}+1$ не превосходит $N$, при указанных условиях время работы алгоритма есть величина $\mathcal{O}\Big(MN^3(1/\varepsilon)^{d/2}\Big)$ и алгоритм реализует схему FPTAS.

### 3.1.5 Рандомизированный алгоритм

Сформулируем рандомизированный алгоритм для задачи 3.

**Алгоритм $\mathcal{A}_7$.**

*Вход алгоритма*: последовательность $\mathcal{Y}$, натуральные числа $T_{\min}$, $T_{\max}$, $M$ и натуральный параметр $k$.

Шаг 1. Сформируем мультимножество $\mathcal{T}$ точек с помощью $k$ независимых случайных выборок (с возвращением) по одному элементу из последовательности $\mathcal{Y}$.

Шаг 2. Для каждого непустого $\mathcal{H} \subseteq \mathcal{T}$ вычислим центроид $\overline{y}(\mathcal{H})$ и с помощью алгоритма $\mathcal{A}_2'$ найдём оптимальное решение $\mathcal{M}^{\overline{y}(\mathcal{H})}$ задачи $2'$ при $x = \overline{y}(\mathcal{H})$.

Шаг 3. В семействе решений, найденных на шаге 2, выберем тот набор $\mathcal{M}_A = \mathcal{M}^{\overline{y}(\mathcal{H})}$ номеров элементов последовательности $\mathcal{Y}$, для которого значение $F(\mathcal{M}^{\overline{y}(\mathcal{H})})$ минимально. Если минимальному значению соответствует несколько наборов, то выберем любой из них.

*Выход алгоритма:* набор $\mathcal{M}_A$.

Свойства алгоритма $\mathcal{A}_7$ устанавливает

**Теорема 3.3.** *Для произвольных вещественного $\delta \in (0,1)$ и натуральных $t \leq k$ алгоритм $\mathcal{A}_7$ находит $\left(1+\frac{1}{\delta t}\right)$-приближённое решение задачи 3 за время $\mathcal{O}\left(2^k\left(dk + N\left(M(T_{\max} - T_{\min} + 1) + d\right)\right)\right)$ с вероятностью не менее $1-(\delta+\alpha)$, где $\alpha = \sum\limits_{i=0}^{t-1} \binom{k}{i} \left(\frac{M}{N}\right)^i \left(1 - \frac{M}{N}\right)^{k-i}$.*

*Доказательство.* Пусть набор $\mathcal{M}^*$ — оптимальное решение задачи 3, а мультимножество $\mathcal{T}'$ состоит из номеров элементов мультимножества $\mathcal{T}$, построенного на шаге 1 работы алгоритма, т.е. $\mathcal{T} = \{y_i \mid i \in \mathcal{T}'\}$.

Предположим, что для мощности пересечения мультимножества $\mathcal{T}'$ и набора $\mathcal{M}^*$ выполнено неравенство $|\mathcal{T}' \cap \mathcal{M}^*| \geq 1$. Тогда одним из подмножеств мультимножества $\mathcal{T}$, рассмотренных на шаге 2, было мультимножество $\mathcal{H} = \{y_i \mid i \in \mathcal{T}' \cap \mathcal{M}^*\}$. Обозначим через $\mathcal{M}'$ полученное на этом же шаге оптимальное решение задачи $2'$, соответствующее мультимножеству $\mathcal{H} = \{y_i \mid i \in \mathcal{T}' \cap \mathcal{M}^*\}$. Из определения шага 3 следует оценка

$$F(\mathcal{M}_A) \leq F(\mathcal{M}'). \tag{3.16}$$

Кроме того, из первого утверждения леммы 3.1 имеем

$$F(\mathcal{M}') = W^{\overline{y}(\mathcal{M}')}(\mathcal{M}') \leq W^{\overline{y}(\mathcal{H})}(\mathcal{M}'). \tag{3.17}$$

Далее, поскольку алгоритм $\mathcal{A}_2'$ находит оптимальное решение задачи $2'$ максимизации функции (3.6), из утверждения 2 леммы 3.1 для функции (3.2) справедлива оценка

$$W^{\overline{y}(\mathcal{H})}(\mathcal{M}') \leq W^{\overline{y}(\mathcal{H})}(\mathcal{M}^*). \tag{3.18}$$

Объединяя (3.16)–(3.18), получим, что при указанном выше предположении

$|\mathcal{T}' \cap \mathcal{M}^*| \geq 1$ справедливо

$$F(\mathcal{M}_A) \leq F(\mathcal{M}') \leq W^{\overline{y}(\mathcal{H})}(\mathcal{M}') \\ \leq W^{\overline{y}(\mathcal{H})}(\mathcal{M}^*) = \sum_{j \in \mathcal{M}^*} \|y_j - \overline{y}(\mathcal{H})\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}^*} \|y_i\|^2. \quad (3.19)$$

Нетрудно показать, что лемму 2.11 можно сформулировать не только для множеств, но и для наборов номеров элементов последовательности, а именно: при условии $|\mathcal{T}' \cap \mathcal{M}^*| \geq t$ с вероятностью более $1 - \delta$ выполнено

$$\sum_{j \in \mathcal{M}^*} \|y_j - \overline{y}(\mathcal{H})\|^2 < \left(1 + \frac{1}{\delta t}\right) \sum_{j \in \mathcal{M}^*} \|y_j - \overline{y}(\mathcal{M}^*)\|^2. \quad (3.20)$$

Наконец, комбинируя (3.19) и (3.20), получим, что при условии $|\mathcal{T}' \cap \mathcal{M}^*| \geq t$ с вероятностью более $1 - \delta$ справедлива цепочка оценок

$$F(\mathcal{M}_A) \leq \sum_{j \in \mathcal{M}^*} \|y_j - \overline{y}(\mathcal{H})\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}^*} \|y_i\|^2 \\ < \left(1 + \frac{1}{\delta t}\right) \sum_{j \in \mathcal{M}^*} \|y_j - \overline{y}(\mathcal{M}^*)\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}^*} \|y_i\|^2 \\ \leq \left(1 + \frac{1}{\delta t}\right) \left( \sum_{j \in \mathcal{M}^*} \|y_j - \overline{y}(\mathcal{M}^*)\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}^*} \|y_i\|^2 \right) = \left(1 + \frac{1}{\delta t}\right) F(\mathcal{M}^*).$$

Это значит, что

$$\Pr\left( F(\mathcal{M}_A) < \left(1 + \frac{1}{\delta t}\right) F(\mathcal{M}^*) \;\middle|\; |\mathcal{T}' \cap \mathcal{M}^*| \geq t \right) > 1 - \delta$$

в терминах условной вероятности. Следовательно, для безусловной вероятности

справедливо неравенство

$$\Pr\left(F(\mathcal{M}_A) < \left(1+\frac{1}{\delta t}\right)F(\mathcal{M}^*)\right) > 1 - \big(\delta + \Pr\big(|\mathcal{T}' \cap \mathcal{M}^*| < t\big)\big),$$

которое, с учётом равенства $\Pr\big(|\mathcal{T}' \cap \mathcal{M}^*| < t\big) = \sum\limits_{i=0}^{t-1} \binom{k}{i} \left(\frac{M}{N}\right)^i \left(1-\frac{M}{N}\right)^{k-i}$, устанавливает вероятностные и аппроксимационные оценки алгоритма $\mathcal{A}_7$.

Оценим временную сложность алгоритма. Трудоёмкость шага 1 определяется мощностью мультимножества $\mathcal{T}$ и равна $\mathcal{O}(k)$. На шаге 2 для каждого из $\mathcal{O}(2^k)$ подмножеств $\mathcal{H}$ мультимножества $\mathcal{T}$ вычисление центра $\overline{y}(\mathcal{H})$ требует $\mathcal{O}(dk)$ операций, а нахождение оптимального решения задачи 2 — $\mathcal{O}\big(N\big(M(T_{\max} - T_{\min} + 1) + d\big)\big)$ операций. Шаг 3 — вычисление наименьшего элемента — требует $\mathcal{O}(2^k)$ операций. Просуммировав затраты на всех шагах, получим анонсированную в формулировке теоремы оценку временной сложности. Теорема доказана.

**Замечание 3.1.** *Трудоёмкость алгоритма $\mathcal{A}_7$ можно оценить величиной $\mathcal{O}\big(2^k\big(dk + N(MN + d)\big)\big)$, поскольку величина $T_{\max} - T_{\min} + 1$ не превосходит $N$.*

**Следствие 3.2.** *Допустим, что $M \geq \beta N$, где $\beta \in (0,1)$ — некоторая константа. Тогда для заданных $\varepsilon > 0$ и $\gamma \in (0,1)$ при фиксированном параметре $k = \max\Big(\Big\lceil \frac{2}{\beta}\Big\lceil \frac{2}{\gamma\varepsilon}\Big\rceil\Big\rceil, \Big\lceil \frac{8}{\beta}\ln\frac{2}{\gamma}\Big\rceil\Big)$ алгоритм $\mathcal{A}_7$ находит $(1+\varepsilon)$-приближённое решение задачи 3 с вероятностью не менее $1-\gamma$ за время $\mathcal{O}(dMN^2)$.*

Доказательство вероятностных и аппроксимационных характеристик алгоритма $\mathcal{A}_7$ при указанном значении параметра $k$ мы здесь опускаем, так как оно аналогично доказательству следствия 2.2 к теореме 2.5.

Сформулируем условия асимптотической точности алгоритма $\mathcal{A}_7$.

**Теорема 3.4.** *Пусть в условиях теоремы* 3.3 *имеют место равенства* $k = \lceil \log_2 N \rceil$, $\delta = (\log_2 N)^{-1/2}$, $t = \left\lceil \frac{kM}{2N} \right\rceil$. *Допустим, что* $M \geq \beta N$, *где* $\beta \in (0,1)$ *— некоторая константа. Тогда алгоритм* $\mathcal{A}_7$ *находит* $(1+\varepsilon_N)$*-приближённое решение задачи* 3 *с вероятностью* $1-\gamma_N$ *за время* $\mathcal{O}\big(dMN^2(T_{\max} - T_{\min} + 1)\big)$, *где*

$$\varepsilon_N \leq \frac{2}{\beta}(\log_2 N)^{-1/2} \xrightarrow[N\to\infty]{} 0,$$

$$\gamma_N \leq (\log_2 N)^{-1/2} + N^{-\frac{\beta}{8\ln 2}} \xrightarrow[N\to\infty]{} 0.$$

*Доказательство.* Доказательство этих качественных свойств алгоритма $\mathcal{A}_7$ поиска подпоследовательности аналогично доказательству оценок погрешности и вероятности несрабатывания алгоритма $\mathcal{A}_4$ для поиска подмножества.

Оценка времени работы алгоритма следует из того, что при $k = \lceil \log_2 N \rceil$ справедлива цепочка равенств

$$\begin{aligned} 2^k\big(dk + N\big(M(T_{\max} - T_{\min} + 1) + d\big)\big) \\ = \mathcal{O}\big(N\big(d\log_2 N + N\big(M(T_{\max} - T_{\min} + 1) + d\big)\big)\big) \\ = \mathcal{O}\Big(dMN^2(T_{\max} - T_{\min} + 1)\Big). \end{aligned}$$

Теорема доказана.

Теорема 3.4 устанавливает условия, при которых алгоритм $\mathcal{A}_7$ асимптотически точен и, согласно замечанию 3.1, имеет трудоёмкость $\mathcal{O}(dMN^3)$.

## 3.2 Задачи многокластерного разбиения

### 3.2.1 Формулировки задач и известные результаты

**Задача 4.** *Дано:* последовательность $\mathcal{Y} = (y_1, \ldots, y_N)$ точек из $\mathbb{R}^d$, натуральные числа $T_{\min}$, $T_{\max}$, $L$ и $M$. *Найти:* непустые непересекающиеся наборы $\mathcal{M}_1, \ldots, \mathcal{M}_L$ номеров элементов последовательности $\mathcal{Y}$ такие, что

$$F(\mathcal{M}_1, \ldots, \mathcal{M}_L) = \sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} \|y_j - \overline{y}(\mathcal{M}_l)\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}} \|y_i\|^2 \to \min, \qquad (3.21)$$

где $\mathcal{M} = \bigcup\limits_{l=1}^{L} \mathcal{M}_l$, $\overline{y}(\mathcal{M}_l) = \frac{1}{|\mathcal{M}_l|} \sum\limits_{j \in \mathcal{M}_l} y_j$, $l = 1, \ldots, L$, — центроид $\left\{ y_j \;\middle|\; j \in \mathcal{M}_l \right\}$, при ограничениях: (i) мощность объединённого набора $\mathcal{M}$ равна $M$, (ii) в последовательности, образованной конкатенацией наборов $\mathcal{M}_1, \ldots, \mathcal{M}_L$, номера упорядочены по возрастанию при условии, что элементы каждого набора образуют возрастающую последовательность, и (iii) номера из объединённого набора $\mathcal{M} = (n_1, \ldots, n_M)$ связаны неравенствами (3.1).

**Задача 5.** *Дано:* последовательность $\mathcal{Y} = (y_1, \ldots, y_N)$ точек из $\mathbb{R}^d$, натуральные числа $T_{\min}$, $T_{\max}$ и $L$. *Найти:* непустые непересекающиеся наборы $\mathcal{M}_1, \ldots, \mathcal{M}_L$ номеров элементов последовательности $\mathcal{Y}$ такие, что минимальна целевая функция (3.21), при ограничениях: (i) в последовательности, образованной конкатенацией наборов $\mathcal{M}_1, \ldots, \mathcal{M}_L$, номера упорядочены по возрастанию при условии, что элементы каждого набора образуют возрастающую последовательность, и (ii) номера из объединённого набора $\mathcal{M} = (n_1, \ldots, n_M)$ связаны неравенствами (3.1), где $M$ зависит от оптимизируемых переменных.

Частный случай задачи 4, в котором $L = 1$, эквивалентен NP-трудной в сильном смысле задаче 3. Поэтому задача 4 также NP-трудна в сильном смысле.

NP-трудность в сильном смысле задачи 5 следует из того, что случай этой задачи, в котором $L = 1$, эквивалентен NP-трудной в сильном смысле [42] задаче разбиения последовательности на два кластера с одним фиксированным центром без ограничений на мощности кластеров.

Как для задачи 4, так и для задачи 5, за исключением их частных случаев, когда $L = 1$, к настоящему времени отсутствовали какие-либо эффективные алгоритмы с оценками точности. В настоящей работе для этих задач предложены 2-приближённые алгоритмы, полиномиальные в случае, когда число кластеров фиксировано.

### 3.2.2 Основы алгоритмов

**Лемма 3.5.** *Пусть*

$$W(\mathcal{M}_1, \ldots, \mathcal{M}_L, x_1, \ldots, x_L) = \sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} \|y_j - x_l\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}} \|y_i\|^2, \tag{3.22}$$

$$G(\mathcal{M}_1, \ldots, \mathcal{M}_L, x_1, \ldots, x_L) = \sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} \Big( 2\langle y_j, x_l \rangle - \|x_l\|^2 \Big),$$

*где* $x_1, \ldots, x_L$ *— точки из* $\mathbb{R}^d$*, а элементы наборов* $\mathcal{M}_l, \ldots, \mathcal{M}_L$ *и* $\mathcal{M}$ *удовлетворяют ограничениям задачи* 4 *(задачи* 5*). Тогда для условных оптимумов функции* (3.22) *справедливы следующие утверждения:*

1) *для любых непустых фиксированных наборов* $\mathcal{M}_1, \ldots, \mathcal{M}_L$ *минимум функции* (3.22) *по переменным* $x_1, \ldots, x_L$ *достигается в точках* $x_l = \overline{y}(\mathcal{M}_l)$, $l = 1, \ldots, L$*, и равен* $F(\mathcal{M}_1, \ldots, \mathcal{M}_L)$*;*

2) *для любого набора* $x = (x_1, \ldots, x_L)$ *фиксированных точек из* $\mathbb{R}^d$ *минимум функции* $W^x(\mathcal{M}_1, \ldots, \mathcal{M}_L)$ *по наборам* $\mathcal{M}_1, \ldots, \mathcal{M}_L$*, удовлетворяющим ограничениям задачи* 4 *(задачи* 5*), достигается на наборах* $\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x$ *номеров элементов последовательности* $\mathcal{Y}$*, для которых максимальна функция* $G^x(\mathcal{M}_1, \ldots, \mathcal{M}_L)$.

*Доказательство.* Первое утверждение леммы легко проверяется дифференцированием и следует также из леммы 2.1. Для доказательства второго утверждения достаточно заметить, что справедливо следующее легко проверяемое равенство

$$W^x(\mathcal{M}_1,\dots,\mathcal{M}_L)=\sum_{j\in\mathcal{N}}\|y_j\|^2-G^x(\mathcal{M}_1,\dots,\mathcal{M}_L), \tag{3.23}$$

в правой части которого сумма не зависит от $\mathcal{M}_1,\dots,\mathcal{M}_L$. Лемма доказана.

Вычислительной базой предлагаемых алгоритмов являются точные полиномиальные алгоритмы решения следующих вспомогательных задач.

**Задача 3′.** *Дано:* последовательность $\mathcal{Y}=(y_1,\dots,y_N)$ и набор $x=(x_1,\dots,x_L)$ точек из $\mathbb{R}^d$, натуральные числа $T_{\min}$, $T_{\max}$ и $M$. *Найти:* непустые непересекающиеся наборы $\mathcal{M}_1,\dots,\mathcal{M}_L$ номеров элементов последовательности $\mathcal{Y}$ такие, что целевая функция $G^x(\mathcal{M}_1,\dots,\mathcal{M}_L)$ максимальна, при тех же, что и в задаче 4, ограничениях на искомые переменные.

**Задача 4′.** *Дано:* последовательность $\mathcal{Y}=(y_1,\dots,y_N)$ и набор $x=(x_1,\dots,x_L)$ точек из $\mathbb{R}^d$, натуральные числа $T_{\min}$, $T_{\max}$. *Найти:* непустые непересекающиеся наборы $\mathcal{M}_1,\dots,\mathcal{M}_L$ номеров элементов последовательности $\mathcal{Y}$ такие, что целевая функция $G^x(\mathcal{M}_1,\dots,\mathcal{M}_L)$ максимальна, при тех же, что и в задаче 5, ограничениях на искомые переменные.

Для изложения алгоритмов решения вспомогательных задач определим функцию

$$g_l^x(n)=2\langle y_n,x_l\rangle-\|x_l\|^2,\ \ n\in\mathcal{N},\ l=1,\dots,L, \tag{3.24}$$

где $x_l$ — точка из набора $x$, $y_n$ — элемент последовательности $\mathcal{Y}$.

В соответствии с определением (3.24) для целевой функции $G^x(\mathcal{M}_1,\dots,\mathcal{M}_L)$

задач $3'$ и $4'$ имеем

$$G^x(\mathcal{M}_1,\ldots,\mathcal{M}_L)=\sum_{l=1}^{L}\sum_{n\in\mathcal{M}_l}g_l^x(n).$$

Кроме того, заметим, что утверждение 2 леммы 3.5 означает равенства

$$\{\mathcal{M}_1^x,\ldots,\mathcal{M}_L^x\}=\arg\min_{\mathcal{M}_1,\ldots,\mathcal{M}_L}W^x(\mathcal{M}_1,\ldots,\mathcal{M}_L)=\arg\max_{\mathcal{M}_1,\ldots,\mathcal{M}_L}G^x(\mathcal{M}_1,\ldots,\mathcal{M}_L). \tag{3.25}$$

В следующей лемме и следствии к ней приведена схема динамического программирования, гарантирующая отыскание оптимального решения $\mathcal{M}_1^x,\ldots,\mathcal{M}_L^x$ задачи $3'$ и (согласно приведенным выше равенствам (3.25)) оптимального решения задачи минимизации функции $W^x(\mathcal{M}_1,\ldots,\mathcal{M}_L)$ при ограничениях задачи 4. Схема следует из результатов [58] и приводится здесь ради полноты изложения.

**Лемма 3.6.** *Пусть выполнены условия задачи* $3'$*. Тогда для любых натуральных* $L$ *и* $M$ *таких, что* $(M-1)T_{\min}\le N-1$ *и* $L\le M$*, оптимальное значение* $G^x_{\max}$ *целевой функции этой задачи находится по формуле*

$$G^x_{\max}=\max_{n\in\{1+(M-1)T_{\min},\ldots,N\}}G^x_{L,\,M}(n), \tag{3.26}$$

*а значения функции* $G^x_{L,\,M}(n)$ *вычисляются по следующим рекуррентным фор-*

*мулам:*

$$G^x_{l,\,m}(n) = g^x_l(n) + \begin{cases} 0, & \text{если } l = 1,\ m = 1, \\ \max\limits_{j\in\gamma_{m-1}(n)} G^x_{1,\,m-1}(j), & \\ & \text{если } l = 1,\ m = 2,\dots,M-(L-1), \\ \max\limits_{j\in\gamma_{m-1}(n)} G^x_{l-1,\,m-1}(j), & \\ & \text{если } l = 2,\dots,L,\ m = l, \\ \max\Bigl\{\max\limits_{j\in\gamma_{m-1}(n)} G^x_{l,\,m-1}(j), \max\limits_{j\in\gamma_{m-1}(n)} G^x_{l-1,\,m-1}(j)\Bigr\}, & \\ & \text{если } l = 2,\dots,L,\ m = l+1,\dots,M-(L-l), \end{cases} \tag{3.27}$$

*где*

$$\gamma_{m-1}(n) = \bigl\{ j \bigm| \max\bigl\{1+(m-2)T_{\min}, n-T_{\max}\bigr\} \le j \le n - T_{\min}\bigr\},\ m = 2,\dots,M, \tag{3.28}$$

*при каждом* $n = 1+(m-1)T_{\min},\dots,N-(M-m)T_{\min}$.

**Следствие 3.3.** *Пусть выполнены условия леммы* 3.6. *Пусть, кроме того,*

$$r^x_{l,\,m}(n) = \begin{cases} 1, & \text{если } l = 1,\ m = 2,\dots,M-(L-1), \\ l-1, & \text{если } l = 2,\dots,L,\ m = l, \\ l-1, & \text{если } \max\limits_{j\in\gamma_{m-1}(n)} G^x_{l,\,m-1}(j) < \max\limits_{j\in\gamma_{m-1}(n)} G^x_{l-1,\,m-1}(j), \\ & l = 2,\dots,L,\ m = l+1,\dots,M-(L-l), \\ l, & \text{если } \max\limits_{j\in\gamma_{m-1}(n)} G^x_{l,\,m-1}(j) \ge \max\limits_{j\in\gamma_{m-1}(n)} G^x_{l-1,\,m-1}(j), \\ & l = 2,\dots,L,\ m = l+1,\dots,M-(L-l), \end{cases}$$

$$I^x_{l,\,m}(n) = \arg \max_{j \in \gamma_{m-1}(n)} G^x_{l,\,m-1}(j),\ l = 1, \ldots, L,\ m = l+1, \ldots, M-(L-l),$$

*при каждом* $n = 1 + (m-1)T_{\min}, \ldots, N - (M-m)T_{\min}$*;*

$$n^x(m) = \begin{cases} \arg \max\limits_{n \in \{1+(M-1)T_{\min}, \ldots, N\}} G^x_{L,\,M}(n), & \text{если } m = M, \\ I^x_{k^x(m),\,m+1}\big(n^x(m+1)\big), & \text{если } m = M-1, \ldots, 1; \end{cases}$$

$$k^x(m) = \begin{cases} L, & \text{если } m = M, \\ r^x_{k^x(m+1),\,m+1}\big(n^x(m+1)\big), & \text{если } m = M-1, \ldots, 1; \end{cases}$$

$$J^x(l) = \begin{cases} 0, & \text{если } l = 0, \\ \big|\big\{m \in \{1, \ldots, M\} \;\big|\; k^x(m) \le l\big\}\big|, & \text{если } l = 1, \ldots, L. \end{cases}$$

*Тогда наборы* $\mathcal{M}^x_1, \ldots, \mathcal{M}^x_L$ *определяются по правилу*

$$\mathcal{M}^x_l = \big\{n \;\big|\; n = n^x(m),\ m = J^x(l-1)+1, \ldots, J^x(l)\big\} \tag{3.29}$$

*при каждом* $l = 1, \ldots, L$*.*

Запишем алгоритм, реализующий приведённую схему, в пошаговом виде.

**Алгоритм $\mathcal{A}'_3$.**

*Вход алгоритма:* последовательность $\mathcal{Y}$, набор $(x_1, \ldots, x_L)$ точек, натуральные числа $T_{\min}$, $T_{\max}$ и $M$.

Шаг 1. Вычислим значения $g^x_l(n)$ для $l = 1, \ldots, L$, $n = 1+(l-1)T_{\min}, \ldots, N-(L-l)T_{\min}$ по формуле (3.24).

Шаг 2. Используя рекуррентные формулы (3.27) и (3.28), вычислим значения $G^x_{l,\,m}(n)$ для каждого $l = 1, \ldots, L$, $m = l, \ldots, M-(L-l)$, $n = 1+(m-1)T_{\min}, \ldots, N-(M-m)T_{\min}$.

Шаг 3. Найдём значение $G^x_{\max}$ максимума целевой функции $G^x$ по формуле (3.26) и оптимальные наборы $\mathcal{M}^x_l$ по формуле (3.29).

*Выход алгоритма:* семейство $\{\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x\}$ наборов.

**Замечание 3.2.** *Перед началом работы алгоритма $\mathcal{A}_3'$ требуется проверка справедливости двух условий леммы* 3.6. *Эти необходимые условия обеспечивают совместность ограничений в задачах* 4 *и* 3′, *а также корректность входных данных алгоритма.*

**Замечание 3.3.** *В* [58] *установлено, что алгоритм $\mathcal{A}_3'$ находит оптимальное решение задачи* 3′ *за время $\mathcal{O}\big(LN\big(M(T_{\max} - T_{\min} + 1) + d\big)\big)$. В этом выражении значение $T_{\max} - T_{\min} + 1$ не превосходит $N$. Поэтому время работы алгоритма оценивается величиной $\mathcal{O}\big(LN(MN + d)\big)$.*

В следующей лемме и следствии к ней приведена схема динамического программирования, гарантирующая отыскание оптимального решения $\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x$ задачи 4′ и (согласно приведенным выше равенствам (3.25)) оптимального решения задачи минимизации функции $W^x(\mathcal{M}_1, \ldots, \mathcal{M}_L)$ при ограничениях задачи 5. Схема следует из результатов [59] и приводится здесь ради полноты изложения.

**Лемма 3.7.** *Пусть выполнены условия задачи* 4′. *Тогда оптимальное значение $G_{\max}^x$ целевой функции этой задачи находится по формуле*

$$G_{\max}^x = \max_{n \in \{1+(L-1)T_{\min}, \ldots, N\}} G_{L,n}^x, \tag{3.30}$$

*а значения функции $G_{L,n}^x$ вычисляются по следующим рекуррентным форму-*

*лам:*

$$G^x_{l,n} = g^x_l(n) + \begin{cases} 0, & \text{если } l = 1,\ n = 1, \dots, T_{\min}, \\ \max\{0, \max\limits_{j \in \gamma_1(n)} G^x_{1,j}\}, & \\ & \text{если } l = 1,\ n = 1 + T_{\min}, \dots, N - (L-1)T_{\min}, \\ \max\limits_{j \in \gamma_{l-1}(n)} G^x_{l-1,j}, & \\ & \text{если } l = 2, \dots, L,\ n = 1 + (l-1)T_{\min}, \dots, lT_{\min}, \\ \max\left\{ \max\limits_{j \in \gamma_{l-1}(n)} G^x_{l-1,j}, \max\limits_{j \in \gamma_l(n)} G^x_{l,j} \right\}, & \\ & \text{если } l = 2, \dots, L,\ n = 1 + lT_{\min}, \dots, N - (L-l)T_{\min}, \end{cases} \tag{3.31}$$

*где*

$$\gamma_l(n) = \left\{ j \;\middle|\; \max\left\{1 + (l-1)T_{\min}, n - T_{\max}\right\} \le j \le n - T_{\min} \right\}, \tag{3.32}$$

*при каждом* $l = 1, \dots, L$, $n = 1 + lT_{\min}, \dots, N - (L-l)T_{\min}$.

**Следствие 3.4.** *Пусть выполнены условия леммы* 3.7. *Пусть, кроме того,*

$$r^x(l,n) = \begin{cases} 0, & \text{если } n = 1, \dots, T_{\min}, \\ 0, & \text{если } \max\limits_{j \in \gamma_1(n)} G^x_{1,j} \le 0,\ n = 1 + T_{\min}, \dots, N - (L-1)T_{\min}, \\ 1, & \text{если } \max\limits_{j \in \gamma_1(n)} G^x_{1,j} > 0,\ n = 1 + T_{\min}, \dots, N - (L-1)T_{\min}, \end{cases}$$

*при* $l = 1$, *и*

$$r^x(l,n) = \begin{cases} l-1, & \text{если } n = 1 + (l-1)T_{\min}, \dots, lT_{\min}, \\ l-1, & \text{если } \max\limits_{j \in \gamma_l(n)} G^x_{l,j} \le \max\limits_{j \in \gamma_{l-1}(n)} G^x_{l-1,j}, \\ & n = 1 + lT_{\min}, \dots, N - (L-l)T_{\min}, \\ l, & \text{если } \max\limits_{j \in \gamma_l(n)} G^x_{l,j} > \max\limits_{j \in \gamma_{l-1}(n)} G^x_{l-1,j}, \\ & n = 1 + lT_{\min}, \dots, N - (L-l)T_{\min}, \end{cases}$$

*при* $l = 2, \dots, L$;

$$I^x(l,n) = \arg \max_{j \in \gamma_l(n)} G^x_{l,j},\ l = 1, \dots, L,\ n = 1 + lT_{\min}, \dots, N - (L-l)T_{\min};$$

$$k^x(m) = \begin{cases} L, & \text{если } m = 1, \\ r^x\big(k^x(m-1), \nu^x(m-1)\big), & \text{если } m = 2, \dots, M^x; \end{cases}$$

$$\nu^x(m) = \begin{cases} \arg \max\limits_{n \in \{1 + (L-1)T_{\min}, \dots, N\}} G^x_{L,n}, & \text{если } m = 1, \\ I^x\big(k^x(m), \nu^x(m-1)\big), & \text{если } m = 2, \dots, M^x, \end{cases}$$

*где* $M^x$ *определяется из условия*

$$M^x = \min\big\{m \in \mathcal{N} \;\big|\; r^x\big(k^x(m), \nu^x(m)\big) = 0\big\};$$

$$J^x(l) = \begin{cases} 0, & \text{если } l = 0, \\ \big|\big\{m \in \{1, \dots, M^x\} \;\big|\; k^x(m) \le l\big\}\big|, & \text{если } l = 1, \dots, L. \end{cases}$$

*Тогда наборы* $\mathcal{M}^x_1, \dots, \mathcal{M}^x_L$ *определяются по правилу*

$$\mathcal{M}^x_l = \big\{n \in \mathcal{N} \;\big|\; n = \nu^x(M^x - m + 1),\ m = J^x(l-1) + 1, \dots, J^x(l)\big\} \qquad (3.33)$$

*при каждом* $l = 1, \ldots, L$.

Мощности оптимальных наборов $\mathcal{M}_l^x$, $l = 1, \ldots, L$, и мощность $M^x$ объединённого набора $\mathcal{M}^x = \bigcup\limits_{l=1}^{L} \mathcal{M}_l^x$ определяются формулами следствия 3.4.

Запишем алгоритм, реализующий приведённую схему, в пошаговом виде.

**Алгоритм $\mathcal{A}_4'$.**

*Вход алгоритма*: последовательность $\mathcal{Y}$, набор $x = (x_1, \ldots, x_L)$ точек, натуральные числа $T_{\min}$, $T_{\max}$.

Шаг 1. Вычислим значения $g_l^x(n)$ для $l = 1, \ldots, L$, $n = 1+(l-1)T_{\min}, \ldots, N-(L-l)T_{\min}$ по формуле (3.24).

Шаг 2. Используя рекуррентные формулы (3.31) и (3.32), вычислим значения $G_{l,n}^x$ для каждого $l = 1, \ldots, L$, $n = 1 + (l-1)T_{\min}, \ldots, N - (L-l)T_{\min}$.

Шаг 3. Найдём значение $G_{\max}^x$ максимума целевой функции $G^x$ по формуле (3.30) и оптимальные наборы $\mathcal{M}_l^x$ по формуле (3.33).

*Выход алгоритма*: семейство $\{\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x\}$ наборов.

**Замечание 3.4.** *В* [59] *установлено, что алгоритм* $\mathcal{A}_4'$ *находит оптимальное решение задачи* $4'$ *за время* $\mathcal{O}\big(LN(T_{\max} - T_{\min} + d)\big)$. *В этом выражении значение* $T_{\max} - T_{\min} + 1$ *не превосходит* $N$. *Поэтому время работы алгоритма оценивается величиной* $\mathcal{O}\big(LN(N + d)\big)$.

### 3.2.3 2-приближённый алгоритм для задачи с ограничениями на мощности кластеров

Суть подхода к поиску решения задачи 4 заключается в следующем. Для каждого упорядоченного набора, содержащего $L$ элементов последовательности $\mathcal{Y}$, находим точное решение вспомогательной задачи $3'$ — семейство наборов

номеров элементов последовательности, которое является допустимым решением исходной задачи 4. Найденное семейство наборов объявляется претендентом на решение исходной задачи и включается в множество допустимых решений. В качестве окончательного решения из построенного множества выбирается семейство наборов, доставляющее наибольшее значение целевой функции задачи $3'$.

Сформулируем алгоритм решения задачи 4, реализующий описанный подход. В приведенной ниже пошаговой записи предполагается, что входные натуральные числа заданы в соответствии с ограничениями задачи и условиями леммы 3.6 (см. замечание 3.2).

**Алгоритм $\mathcal{A}_8$.**

*Вход алгоритма:* последовательность $\mathcal{Y}$, натуральные числа $T_{\min}$, $T_{\max}$, $M$ и $L$.

Шаг 1. Для каждого набора $x = (x_1, \ldots, x_L) \in \mathcal{Y}^L$, сформированного из элементов последовательности $\mathcal{Y}$, с помощью алгоритма $\mathcal{A}'_3$ найдём оптимальное решение $\{\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x\}$ задачи $3'$.

Шаг 2. В множестве $\bigl\{\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x \bigm| x \in \mathcal{Y}^L\bigr\}$ в качестве решения выберем то семейство $\{\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A\} = \{\mathcal{M}_1^{x(A)}, \ldots, \mathcal{M}_L^{x(A)}\}$ наборов, для которого функция $G^x(\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x)$ имеет наибольшее значение. Если наибольшему значению соответствует несколько семейств, то выберем любое из них.

*Выход алгоритма:* семейство $\{\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A\}$ наборов.

Свойства изложенного алгоритмического решения устанавливает

**Теорема 3.5.** *Алгоритм $\mathcal{A}_8$ находит 2-приближённое решение задачи 4 за время $\mathcal{O}\bigl(LN^{L+1}\bigl(M(T_{\max} - T_{\min} + 1) + d\bigr)\bigr)$. Оценка 2 точности алгоритма достижима.*

*Доказательство.* Пусть $\{\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*\}$ — оптимальное решение задачи 4, а $\{\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A\}$ — решение, полученное в результате работы алгоритма $\mathcal{A}_8$.

Оптимальному решению $\{\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*\}$ задачи 4 соответствует набор $\{\overline{y}(\mathcal{M}_1^*), \ldots, \overline{y}(\mathcal{M}_L^*)\}$ центроидов. Рассмотрим точку $t_l = \arg\min\limits_{j \in \mathcal{M}_l^*} \|y_j - \overline{y}(\mathcal{M}_l^*)\|$, $l = 1, \ldots, L$, из мультимножества $\{y_i \mid i \in \mathcal{M}_l^*\}$, ближайшую к центроиду этого мультимножества. Эта точка и само мультимножество $\{y_i \mid i \in \mathcal{M}_l^*\}$ удовлетворяют условиям леммы 2.2. Поэтому, применяя неравенство леммы 2.2 к каждому из мультимножеств $\{y_i \mid i \in \mathcal{M}_l^*\}$, $l = 1, \ldots, L$, найдём оценку для величины

$$
\begin{aligned}
W(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*, t_1, \ldots, t_L) &= \sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l^*} \|y_j - t_l\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}^*} \|y_i\|^2 \\
&\leq 2 \sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l^*} \|y_j - \overline{y}(\mathcal{M}_l^*)\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}^*} \|y_i\|^2 \\
&\leq 2 \sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l^*} \|y_j - \overline{y}(\mathcal{M}_l^*)\|^2 + 2 \sum_{i \in \mathcal{N} \setminus \mathcal{M}^*} \|y_i\|^2 = 2F(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*), \quad (3.34)
\end{aligned}
$$

где $\mathcal{M}^* = \bigcup\limits_{l=1}^{L} \mathcal{M}_l^*$.

С другой стороны, заметим, что набор $t = (t_1, \ldots, t_L)$ точек, ближайших к центроидам $\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*$, является одним из наборов $(x_1, \ldots, x_L) \in \mathcal{Y}^L$, рассмотренных на шаге 1 алгоритма $\mathcal{A}_8$. Пусть $\{\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t\}$ — оптимальное решение задачи $3'$ при $x = t$, полученное на шаге 1 алгоритма $\mathcal{A}_8$. Тогда в соответствии с утверждением 2 леммы 3.5, т.е. согласно (3.25), семейство $\{\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t\}$ доставляет минимум функции $W^x(\mathcal{M}_1, \ldots, \mathcal{M}_L)$ при $x = t$. Поэтому

$$
W(\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t, t_1, \ldots, t_L) \leq W(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*, t_1, \ldots, t_L). \quad (3.35)
$$

Далее, по определению шага 2 в соответствии с (3.23) справедлива оценка

$$W(\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A, x_1^A, \ldots, x_L^A) \leq W(\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t, t_1, \ldots, t_L), \tag{3.36}$$

где $(x_1^A, \ldots, x_L^A) = x(A)$. Кроме того, из первого утверждения леммы 3.5 следует неравенство

$$F(\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A) \leq W(\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A, x_1^A, \ldots, x_L^A). \tag{3.37}$$

Наконец, объединяя (3.34)–(3.37), получим цепочку оценочных неравенств

$$\begin{aligned} F(\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A) \leq W(\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A, x_1^A, \ldots, x_L^A) \leq W(\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t, t_1, \ldots, t_L) \\ \leq W(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*, t_1, \ldots, t_L) \leq 2F(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*), \end{aligned}$$

которая завершает доказательство оценки точности алгоритма.

Оценка трудоёмкости алгоритма следует из того, что на шаге 1 для каждого из $N^L$ наборов $(x_1, \ldots, x_L) \in \mathcal{Y}^L$ алгоритм $\mathcal{A}_3'$ находит оптимальное решение задачи $3'$ за время $\mathcal{O}\big(LN\big(M(T_{\max} - T_{\min} + 1) + d\big)\big)$, а на шаге 2 выбор наименьшего элемента осуществляется за $\mathcal{O}(N^L)$ операций. Достижимость оценки точности алгоритма $\mathcal{A}_8$ следует из достижимости оценки точности 2-приближённого алгоритма для частного случая (когда $L = 1$) задачи 4 (см. [56]).

**Замечание 3.5.** *Согласно замечанию* 3.3*, время работы алгоритма* $\mathcal{A}_8$ *оценивается величиной* $\mathcal{O}\big(LN^{L+1}(MN + d)\big)$*; при фиксированном* $L$ *алгоритм полиномиален.*

### 3.2.4 2-приближённый алгоритм для задачи с оптимизируемыми мощностями кластеров

Суть подхода к поиску решения задачи 5 заключается в следующем. Для каждого упорядоченного набора, содержащего $L$ элементов последовательности $\mathcal{Y}$, находим точное решение вспомогательной задачи $4'$ — семейство наборов номеров элементов последовательности, которое является допустимым решением исходной задачи 5. Найденное семейство наборов объявляется претендентом на решение исходной задачи и включается в множество её допустимых решений. В качестве окончательного решения из построенного множества выбирается семейство наборов, доставляющее наименьшее значение целевой функции задачи 5.

Сформулируем алгоритм решения задачи 5, реализующий приведённый подход.

**Алгоритм $\mathcal{A}_9$.**

*Вход алгоритма*: последовательность $\mathcal{Y}$, натуральные числа $T_{\min}$, $T_{\max}$ и $L$.

Шаг 1. Для каждого набора $x = (x_1, \ldots, x_L) \in \mathcal{Y}^L$, сформированного из элементов последовательности $\mathcal{Y}$, с помощью алгоритма $\mathcal{A}'_4$ найдём оптимальное решение $\{\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x\}$ задачи $4'$.

Шаг 2. Среди решений, найденных на шаге 2, выберем то семейство $\{\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A\}$ наборов, для которого значение $F(\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A)$ минимально. Если минимальному значению соответствует несколько семейств, то выберем любое из них.

*Выход алгоритма:* семейство $\{\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A\}$ наборов.

Согласно пошаговой записи алгоритм $\mathcal{A}_9$ находит решение в виде

$$\{\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A\} = \arg\min_{x \in \mathcal{Y}^L} F(\mathcal{M}_1^x, \ldots, \mathcal{M}_L^x). \tag{3.38}$$

**Теорема 3.6.** *Алгоритм $\mathcal{A}_9$ находит 2-приближённое решение задачи 5 за время $\mathcal{O}\big(LN^{L+1}(T_{\max} - T_{\min} + d)\big)$.*

*Доказательство.* Частично доказательство повторяет доказательство теоремы 3.5 и приводится здесь для удобства восприятия.

Пусть $\{\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*\}$ — оптимальное решение задачи 5, а $\{\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A\}$ — решение, полученное в результате работы алгоритма $\mathcal{A}_9$.

Оптимальному решению $\big\{\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*\big\}$ задачи 5 соответствует набор $\big\{\overline{y}(\mathcal{M}_1^*), \ldots, \overline{y}(\mathcal{M}_L^*)\big\}$ центроидов. Рассмотрим точку $t_l = \arg\min\limits_{i \in \mathcal{M}_l^*} \|y_i - \overline{y}(\mathcal{M}_l^*)\|$, $l = 1, \ldots, L$, из мультимножества $\big\{y_i \;\big|\; i \in \mathcal{M}_l^*\big\}$, ближайшую к центроиду этого мультимножества. Эта точка и само мультимножество $\big\{y_i \;\big|\; i \in \mathcal{M}_l^*\big\}$ удовлетворяют условиям леммы 2.2. Поэтому, применяя неравенство леммы 2.2 к каждому из мультимножеств $\big\{y_i \;\big|\; i \in \mathcal{M}_l^*\big\}$, найдём оценку для величины

$$\begin{aligned} W(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*, t_1, \ldots, t_L) &= \sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l^*} \|y_j - t_l\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}^*} \|y_i\|^2 \\ &\leq 2 \sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l^*} \|y_j - \overline{y}(\mathcal{M}_l^*)\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}^*} \|y_i\|^2 \\ &\leq 2 \sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l^*} \|y_j - \overline{y}(\mathcal{M}_l^*)\|^2 + 2 \sum_{i \in \mathcal{N} \setminus \mathcal{M}^*} \|y_i\|^2 = 2F(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*), \end{aligned} \tag{3.39}$$

где $\mathcal{M}^* = \bigcup\limits_{l=1}^{L} \mathcal{M}_l^*$.

С другой стороны, заметим, что набор $t = (t_1, \ldots, t_L)$ точек, ближайших к центроидам $\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*$, является одним из наборов $(x_1, \ldots, x_L) \in \mathcal{Y}^L$, рассмот-

ренных на шаге 1 алгоритма $\mathcal{A}_9$. Пусть $\{\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t\}$ — оптимальное решение задачи $4'$ при $x = t$, полученное на шаге 1 алгоритма $\mathcal{A}_9$. Тогда в соответствии с утверждением 2 леммы 3.5, т.е. согласно (3.25), семейство $\{\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t\}$ доставляет минимум функции $W^x(\mathcal{M}_1, \ldots, \mathcal{M}_L)$ при $x = t$. Поэтому

$$W(\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t, t_1, \ldots, t_L) \le W(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*, t_1, \ldots, t_L). \tag{3.40}$$

Далее, по определению шага 2 в соответствии с (3.38) справедлива оценка

$$F(\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A) \le F(\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t). \tag{3.41}$$

Кроме того, из первого утверждения леммы 3.5 следует неравенство

$$F(\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t) \le W(\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t, t_1, \ldots, t_L). \tag{3.42}$$

Наконец, объединяя (3.39)–(3.42), получим цепочку оценочных неравенств

$$\begin{aligned} F(\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A) \le F(\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t) &\le W(\mathcal{M}_1^t, \ldots, \mathcal{M}_L^t, t_1, \ldots, t_L) \\ &\le W(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*, t_1, \ldots, t_L) \le 2F(\mathcal{M}_1^*, \ldots, \mathcal{M}_L^*), \end{aligned}$$

которая завершает доказательство оценки точности алгоритма.

Оценка трудоёмкости алгоритма следует из того, что на шаге 1 для каждого из $N^L$ наборов $(x_1, \ldots, x_L) \in \mathcal{Y}^L$ алгоритм $\mathcal{A}_4'$ находит оптимальное решение задачи $4'$ за время $\mathcal{O}\big(LN(T_{\max} - T_{\min} + d)\big)$, а на шаге 2 вычисление значений $F(\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A)$ для каждого из $N^L$ семейств $\{\mathcal{M}_1^A, \ldots, \mathcal{M}_L^A\}$ осуществляется за $\mathcal{O}(dN)$ операций, выбор наименьшего элемента — за $\mathcal{O}(N^L)$ операций. Теорема доказана.

**Замечание 3.6.** *Согласно замечанию* 3.4*, время работы алгоритма* $\mathcal{A}_9$ *оцени-*

*вается величиной* $\mathcal{O}\big(LN^{L+1}(N+d)\big)$*; при фиксированном* $L$ *алгоритм полиномиален.*

# Заключение

В диссертации исследованы актуальные квадратичные задачи кластеризации конечных множества и последовательности точек евклидова пространства. Основные результаты диссертационной работы заключаются в следующем:

1. Для квадратичной задачи разбиения конечного множества точек евклидова пространства на два кластера при фиксированном центре одного из кластеров:

(а) построен 2-приближённый полиномиальный алгоритм для случая задачи без ограничений на мощности кластеров;

(б) предложен рандомизированный алгоритм (ориентированный на случай задачи с ограничениями на мощности кластеров), который при заданных относительной ошибке и вероятности несрабатывания для установленных значений параметров находит приближённое решение за полиномиальное время; найдены условия, при которых этот алгоритм асимптотически точен.

Алгоритм из п. 1 (а) имеет меньшую трудоёмкость по сравнению с лучшим из известных алгоритмов при той же, как и у известного алгоритма, точности. Алгоритм из п. 1 (б) является первым алгоритмом рандомизированного типа, предложенным для задачи из этого пункта.

2. Для квадратичных евклидовых задач 2-кластеризации конечных множества и последовательности (с ограничениями на выбор элементов, входящих в кластеры) при фиксированном центре одного из кластеров и дополнительном

ограничении на мощности кластеров:

(а) построены точные алгоритмы для случая целочисленных входов задачи; при фиксированной размерности пространства алгоритмы псевдополиномиальны;

(б) показано, что для общих случаев задач не существует полностью полиномиальных приближённых схем (FPTAS), если P $\neq$ NP; такие схемы построены для случаев задач, в которых размерность пространства фиксирована.

Алгоритм кластеризации множества из п. 2 (а) является новым решением задачи, послужившим важным промежуточным результатом, на котором основана идея построения оригинальных аппроксимационных схем из п. 2 (б). Алгоритм разбиения последовательности из п. 2 (а) построен впервые. Факт несуществования схемы FPTAS для общих случаев задач из п. 2 также установлен впервые; результаты по построению приближённых схем для указанных в п. 2 (б) случаев задач приоритетны.

3. Для квадратичной евклидовой задачи многокластерного разбиения конечной последовательности точек с ограничениями на выбор внутрикластерных элементов при фиксированном центре одного из кластеров построены 2-приближённые алгоритмы, ориентированные как на случай задачи без ограничений на мощности кластеров, так и на случай с ограничениями; алгоритмы полиномиальны при фиксированном числе кластеров.

На настоящее время результаты п. 3 являются единственными алгоритмами с гарантированными оценками точности, предложенными для задач из этого пункта.

Несмотря на то, что в диссертации получен ряд результатов для рассматриваемых труднорешаемых задач, эти задачи по-прежнему остаются слабоизученными в алгоритмическом плане. Поэтому перспективным направлением дальнейших исследований представляется разработка более быстрых эффективных

алгоритмов с оценками качества, в том числе линейных и сублинейных, что особенно важно в плане решения проблемы Big Data.

# Литература

[1] Steinhaus H. Sur la division des corp materiels en parties // Bull. Acad. Polon. Sci. — 1956. — Vol. 3, № 12. — P. 801–804.

[2] Lloyd S. Least squares quantization in PCM // IEEE Trans. Inform. Theory. — 1982. — Vol. 28, № 2. — P. 129–137.

[3] Ball G., Hall D. ISODATA, a novel method of data analysis and pattern classification // Technical report NTIS AD 699616. Stanford Research Institute, Stanford, CA. 1965.

[4] MacQueen J. Some methods for classification and analysis of multivariate observations // Proc. 5-th Berkeley Symp. on Mathematics, Statistics and Probability. — 1967. — Vol. 1. — P. 281–297.

[5] Jain A.K. Data clustering: 50 years beyond k-means // Pattern Recognit. Lett. — 2010. — Vol. 31, № 8. — P. 651–666.

[6] Fisher W.D. On Grouping for Maximum Homogeneity // J. Amer. Statist. Assoc. — 1958. — Vol. 53, № 284. — P. 789–798.

[7] Aloise D., Deshpande A., Hansen P., Popat P. NP-hardness of Euclidean sum-of-squares clustering // Mach. Learn. — 2009. — Vol. 75, № 2. — P. 245–248.

[8] Rao M.R. Cluster Analysis and Mathematical Programming // J. Amer. Statist. Assoc. — 1971. — Vol. 66. — P. 622–626.

[9] Inaba M., Katoh N., Imai H. Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering // Proc. 10-th Annual Symp. on Comput. geometry. — 1994. — P. 332-339.

[10] Matousek J. On approximate geometric k-clustering // Discrete Comput. Geom. — 2000. — Vol. 24, № 1. — P. 61–84.

[11] Har-Peled S., Mazumdar S. On coresets for k-means and k-median clustering // Proc. 36-th Annual ACM Symp. on Theory of computing (STOC'04). — 2004. — P. 291–300.

[12] Badoiu M., Har-Peled S., Indyk P. Approximate clustering via Coresets // Proc. 34-th Annual ACM Symp. on Theory of Computing (STOC'02). — 2002. — P. 250–257.

[13] De La Vega W.F., Karpinski M., Kenyon C., Rabani Y. Approximation schemes for clustering problems // Proc. 35-th Annual ACM Symp. on Theory of Computing (STOC'03). — 2003. — P. 50–58.

[14] Kumar A., Sabharwal Y., Sen S. A simple linear time $(1 + \varepsilon)$-approximation algorithm for K-means clustering in any dimensions // Proc. 45-th Annual IEEE Symp. on Foundations of Computer Science (FOCS'04). — 2004. — P. 454–462.

[15] Kumar A., Sabharwal Y., Sen S. Linear-time approximation schemes for clustering problems in any dimensions // J. ACM. — 2010. — Vol. 57, № 2. — P. 1–32.

[16] Kanungo T., Mount D., Netanyahu N.S, Piatko C.D., Silverman R., Wu A.Y. A local search approximation algorithm for k-means clustering // Comput. Geom. — 2004. — Vol. 28, № 2–3. — P. 89–112.

[17] Song M., Rajasekaran S. Fast Algorithms for Constant Approximation k-Means Clustering // Lect. Notes Comput. Sci. — 2005. — Vol. 3827. — P. 1029–1038.

[18] Awasthi P., Charikar M., Krishnaswamy R., Sinop A.K. The hardness of approximation of euclidean k-means // Proc. 31-st Int. Symp. on Computational Geometry (SoCG'15). — 2015. — P. 754–767.

[19] Lee E., Schmidt M., Wright J. Improved and simplified inapproximability for k-means // Inform. Process. Lett. — 2017. — Vol. 120. — P. 40–43.

[20] Mahajan M., Nimbhorkar P., Varadarajan K. The planar k-means problem is NP-hard // Theor. Comput. Sci. — 2012. — Vol. 442. — P. 13–21.

[21] Aloise D., Hansen, P. On the Complexity of Minimum Sum-of-Squares Clustering // Les Cahiers du GERAD. — 2007. — Vol. G-2007-50.

[22] Ostrovsky R., Rabani Y. Polynomial-time approximation schemes for geometric min-sum median clustering // J. ACM. — 2002. — Vol. 49, № 2. — P. 139–156.

[23] Feldman D., Monemizadeh M., Sohler C. A PTAS for k-means clustering based on weak coresets // Proc. 23-rd Annual Symp. on Computational geometry (SoCG'07). — 2007. — P. 11–18.

[24] Har-Peled S., Kushal A. Smaller coresets for k-median and k-means clustering // Discrete Comput. Geom. — 2007. — Vol. 37, № 1. — P. 3–19.

[25] Гимади Э.Х., Кельманов А.В., Кельманова М.А., Хамидуллин С.А. Апостериорное обнаружение в числовой последовательности квазипериодического фрагмента при заданном числе повторов // Сиб. журн. индустр. математики. — 2006. — Т. 9, № 1(25). — С. 55–74.

[26] 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 Recognit. Image Anal. — 2008. — Vol. 18, № 1. — P. 30—42.

[27] Бабурин А.Е., Гимади Э.Х., Глебов Н.И., Пяткин А.В. Задача отыскания подмножества векторов с максимальным суммарным весом // Дискрет. анализ и исслед. опер. — 2007. — Т. 14, № 1. — С. 32–42.

[28] Кельманов А.В., Пяткин А.В. О сложности одного из вариантов задачи выбора подмножества «похожих» векторов // Доклады РАН. — 2008. — Т. 421, № 5. — С. 590–592.

[29] Кельманов А.В., Пяткин А.В. Об одном варианте задачи выбора подмножества векторов // Дискрет. анализ и исслед. опер. — 2008. — Т. 15, № 5. — С. 20–34.

[30] De Waal T., Pannekoek J., Scholtus S. Handbook of statistical data editing and imputation. — Hoboken, New Jersey: John Wiley and Sons, Inc., 2011. — 456 p.

[31] Osborne J.W. Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data. — Los Angeles: SAGE Publication, Inc., 2013. — 296 p.

[32] Greco L., Alessio F. Robust methods for data reduction. — London: Chapman and Hall/CRC, 2015. — 297 p.

[33] Bishop C.M. Pattern recognition and machine learning. — New York: Springer-Verlag, 2006. — 738 p.

[34] James G., Witten D., Hastie T., Tibshirani R. An introduction to statistical learning. — New York: Springer-Verlag, 2013. — 426 p.

[35] Hastie T., Tibshirani R. and Friedman J. The elements of statistical learning. — New York: Springer-Verlag, 2009. — 763 p.

[36] Aggarwal C.C. Data mining: The textbook. — Springer International Publishing, 2015. — 734 p.

[37] Fu T. A review on time series data mining // Engineering Applications of Artificial Intelligence. — 2011. — Vol. 24, № 1. — P. 164–181.

[38] Kuenzer C., Dech S., Wagner W. Remote sensing time series. Remote Sensing and Digital Image Processing, Vol. 22. — Springer International Publishing, 2015.

[39] Liao T.W. Clustering of time series data — a survey // Pattern Recognit. — 2005. — Vol. 38, № 11. — P. 1857–1874.

[40] Кельманов А.В. Проблема off-line обнаружения повторяющегося фрагмента в числовой последовательности // Тр. Ин-та математики и механики УрО РАН. — 2008. — Т. 14, № 2. — С. 81–88.

[41] Кельманов А.В., Пяткин А.В. О сложности некоторых задач поиска подмножеств векторов и кластерного анализа // Журн. вычисл. математики и мат. физики. — 2009. — Т. 49, № 11. — С. 2059–2065.

[42] Кельманов А.В., Пяткин А.В. О сложности некоторых задач кластерного анализа векторных последовательностей // Дискрет. анализ и исслед. опер. — 2013. — Т. 20, № 2. — С. 47–57.

[43] Гимади Э.Х., Пяткин А.В., Рыков И.А. О полиномиальной разрешимости некоторых задач выбора подмножеств векторов в евклидовом пространстве фиксированной размерности // Дискрет. анализ и исслед. опер. — 2008. — Т. 15, № 6. — С. 11–19.

[44] Шенмайер В.В. Решение некоторых задач поиска подмножества векторов с использованием диаграмм Вороного // Дискрет. анализ и исслед. опер. — 2016. — Т. 23, № 4. — С. 102–115.

[45] Долгушев А.В., Кельманов А.В. Приближенный алгоритм решения одной задачи кластерного анализа // Дискрет. анализ и исслед. опер. — 2011. — Т. 18, № 2. — С. 29–40.

[46] Кельманов А.В., Романченко С.М. FPTAS для одной задачи поиска подмножества векторов // Дискрет. анализ и исслед. опер. — 2014. — Т. 21, № 3. — С. 41–52.

[47] Гимади Э.Х., Глазков Ю.В., Рыков И.А. О двух задачах выбора подмножества векторов с целочисленными координатами в евклидовом пространстве с максимальной нормой суммы размерности // Дискрет. анализ и исслед. опер. — 2008. — Т. 15, № 4. — С. 30—43.

[48] Гимади Э.Х., Рыков И.А. Рандомизированный алгоритм отыскания подмножества векторов с максимальной евклидовой нормой их суммы // Дискретн. анализ и исслед. опер. — 2015. — Т. 22, № 3. — С. 5–17.

[49] Gimadi E., Rykov I. Efficient randomized algorithm for a vector subset problem // Lect. Notes Comput. Sci. — 2016. — Vol. 9869. — P. 148–158.

[50] Долгушев А.В., Кельманов А.В., Шенмайер В.В. Приближенная полиномиальная схема для одной задачи кластерного анализа // Интеллектуализация обработки информации: 9-я международная конференция. Сборник докладов. (Республика Черногория, г. Будва, 16–22 сентября, 2012). — М.: Торус Пресс, 2012. — С. 242-244.

[51] Долгушев А.В., Кельманов А.В., Шенмайер В.В. Полиномиальная аппроксимационная схема для одной задачи разбиения конечного множества на два кластера // Тр. Ин-та математики и механики УрО РАН. — 2015. — Т. 21, № 3. — С. 100–109.

[52] Wirth I. Algorithms + data structures = programs. — New Jersey: Prentice Hall, 1976. — 366 p.

[53] Марков А.А. Исчисление вероятностей. — СПб: Тип. Имп. Акад. наук, 1900. — 382 стр.

[54] Motwani R., Raghavan P. Randomized algorithms. — New York: Cambridge University Press, 1995. — 476 p.

[55] Vazirani V.V. Approximation Algorithms. — Berlin, Heidelberg, New York: Springer-Verlag, 2003. — 380 p.

[56] Кельманов А.В., Хамидуллин С.А. Приближённый полиномиальный алгоритм для одной задачи разбиения последовательности // Дискрет. анализ и исслед. опер. — 2014. — Т. 21, № 1. — С. 53–66.

[57] Кельманов А.В., Хамидуллин С.А. Апостериорное обнаружение заданного числа одинаковых подпоследовательностей в квазипериодической последовательности // Журн. вычисл. математики и мат. физики. — 2001. — Т. 41, № 5. — С. 807–820.

[58] Кельманов А.В., Михайлова Л.В. Совместное обнаружение в квазипериодической последовательности заданного числа фрагментов из эталонного набора и ее разбиение на участки, включающие серии одинаковых фрагментов // Журн. вычисл. математики и мат. физики. — 2006. — Т. 46, № 1. — С. 172–189.

[59] Кельманов А.В., Михайлова Л.В. Апостериорное обнаружение квазипериодических фрагментов из эталонного набора в числовой последовательности и ее разбиение на участки, включающие серии одинаковых фрагментов // Журн. вычисл. математики и мат. физики. — 2008. — Т. 48, № 5. — С. 899–915.

## Публикации автора по теме диссертации

Статьи в журналах

[60] Кельманов А.В., Хандеев В.И. Полиномиальный алгоритм с оценкой точности 2 для решения одной задачи кластерного анализа // Дискрет. анализ и исслед. опер. — 2013. — Т. 20, № 4. — С. 36–45. РИНЦ, RSCI.
Kel'manov A.V., Khandeev V.I. A 2-Approximation Polynomial Algorithm for a Clustering Problem // J. Appl. Ind. Math. — 2013. — Vol. 7, № 4. — P. 515–521. Scopus. DOI: 10.1134/S1990478913040066.

[61] Кельманов А.В., Хандеев В.И. Рандомизированный алгоритм для одной задачи двухкластерного разбиения множества векторов // Журн. вычисл. математики и мат. физики. — 2015. — Т. 55, № 2. — С. 335–344. РИНЦ, RSCI. DOI: 10.7868/S0044466915020131.
Kel'manov A.V., Khandeev V.I. A Randomized Algorithm for Two-Cluster Partition of a Set of Vectors // Comput. Math. Math. Phys. — 2015. — Vol. 55, № 2. — P. 330–339. Scopus, WoS. DOI: 10.1134/S096554251502013X.

[62] Кельманов А.В., Хандеев В.И. Точный псевдополиномиальный алгоритм для одной задачи двухкластерного разбиения множества векторов // Дискрет. анализ и исслед. опер. — 2015. — Т. 22, № 4. — С. 50–62. РИНЦ, RSCI. DOI: 10.17377/daio.2015.22.463.

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. Ind. Math. — 2015. — Vol. 9, № 4. — P. 497–502. Scopus. DOI: 10.1134/S1990478915040067.

[63] Кельманов А.В., Хандеев В.И. Полностью полиномиальная аппроксимационная схема для специального случая одной квадратичной евклидовой задачи 2-кластеризации // Журн. вычисл. математики и мат. физики. — 2016. — Т. 56, № 2. — С. 332–340. РИНЦ, RSCI. DOI: 10.7868/S0044466916020113.
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. — 2016. — Vol. 56, № 2. — P. 334–341. Scopus, WoS. DOI: 10.1134/S0965542516020111.

[64] Кельманов А.В., Хамидуллин С.А., Хандеев В.И. Точный псевдополиномиальный алгоритм для одной задачи разбиения последовательности // Автоматика и телемеханика. — 2017. — № 1. — С. 80–90. РИНЦ, RSCI.
Kel'manov A.V., Khamidullin S.A., Khandeev V.I. Exact Pseudopolynomial Algorithm for one Sequence Partitioning Problem // Automation and Remote Control. — 2017. — Vol. 78, № 1. — P. 66–73. Scopus, WoS. DOI: 10.1134/S0005117917010052.

[65] Кельманов А.В., Хамидуллин С.А., Хандеев В.И. Полностью полиномиальная аппроксимационная схема для одной задачи двухкластерного разбиения последовательности // Дискрет. анализ и исслед. опер. — 2016. — Т. 23, № 2. — С. 21–40. РИНЦ, RSCI. DOI: 10.17377/daio.2016.23.511.
Kel'manov A.V., Khamidullin S.A., Khandeev V.I. A Fully Polynomial-Time Approximation Scheme for a Sequence 2-Cluster Partitioning Problem // J.

Appl. Ind. Math. — 2016. — Vol. 10, № 2. — P. 209–219. Scopus. DOI: 10.1134/S199047891602006X.

[66] Кельманов А.В., Михайлова Л.В., Хамидуллин С.А., Хандеев В.И. Приближенный алгоритм для задачи разбиения последовательности на кластеры с ограничениями на их мощность // Тр. Ин-та математики и механики УрО РАН. — 2016. — Т. 22, № 3. — С. 144–152. РИНЦ, RSCI. DOI: 10.21538/0134-4889-2016-22-3-144-152.

[67] Kel'manov A.V., Mikhailova L.V., Khamidullin S.A., Khandeev V.I. An Approximation Algorithm for a Problem of Partitioning a Sequence into Clusters with Restrictions on Their Cardinalities // Lect. Notes Comput. Sci. — 2016. — Vol. 9869. — P. 171–181. Scopus. DOI: 10.1007/978-3-319-44914-2_14.

[68] Кельманов А.В., Михайлова Л.В., Хамидуллин С.А., Хандеев В.И. Приближенный алгоритм для задачи разбиения последовательности на кластеры // Журн. вычисл. математики и мат. физики. — 2017. Т. 57, № 8. — С. 149–157. RSCI (ядро РИНЦ). DOI: 10.7868/S0044466917080087.

Тезисы докладов

[69] Кельманов А.В., Хандеев В.И. 2-Приближенный полиномиальный алгоритм для одной задачи кластерного анализа // Материалы V Всероссийской конференции «Проблемы оптимизации и экономические приложения». (Омск, 2–6 июля, 2012). — 2012. — С. 135.

[70] Кельманов А.В., Хандеев В.И. Полиномиальный алгоритм с оценкой точности 2 для решения одной задачи кластерного анализа // Интеллектуализация обработки информации: 9-я международная конференция. Сборник

докладов. (Республика Черногория, г. Будва, 16–22 сентября, 2012). — М.: Торус Пресс. — 2012. — С. 279–282.

[71] Кельманов А.В., Хандеев В.И. Рандомизированный алгоритм для одной задачи кластерного анализа // Материалы международной конференции «Дискретная оптимизация и исследование операций» (DOOR-2013). (Новосибирск, Академгородок, 24–28 июня, 2013). — Новосибирск: Изд-во Института математики СО РАН. — 2013. — С. 160.

[72] Кельманов А.В., Хандеев В.И. Рандомизированный алгоритм для одной NP-трудной задачи кластерного анализа // Математические методы распознавания образов: 16-я Всероссийская конференция (ММРО-16). Тезисы докладов. (г. Казань, 6–12 сентября, 2013). — М.: Торус Пресс. — 2013. — С. 35.

[73] Kel'manov A.V., Khandeev V.I. A randomized algorithm for a clustering problem // Proceedings of IV International Conference «Optimization and applications» (OPTIMA-2013). (Petrovac, Montenegro, September 22–28, 2013). — 2013. — P. 86.

[74] Kel'manov A.V., Khandeev V.I. An exact pseudopolynomial algorithm for a two-cluster partitioning problem // Abstracts of the 16th Baikal International School-Seminar «Methods of Optimization and Their Applications». (Olkhon Island, Baikal, June 30 – July 6, 2014). — Irkutsk: Melentiev Energy Systems Institute SB RAS. — 2014. — P. 51.

[75] Kel'manov A.V., Khandeev V.I. An exact pseudopolynomial algorithm for a bi-partitioning problem // Abstracts of the V International Conference «Optimization and Applications» (OPTIMA-2014). (Petrovac, Montenegro, September 28 – October 4, 2014). — 2014. — P. 108–109.

[76] Kel'manov A.V., Khandeev V.I. An exact pseudopolynomial algorithm for a vectors set bi-partitioning problem // Abstracts of the 10th International Conference «Intelligent Information Processing» (IIP-2014). (Greece, Crete, October 4–11, 2014). — 2014. — P. 94–95.

[77] Кельманов А.В., Хамидуллин С.А., Хандеев В.И. Точный псевдополиномиальный алгоритм для одной задачи бикластеризации последовательности // XV всероссийская конференция «Математическое программирование и приложения». Тезисы докладов. (г. Екатеринбург, 2–6 марта, 2015). — Издательство Института математики и механики УрО РАН. — 2015. — С. 139–140.

[78] Кельманов А.В., Хандеев В.И. FPTAS для одной задачи двухкластерного разбиения множества векторов // XV всероссийская конференция «Математическое программирование и приложения». Тезисы докладов. (г. Екатеринбург, 2–6 марта, 2015). — Издательство Института математики и механики УрО РАН. — 2015. — С. 141–142.

[79] Kel'manov A.V., Khandeev V.I. FPTAS for special case of a quadratic Euclidean bi-partitioning problem // Abstract of the 28th Conference of the European Chapter on Combinatorial Optimization (ECCO XXVIII - 2015). (Italy, Catania, May 28–30, 2015). — 2015. — P. 30.

[80] Кельманов А.В., Хамидуллин С.А., Хандеев В.И. FPTAS для специального случая одной квадратичной евклидовой задачи би-кластеризации последовательности // Материалы VI международной конференции «Проблемы оптимизации и экономические приложения». (Омск, 28 июня – 4 июля, 2015). — 2015. — С. 138.

[81] Кельманов А.В., Хамидуллин С.А., Хандеев В.И. Полностью полиноми-

альная приближенная схема для одной задачи 2-кластеризации последовательности // Математические методы распознавания образов: 17-я Всероссийская конференция (ММРО-17). Тезисы докладов. (г. Светлогорск, 19–25 сентября, 2015). — М.: Торус Пресс. — 2015. — С. 104–105.

[82] Кельманов А.В., Хандеев В.И. Полностью полиномиальная приближенная схема для одной квадратичной задачи 2-кластеризации // Математические методы распознавания образов: 17-я Всероссийская конференция (ММРО-17). Тезисы докладов. (г. Светлогорск, 19–25 сентября, 2015) — М.: Торус Пресс. — 2015. — С. 106–107.

[83] Kel'manov A.V., Khandeev V.I. Fully polynomial-time approximation scheme for a special case of a quadratic Euclidean 2-clustering problem // Abstracts of the VI International Conference «Optimization and Applications» (OPTIMA-2015). (Petrovac, Montenegro, September 27 – October 3, 2015). — 2015. — P. 94–95.

[84] Kel'manov A.V., Khamidullin S.A., Khandeev V.I. Fully polynomial-time approximation scheme for a sequence 2-clustering problem // Abstracts of the VI International Conference «Optimization and Applications» (OPTIMA-2015). (Petrovac, Montenegro, September 27 – October 3, 2015). — 2015. — P. 96–97.

[85] Kel'manov A.V., Khamidullin S.A., Khandeev V.I., Mikhailova L.V. An approximation algorithm for one NP-hard problem of partitioning a sequence into clusters with restrictions on their cardinalities // Abstracts of the VII International Conference «Optimization and Applications» (OPTIMA-2016). (Petrovac, Montenegro, September 25 – October 3, 2016). — 2016. — P. 80–81.

[86] Kel'manov A.V., Khamidullin S.A., Khandeev V.I., Mikhailova L.V. An

approximation algorithm for a problem of partitioning a sequence into clusters // Abstracts of the VII International Conference «Optimization and Applications» (OPTIMA-2016). (Petrovac, Montenegro, September 25 – October 3, 2016). — 2016. — P. 82.

[87] Kel'manov A.V., Khamidullin S.A., Khandeev V.I., Mikhailova L.V. An approximation algorithm for one np-hard problem of partitioning a sequence into clusters with restrictions on their cardinalities // Book of abstracts of the 11th International Conference «Intelligent data Processing» (IDP-2016). (Barcelona, Spain, October 10–14, 2016). — 2016. — P. 72–73.

[88] Kel'manov A.V., Khamidullin S.A., Khandeev V.I., Mikhailova L.V. An approximation algorithm for a problem of partitioning a sequence into clusters // Book of abstracts of the 11th International Conference «Intelligent data Processing» (IDP-2016). (Barcelona, Spain, October 10–14, 2016). — 2016. — P. 74–75.