БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ЖУРНАЛ
БЕЛОРУССКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

МАТЕМАТИКА ИНФОРМАТИКА

JOURNAL
OF THE BELARUSIAN STATE UNIVERSITY

MATHEMATICS and INFORMATICS

Издается с января 1969 г.
(до 2017 г. – под названием «Вестник БГУ.
Серия 1, Физика. Математика. Информатика»)

Выходит три раза в год

3 2019

МИНСК
БГУ

РЕДАКЦИОННАЯ КОЛЛЕГИЯ

EDITORIAL BOARD

Вещественный, комплексный и функциональный анализ

Real, complex and functional analysis

УДК 517.9

РАСШИРЕНИЯ НЕЗАМЫКАЕМЫХ ОПЕРАТОРОВ И ЗАДАЧА УМНОЖЕНИЯ РАСПРЕДЕЛЕНИЙ

А. Б. АНТОНЕВИЧ[1)]***, Ч. ДОЛИЧАНИН***[2)]

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*
[2)]*Нови-Пазарский университет, Вука Караджича, бб, 36300, г. Нови Пазар, Сербия*

Предложены конструкции расширений незамыкаемых операторов и приведены примеры приложений этих конструкций. Исходным объектом является отображение f, заданное на множестве $D(f) \subset X$, при этом априорный выбор множества X есть дополнительное искусственно внесенное ограничение. Основная идея построений связана с тем, что $D(f)$ можно рассматривать как подмножество в некотором множестве, более широком, чем X, и тогда область определения расширения также лежит в этом более широком множестве. Частным случаем изучаемых вопросов является задача умножения распределений (обобщенных функций), для решения которой вводятся так называемые новые обобщенные функции. Показано, что сложность этой задачи определяется незамыкаемостью исходной операции и что построение новых обобщенных функций базируется на тех же идеях, что и построение расширений незамыкаемых операторов.

Ключевые слова: расширение оператора; замкнутый оператор; умножение распределений.

Образец цитирования:
Антоневич АБ, Доличанин Ч. Расширения незамыкаемых операторов и задача умножения распределений. *Журнал Белорусского государственного университета. Математика. Информатика.* 2019;3:6–17.
https://doi.org/10.33581/2520-6508-2019-3-6-17

For citation:
Antonevich AB, Dolicanin C. Extensions of nonclosable operators and multiplication of distributions. *Journal of the Belarusian State University. Mathematics and Informatics.* 2019;3:6–17. Russian.
https://doi.org/10.33581/2520-6508-2019-3-6-17

Авторы:
Анатолий Борисович Антоневич – доктор физико-математических наук, профессор; профессор кафедры функционального анализа и аналитической экономики механико-математического факультета.
Чемал Доличанин – доктор математических наук; профессор отделения математических наук.

Authors:
Anatolij B. Antonevich, doctor of science (physics and mathematics), full professor; professor at the department of functional analysis and analytical economics, faculty of mechanics and mathematics.
antonevich@bsu.by
http://orcid.org/0000-0002-2960-9640
Cemal Dolicanin, doctor of science (mathematics); professor of mathematics.
cdolicanin@np.ac.rs
http://orcid.org/0000-0003-4830-1454

EXTENSIONS OF NONCLOSABLE OPERATORS AND MULTIPLICATION OF DISTRIBUTIONS

A. B. ANTONEVICH [a], ***C. DOLICANIN*** [b]

[a] *Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*
[b] *State University of Novi Pazar, Vuka Karadzica, bb, Novi Pazar 36300, Serbia*
Corresponding author: A. B. Antonevich (antonevich@bsu.by)

In the paper some new constructions of extensions of nonclosable operators is proposed and several examples of applications are given. One of particular cases of the problem under consideration is the question on multiplication of distributions, a solution to which can be given by introduction of the so-called new generalized functions. It was demonstrated that the main obstacle for multiplication of distributions is nonclosablility of classical multiplication and the construction of new generalized functions is based on the ideas similar to that used under construction of the extension of nonclosable operators.

Keywords: extension of operator; closed operator; multiplication of distributions.

Введение

Для линейного оператора наиболее естественным расширением является так называемое замыкание, но оно существует не для всех операторов. В связи с этим в [1] был рассмотрен вопрос о том, *какой оператор может играть роль замыкания в случае незамыкаемого оператора*. В данной работе, которая есть дальнейшее развитие [1], предложена новая конструкция расширения произвольных отображений, в том числе и незамыкаемых операторов, и приведены примеры приложений этой конструкции.

Одним из частных случаев проблемы расширения отображений является задача умножения обобщенных функций (распределений). В классической теории определено умножение распределения только на гладкую функцию и доказано, что невозможно корректно задать произведение произвольных распределений [2]. Вместе с тем в приложениях возникают уравнения, в которые входят произведения распределений, и вопрос о том, какой смысл можно придать таким произведениям, продолжает привлекать внимание многих специалистов (см., например, [3–8]).

Решение задачи об умножении распределений заключается в следующем. По заданному пространству распределений E может быть построена коммутативная ассоциативная дифференциальная алгебра $G(E)$, элементы которой сохраняют часть свойств распределений и называются *новыми обобщенными функциями* или *мнемофункциями*. Такая алгебра есть расширение исходного пространства E, при построении которого происходит дробление элементов из E на более мелкие (каждому элементу u из E соответствует обширное семейство ассоциированных с ним элементов из $G(E)$) и добавление элементов, не ассоциированных с точками из E, т. е. качественно отличающихся от распределений.

В работе показано, что сложность задачи расширения операции умножения на все пространство распределений определяется незамыкаемостью исходной операции и что построение алгебр мнемофункций базируется на тех же идеях, что и построение расширений незамыкаемых операторов.

Еще одним частым случаем задачи о расширении операторов является вопрос о построении оператора в заданном функциональном пространстве, соответствующего формальному дифференциальному выражению с обобщенными коэффициентами. Особенность этой задачи заключается в том, что строится оператор, соответствующий дифференциальному выражению в целом, и при этом может быть, что отдельные слагаемые не определены. Для конкретных дифференциальных выражений с обобщенными коэффициентами такие расширения рассматривались, например, в [9; 10].

Замыкание линейного оператора

Умножение на заданное распределение u есть линейный оператор в пространстве распределений $\mathcal{D}'(\mathbb{R})$, определенный на всюду плотном подпространстве $C^\infty(\mathbb{R})$, состоящем из гладких функций [11]. Поэтому задача о задании произведения uv для распределений v, не являющихся гладкими функциями, есть частный случай классической задачи функционального анализа о построении расширений (продолжений) операторов. У неограниченного линейного оператора, определенного на всюду плотном подпространстве $X_0 \subset X$ в банаховом пространстве X и действующего в банахово пространство Y,

существует много различных продолжений, и все они – неограниченные операторы. Выделяется случай, когда существует «хорошее» расширение, а именно являющееся замкнутым оператором.

Напомним, что линейный оператор A, определенный на всюду плотном подпространстве $D(A)$ в банаховом пространстве X, называется *замкнутым*, если для последовательности точек $x_n \in D(A)$ из того, что $x_n \to x \in X$ и $Ax_n \to y \in Y$, следует, что $x \in D(A)$ и $Ax = y$.

Для каждого $x \in X$ существует много последовательностей $x_n \in X_0$, для которых $x_n \to x$, но при этом последовательность Ax_n может не сходиться. Обозначим через X_A множество таких $x \in X$, что имеется хотя бы одна последовательность $x_n \in X_0$ такая, что $x_n \to x$, и при этом последовательность Ax_n также сходится. С помощью этой последовательности кажется естественным задать значение искомого расширения в точке $x \in X_A$ формулой

$$X_A \ni x \to \lim_{n\to\infty} Ax_n \in Y. \tag{1}$$

Однако в общем случае данное определение некорректно, так как правая часть (1) может зависеть от выбора последовательности x_n. Оператор, у которого предел не зависит от такого выбора, называется *замыкаемым*. Свойство замыкаемости оператора обычно формулируется как условие: из того, что $x_n \to 0$ и существует $\lim\limits_{n\to\infty} Ax_n = y$, следует, что $y = 0$. При выполнении этого условия формула (1) задает замкнутый оператор $\overline{A}$ с областью определения $D(\overline{A}) = X_A$, называемый *замыканием* оператора A.

Пример 1. Пусть $X = L_1[0, 1]$, $Y = L_1[0, 1] \oplus \mathbb{C}$. Рассмотрим оператор A с областью определения $D(A) = X_0 = C^1[0, 1] \subset L_1[0, 1]$, действующий в Y по формуле $Au = (u', u(0))$. Здесь задача о нахождении решения уравнения $Au = (f, C)$ есть вырожденный случай задачи Коши:

$$u'(t) = f(t),\ u(0) = C. \tag{2}$$

Необходимость расширения оператора связана с тем, что при заданной области определения задача (2) разрешима только для непрерывных f. Чтобы расширить множество тех $f \in L_1[0, 1]$, для которых существует решение, надо продолжить оператор на более широкую область определения.

Рассматриваемый оператор A замыкаем, область определения замыкания $\overline{A}$ есть подпространство $W_1^1[0, 1]$, состоящее из функций, представимых в виде

$$u(x) = u(0) + \int_0^x y(s)\,ds,\ \ y \in L_1[0, 1].$$

При таком представлении функция y называется *сильной производной* и замыкание действует по формуле $\overline{A}u = (u', u(0))$, где u' есть сильная производная.

Полезность указанной конструкции подтверждается следующим утверждением.

Предложение. *Задача Коши $\overline{A}u = (f, C)$ имеет, и притом единственное, решение для любой функции $f \in L_1[0, 1]$ и любого C, т. е. у оператора $\overline{A}$ существует ограниченный обратный, определенный на всем $L_1[0, 1] \oplus \mathbb{C}$ и действующий по формуле*

$$\overline{A}^{-1}(f, C) = C + \int_0^x f(s)\,ds. \tag{3}$$

Таким образом, особая роль замыкания в примере 1 проявляется в том, что при расширении оператора A на подпространство меньшее, чем $W_1^1[0, 1]$, не выполнено утверждение о существовании решения задачи Коши для любого $f \in L_1[0, 1]$, а при расширении оператора на подпространство более широкое, чем $W_1^1[0, 1]$, нарушается единственность решения задачи Коши.

Расширения отображений и расслоения

Сделаем несколько замечаний общего характера о расширениях отображений. Обычно говорят, что f есть отображение из множества X в множество Y, определенное на подмножестве $D(f) \subset X$, если каждому элементу $x \in D(f)$ поставлен в соответствие ровно один элемент $f(x) \in Y$. При этом

продолжением (расширением) f называют отображение F, определенное на подмножестве $D(F) \subset X$, содержащем $D(f)$ такое, что $F(x) = f(x)$ для $x \in D(f)$.

Основная идея приведенных ниже построений связана с тем, что исходным объектом является отображение f, определенное только на $D(f) \subset X$, при этом априорный выбор множества X есть дополнительное искусственно внесенное ограничение. Заданную область определения можно рассматривать как подмножество не в X, а в более широком множестве, которому будет принадлежать и область определения расширения.Такое более широкое множество естественно возникает, если исходить из определения отображения, используемого, например, в [12].

Отношением между множествами X и Y называется любое подмножество G из декартова произведения $X \times Y$. Его область определения $D(G)$ есть проекция G на X. Отношение *функционально по x*, если из того, что $(x, y_1) \in G$, $(x, y_2) \in G$, следует, что $y_1 = y_2$.

Отображением из X в Y называется отношение $G \subset X \times Y$, функциональное по x. Иначе говоря, отображение отождествляется с его графиком при обычном понятии отображения, т. е.

$$G = G(f) = \{(x, f(x)) : x \in D(f)\} \subset X \times Y.$$

Такая точка зрения оправдана тем, что данное определение придает точный теоретико-множественный смысл выражению «поставлен в соответствие».

Условие функциональности по x означает, что проектирование $p : (x, y) \to x$ на X устанавливает биекцию между G и $D(G) = D(f)$, позволяющую отождествить эти множества. После такого отождествления областью определения можно считать само множество $G(f)$, и тогда на $G(f)$ действие отображения f задается как проектирование на вторую координату: $f : (x, y) \to y$. Представление отображения в таком виде будем называть его *нормальной формой*.

Если для заданного отображения f подмножество (т. е. отношение) $G_1 \subset X \times Y$ содержит $G(f)$ и f_1 действует как проектирование на вторую координату на G_1, то f_1 есть расширение исходного отображения f, заданного в нормальной форме.

Если X и Y – топологические пространства, то наиболее естественным и канонически определенным является продолжение на замыкание $\overline{G(f)}$.

Определение 1. *Замыканием отображения f*, действующего из топологического пространства X в топологическое пространство Y, будем называть отображение $\overline{f}$, определенное на подмножестве $\overline{G(f)} \subset X \times Y$ и действующее как проектирование на вторую координату.

Для описания возникающих соотношений удобно использовать терминологию из теории расслоенных пространств [13].

Тройка (E, B, p), где E и B есть заданные топологические пространства, $p : E \to B$ – сюръективное непрерывное отображение, называется *расслоением* с пространством расслоения E, базой B и проекцией p.

Подмножество $p^{-1}(b) \subset E$ называется *слоем над точкой b* и обозначается E_b. При этом пространство E представляется в виде объединения непересекающихся слоев. Будем говорить, что точки из слоя над b *ассоциированы с b*.

Сечением расслоения называется отображение $S : B \to E$, правое обратное к проекции, т. е. такое, что $p(S(x)) = x$ для всех $x \in B$.

Выделяется случай, когда для каждого слоя существует гомеоморфизм с некоторым пространством F, которое называют *типовым слоем*. Таких гомеоморфизмов много, причем обычно среди них нет канонически заданного.

Расслоение будем называть *векторным*, если типовой слой F является векторным пространством (стандартное определение векторного расслоения требует выполнения ряда дополнительных условий [13], но в рассматриваемых ниже вопросах существенно только указанное свойство).

Используем эту терминологию в случае, когда $G \subset X \times Y$ – произвольное отношение между топологическими пространствами. Тогда отображение p, действующее как проектирование на X, задает на G структуру расслоения над $D(G) = p(G)$, при котором слои могут быть разной структуры. С этой точки зрения график отображения $G(f)$ есть расслоение, в котором каждый слой состоит ровно из одной точки,

а область определения $\overline{G(f)}$ нельзя отождествить с подмножеством исходного пространства X, так как, если замыкание графика $\overline{G(f)}$ рассматривать как расслоение, могут возникнуть слои, содержащие несколько различных точек. В этом заключается принципиальное отличие продолжения отображения в смысле определения 1 от классического.

Пример 2. Пусть комплекснозначная функция f определена на отрезке $[-1, 1]$, непрерывна при $x \neq 0$, а в точке 0 непрерывна слева и существует предел справа $f(+0)$. В этом примере замыканием f в смысле определения 1 является отображение $\tilde{f} : \overline{G(f)} \to \mathbb{C}$, действующее как проектирование на вторую координату в $[-1, 1] \times \mathbb{C}$. Если $f(+0) \neq f(0)$, то замыкание $\overline{G(f)}$ графика $G(f)$ получается присоединением к графику точки $(0, f(+0))$. Это множество, как отношение на $[-1, 1] \times \mathbb{C}$, не является функциональным по x. Если его рассматривать как расслоение над $[-1, 1]$, то слой над точкой 0 состоит из двух точек: $(0, f(0))$ и $(0, f(+0))$. Здесь проявляется эффект дробления: переход от $[-1, 1]$ к расширенной области определения $\overline{G(f)}$ заключается в том, что точка 0 распадается на две точки.

Рассматривая пример 2, можно отметить, что одно из решений вопроса о расширении области определения числовых функций содержится в теории Гельфанда коммутативных банаховых алгебр [14] и в ряде случаев переход к замыканию функций в смысле определения 1 совпадает с расширением области определения, построенным в теории Гельфанда.

Поскольку нас в первую очередь интересуют расширения линейных операторов, отметим специфику этого случая, когда рассматриваемые подмножества E в прямой сумме $X \oplus Y$ банаховых пространств являются векторными подпространствами. Тогда база расслоения $B = p(E) = D(E)$ также является векторным пространством. Кроме того, при проектировании на первую координату слой $p^{-1}(0) = \{(0, y) \in G\}$ над точкой $0 \in X$ есть векторное подпространство в $X \oplus Y$, которое естественно вкладывается в Y с помощью проектирования на вторую координату. Если $b \neq 0$, то слой $p^{-1}(b) = \{(b, y) \in E\}$ не является векторным пространством, но если выбрать точку $(b, y_0) \in p^{-1}(b)$, то отображение $p^{-1}(b) \ni (b, y) \to (0, y - y_0)$ есть биекция с векторным пространством $p^{-1}(0)$, позволяющая задать на $p^{-1}(b)$ структуру векторного пространства. Тем самым получаем, что проектирование на первую координату задает на подпространстве E структуру векторного расслоения с типовым слоем $F = p^{-1}(0)$. При этом биекция с типовым слоем не задается канонически, и поэтому именно понятие векторного расслоения адекватно описывает взаимоотношение между векторным подпространством E из $X \oplus Y$ и его проекцией B.

Замыкание незамыкаемого оператора

Перейдем к основному объекту исследования в данной работе – построению расширений незамыкаемых операторов. Одна из конструкций расширения, которое может играть роль замыкания в случае незамыкаемого оператора, была предложена в [1]. Она по существу уже содержится в конструкции классического замыкания, но несколько отличается от привычных результатов теории операторов.

На первом шаге построения замыкания рассматривается векторное пространство $\tilde{G}_A$, состоящее из последовательностей $x_n \in X_0$ таких, что x_n сходится в X и последовательность образов Ax_n сходится в Y. Две последовательности x_n и $\tilde{x}_n$ называются эквивалентными, если $x_n - \tilde{x}_n \to 0$ и $A(x_n - \tilde{x}_n) \to 0$.

Пусть G_A есть векторное пространство, состоящее из классов эквивалентных последовательностей из $\tilde{G}_A$. Иначе говоря, если ввести подпространство

$$N_0 = \{(x_n) \in \tilde{G}_A : x_n \to 0,\ Ax_n \to 0\}, \tag{4}$$

то G_A по определению есть фактор-пространство: $G_A = \tilde{G}_A / N_0$.

На пространстве G_A определен оператор $\tilde{A} : G_A \to Y$, который ставит в соответствие классу эквивалентности, содержащему последовательность x_n, предел последовательности Ax_n:

$$G_A \ni [(x_n)] \to \lim_{n \to \infty} Ax_n \in Y. \tag{5}$$

Эта часть конструкции не использует замыкаемость и применима к любому линейному оператору, что позволяет определить замыкание оператора следующим образом.

Определение 2. *Замыканием оператора A* с областью определения $X_0 \subset X$ будем называть оператор $\tilde{A}$, определенный на построенном фактор-пространстве $G_A = \tilde{G}_A / N_0$ и действующий по формуле (5).

При анализе введенного определения замыкания оператора будем исходить из геометрического смысла описанной конструкции. Покажем, что это частный случай общего определения 1.

Пусть

$$G(A) = \{(x,\ Ax) : x \in D(A)\} \subset X \oplus Y$$

есть график оператора. Тогда $\tilde{G}_A$ есть множество всех лежащих в $G(A)$ последовательностей Коши в смысле нормы из $X \oplus Y$, а пространство G_A есть пополнение графика $G(A)$ и в силу полноты $X \oplus Y$ есть его замыкание $\overline{G(A)}$. При этом оператор $\tilde{A}$ действует из $G_A = \overline{G(A)}$ в Y как проектирование на вторую координату.

Проекция подпространства $\overline{G(A)}$ на X есть введенное ранее подпространство X_A. Условие замыкаемости оператора эквивалентно тому, что подпространство $\overline{G(A)} \subset X \oplus Y$ есть линейное отношение, функциональное по x, т. е. является графиком некоторого оператора, и эквивалентно условию, что проектирование $\overline{G(A)}$ на X является инъективным отображением. Это позволяет, как и для произвольных отображений, отождествить точку из $\overline{G(A)}$ с его первой проекцией, после чего получаем классическое замыкание – оператор, определенный на подпространстве X_A в X.

В случае незамыкаемого оператора проектирование p подпространства $\overline{G(A)}$ на первую координату не является инъективным, так как подпространство $M_A = p^{-1}(0)$ отлично от нулевого. Будем называть его *мерой незамыкаемости* оператора A. Это подпространство состоит из классов эквивалентных последовательностей $x_n \in X_0$ таких, что $x_n \to 0$, $Ax_n \to y$.

Таким образом, пространство $\overline{G(A)}$ имеет структуру векторного расслоения над X_A с типовым слоем $F = M_A$ и распадается на слои $(G_A)_x = p^{-1}(x)$. Будем говорить, что элемент $g \in G$ *ассоциирован* с x, если g принадлежит слою G_x.

Если последовательность x_n из $\tilde{G}_A$ сходится к нулю в X, то ее (и соответствующий класс эквивалентности из G) естественно называть *бесконечно малым элементом*. В этой терминологии в случае замыкаемого оператора бесконечно малым является только нулевой элемент, а в случае незамыкаемого A существует ненулевое подпространство M_A, состоящее из бесконечно малых.

Как показано выше, разность двух элементов из слоя над заданной точкой $x \in X_A$ является бесконечно малым элементом, в частности если задать один элемент g из этого слоя, то слой состоит из элементов вида $g + h$, где h – бесконечно малый элемент. Если $x \in X_0$, то элемент из слоя можно задать канонически, поставив в соответствие точке x класс, содержащий стационарную последовательность $x_n = x \in X_0$. Тем самым X_0 вкладывается в G_A, а образом при этом вложении является график исходного оператора $G(A)$, и действие $\tilde{A}$ на нем совпадает с действием A.

Следовательно, $\tilde{A}$ есть частный случай расширения отображения в смысле определения 1: сначала A представляется в нормальной форме, как оператор с областью определения $G(A) \subset X \oplus Y$, действующий как проектирование на второй сомножитель, а расширение $\tilde{A}$ определено на $\overline{G(A)}$ и также действует как проектирование на второй сомножитель.

Поясним, в каком смысле построенный оператор является замкнутым, так как к нему неприменимо классическое определение.

Лемма 1. *Линейный оператор, определенный на всюду плотном подпространстве $D(B)$ в банаховом пространстве X и действующий в банахово пространство Y, замкнут тогда и только тогда, когда его область определения $D(B)$ является полным пространством относительно нормы графика*

$$\|x\|_A = \|x\|_X + \|Bx\|_Y.$$

Доказательство следует непосредственно из определений.

Согласно лемме оператор будем называть *замкнутым*, если его область определения является полным пространством относительно нормы графика.

У построенного оператора $\tilde{A}$ область определения есть полное пространство $\overline{G(A)}$, и этот оператор является замкнутым в смысле введенного определения.

Поясним, в каком смысле построенный оператор разрывен. Точке $g = (x, y) \in G$ поставим в соответствие число $q(g) = \|x\|_X$. В случае замыкаемого оператора это норма на G, которая переходит в норму

на X_A при изоморфизме между G и X_A, задаваемом проекцией. Если оператор незамыкаем, то $q(g)$ есть полунорма, так как $q(g)=0$ для элементов $g \in M_A$, т. е. ядро полунормы

$$\ker q = \{g \in G : q(g) = 0\}$$

состоит из бесконечно малых элементов.

Если $\tilde{A}$ рассматривать как оператор из (неотделимого) топологического векторного пространства (G, q) в Y, то он разрывен. Действительно, если линейный оператор B непрерывен, как отображение из некоторого пространства Z с полунормой q в Y, то при действии B ядро полунормы

$$\ker q = \{z \in Z : q(z) = 0\}$$

переходит в нуль. В исследуемом случае незамыкаемого оператора ядро полунормы есть ненулевое подпространство M_A, при действии оператора $\tilde{A}$ оно не переходит в нуль, так как биективно отображается на себя. Значит, построенный оператор разрывен.

В результате приведенных рассуждений получено следующее описание структуры пространства G_A и действия оператора $\tilde{A}$.

Теорема 1. *Пространство G_A, на котором определен оператор $\tilde{A}$, естественно реализуется как замкнутое векторное подпространство в $X \oplus Y$, являющееся замыканием графика $G(A)$; проектирование на первую координату задает на G_A структуру векторного расслоения над X_A, в котором элементы слоя отличаются на бесконечно малые. При такой реализации действие оператора $\tilde{A}$ совпадает с проектированием G_A на вторую координату в прямой сумме $X \oplus Y$, и этот оператор является замкнутым.*

Таким образом, область определения замыкания оператора получается дроблением точек из X_A: к каждой точке из X_A добавляются бесконечно малые элементы.

Пример 3. Пусть оператор A определен на $C^1[0, 1] \subset L_1[0, 1]$, действует в прямую сумму $Y = L_1[0, 1] \oplus \oplus \mathbb{C} \oplus \mathbb{C}$ по формуле $Au = (u', u(0), u'(0))$. Этот оператор соответствует переопределенной задаче Коши: найти функцию $u \in C^1[0, 1]$, удовлетворяющую условиям

$$u'(x) = f(x),\ u(0) = C,\ u'(0) = C_1. \tag{6}$$

Задача (6) разрешима только для правых частей, удовлетворяющих условию $f \in C[0, 1]$ и $C_1 = f(0)$. Для таких правых частей ее решение задается формулой (3).

Как отмечалось, множество правых частей, для которых задача разрешима, увеличивается при расширении области определения оператора. Возникает естественный вопрос: можно ли расширить исходное пространство так, чтобы поставленная задача была разрешима для любой правой части из Y (подобно тому как это было сделано в примере 2)? Покажем, что положительный ответ дает замыкание оператора A в смысле определения 2.

Прежде всего обратим внимание на то, что рассматриваемый оператор незамыкаем. Действительно, последовательность

$$u_n(x) = \begin{cases} x(1 - nx)^2, & 0 \le x \le \dfrac{1}{n}, \\ 0, & x > \dfrac{1}{n}, \end{cases}$$

принадлежит области определения, сходится к нулю в $L_1[0, 1]$, и при этом последовательность образов Au_n имеет ненулевой предел: $u_n' \to 0$ в $L_1[0, 1]$, $u_n(0) = 0 \to 0$, но $u_n'(0) = 1 \to 1$. Мерой незамыкаемости этого оператора является одномерное подпространство $M = \{(0, 0, \xi)\} \in L_1[0, 1] \oplus \mathbb{C} \oplus \mathbb{C}$.

Рассмотрим замыкание оператора A в смысле определения 2.

Пространство $\tilde{G}$ есть множество последовательностей $u_n \in C^1$ таких, что выполнены четыре условия:

i) $u_n \to u$ в $L_1[0, 1]$;

ii) $u_n' \to y$ в $L_1[0, 1]$;

iii) $u_n(0) \to \xi$;

iv) $u_n'(0) \to \eta$.

Из первых двух условий следует, что функция u абсолютно непрерывна и $u_n(0) \to u(0)$, т. е. $\xi = u(0)$. Но при этом у предельной функции u производная $u'(0)$ может не существовать и, даже если она существует, может быть, что последовательность $u_n'(0)$ не сходится к $u'(0)$.

Согласно предложенной конструкции две последовательности из $\tilde{G}$ называются *эквивалентными*, если для них все указанные выше пределы совпадают, пространство G_A по определению есть факторпространство по этому отношению эквивалентности.

Указанное отношение эквивалентности содержит дополнительное условие по сравнению с отношением эквивалентности из примера 2. В результате каждый класс эквивалентности, построенный в примере 2, содержит много разных новых классов эквивалентности. Именно поэтому возникает расслоение – каждой функции $u \in W_1^1[0, 1]$ соответствует семейство элементов пространства G_A, которое естественно параметризуется числом η и имеет структуру одномерного векторного пространства.

Действительно, каждый такой класс эквивалентности $\tilde{u} = \left[(u_n)\right]$ состоит из последовательностей непрерывно дифференцируемых функций u_n, сходящихся к абсолютно непрерывной функции u специальным образом. Он задается функцией u и числом $\eta = \lim u_n'(0)$, которое будем интерпретировать как значение производной $\tilde{u}$ в точке 0 и обозначать $\tilde{u}'(0)$. Построенное пространство обозначим $\widetilde{W_1^1[0, 1]}$, которое согласно сказанному является векторным расслоением над $W_1^1[0, 1]$.

Таким образом, элемент $\left[(u_n)\right]$ из $\widetilde{W_1^1[0, 1]}$ не определяется однозначно абсолютно непрерывной функцией u из условия i), а «помнит» о способе своего возникновения из непрерывно дифференцируемых функций, а именно сохраняет информацию о поведении значений $u_n'(0)$. Это поясняет смысл термина *мнемофункция* (от греч. $\mu\nu\acute{\eta}\mu\eta$ – память), используемого для аналогичных объектов, возникающих при построении расширений пространств распределений. Оператор $\tilde{A}$ действует по формуле, внешне такой же, как формула исходного оператора: $\tilde{A}\tilde{u} = (\tilde{u}', \tilde{u}(0), \tilde{u}'(0))$, где $\tilde{u}'$ есть сильная производная от u. Здесь принципиально новым является то, что на построенном пространстве заданы величины $\tilde{u}'(0)$, которые не определены на $W_1^1[0, 1]$.

Для пояснения естественности введения пространства $\widetilde{W_1^1[0, 1]}$ отметим, что оно возникает при рассмотрении так называемых сингулярно возмущенных задач. Примером может служить задача Коши

$$\frac{1}{n}u''(x) + u'(x) = f(x),\ u(0) = C,\ u'(0) = C_1 \tag{7}$$

для уравнения с малым параметром при второй производной, где дифференцирование есть сильная производная.

Пусть u_n – решения задачи (7). Эта последовательность состоит из функций, производные которых абсолютно непрерывны, и сходится к абсолютно непрерывной функции u, являющейся решением задачи Коши (2). При этом $u_n'(0) = C_1$ и, в частности, сходятся к C_1.

Таким образом, при любых $(f, C, C_1) \in Y$ последовательность u_n задает элемент $\tilde{u}$ построенного пространства $\widetilde{W_1^1[0, 1]}$, являющийся решением задачи (7). При другом значении C_1 последовательность решений также сходится к решению задачи (2), но эти последовательности не нужно отождествлять, так как они по-разному сходятся к u и задают разные элементы из $\widetilde{W_1^1[0, 1]}$. Из сказанного вытекает следующая теорема.

Теорема 2. *В построенном расширенном пространстве* $\widetilde{W_1^1[0, 1]}$ *решение переопределенной задачи Коши (6) существует и единственно при любой правой части* $(f, C, C_1) \in Y$.

Расширенное замыкание линейного оператора

Конструкция замыкания задает расширение оператора, определенное на элементах некоторого векторного расслоения над частью пространства X. Естественно возникает задача о задании расширения оператора на расширение всего пространства X.

Опишем одно из решений этой задачи, в котором наряду с расширением начального пространства X происходит расширение и финального пространства Y. Оно основано на отказе от требования сходимости последовательностей Ax_n.

Пусть $\hat{X}$ есть векторное пространство, состоящее из всех последовательностей Коши $x_n \in X_0$. Построим фактор-пространство $\hat{G}^* = \hat{G}/\hat{G}_0$, где $\hat{G}_0$ есть подпространство (4). Аналогично предыдущему построенное пространство $\hat{G}^*$ является векторным расслоением с типовым слоем M_A, но уже над всем X.

Пусть $\hat{Y}$ есть пространство всех последовательностей (y_n) в Y и пусть

$$Y^* = \hat{Y}/\hat{Y}_0, \text{ где } \hat{Y}_0 = \{(y_n) \in Y : y_n \to 0\}.$$

Заметим, что Y^* является расширением исходного пространства Y, так как последнее естественно вкладывается в Y^*: точке $y \in Y$ ставится в соответствие класс эквивалентности, состоящий из последовательностей, сходящихся к y.

Определение 3. *Расширенным замыканием оператора A* будем называть оператор $\hat{A}$, действующий из векторного расслоения $\hat{G}^*$ над X в расширенное пространство Y^* по формуле

$$\hat{A}([x_n, Ax_n]) = [(Ax_n)] \in Y^*.$$

В конкретных приложениях естественны модификации предложенной конструкции, которые определяются постановкой исходной задачи. Например, бывает более подходящим рассмотрение фактор-пространства $\tilde{G}_A/N$ по подпространству, меньшему N_0.

Другая модификация нужна в случае, когда исходный оператор A определен на подпространстве $X_0 \subset X$ и действует в X. Тогда можно модифицировать конструкцию так, чтобы расширение оператора также действовало из некоторого расширенного пространства в себя.

Подводя итог сказанному, можно заключить, что расширенное замыкание действует в новых более широких пространствах, построение которых включает два типа преобразований:

• дробление начального пространства X, при котором точке $x \in X$ соответствует слой над x – обширное семейство элементов из нового пространства, ассоциированных с x и отличающихся друг от друга на бесконечно малые;

• добавление к финальному пространству Y новых идеальных элементов, не ассоциированных с элементами исходного пространства.

Отметим, что описанные конструкции есть частный случай широко распространенного метода построения новых пространств: сначала задается некоторое пространство последовательностей из исходного пространства, а искомое пространство получается из него с помощью специального отношения эквивалентности. В качестве примеров расширений, полученных только добавлением новых идеальных элементов, можно указать пополнение нормированного пространства и так называемый секвенциальный подход к построению обобщенных функций, описанный в [15].

Наиболее известным примером расширения пространств, при котором наряду с добавлением качественно новых идеальных элементов происходит дробление точек исходного пространства, является нестандартное расширение поля $\mathbb{R}$ [16].

Напомним описание одного из нестандартных расширений поля $\mathbb{R}$ и покажем, что его построение аналогично рассмотренным выше конструкциям, тем самым еще раз подтвердив их содержательность.

Пусть $\hat{\mathbb{R}} = \mathbb{R}^{\mathbb{N}}$ есть алгебра, состоящая из всех числовых последовательностей (x_n), при этом $\mathbb{R}$ вкладывается в эту алгебру с помощью постоянных последовательностей. *Нестандартное расширение* $\mathbb{R}$ определяется как фактор-пространство $\mathbb{R}^* = \mathbb{Z}/J$ по некоторому максимальному идеалу. Замечательные свойства построенного множества заключаются в том, что $\mathbb{R}^*$ есть линейно упорядоченное поле. При этом появляются бесконечно малые и бесконечно большие элементы: $\gamma \in \mathbb{R}^*$ является *бесконечно малым*, если $-h \prec \gamma \prec h$ для любого положительного $h \in \mathbb{R}$, *бесконечно большие* элементы могут быть описаны как обратные к бесконечно малым.

Выделяется подмножество $\overline{\mathbb{R}}$, состоящее из конечных нестандартных чисел, т. е. представимых в виде суммы вещественного числа и бесконечно малого.

Таким образом, переход от $\mathbb{R}$ к $\mathbb{R}^*$ заключается в добавлении бесконечно больших величин и дроблении точек из $\mathbb{R}$ с помощью введения бесконечно малых. При этом подмножество $\overline{\mathbb{R}}$ имеет структуру векторного расслоения над $\mathbb{R}$, в котором типовым слоем является пространство бесконечно малых.

Оператор умножения на распределение

Оператор U, действующий как умножение на заданное распределение u, обычно незамыкаем, что и определяет сложности, связанные с заданием умножения в пространстве распределений.

К таким операторам U применимы описанные выше конструкции. Для примера рассмотрим замыкание оператора U, действующего как умножение на функцию Хевисайда:

$$\Theta(x)=\begin{cases}0,\ x\le 0,\\ 1,\ x>0.\end{cases}$$

Лемма 2. *Оператор U умножения на Θ незамыкаем в пространстве распределений. Мерой его незамыкаемости M_U является подпространство, состоящее из линейных комбинаций δ-функции и ее производных.*

Произведение $\Theta\delta$ встречается в ряде задач, но это формальное выражение, которое не определено в теории распределений и в силу незамыкаемости оператора умножения на Θ не может быть задано каноническим образом. Согласно описанной конструкции замыкание оператора U определено на векторном расслоении G_Θ, и в этом смысле умножение на Θ задано на точках ν из слоя над точкой δ в этом расслоении. Рассмотрим, как устроен данный слой и какие значения может принимать произведение $\Theta\nu$ для разных ν.

Возьмем функцию $\gamma\in\mathcal{D}(\mathbb{R})$ такую, что $\operatorname{supp}\gamma\in(-1,1)$ и выполняется

$$\int\gamma(x)\,dx=1.$$

Рассмотрим две последовательности гладких функций: $\gamma_n(x)=n\gamma(nx)$ и $h_n(x)=n\gamma\big(n(x-1)\big)-$ $-n\gamma\big(n(x+1)\big)$. Обозначим

$$a(\gamma)=\int_0^\infty\gamma(x)\,dx.$$

Теорема 3. *Пусть G_Θ есть векторное расслоение, построенное по оператору умножения на Θ. Слой G_δ над точкой δ состоит из классов эквивалентности ν, порожденных последовательностями гладких функций вида*

$$\nu_n=\gamma_n+\sum_{k=0}^{m}C_k h_n^{(k)}.$$

При этом каждая такая последовательность задает свой класс эквивалентности и действие замыкания оператора выражается формулой

$$\tilde\Theta\nu=\big(a(\gamma)+C_0\big)\delta+\sum_{k=1}^{m}C_k\delta^{(k)}.$$

Произведение $\Theta\delta'$ также не определено в теории распределений. Для δ' стандартные аппроксимирующие последовательности имеют вид

$$w_n(x)=n^2\gamma'(nx).$$

Для таких последовательностей предел произведения Θw_n, как правило, не существует, поэтому в общем случае последовательность w_n не порождает точку из области определения замыкания оператора, но порождает точку из области определения расширенного замыкания. При этом последовательность произведений Θw_n в пространстве $\mathcal{D}'(\mathbb{R})$ ведет себя как δ-функция с бесконечно большим коэффициентом и задает бесконечно большой элемент из расширенного пространства.

Дальнейшее развитие изложенных идей приводит к построению алгебр мнемофункций, в которых определено умножение на любой элемент. Опишем для примера предложенную Ю. В. Егоровым в [6] конструкцию наиболее простой из таких алгебр, укладывающуюся в описанную схему построения расширений.

Пусть $\widetilde{C^\infty(\mathbb{R})}$ есть алгебра, состоящая из всех последовательностей бесконечно дифференцируемых функций на $\mathbb{R}$, и пусть

$$J=\Big\{(f_n)\in\widetilde{C^\infty(\mathbb{R})}:\forall a>0\ \exists N\text{ такое, что }f_n(x)=0\text{ для всех }n\ge N\text{ и }x\in[-a,a]\Big\}.$$

Подмножество J является идеалом в $\widetilde{C^\infty(\mathbb{R})}$. *Алгебра новых обобщенных функций по Егорову* определяется как фактор-алгебра

$$G_E:=\widetilde{C^\infty(\mathbb{R})}/J.$$

Положим G_a – подпространство в G_E, порожденное последовательностями, сходящимися в $\mathcal{D}'(\mathbb{R})$, и F – подпространство в G_E, порожденное бесконечно малыми, т. е. последовательностями, сходящимися к нулю. Тогда отображение p: $G_a\ni\big[(f_n)\big]\to\lim f_n\in\mathcal{D}'(\mathbb{R})$ задает на G_a структуру векторного расслоения над $\mathcal{D}'(\mathbb{R})$ с типовым слоем F, состоящим из бесконечно малых. При этом в G_E есть обширное множество идеальных элементов, не ассоциированных с распределениями.

Заключение

В статье рассмотрены конструкции новых пространств, возникающих при построении расширений отображений. Поясним естественность введения таких пространств с точки зрения приложений.

Пусть изучается эксперимент по воздействию на некоторую систему. Обычно говорят, что Z есть множество состояний системы, а W есть множество результатов воздействия, если каждому элементу из Z соответствует однозначно определенный результат воздействия, принадлежащий W. В первоначальной модели предполагается, что состояния системы описываются элементами пространства X, а результаты воздействия – элементами пространства Y, а именно для некоторых «простых» состояний (из всюду плотного подпространства $X_0 = D(A)$) задан оператор A, определяющий результат воздействия на систему: при состоянии $x \in X_0$ получаем на выходе $Ax \in Y$.

Задача заключается в описании результата воздействия для более сложных состояний системы, т. е. построении расширений оператора A.

Для каждой точки $x \in X$ существует последовательность $x_n \in X_0$, сходящаяся к x, т. е. x есть предел простых состояний. С такой последовательностью связана последовательность результатов эксперимента $Ax_n \in Y$ и ее предел y, если он существует. Если оператор A незамыкаемый, то по элементу $x \in X$ нельзя однозначно определить y. Это означает, что в действительности элемент x не задает состояние системы, а для получения однозначного ответа о результате нужна дополнительная информация о том, как именно этот x получен из простых состояний. Иначе говоря, для рассматриваемых систем происходит уточнение постановки задачи – в новой модели состояние системы описывается одним из ассоциированных с x элементов расширенного пространства.

С этой точки зрения переход к расширению $\hat{Y}$ пространства Y требуется в ситуации, когда для уточненного состояния, определяемого последовательностью $x_n \to x$, последовательность Ax_n не сходится в Y и результат эксперимента не задается точкой из Y, но он описывается классом эквивалентности, содержащим последовательность Ax_n. Иными словами, множеством результатов является пространство Y^*.

В частности, описанная ситуация имеет место в задачах, содержащих произведение обобщенных функций. В первоначальной модели явления считается, что состояния системы задаются распределениями, в уточненной модели состояния системы и результаты экспериментов описываются мнемофункциями, а результат воздействия – расширенным замыканием исходного оператора.

Библиографические ссылки

1. Burachewski A, Radyno Ya, Antonevich A. On closability of nonclosable operators. *Panamerican Mathematical Journal.* 1997; 7(4):37–51.

2. Schwartz L. Sur l'impossibilité de la multiplication des distributions. *Comptes rendus de l'Académie des Sciences.* 1954;239:847–848.

3. Иванов ВК. Гиперраспределения и умножение распределений Шварца. *Доклады АН СССР.* 1972;204(5):1045–1048.

4. Colombeau JF, editor. *New generalized functions and multiplication of distributions*. Amsterdam: North-Holland; 1984. 375 p. (North-Holland Mathematics Studies; volume 84).

5. Rosinger EE. *Generalized solutions of nonlinear partial differential equations*. Amsterdam: Elsevier Science; 1987. 409 p. (North-Holland Mathematics Studies; volume 146).

6. Егоров ЮВ. К теории обобщенных функций. *Успехи математических наук.* 1990;45(5):3–40.

7. Oberguggenberger M. *Multiplication of distributions and application to partial differential equations (Pitman Research Notes in Mathematics).* Harlow: Longman Higher Education; 1992. 336 p.

8. Rosinger EE. *Singularities and differential algebras of generalized functions: a basic dichotomic shealf theoretic singularity test.* [S. l.]: Lambert Academic Publishing; 2013. 192 p.

9. Кот МГ. Асимптотика собственных значений операторов, аппроксимирующих дифференциальные уравнения с δ-образными коэффициентами. *Журнал Белорусского государственного университета. Математика. Информатика.* 2017;1:4–10.

10. Шкадинская ЕВ. Об уравнениях, содержащих производную дельта-функции. *Журнал Белорусского государственного университета. Математика. Информатика*. 2017;3:19–26.

11. Владимиров ВС. *Обобщенные функции в математической физике.* Москва: Наука; 1979.

12. Дьедонне Ж. *Основы современного анализа.* Москва: Мир; 1964.

13. Мищенко АС. *Векторные расслоения и их применения.* Москва: Наука; 1984.

14. Бурбаки Н. *Спектральная теория.* Москва: Мир; 1973.

15. Антосик П, Микусинский Я, Сикорский Р. *Теория обобщенных функций. Секвенциальный подход.* Москва: Мир; 1976.

16. Девис М. *Прикладной нестандартный анализ.* Москва: Мир; 1980.

References

1. Burachewski A, Radyno Ya, Antonevich A. On closability of nonclosable operators. *Panamerican Mathematical Journal.* 1997; 7(4):37–51.

2. Schwartz L. Sur l'impossibilité de la multiplication des distributions. *Comptes rendus de l'Académie des Sciences.* 1954;239:847–848.

3. Ivanov VK. [Hyperdistributions and the multiplication of Schwartz distributions]. *Doklady AN SSSR.* 1972;204(5):1045–1048. Russian.

4. Colombeau JF, editor. *New generalized functions and multiplication of distributions*. Amsterdam: North-Holland; 1984. 375 p. (North-Holland Mathematics Studies; volume 84).

5. Rosinger EE. *Generalized solutions of nonlinear partial differential equations*. Amsterdam: Elsevier Science; 1987. 409 p. (North-Holland Mathematics Studies; volume 146).

6. Egorov YuV. [A contribution to the theory of generalized functions]. *Uspekhi matematicheskikh nauk.* 1990;45(5):3–40. Russian.

7. Oberguggenberger M. *Multiplication of distributions and application to partial differential equations (Pitman Research Notes in Mathematics)*. Harlow: Longman Higher Education; 1992. 336 p.

8. Rosinger EE. *Singularities and differential algebras of generalized functions: a basic dichotomic shealf theoretic singularity test.* [S. l.]: Lambert Academic Publishing; 2013. 192 p.

9. Kot MG. Asymptotics of the eigenvalues of approximating differential equations with δ-different coefficients. *Journal of the Belarusian State University. Mathematics and Informatics.* 2017;1:4–10. Russian.

10. Shkadzinskaya AV. On equations containing derivative of the delta-function. *Journal of the Belarusian State University. Mathematics and Informatics*. 2017;3:19–26. Russian.

11. Vladimirov VS. *Obobshchennye funktsii v matematicheskoi fizike* [Generalized functions in mathematical physics]. Moscow: Nauka; 1979. Russian.

12. Dieudonne J. *Foundations of modern analysis.* New York: Academic Press; 1960.

Russian edition: Dieudonne J. *Osnovy sovremennogo analiza*. Moscow: Mir; 1964.

13. Mishchenko AS. *Vektornye rassloeniya i ikh primeneniya* [Vector bundles and their applications]. Moscow: Nauka; 1984. Russian.

14. Bourbaki N. *Theories Spectrales. Chapitres 1 et 2*. Paris: Hermann; 1967.

Russian edition: Burbaki N. *Spektral'naya teoriya.* Moscow: Mir; 1973.

15. Antosik P, Mikusinski Ja, Sikorski R. *Theory of distributions. The sequential approach.* Amsterdam: Elsevier Scientific; 1973.

Russian edition: Antosik P, Mikusinskii Ya, Sikorskii R. *Teoriya obobshchennykh funktsii. Sekventsial'nyi podkhod.* Moscow: Mir; 1976.

16. Davis M. *Applied nonstandard analysis.* New York: Wiley; 1977.

Russian edition: Devis M. *Prikladnoi nestandartnyi analiz.* Moscow: Mir; 1980.

Статья поступила в редколлегию 22.09.2019.
Received by editorial board 22.09.2019.

УДК 517.5

СУММЫ ФЕЙЕРА РАЦИОНАЛЬНОГО РЯДА ФУРЬЕ – ЧЕБЫШЕВА И АППРОКСИМАЦИИ ФУНКЦИИ $|x|^s$

П. Г. ПОЦЕЙКО[1)]***, Е. А. РОВБА***[1)]

[1)]*Гродненский государственный университет им. Янки Купалы, ул. Э. Ожешко, 22, 230023, г. Гродно, Беларусь*

Изучаются аппроксимативные свойства сумм Фейера рядов Фурье по системе алгебраических дробей Чебышева – Маркова и приближения суммами Фейера функции $|x|^s$, $0 < s < 2$, на отрезке $[-1, 1]$. Рассматривается одна ортогональная система алгебраических дробей Чебышева – Маркова и вводятся суммы Фейера соответствующих рациональных рядов Фурье – Чебышева. Устанавливаются порядок приближений последовательностями сумм Фейера непрерывных на отрезке функций в терминах модуля непрерывности и достаточные условия на параметр, обеспечивающие равномерную сходимость. Находятся оценки поточечных и равномерных приближений функции $|x|^s$, $0 < s < 2$, на отрезке $[-1, 1]$, асимптотические выражения при $n \to \infty$ мажоранты равномерных приближений, а также оптимальное значение параметра, при котором обеспечивается наибольшая скорость приближений исследуемой функции суммами Фейера рациональных рядов Фурье – Чебышева.

Ключевые слова: ряд Фурье – Чебышева; частичные суммы; суммы Фейера; модуль непрерывности; равномерная сходимость; асимптотические оценки; точные константы.

FEJER MEANS OF RATIONAL FOURIER – CHEBYSHEV SERIES AND APPROXIMATION OF FUNCTION $|x|^s$

P. G. PATSEIKA[a]***, Y. A. ROUBA***[a]

[a]*Yanka Kupala State University of Grodno, 22 Ažeška Street, Hrodna 230023, Belarus*
Corresponding author: P. G. Patseika (pahamatby@gmail.com)

Approximation properties of Fejer means of Fourier series by Chebyshev – Markov system of algebraic fractions and approximation by Fejer means of function $|x|^s$, $0 < s < 2$, on the interval $[-1, 1]$ are studied. One orthogonal system of Chebyshev – Markov algebraic fractions is considers, and Fejer means of the corresponding rational Fourier – Chebyshev

Образец цитирования:
Поцейко ПГ, Ровба ЕА. Суммы Фейера рационального ряда Фурье – Чебышева и аппроксимации функции $|x|^s$. *Журнал Белорусского государственного университета. Математика. Информатика.* 2019;3:18–34.
https://doi.org/10.33581/2520-6508-2019-3-18-34

For citation:
Patseika PG, Rouba YA. Fejer means of rational Fourier – Chebyshev series and approximation of function $|x|^s$. *Journal of the Belarusian State University. Mathematics and Informatics.* 2019;3:18–34. Russian.
https://doi.org/10.33581/2520-6508-2019-3-18-34

Авторы:
Павел Геннадьевич Поцейко – аспирант кафедры фундаментальной и прикладной математики факультета математики и информатики. Научный руководитель – Е. А. Ровба.
Евгений Алексеевич Ровба – доктор физико-математических наук, профессор; заведующий кафедрой фундаментальной и прикладной математики факультета математики и информатики.

Authors:
Pavel G. Patseika, postgraduate student at the department of fundamental and applied mathematics, faculty of mathematics and informatics.
pahamatby@gmail.com
http://orcid.org/0000-0001-7835-0500
Yauheni A. Rouba, doctor of science (physics and mathematics), full professor; head of the department of fundamental and applied mathematics, faculty of mathematics and informatics.
rovba.ea@gmail.com

series is introduce. The order of approximations of the sequence of Fejer means of continuous functions on a segment in terms of the continuity module and sufficient conditions on the parameter providing uniform convergence are established. A estimates of the pointwise and uniform approximation of the function $|x|^s$, $0 < s < 2$, on the interval $[-1, 1]$, the asymptotic expressions under $n \to \infty$ of majorant of uniform approximations, and the optimal value of the parameter, which provides the highest rate of approximation of the studied functions are sums of rational use of Fourier – Chebyshev are found.

Keywords: Fourier – Chebyshev series; partial sums; Fejer means; modulus of continuity; uniform convergence; asymptotic estimates; exact constants.

Введение

Метод приближений средними арифметическими рядов Фурье 2π-периодических функций имеет богатую историю и ведет свое начало с работ Л. Фейера [1], А. Лебега [2] и др. К настоящему времени метод средних арифметических Фейера тригонометрических рядов Фурье достаточно хорошо изучен и нашел широкое применение в полиномиальной аппроксимации (см., например, [3–6]). А. В. Ефимов получил выражение главного члена уклонения функции от ее сумм Фейера и сумм Фурье, а также установил асимптотически точные равенства для верхних граней этих уклонений, распространенных на классы H_2^α и $W^r H_2^\alpha$ в непрерывной метрике [7]. Г. К. Лебедь и А. А. Авдеенко пришли к аналогичным результатам в интегральной метрике [8].

В 1956 г. М. М. Джрбашян ввел рациональные ряды Фурье, обобщающие соответствующие классические тригонометрические ряды [9]. Одним из основных результатов этой работы было компактное представление ядра Дирихле рациональных рядов Фурье. Основываясь на таком представлении, В. Н. Русак предложил рациональные операторы типа Фейера, Джексона, Валле Пуссена [10] (см. также [11]).

Рациональные операторы Джексона и Валле Пуссена нашли широкое применение в теории рациональных приближений как с фиксированными, так и со свободными полюсами. С их помощью были найдены новые классы функций, отражающие особенности рациональной аппроксимации (см., например, [12–15]). Рациональные операторы Фейера такого применения не нашли и практически не использовались.

На отрезке $[-1, 1]$ рациональные интегральные операторы типа Фейера на основании частичных сумм рядов Фурье – Чебышева по системе рациональных функций, введенной М. М. Джрбашяном и А. А. Китбаляном как метод рациональных приближений с фиксированными полюсами, были построены и исследованы в [16; 17].

Приближения функций, удовлетворяющих условию Липшица, посредством интегральных рациональных операторов типа Фейера были изучены на вещественной оси [12; 13] и на отрезке [18; 19].

Задача аппроксимации функции $|x|$ на отрезке $[-1, 1]$ ведет свою богатую историю с начала XX в., когда полиномиальная аппроксимация этого примера негладкой функции заинтересовала А. Лебега, Д. Джексона и С. Н. Бернштейна [20]. Проблеме посвящен ряд исследований. Новый импульс в ее изучении придала работа Д. Ньюмена [21] о рациональной аппроксимации функции $|x|$ на отрезке $[-1, 1]$. Тема была продолжена во многих трудах (см., например, [22; 23]), и окончательный результат был получен Г. Шталем [24].

Начало исследованию приближений функции $|x|^s$, $s > 0$, также положено С. Н. Бернштейном [25]. К настоящему времени имеется достаточно большое число работ, посвященных как наилучшим приближениям этой функции (см., например, [26–29]), так и конкретным методам приближений (см., например, [30; 31]).

В [32] авторами были построены и исследованы ряды Фурье по одной системе алгебраических дробей Чебышева – Маркова, которая является обобщением классической системы полиномов Чебышева первого рода. В частности, построен интеграл Дирихле и изучены его аппроксимативные свойства в приближениях индивидуальных функций.

В настоящей работе на основании вышеизложенных результатов изучаются аппроксимативные свойства сумм Фейера указанных рациональных рядов Фурье – Чебышева. Ставится задача получить аналоги теорем о равномерной сходимости последовательностей сумм Фейера для непрерывных на отрезке функций (см., например, [33]), а также исследовать приближения индивидуальных функций рассматриваемым методом суммирования.

Суммы Фейера рациональных рядов Фурье – Чебышева на отрезке и их аппроксимативные свойства

Пусть задана частичная сумма порядка $2n$ ряда Фурье по системе алгебраических дробей Чебышева – Маркова для четной функции $f \in C[-1, 1]$:

$$s_{2n}(f, x) = \frac{c_0}{2} + \sum_{k=0}^{n} c_{2k} M_{2k}(x), \; n = 0, 1, \ldots. \tag{1}$$

Аппроксимативные свойства частичных сумм (1) исследованы нами в [32]. Имеет место следующая теорема.

Теорема 1 [32]. *Для частичных сумм (1) справедливо представление*

$$s_{2n}(f, x) = \frac{1}{\pi} \int_{-\pi/2}^{\pi/2} f(\cos v) \frac{\sin\big((2n+1)\varphi(u, v)\big)}{\sin \varphi(u, v)} \lambda(v) dv, \;\; x = \cos u, \tag{2}$$

где

$$\varphi(u, v) = \int_{u}^{v} \lambda(y) dy, \;\; \lambda(y) = \frac{1 - \alpha^4}{1 + 2\alpha^2 \cos 2y + \alpha^4}, \;\; \alpha \in [0, 1), \tag{3}$$

причем оператор $s_{2n} : f \to \mathbb{R}_{2n}(a)$, где $\mathbb{R}_{2n}(a)$ – множество рациональных функций вида $\dfrac{p_{2n}(x)}{\big(1 + a^2 x^2\big)^n}$, $p_{2n}(x) \in \mathbb{P}_{2n}$, является точным на константах.

Составим среднее арифметическое частичных сумм (1)

$$\sigma_{2n}(f, x) = \frac{1}{n+1} \sum_{k=0}^{n} s_{2k}(f, x), \;\; x \in [-1, 1], \;\; n = 0, 1, \ldots. \tag{4}$$

Выражения (4) естественно назвать суммами Фейера рядов Фурье по системе алгебраических дробей Чебышева – Маркова.

Теорема 2. *Если функция f определена и абсолютно суммируема с весом*

$$\rho(x, a) = \frac{\sqrt{1 + a^2}}{\big(1 + a^2 x^2\big)\sqrt{1 - x^2}}, \;\; -1 < x < 1, \;\; a > 0,$$

на отрезке $[-1, 1]$, то для сумм Фейера справедливо представление

$$\sigma_{2n}(f, x) = \frac{1}{\pi(n+1)} \int_{-\pi/2}^{\pi/2} f(\cos v) \frac{\sin^2\big((n+1)\varphi(u, v)\big)}{\sin^2 \varphi(u, v)} \lambda(v) dv, \;\; x = \cos u, \tag{5}$$

здесь $\varphi(u, v)$, $\lambda(v)$ из (3). Кроме этого, оператор $\sigma_{2n} : f \to \mathbb{R}_{2n}(a)$ является положительным и точным для единицы.

Доказательство. Для доказательства первого утверждения теоремы подставим (2) в (4). Тогда для $n = 0, 1, \ldots$ получим

$$\sigma_{2n}(f, x) = \frac{1}{\pi(n+1)} \int_{-\pi/2}^{\pi/2} \frac{f(\cos v)}{\sin \varphi(u, v)} \sum_{k=0}^{n} \sin\big((2k+1)\varphi(u, v)\big) \lambda(v) dv.$$

Отсюда приходим к (5).

Второе утверждение теоремы следует из условия существования ряда Фурье по системе рациональных дробей Чебышева – Маркова для непрерывной на отрезке $[-1, 1]$ четной функции f, полученного в [32], а также представления (4).

Из (5) вытекает, что оператор $\sigma_{2n}(\cdot, x)$ положительный. Его точность на единице следует из точности на единице частичных сумм $s_{2n}(f, x)$ и соотношения (4), что и доказывает теорему 2.

Лемма 1. *Для сумм Фейера (4) имеет место представление*

$$\sigma_{2n}(f,\,x)=\frac{1}{\pi(n+1)\lambda(u)}\int\limits_{-\pi/2}^{\pi/2}f(\cos v)\frac{\sin^2\left((n+1)\varphi(u,\,v)\right)}{\sin^2(v-u)}dv,\quad x=\cos u, \tag{6}$$

здесь $\lambda(u)$ *из (3).*

Доказательство. Из [9, с. 14] следует, что для $\varphi(u,\,v)$ справедливо

$$\exp\left[in\varphi(u,\,v)\right]=\sqrt{\frac{\pi_n(\zeta)}{\pi_n(\xi)}},\ \pi_n(y)=\left(\frac{y^2+\alpha^2}{1+\alpha^2y^2}\right)^n,\ \xi=e^{iu},\ \zeta=e^{iv},\ x=\cos u.$$

Следовательно,

$$\sin^2\varphi(u,\,v)=\left(\frac{1}{2i}\right)^2\left[\sqrt{\frac{\zeta^2+\alpha^2}{1+\alpha^2\zeta^2}\frac{1+\alpha^2\xi^2}{\xi^2+\alpha^2}}-\sqrt{\frac{\xi^2+\alpha^2}{1+\alpha^2\xi^2}\frac{1+\alpha^2\zeta^2}{\zeta^2+\alpha^2}}\right]^2=\sin^2(v-u)\lambda(u)\lambda(v).$$

Подставив последнее выражение в (5), придем к (6), что и доказывает лемму 1.

Замечание 1. Положив в (6) $\alpha=0$, получим

$$\sigma_{2n}(f,\,x)=\frac{1}{\pi(n+1)}\int\limits_{-\pi/2}^{\pi/2}f(\cos v)\frac{\sin^2\left((n+1)(v-u)\right)}{\sin^2(v-u)}dv,\ x=\cos u.$$

Другими словами, при переходе к полиномиальному случаю выражение (6) представляет собой классические суммы Фейера рядов Фурье – Чебышева при условии четности функции *f*.

Равномерная сходимость сумм Фейера для непрерывных на отрезке $[-1,\,1]$ функций

Изучим поведение последовательности сумм Фейера (4) для $n\to\infty$ при приближении функций $f\in C[-1,\,1]$, а также определим достаточные условия, которым должен удовлетворять параметр α для равномерной сходимости этой последовательности.

Отметим, что в данном случае при каждом значении индекса *n* могут выбираться соответствующие значения параметра α, т. е., вообще говоря, $\alpha=\alpha_n$, $n=0,\,1,\,\ldots$. Это обстоятельство будем учитывать в дальнейшем.

Рассмотрим последовательность сумм Фейера

$$\left\{\sigma_{2n}(f,\,x,\,\alpha_n)\right\}_{n=0}^{n=+\infty}. \tag{7}$$

Теорема 3. *Для всякой четной функции* $f\in C[-1,\,1]$ *справедливо неравенство*

$$\left|f(x)-\sigma_{2n}(f,\,x,\,\alpha_n)\right|\le 4\left(\omega_f\left(\frac{\ln\left((n+1)\lambda(u)\right)\sqrt{1-x^2}}{(n+1)\lambda(u)}\right)+\omega_f\left(\frac{|x|}{(n+1)\lambda(u)}\right)\right), \tag{8}$$

где ω_f *– модуль непрерывности функции f на отрезке* $[-1,\,1]$, $\lambda(u)$ *из (3),* $x=\cos u$, $u\in[0,\,\pi]$.

Доказательство. Воспользуемся представлением (6). Из π-периодичности подынтегральной функции следует, что

$$\sigma_{2n}(f,\,x)=\frac{1}{\pi(n+1)\lambda(u)}\int\limits_{|v-u|\le\pi/2}f(\cos v)\frac{\sin^2\left((n+1)\varphi(u,\,v)\right)}{\sin^2(v-u)}dv,\ x=\cos u.$$

Учитывая точность оператора $\sigma_{2n}(\cdot,\,x)$ для единицы, из последнего соотношения получим

$$f(x)-\sigma_{2n}(f,\,x,\,\alpha_n)=\frac{1}{\pi(n+1)\lambda(u)}\int\limits_{|v-u|\le\pi/2}\left(f(\cos u)-f(\cos v)\right)\frac{\sin^2\left((n+1)\varphi(u,\,v)\right)}{\sin^2(v-u)}dv. \tag{9}$$

После замены переменной $v - u = t$ имеем

$$f(x) - \sigma_{2n}(f, x, \alpha_n) = \frac{1}{\pi(n+1)\lambda(u)}\left(\int\limits_{-\pi/2}^{0} + \int\limits_{0}^{\pi/2}\right)\left(f(\cos u) - f\left(\cos(u+t)\right)\right)\frac{\sin^2\left((n+1)\varphi(+u, t)\right)}{\sin^2 t}dt,$$

где

$$\varphi(+u, t) = \int\limits_{0}^{t}\frac{1-\alpha^4}{1+2\alpha^2\cos 2(y+u)+\alpha^4}dy,\ \alpha \in [0, 1).$$

Выполнив в первом интеграле еще одну замену: $t \sim -t$, приходим к выражению

$$f(x) - \sigma_{2n}(f, x, \alpha_n) = \frac{1}{\pi}\left[I_n(u) + I_n(-u)\right], \tag{10}$$

где

$$I_n(\pm u) = \frac{1}{(n+1)\lambda(u)}\int\limits_{0}^{\pi/2}\left(f(\cos u) - f\left(\cos(u \pm t)\right)\right)\frac{\sin^2\left((n+1)\varphi(\pm u, t)\right)}{\sin^2 t}dt.$$

Для дальнейших рассуждений воспользуемся методом А. Ф. Тимана [34, с. 269]. Заметив, что

$$\left|f(\cos u) - f\left(\cos(u \pm t)\right)\right| \le \omega_f\left(\left|\sin t \sin u\right|\right) + 2\omega_f\left(\left|\sin^2\frac{t}{2}\cos u\right|\right),$$

имеем

$$\left|I_n(\pm u)\right| \le \frac{1}{(n+1)\lambda(u)}\left[\int\limits_{0}^{\pi/2}\omega_f\left(\left|\sin t \sin u\right|\right)\frac{\sin^2\left((n+1)\varphi(\pm u, t)\right)}{\sin^2 t}dt + \right.$$

$$\left. + 2\int\limits_{0}^{\pi/2}\omega_f\left(\left|\sin^2\frac{t}{2}\cos u\right|\right)\frac{\sin^2\left((n+1)\varphi(\pm u, t)\right)}{\sin^2 t}dt\right] \le$$

$$\le \frac{1}{(n+1)\lambda(u)}\left[\omega_f\left(\frac{\ln\left((n+1)\lambda(u)\right)\sin u}{(n+1)\lambda(u)}\right)\int\limits_{0}^{\pi/2}\left(\sin t\frac{(n+1)\lambda(u)}{\ln\left((n+1)\lambda(u)\right)}+1\right)\frac{\sin^2\left((n+1)\varphi(\pm u, t)\right)}{\sin^2 t}dt + \right.$$

$$\left. + 2\omega_f\left(\frac{\cos u}{(n+1)\lambda(u)}\right)\int\limits_{0}^{\pi/2}\left(\sin^2\frac{t}{2}(n+1)\lambda(u)+1\right)\frac{\sin^2\left((n+1)\varphi(\pm u, t)\right)}{\sin^2 t}dt\right] \le$$

$$\le \omega_f\left(\frac{\ln\left((n+1)\lambda(u)\right)\sin u}{(n+1)\lambda(u)}\right)(I_1 + I_2) + 2\omega_f\left(\frac{\cos u}{(n+1)\lambda(u)}\right)(I_3 + I_2), \tag{11}$$

где

$$I_1 = \frac{1}{\ln\left((n+1)\lambda(u)\right)}\int\limits_{0}^{\pi/2}\frac{\sin^2\left((n+1)\varphi(\pm u, t)\right)}{\sin t}dt;$$

$$I_2 = \frac{1}{(n+1)\lambda(u)}\int\limits_{0}^{\pi/2}\frac{\sin^2\left((n+1)\varphi(\pm u, t)\right)}{\sin^2 t}dt;$$

$$I_3 = \frac{1}{4}\int\limits_{0}^{\pi/2}\frac{\sin^2\left((n+1)\varphi(\pm u, t)\right)}{\cos^2\frac{t}{2}}dt.$$

Учитывая, что $\varphi(\pm u, t) \le \lambda(u)t$, $0 < t < \dfrac{\pi}{2}$, разобьем интеграл I_1 на два интеграла по промежуткам $\left[0, \dfrac{\pi}{2(n+1)\lambda(u)}\right]$ и $\left[\dfrac{\pi}{2(n+1)\lambda(u)}, \dfrac{\pi}{2}\right]$. Применив к первому из них неравенства $\sin\big((n+1)\lambda(u)t\big) \le$ $\le (n+1)\lambda(u)t$, $\sin t \ge \dfrac{2}{\pi}t$, а ко второму – неравенство $\sin\big((n+1)\lambda(u)t\big) \le 1$, получим

$$I_1 \le \left(\frac{\pi}{2}\right)^3 \frac{1}{2\ln\big((n+1)\lambda(u)\big)} + \frac{\pi}{2}. \tag{12}$$

Рассуждая аналогично, оценим интегралы I_2 и I_3:

$$I_2 \le \frac{\pi}{2} + \frac{2}{\pi}, \tag{13}$$

$$I_3 \le \frac{1}{2}. \tag{14}$$

Подставив (12)–(14) в (11), имеем

$$\big|I_n(\pm u)\big| \le \omega_f\left(\frac{\ln\big((n+1)\lambda(u)\big)\sin u}{(n+1)\lambda(u)}\right)\left[\frac{1}{2\ln\big((n+1)\lambda(u)\big)}\left(\frac{\pi}{2}\right)^3 + \pi + \frac{2}{\pi}\right] +$$

$$+ \omega_f\left(\frac{\cos u}{(n+1)\lambda(u)}\right)\left[1 + \pi + \frac{4}{\pi}\right].$$

С учетом последнего соотношения при достаточно больших n в (10) получим

$$\big|f(x) - \sigma_{2n}(f, x, \alpha_n)\big| \le \omega_f\left(\frac{\ln\big((n+1)\lambda(u)\big)\sin u}{(n+1)\lambda(u)}\right)\left[\frac{\pi^2}{8} + 2 + \frac{4}{\pi^2}\right] +$$

$$+ \omega_f\left(\frac{\cos u}{(n+1)\lambda(u)}\right)\left[\frac{2}{\pi} + 2 + \frac{8}{\pi^2}\right].$$

Воспользовавшись свойством модуля непрерывности, а также вычислив

$$\frac{\pi^2}{8} + 2 + \frac{4}{\pi^2} \approx 3{,}63, \quad \frac{2}{\pi} + 2 + \frac{8}{\pi^2} \approx 3{,}44,$$

приходим к оценке (8). Теорема 3 доказана.

Следствие 1. *Если выполняется условие*

$$\lim_{n\to\infty} \frac{n+1}{\ln(n+1)}(1 - \alpha_n) = \infty, \tag{15}$$

то последовательность сумм Фейера (7) сходится к $f(x)$ *равномерно на всем отрезке* $[-1, 1]$.

Заметим, что здесь и далее для каждого индекса n может выбираться соответствующее α_n. Мы не будем указывать эту зависимость, так как все приведенные оценки являются равномерными относительно $\alpha \in [0, 1)$.

О приближениях функции $|x|^s$ суммами Фейера

Следующий этап наших исследований – изучение приближений функции $|x|^s$ на отрезке $[-1, 1]$ суммами Фейера. Введем обозначения

$$\varepsilon_{2n}(x, \alpha) = |x|^s - \sigma_{2n}\big(|\cdot|^s, x\big), \ x \in [-1, 1],$$

$$\varepsilon_{2n}(\alpha) = \Big\| |x|^s - \sigma_{2n}\big(|\cdot|^s, x\big)\Big\|_{C[-1, 1]}, \ n \in \mathbb{N}.$$

Теорема 4. *Для приближений функции* $|x|^s$, $0 < s < 2$, *на отрезке* $[-1, 1]$ *суммами Фейера (4) справедливы следующие соотношения:*

1) $$\varepsilon_{2n}(x, \alpha) = \frac{1}{2^{s-2}\pi(n+1)\lambda(u)} \sin\frac{\pi s}{2} \int_0^1 \frac{(1-t^2)^s t^{1-s}\left[(-1)^n \chi_{n+1}(t)\cos\psi_{n+1}(u, t) + \cos\left(2\arg\frac{\xi}{1+t^2\xi^2}\right)\right]}{1 + 2t^2\cos 2u + t^4} dt, \quad (16)$$

где

$$\psi_{n+1}(u, t) = 2\arg\frac{\xi}{1+t^2\xi^2} + (n+1)\arg\frac{\xi^2+\alpha^2}{1+\alpha^2\xi^2}, \ \xi = e^{iu}, \ x = \cos u;$$

2) $$|\varepsilon_{2n}(x, \alpha)| \le \frac{1}{2^{s-2}\pi(n+1)\lambda(u)} \sin\frac{\pi s}{2} \int_0^1 \frac{(1-t^2)^s t^{1-s}\sqrt{1 + 2(-1)^n \chi_{n+1}(t)\cos\gamma_{n+1}(u) + \chi_{n+1}^2(t)}}{1 + 2t^2\cos 2u + t^4} dt, \quad (17)$$

где

$$\chi_{n+1}(t) = \left(\frac{t^2-\alpha^2}{1-\alpha^2 t^2}\right)^{n+1}, \ \gamma_{n+1}(u) = (n+1)\arg\frac{\xi^2+\alpha^2}{1+\alpha^2\xi^2}, \ \xi = e^{iu}, \ x = \cos u, \ x \in [-1, 1];$$

3) $$\varepsilon_{2n}(\alpha) \le \varepsilon_{2n}^*(\alpha), \quad (18)$$

где

$$\varepsilon_{2n}^*(\alpha) = \frac{1}{2^{s-2}\pi(n+1)} \sin\frac{\pi s}{2}\left(I_1(\alpha, n) + I_2(\alpha, n)\right); \quad (19)$$

$$I_1(\alpha, n) = \frac{1-\alpha^2}{1+\alpha^2} \int_\alpha^1 (1-t^2)^{s-1} t^{1-s} \frac{1-\chi_{n+1}(t)}{1-t^2} dt;$$

$$I_2(\alpha, n) = \frac{1+\alpha^2}{1-\alpha^2} \int_0^\alpha \frac{(1-t^2)^s t^{1-s}}{(1+t^2)^2}\left(1 - |\chi_{n+1}(t)|\right) dt, \ \alpha \in [0, 1), \ n \in \mathbb{N}.$$

Неравенство (17) является точным. Равенство достигается в точке $x = 0$, *что соответствует значению параметра* $u = \frac{\pi}{2}$, *а также на концах отрезка, где* $u = 0$.

Доказательство. Вывод интегрального представления (16) и оценки (17) опустим в связи с тем, что он аналогичен таковому соответствующего результата в [36].

Для доказательства точности оценки (17) положим в ней $x = 0$ и $x = \pm 1$. Тогда

$$|\varepsilon_{2n}(0, \alpha)| \le \frac{1}{2^{s-2}\pi(n+1)} \sin\frac{\pi s}{2} \frac{1-\alpha^2}{1+\alpha^2} \int_0^1 (1-t^2)^{s-1} t^{1-s} \frac{1-\chi_{n+1}(t)}{1-t^2} dt,$$

$$|\varepsilon_{2n}(\pm 1, \alpha)| \le \frac{1}{2^{s-2}\pi(n+1)} \sin\frac{\pi s}{2} \frac{1+\alpha^2}{1-\alpha^2} \int_0^1 \frac{(1-t^2)^s t^{1-s}}{(1+t^2)^2}\left(1 - |\chi_{n+1}(t)|\right) dt.$$

Подставив аналогичные значения в соотношение (16), видим, что последние неравенства обращаются в равенства.

Для доказательства оценки (18) исследуем величину $\varepsilon_{2n}(\alpha)$. Имеем

$$\varepsilon_{2n}(\alpha) = \max_{x\in[-1,1]} |\varepsilon_{2n}(x, \alpha)| = \max_{x\in[-1,1]} \left||x|^s - \sigma_{2n}(|\cdot|^s, x)\right| = \max_{x\in[-1,1]} \left|\frac{1}{n+1}\sum_{k=0}^n \delta_{2k}(x, \alpha)\right|, \quad (20)$$

где $\delta_{2k}(x, \alpha) = |x|^s - s_{2k}(|\cdot|^s, x)$, $k = 0, \ldots, n$, – приближения функции $|x|^s$, $0 < s < 2$, на отрезке $[-1, 1]$ частичными суммами (2). Для величины $\delta_{2k}(x, \alpha)$ справедливо представление [36]

$$\delta_{2k}(x, \alpha) = \frac{(-1)^k}{\pi \cdot 2^{s-2}} \sin\frac{\pi s}{2} \int_0^1 \frac{\left(1-t^2\right)^s t^{1-s}}{1-\alpha^2 t^2} \sqrt{\frac{1+2\alpha^2\cos 2u+\alpha^4}{1+2t^2\cos 2u+t^4}}\, \chi_k(t)\cos\eta_k\, dt,$$

$$\eta_k = \eta_k(x, t, \alpha) = \arg\frac{\xi^2+\alpha^2}{1+t^2\xi^2} + k\arg\frac{\xi^2+\alpha^2}{1+\alpha^2\xi^2}.$$

Подставив представление для $\delta_{2k}(x, \alpha)$ в (20), имеем

$$\varepsilon_{2n}(\alpha) = \frac{1}{2^{s-2}\pi(n+1)} \sin\frac{\pi s}{2} \max_{x\in[-1,1]} \left| \int_0^1 \frac{\left(1-t^2\right)^s t^{1-s}}{1-\alpha^2 t^2} \sqrt{\frac{1+2\alpha^2\cos 2u+\alpha^4}{1+2t^2\cos 2u+t^4}} \sum_{k=0}^{n} (-1)^k \cos\eta_k(x, t, \alpha)\chi_k(t)dt \right| \le$$

$$\le \frac{1}{2^{s-2}\pi(n+1)} \sin\frac{\pi s}{2} \max_{x\in[-1,1]} I_3(x, \alpha),$$

здесь

$$I_3(x, \alpha) = \int_0^1 \frac{\left(1-t^2\right)^s t^{1-s}}{1-\alpha^2 t^2} \sqrt{\frac{1+2\alpha^2\cos 2u+\alpha^4}{1+2t^2\cos 2u+t^4}} \sum_{k=0}^{n} \left|\chi_k(t)\right| dt, \;\; x = \cos u.$$

Учитывая, что $\cos 2u = 2x^2 - 1$, из последнего соотношения получим

$$I_3(x, \alpha) = \left(1-\alpha^2\right) \int_0^1 \frac{\left(1-t^2\right)^s t^{1-s}}{1-\alpha^2 t^2} \sqrt{\frac{1+A^2x^2}{1+T^2x^2}} \sum_{k=0}^{n} \left|\chi_k(t)\right| dt, \tag{21}$$

где $A = \dfrac{2\alpha}{1-\alpha^2}$; $T = \dfrac{2t}{1-t^2}$. Исследовав функцию $\gamma(x) = \sqrt{\dfrac{1+A^2x^2}{1+T^2x^2}}$ по аналогии с [32], заключаем, что при $0 < t < \alpha$ она возрастает, а значит, достигает максимального значения при $x = 1$, что соответствует значению параметра $u = 0$. В то же время при $\alpha < t < 1$ функция $\gamma(x)$ убывает и, следовательно, максимальное ее значение будет уже при $x = 0$, что соответствует значению параметра $u = \dfrac{\pi}{2}$. Разбивая интеграл в правой части (21) на два интеграла по промежуткам $[0, \alpha]$ и $[\alpha, 1]$, найдем

$$\varepsilon_{2n}(\alpha) \le \frac{1}{2^{s-2}\pi(n+1)} \sin\frac{\pi s}{2} \left[\left(1-\alpha^2\right) \int_\alpha^1 \frac{\left(1-t^2\right)^{s-1} t^{1-s}}{1-\alpha^2 t^2} \sum_{k=0}^{n} \chi_k(t)\, dt + \right.$$

$$\left. + \left(1+\alpha^2\right) \int_0^\alpha \frac{\left(1-t^2\right)^s t^{1-s}}{\left(1-\alpha^2 t^2\right)\left(1+t^2\right)} \sum_{k=0}^{n} \left|\chi_k(t)\right| dt \right].$$

Заметив, что суммы в каждом из интегралов представляют собой суммы членов геометрических прогрессий с соответствующими знаменателями, придем к оценке (18).

Неравенства (17) и (18) получены в предположении, что $x \in (0, 1)$. Из приведенных выше рассуждений вытекает, что они также будут верны и на промежутке $(-1, 0)$. Справедливость неравенства (17) в точках $x = \pm 1$, а также при $x = 0$ следует из непрерывности левой и правой частей этого неравенства относительно переменной x на $[-1, 1]$. Теорема 4 доказана полностью.

Замечание 2. Учитывая, что $\cos\gamma_{n+1}(u) = M_{2n+2}(x)$, $x = \cos u$, есть алгебраическая дробь Чебышева – Маркова порядка $2n + 2$, оценку (17) можно переписать в виде

$$\left|\varepsilon_{2n}(x, \alpha)\right| \le \frac{1}{2^{s-2}\pi(n+1)\lambda(u)} \sin\frac{\pi s}{2} \int_0^1 \frac{\left(1-t^2\right)^s t^{1-s}}{1+2t^2\cos 2u+t^4} \sqrt{1+(-1)^n 2\chi_{n+1}(t) M_{2n+2}(x) + \chi_{n+1}^2(t)}\, dt.$$

Исследование приближений функции $|x|^s$, $0 < s < 2$, суммами Фейера в полиномиальном случае

В формулировке теоремы 4 положим $\alpha = 0$. Тогда $\varepsilon_{2n}(x, 0) = \varepsilon_{2n}(x)$ и $\varepsilon_{2n}(0) = \varepsilon_{2n}$ есть соответственно поточечные и равномерные приближения функции $|x|^s$, $0 < s < 2$, на отрезке $[-1, 1]$ суммами Фейера рядов Фурье по системе многочленов Чебышева первого рода $T_{2n}(x)$. В этом случае

$$|\varepsilon_{2n}(x)| \le \frac{1}{2^{s-2}\pi(n+1)} \sin\frac{\pi s}{2} \int_0^1 \frac{(1-t^2)^s t^{1-s} \sqrt{1+(-1)^n 2t^{2n+2} T_{2n+2}(x) + t^{4n+4}}}{1+2t^2\cos 2u + t^4} dt, \ x \in [-1, 1], \tag{22}$$

$$\varepsilon_{2n} \le \frac{1}{2^{s-2}\pi(n+1)} \sin\frac{\pi s}{2} \int_0^1 (1-t^2)^{s-1} t^{1-s} \frac{1-t^{2n+2}}{1-t^2} dt, \ n \in \mathbb{N}. \tag{23}$$

Оценка (22) точна. Равенство достигается при $x = 0$, а также на концах отрезка.

Поскольку

$$|\varepsilon_{2n}(0)| = \frac{1}{2^{s-2}\pi(n+1)} \sin\frac{\pi s}{2} \int_0^1 (1-t^2)^{s-1} t^{1-s} \frac{1-t^{2n+2}}{1-t^2} dt, \ n \in \mathbb{N},$$

заключаем, что в (23) имеет место знак равенства, т. е.

$$\varepsilon_{2n} = \frac{1}{2^{s-2}\pi(n+1)} \sin\frac{\pi s}{2} \int_0^1 (1-t^2)^{s-1} t^{1-s} \frac{1-t^{2n+2}}{1-t^2} dt, \ n \in \mathbb{N}. \tag{24}$$

Представляет интерес найти асимптотическую оценку равномерных приближений функции $|x|^s$, $0 < s < 2$, суммами Фейера полиномиальных рядов Фурье – Чебышева.

Теорема 5. *Для равномерных приближений функции $|x|^s$, $0 < s < 2$, суммами Фейера полиномиальных рядов Фурье – Чебышева при $n \to \infty$ имеют место асимптотические равенства*

$$\varepsilon_{2n} \sim \frac{1}{\pi} \begin{cases} \dfrac{1}{2^{s-1}(1-s)} \sin\dfrac{\pi s}{2} \dfrac{\Gamma(s)}{(n+1)^s}, & s \in (0, 1), \\ \dfrac{\ln(n+1)}{n+1}, & s = 1, \\ \dfrac{1}{2^{s-1}(1-s)} \sin\dfrac{\pi s}{2} \dfrac{\Gamma(s)\Gamma\left(1-\frac{s}{2}\right)}{\Gamma\left(\frac{s}{2}\right)(n+1)}, & s \in (1, 2), \end{cases} \tag{25}$$

где $\Gamma(s)$ – гамма-функция Эйлера.

Доказательство. Исследуем интеграл в (24):

$$I_4 = \int_0^1 \left(\frac{1-t^2}{t}\right)^{s-1} \frac{1-t^{2n+2}}{1-t^2} dt, \quad 0 < s < 2, n \in \mathbb{N}.$$

Рассмотрим случай $s \in (0, 1]$. Воспользуемся методом, предложенным в [37]. Продифференцируем последний интеграл по параметру n. Имеем

$$\frac{\partial I_4}{\partial n} = -2\int_0^1 \left(\frac{1-t^2}{t}\right)^{s-1} \frac{\ln t}{1-t^2} e^{(2n+2)\ln t} dt, \ 0 < s \le 1, \ n \in \mathbb{N}.$$

Для исследования асимптотического поведения интеграла справа применим метод Лапласа [38–40]. Функция $\ln t$ возрастает в промежутке $0 < t < 1$, следовательно, достигает своего максимального значения при $t = 1$. Используя разложение $\ln t = (t-1) + o(t-1)$ и асимптотическое соотношение

$$\left(\frac{1-t^2}{t}\right)^{s-1} \frac{\ln t}{1-t^2} \sim -2^{s-2}(1-t)^{s-1},$$

справедливые при $t \to 1$, находим, что при достаточно малом $\varepsilon > 0$ и $n \to \infty$

$$\frac{\partial I_4}{\partial n} \sim 2^{s-1} \int\limits_{1-\varepsilon}^{1} (1-t)^{s-1} e^{(2n+2)(t-1)}\, dt.$$

В последнем интеграле выполним замену $1 - t = u$. Тогда

$$\frac{\partial I_4}{\partial n} \sim 2^{s-1} \int\limits_{0}^{\varepsilon} u^{s-1} e^{-(2n+2)u} du, \;\; n \to \infty.$$

Положив в интеграле справа $(2n+2)u = t$, получим

$$\frac{\partial I_4}{\partial n} \sim \frac{\Gamma(s)}{2(n+1)^s}, \;\; n \to \infty.$$

Чтобы прийти к асимптотике интеграла I_4, необходимо в последнем асимптотическом равенстве произвести интегрирование по параметру n. В итоге при $n \to \infty$ для интеграла I_4 имеем

$$I_4 \sim \begin{cases} \dfrac{\Gamma(s)}{2(1-s)(n+1)^{s-1}} + C_1, \;\; s \in (0,1), \\ \dfrac{1}{2}\ln(n+1), \;\; s = 1, \end{cases} \tag{26}$$

где C_1 – некоторая константа, не зависящая от n.

Пусть теперь $s \in (1, 2)$. Тогда

$$I_4 = \int\limits_0^1 \frac{\left(1-t^2\right)^{s-2}}{t^{s-1}} dt - \int\limits_0^1 \frac{\left(1-t^2\right)^{s-2}}{t^{s-1}} t^{2n+2}\, dt.$$

Применяя те же методики исследования, найдем

$$I_4 \sim \frac{\Gamma(s)\Gamma\left(1-\frac{s}{2}\right)}{2(s-1)\Gamma\left(\frac{s}{2}\right)} + \frac{\Gamma(s)}{2(s-1)(n+1)^{s-1}}, \;\; s \in (1, 2), \;\; n \to \infty. \tag{27}$$

Из (26) и (27) следует (25). Теорема 5 доказана.

Асимптотика мажоранты равномерных приближений функции $|x|^s$, $0 < s < 2$, рациональными суммами Фейера

На этом этапе исследования найдем асимптотическое выражение при $n \to \infty$ величины (19). С этой целью в интегралах $I_1(\alpha, n)$ и $I_2(\alpha, n)$ выполним замену переменной интегрирования: $t^2 = \dfrac{1-u}{1+u}$, $dt = -\dfrac{du}{(1+u)\sqrt{1-u^2}}$. Тогда

$$\varepsilon_{2n}^*(\alpha) = \frac{1}{\pi(n+1)} \sin\frac{\pi s}{2}\left[I_1(\alpha, n) + I_2(\alpha, n)\right], \tag{28}$$

где

$$I_1(\alpha, n) = \beta\int_0^{\beta} \frac{u^{s-1}}{\left(1-u^2\right)^{s/2}}\left(1-\left(\frac{\beta-u}{\beta+u}\right)^{n+1}\right)\frac{du}{u};$$

$$I_2(\alpha, n) = \frac{1}{\beta}\int_{\beta}^{1} \frac{u^{s}}{\left(1-u^2\right)^{s/2}}\left(1-\left(\frac{u-\beta}{\beta+u}\right)^{n+1}\right)du,\ \beta = \frac{1-\alpha^2}{1+\alpha^2}.$$

Теорема 6. *Для мажоранты* $\varepsilon^*_{2n}(\alpha)$ *равномерных приближений функции* $|x|^s$, $0 < s < 2$, *на отрезке* $[-1, 1]$ *суммами Фейера рядов Фурье по системе алгебраических дробей Чебышева – Маркова при* $n \to \infty$ *справедливы асимптотические равенства*

$$\varepsilon^*_{2n}(\alpha) \sim \frac{1}{\pi}\sin\frac{\pi s}{2}\begin{cases}\left(\frac{\beta}{2}\right)^s \frac{2\Gamma(s)}{(1-s)(n+1)^s} + \nu_n(\beta, s), & s \in (0, 1),\\ \beta\frac{\ln(n+1)}{n+1} + \nu_n(\beta, 1), & s = 1,\\ \frac{\beta}{n+1}\int_0^{\beta}\frac{u^{s-2}du}{\left(1-u^2\right)^{s/2}} + \nu_n(\beta, s), & s \in (1, 2),\end{cases} \tag{29}$$

где

$$\nu_n(\beta, s) = \frac{1}{\beta(n+1)}\int_0^{\arccos\beta}\cos^s\theta\sin^{1-s}\theta d\theta - \frac{1}{2}\Gamma\left(1-\frac{s}{2}\right)\left(\frac{1-\beta}{1+\beta}\right)^{n+1}\frac{\left(1-\beta^2\right)^{1-\frac{s}{2}}}{\left(\beta(n+1)\right)^{2-\frac{s}{2}}}.$$

Доказательство. Исследуем каждый из интегралов, входящих в (28), по отдельности. Изучим их асимптотическое поведение при $n \to \infty$. Дальнейшему доказательству теоремы 6 предпошлем две леммы.

Лемма 2. *Справедливы асимптотические равенства* ($n \to \infty$)

$$I_1(\alpha, n) \sim \begin{cases}\left(\frac{\beta}{2}\right)^s \frac{2\Gamma(s)}{(1-s)(n+1)^{1-s}}, & s \in (0, 1),\\ \beta\ln(n+1), & s = 1,\\ \beta\int_0^{\beta}\frac{u^{s-2}du}{\left(1-u^2\right)^{s/2}} - \beta\left(\frac{\beta}{2(n+1)}\right)^{s-1}\Gamma(s-1), & s \in (1, 2).\end{cases} \tag{30}$$

Доказательство. Пусть $s \in (0, 1]$. Тогда воспользуемся методиками исследования подобных интегралов, изложенными в [37]. Продифференцируем интеграл $I_1(\alpha, n)$ по параметру n:

$$\frac{\partial I_1(\alpha, n)}{\partial n} = -\beta\int_0^{\beta}\frac{u^{s-1}}{\left(1-u^2\right)^{s/2}}\left(\frac{\beta-u}{\beta+u}\right)^{n+1}\ln\frac{\beta-u}{\beta+u}\cdot\frac{du}{u}.$$

Для изучения асимптотического поведения интеграла справа в последнем соотношении применим метод Лапласа (см., например, [38–40]). Перепишем интеграл в виде

$$\frac{\partial I_1(\alpha, n)}{\partial n} = -\beta\int_0^{\beta} f(u)e^{(n+1)S(u)}du,\ S(u) = \ln\frac{\beta-u}{\beta+u},\ f(u) = \frac{u^{s-1}}{\left(1-u^2\right)^{s/2}}\left(\frac{1}{u}\ln\frac{\beta-u}{\beta+u}\right).$$

Функция $S(u)$ убывает на промежутке $0 < u < \beta$, $0 < \beta \le 1$, поскольку $S'(u) < 0$, и, следовательно, достигает своего максимального значения при $u = 0$. Используя разложения $S(u) = \dfrac{-2u}{\beta} + o(u)$, а также $f(u) \sim -\dfrac{2}{\beta} u^{s-1}$, справедливые при $u \to 0$, для бесконечно малого $\varepsilon > 0$ и $n \to \infty$ получим

$$\frac{\partial I_1(\alpha, n)}{\partial n} \sim 2\int_0^{\varepsilon} u^{s-1} e^{-2(n+1)u/\beta}\, du.$$

Выполнив замену $\dfrac{2(n+1)u}{\beta} = v$ в интеграле справа, имеем

$$\frac{\partial I_1(\alpha, n)}{\partial n} \sim 2\left(\frac{\beta}{2(n+1)}\right)^s \int_0^{\frac{2(n+1)\varepsilon}{\beta}} v^{s-1} e^{-v} dv \sim 2\left(\frac{\beta}{2(n+1)}\right)^s \Gamma(s),\ n \to \infty.$$

После интегрирования в последнем соотношении по параметру n найдем

$$I_1(\alpha, n) \sim \begin{cases} \left(\dfrac{\beta}{2}\right)^s \dfrac{2\Gamma(s)}{(1-s)(n+1)^{1-s}},\ s \in (0, 1), \\ \beta \ln(n+1),\ s = 1,\ n \to \infty. \end{cases} \tag{31}$$

Пусть теперь $s \in (1, 2)$. В этом случае

$$I_1(\alpha, n) = \beta\int_0^{\beta} \frac{u^{s-2}}{\left(1-u^2\right)^{s/2}} du - \beta\int_0^{\beta} \frac{u^{s-2}}{\left(1-u^2\right)^{s/2}} \left(\frac{\beta - u}{\beta + u}\right)^{n+1} du.$$

В правой части последнего равенства первый интеграл не зависит от n. Применив для исследования второго интеграла соответствующие методики нахождения асимптотик, получим

$$I_1(\alpha, n) \sim \beta\int_0^{\beta} \frac{u^{s-2}}{\left(1-u^2\right)^{s/2}} du - \beta\left(\frac{\beta}{2(n+1)}\right)^{s-1} \Gamma(s-1),\ 1 < s < 2,\ n \to \infty. \tag{32}$$

Из асимптотических соотношений (31) и (32) приходим к (30). Лемма 2 доказана.

Перейдем к рассмотрению интеграла $I_2(\alpha, n)$.

Лемма 3. *Справедливо асимптотическое равенство*

$$I_2(\alpha, n) \sim \frac{1}{\beta} \int_0^{\arccos\beta} \cos^s\theta \sin^{1-s}\theta d\theta - \frac{1}{2\beta}\Gamma\left(1 - \frac{s}{2}\right)\left(\frac{1-\beta}{1+\beta}\right)^{n+1} \left(\frac{1-\beta^2}{\beta(n+1)}\right)^{1-\frac{s}{2}},\ n \to \infty. \tag{33}$$

Доказательство. В интеграле $I_2(\alpha, n)$ выполним замену переменной по формуле $u = \cos\theta$. Тогда

$$I_2(\alpha, n) = \frac{1}{\beta} \int_0^{\arccos\beta} \cos^s\theta \sin^{1-s}\theta d\theta - \frac{1}{\beta} \int_0^{\arccos\beta} \cos^s\theta \sin^{1-s}\theta \left(\frac{\cos\theta - \beta}{\cos\theta + \beta}\right)^{n+1} d\theta. \tag{34}$$

Первый интеграл в последнем соотношении не зависит от n и существует при любых $0 < \beta \le 1$, $0 < s < 2$. Для исследования асимптотического поведения при $n \to \infty$ второго интеграла воспользуемся методом Лапласа. Запишем

$$I_5(\alpha, n) = \frac{1}{\beta} \int_0^{\arccos\beta} \cos^s\theta \sin^{1-s}\theta \left(\frac{\cos\theta - \beta}{\cos\theta + \beta}\right)^{n+1} d\theta = \frac{1}{\beta} \int_0^{\arccos\beta} f(\theta) e^{(n+1)S(\theta)} d\theta,$$

где $f(\theta) = \cos^s\theta \sin^{1-s}\theta$, $S(\theta) = \ln\dfrac{\cos\theta - \beta}{\cos\theta + \beta}$. Функция $S(\theta)$ убывает на промежутке $0 < \theta < \arccos\beta$, $0 < \beta \le 1$, поскольку $S'(\theta) < 0$, и, следовательно, достигает своего максимального значения при $\theta = 0$. Используя разложение

$$S(\theta) = \ln\frac{1-\beta}{1+\beta} - \frac{\beta}{1-\beta^2}\theta^2 + o\left(\theta^2\right),$$

а также асимптотическое равенство $f(\theta) \sim \theta^{1-s}$, справедливые при $\theta \to 0$, для бесконечно малого $\varepsilon > 0$ и $n \to \infty$ находим, что

$$I_5(\alpha, n) \sim \frac{1}{\beta}\left(\frac{1-\beta}{1+\beta}\right)^{n+1} \int_0^{\varepsilon} \theta^{1-s} e^{-\frac{(n+1)\beta}{1-\beta^2}\theta^2} d\theta.$$

После замены переменной по формуле $\frac{(n+1)\beta}{1-\beta^2}\theta^2 = u^2$ придем к соотношению

$$I_5(\alpha, n) \sim \frac{1}{\beta}\left(\frac{1-\beta}{1+\beta}\right)^{n+1}\left(\frac{1-\beta^2}{\beta(n+1)}\right)^{1-\frac{s}{2}} \int_0^{\varphi(\varepsilon, n)} u^{1-s}e^{-u^2}du,\ n \to \infty,$$

где $\varphi(\varepsilon, n) = \varepsilon\sqrt{\frac{(n+1)\beta}{1-\beta^2}} \to \infty$, $n \to \infty$. Учитывая, что

$$\int_0^{+\infty} u^{1-s}e^{-u^2}du = -\frac{s}{4}\Gamma\left(-\frac{s}{2}\right) = \frac{1}{2}\Gamma\left(1-\frac{s}{2}\right),$$

окончательно имеем

$$I_5(\alpha, n) \sim \frac{1}{2\beta}\Gamma\left(1-\frac{s}{2}\right)\left(\frac{1-\beta}{1+\beta}\right)^{n+1}\left(\frac{1-\beta^2}{\beta(n+1)}\right)^{1-\frac{s}{2}},\ n \to \infty.$$

Подставив асимптотическое выражение интеграла $I_5(\alpha, n)$ в соотношение (34), придем к (33). Лемма 3 доказана.

Объединяя в (28) соотношения (30) и (33), получим асимптотические равенства (29), что доказывает теорему 6.

Следствие 2. *Положив в формулировке теоремы 6 значение* α *равным нулю, приходим к асимптотическим равенствам (25). Значит, в полиномиальном случае оценка (18) точна. Равенство достигается при* $x = 0$.

Представляет интерес минимизировать правую часть соотношения (29) посредством выбора оптимального для этой задачи параметра $\beta = \beta^*$. Другими словами, искать оценку наилучшего равномерного приближения функции $|x|^s$, $0 < s < 2$, суммами Фейера (4). С этой целью положим

$$\varepsilon_{2n} = \inf_{0 \le \alpha < 1} \varepsilon_{2n}(\alpha),\ \varepsilon^*_{2n} = \inf_{0 \le \alpha < 1} \varepsilon^*_{2n}(\alpha).$$

Очевидно следующее соотношение (см. (18)):

$$\varepsilon_{2n} \le \varepsilon^*_{2n},\ n \in \mathbb{N}.$$

Теорема 7. *Для заданного* $0 < s < 2$ *при* $n \to \infty$ *справедливы асимптотические равенства*

$$\varepsilon^*_{2n} \sim \begin{cases} 2^{1-s}\Gamma(s)\dfrac{1+s}{s}\left[\left(\dfrac{1}{\pi}\sin\dfrac{\pi s}{2}\right)^{2s+1}\dfrac{s}{1-s}\Gamma^{2s}\left(1-\dfrac{s}{2}\right)\right]^{\frac{1}{s+1}}\dfrac{1}{(n+1)^{\frac{2s}{s+1}}},\ s \in (0, 1), \\ \dfrac{2}{\pi}\dfrac{\sqrt{\ln(n+1)}}{n+1},\ s = 1, \\ \dfrac{1}{\pi}\sin\dfrac{\pi s}{2}\inf\limits_{0<\beta\le 1}\left[\beta\displaystyle\int_0^{\beta}\frac{u^{s-2}du}{\left(1-u^2\right)^{s/2}} + \frac{1}{\beta}\int_{\beta}^{1}\frac{u^s du}{\left(1-u^2\right)^{s/2}}\right]\frac{1}{n+1},\ s \in (1, 2). \end{cases} \tag{35}$$

Доказательство. Исследуем равенства (29). При выполнении условия (15) второе слагаемое в выражении для $\nu_n(\beta, s)$ имеет заведомо больший порядок малости в сравнении с первым, и, следовательно,

$$\nu_n(\beta, s) \sim \frac{1}{\beta(n+1)} \int_0^{\arccos\beta} \cos^s\theta \sin^{1-s}\theta d\theta, \ 0 < s < 2, \ n \to \infty. \tag{36}$$

Рассмотрим случай $s \in (0, 1]$. В соотношении (36) положим $\beta = \beta(n) \to 0$ с сохранением условия (15) и получим

$$\nu_n(\beta, s) \sim \frac{2^{1-s}\Gamma(s)}{\Gamma^2\left(\frac{s}{2}\right)} \cdot \frac{\pi}{\sin\frac{\pi s}{2}} \cdot \frac{1}{\beta(n+1)}, \ n \to \infty.$$

Следовательно, при $n \to \infty$ для мажоранты $\varepsilon_{2n}^*(\alpha)$ справедливы асимптотические равенства

$$\varepsilon_{2n}^*(\alpha) \sim \begin{cases} \frac{1}{\pi}\sin\frac{\pi s}{2}\frac{2^{1-s}\beta^s\Gamma(s)}{(1-s)(n+1)^s} + \frac{2^{1-s}\Gamma(s)}{\Gamma^2\left(\frac{s}{2}\right)}\frac{1}{\beta(n+1)}, & s \in (0, 1), \\ \frac{1}{\pi}\left(\beta\frac{\ln(n+1)}{n+1} + \frac{1}{\beta(n+1)}\right), & s = 1. \end{cases}$$

При каждом заданном $0 < s \le 1$ соответствующие значения величины $\varepsilon_{2n}^*(\alpha)$ имеют строгий минимум при $0 < \beta \le 1$. Решая экстремальную задачу

$$\varepsilon_{2n}^*(\alpha) \underset{0<\beta\le 1}{\to} \min,$$

находим, что оптимальными при заданном s, $0 < s \le 1$, будут значения

$$\beta^* = \begin{cases} \left(\frac{\pi(1-s)}{s\Gamma^2\left(\frac{s}{2}\right)\sin\frac{\pi s}{2}}\right)^{\frac{1}{s+1}} \frac{1}{(n+1)^{\frac{1-s}{1+s}}}, & s \in (0, 1), \\ \frac{1}{\sqrt{\ln(n+1)}}, & s = 1. \end{cases}$$

При этом

$$\varepsilon_{2n}^* \sim \begin{cases} 2^{1-s}\Gamma(s)\frac{1+s}{s}\left(\left(\frac{1}{\pi}\sin\frac{\pi s}{2}\right)^{2s+1}\frac{s}{1-s}\Gamma^{2s}\left(1-\frac{s}{2}\right)\right)^{\frac{1}{s+1}}\frac{1}{(n+1)^{\frac{2s}{s+1}}}, & s \in (0, 1), \\ \frac{2}{\pi}\frac{\sqrt{\ln(n+1)}}{n+1}, \ s = 1, \ \alpha^* = \sqrt{\frac{1-\beta^*}{1+\beta^*}}. \end{cases} \tag{37}$$

Пусть теперь $s \in (1, 2)$. В этом случае из (29) и (36) получим

$$\varepsilon_{2n}^*(\alpha) \sim \frac{1}{\pi}\sin\frac{\pi s}{2}\left[\beta\int_0^\beta \frac{u^{s-2}du}{\left(1-u^2\right)^{s/2}} + \frac{1}{\beta}\int_\beta^1 \frac{u^s du}{\left(1-u^2\right)^{s/2}}\right]\frac{1}{n+1}, \ n \to \infty.$$

Отсюда

$$\varepsilon_{2n}^* \sim \frac{1}{\pi}\sin\frac{\pi s}{2}\inf_{0<\beta\le 1}\left[\beta\int_0^\beta \frac{u^{s-2}du}{\left(1-u^2\right)^{s/2}} + \frac{1}{\beta}\int_\beta^1 \frac{u^s du}{\left(1-u^2\right)^{s/2}}\right]\frac{1}{n+1}, \ n \to \infty. \tag{38}$$

Из соотношений (37) и (38) следует (35). Теорема 7 доказана.

Замечание 3. Сравнивая результаты теорем 5 и 7, приходим к выводу, что при $s \in (0, 1]$ специальным выбором параметра α возможно добиться скорости приближений рациональными суммами Фейера большей в сравнении с полиномиальным случаем. Данный результат отражает особенности рациональной аппроксимации непрерывных функций со степенными особенностями. Если же $s \in (1, 2)$, то ситуация иная. Из (35) следует, что оптимальное значение параметра не увеличивает скорость убывания мажоранты ε^*_{2n}. Однако можно заметить, что при заданном $s \in (1, 2)$ найденное оптимальное значение параметра будет уменьшать константу.

Замечание 4. Нетрудно получить асимптотические разложения для оптимального параметра α^*, например при $s = 1$:

$$\alpha^* = 1 - \frac{1}{\sqrt{\ln(n+1)}} + o\left(\frac{1}{\sqrt{\ln(n+1)}}\right), \ n \to \infty.$$

Замечание 5. В [32] получена точная асимптотическая оценка равномерного приближения функции $|x|$ частичными суммами рациональных рядов Фурье – Чебышева на отрезке $[-1, 1]$. Точность была установлена благодаря тому, что максимальное уклонение частичных сумм от функции $|x|$ достигалось в точках x, равных 0, ±1. В случае приближений суммами Фейера ситуация изменилась. По нашему мнению, при заданном s, $0 < s < 2$, максимальное уклонение достигается в некоторой точке x^*_n из окрестности нуля, и, видимо, при $n \to \infty$ имеют место равенства (35) для величин ε_{2n}.

Заключение

В работе изучены аппроксимационные свойства сумм Фейера рядов Фурье по одной системе алгебраических дробей Чебышева – Маркова. Найдено интегральное представление для исследуемых сумм (теорема 2), получены оценки сверху для равномерных приближений посредством сумм Фейера непрерывных на отрезке функций в терминах модулей непрерывности (теорема 3). Проведено исследование приближений функции $|x|^s$, $0 < s < 2$, изучаемыми суммами Фейера. В частности, установлены оценки поточечных и равномерных приближений (теорема 4), найдена асимптотическая оценка мажоранты соответствующих приближений (теорема 6), получено оптимальное значение параметра, обеспечивающее максимальную скорость приближений исследуемой функции суммами Фейера (теорема 7). Следствием полученных результатов являются асимптотические оценки приближений функции $|x|^s$, $0 < s < 2$, суммами Фейера полиномиальных рядов Фурье – Чебышева (теорема 5).

Библиографические ссылки

1. Fejér L. Untersuchungen über Fouriersche Reihen. *Mathematische Annalen.* 1904;58(1–2):51–69. DOI: 10.1007/BF01447779.

2. Lebesgue H. Sur les intégrales singulières. *Annales de la faculté des sciences de Toulouse. 3e série.* 1909;1:25–117.

3. Bernstein S. *Sur l'ordre de la meilleure approximation des fonctions continues par des polynomes de degré donné.* Bruxelles: Hayez, imprimeur des Academies Royales; 1912. 104 p.

4. Никольский СМ. Об асимптотическом поведении остатка при приближении функций, удовлетворяющих условию Липшица, суммами Фейера. *Известия АН СССР. Серия математическая.* 1940;4(6):501–508.

5. Zygmund A. On the degree of approximation of functions by Fejer means. *Bulletin of the American Mathematical Society.* 1945;51(4):274–278.

6. Новиков ОА, Ровенская ОГ. Приближение классов интегралов Пуассона суммами Фейера. *Компьютерные исследования и моделирование.* 2015;7(4):813–819. DOI: 10.20537/2076-7633-2015-7-4-813-819.

7. Ефимов АВ. О приближении некоторых классов непрерывных функций суммами Фурье и суммами Фейера. *Известия АН СССР. Серия математическая*. 1958;22(1):81–116.

8. Лебедь ГК, Авдеенко АА. О приближении периодических функций суммами Фейера. *Известия АН СССР. Серия математическая*. 1971;35(1):83–92.

9. Джрбашян ММ. К теории рядов Фурье по рациональным функциям. *Известия Академии наук Армянской ССР. Серия физико-математическая.* 1956;9(7):1–27.

10. Русак ВН. *Рациональные функции как аппарат приближения.* Минск: БГУ им. В. И. Ленина; 1979. 179 с.

11. Petrushev PP, Popov VA. *Rational approximation of real functions*. Cambridge: Cambridge University Press; 1987. 386 p.

12. Русак ВН. Точные порядки наилучших рациональных приближений на классах функций, представимых в виде свертки. *Доклады Академии наук СССР.* 1984;279(4):810–812.

13. Русак ВН. Точные порядковые оценки для наилучших рациональных приближений на классах функций, представимых в виде свертки. *Математический сборник.* 1985;128(4):492–515.

14. Пекарский АА. Чебышевские рациональные приближения в круге, на окружности и на отрезке. *Математический сборник.* 1987;133(1):86–102.

15. Смотрицкий КА. Аппроксимация рациональными операторами Валле Пуссена на отрезке. *Труды Института математики НАН Беларуси*. 2001;9:136–139.

16. Ровба ЕА. Рациональные интегральные операторы на отрезке. *Вестник БГУ. Серия 1. Физика. Математика. Информатика.* 1996;1:34–39.

17. Смотрицкий КА. О приближении выпуклых функций рациональными интегральными операторами на отрезке. *Вестник БГУ. Серия 1. Физика. Математика. Информатика.* 2005;3:64–70.

18. Ровба ЕА. Приближение функций, дифференцируемых в смысле Римана – Лиувилля, рациональными операторами. *Доклады Национальной академии наук Беларуси.* 1996;40(6):18–22.

19. Ровба ЕА. О приближении рациональными операторами Фейера и Джексона функций ограниченной вариации. *Доклады Национальной академии наук Беларуси*. 1998;42(4):13–17.

20. Bernstein S. Sur la meilleure approximation de $|x|$ par des polynomes de degres donnes. *Acta Mathematica.* 1914;37:1–57. DOI: 10.1007/BF02401828.

21. Newman DJ. Rational approximation to $|x|$. *Michigan Mathematical Journal.* 1964;11(1):11–14. DOI: 10.1307/mmj/1028999029.

22. Буланов АП. Асимптотика для наименьших уклонений $|x|$ от рациональных функций. *Математический сборник.* 1968; 76(118-2):288–303.

23. Вячеславов НС. О приближении функции $|x|$ рациональными функциями. *Математические заметки.* 1974;16(1):163–171.

24. Шталь Г. Наилучшие равномерные рациональные аппроксимации $|x|$ на $[-1, 1]$. *Математический сборник*. 1992;183(8): 85–118.

25. Бернштейн СН. О наилучшем приближении $|x|^p$ при помощи многочленов весьма высокой степени. *Известия АН СССР. Серия математическая*. 1938;2(2):169–190.

26. Freud G, Szabados J. Rational approximation to x^a. *Acta Mathematica Academiae Scientiarum Hungaricae.* 1967;18(3–4): 393–399. DOI: 10.1007/BF02280298.

27. Гончар АА. О скорости рациональной аппроксимации непрерывных функций с характерными особенностями. *Математический сборник*. 1967;73(4):630–638.

28. Вячеславов Н. Об аппроксимации x^a рациональными функциями. *Известия АН СССР. Серия математическая*. 1980; 44(1):92–109.

29. Stahl HR. Best uniform rational approximation of x^a on [0, 1]. *Bulletin of the American Mathematical Society.* 1993;28(1):116–122.

30. Revers M. On the asymptotics of polynomial interpolation to x^a at the Chebyshev nodes. *Journal of Approximation Theory.* 2013;65:70–82. DOI: 10.1016/j.jat.2012.09.005.

31. Райцин РА. Асимптотические свойства равномерных приближений функций с алгебраическими особенностями частичными суммами ряда Фурье – Чебышева. *Известия высших учебных заведений. Математика.* 1980;3:45–49.

32. Rouba Y, Patseika P, Smatrytski K. On one system of rational Chebyshev – Markov fractions. *Analysis Mathematica.* 2018; 44(1):115–140. DOI: 10.1007/s10476-018-0110-7.

33. Натансон ИП. *Конструктивная теория функций.* Москва: ГИТТЛ; 1949. 684 с.

34. Тиман АФ. *Теория приближений функций действительного переменного.* Москва: ГИФМЛ; 1960. 624 с.

35. Титчмарш Е. *Теория функций*. Москва: Наука; 1980. 463 с.

36. Ровба ЕА, Поцейко ПГ. Аппроксимация функции $|x|^s$ на отрезке $[-1, 1]$ частичными суммами рационального ряда Фурье – Чебышева. *Веснік Гродзенскага дзяржаўнага ўніверсітэта імя Янкі Купалы. Серыя 2. Матэматыка. Фізіка. Інфарматыка, вылічальная тэхніка і кіраванне.* 2019;9(3):16–28.

37. Сидоров ЮВ, Федорюк МВ, Шабунин МИ. *Лекции по теории функций комплексного переменного.* Москва: Наука; 1989. 480 с.

38. Евграфов МА. *Асимптотические оценки и целые функции*. Москва: Наука; 1979. 320 с.

39. Федорюк МВ. *Асимптотика. Интегралы и ряды*. Москва: Наука; 1987. 544 с.

40. Copson ET. *Asymptotic Expansions*. Cambridge: Cambridge University Press; 1965. 124 p. (Cambridge Tracts in Mathematics and Mathematical Physics; no. 55).

References

1. Fejér L. Untersuchungen über Fouriersche Reihen. *Mathematische Annalen.* 1904;58(1–2):51–69. DOI: 10.1007/BF01447779.

2. Lebesgue H. Sur les intégrales singulières. *Annales de la faculté des sciences de Toulouse. 3e série.* 1909;1:25–117.

3. Bernstein S. *Sur l'ordre de la meilleure approximation des fonctions continues par des polynomes de degré donné.* Bruxelles: Hayez, imprimeur des Academies Royales; 1912. 104 p.

4. Nikol'skii SM. [On the asymptotic behavior of the remainder under approximation of functions satisfying the Lipschitz condition, by fejér sums]. *Izvestiya AN SSSR. Seriya matematicheskaya.* 1940;4(6):501–508. Russian.

5. Zygmund A. On the degree of approximation of functions by Fejer means. *Bulletin of the American Mathematical Society.* 1945;51(4):274–278.

6. Novikov OA, Rovenska OG. Approximation of classes of Poisson integrals by Fejer sums. *Komp'yuternye issledovaniya i modelirovanie.* 2015;7(4):813–819. Russian. DOI: 10.20537/2076-7633-2015-7-4-813-819.

7. Efimov AV. [On the approximation of some classes of continuous functions by Fourier sums and Fejer sums]. *Izvestiya AN SSSR. Seriya matematicheskaya.* 1958;22(1):81–116. Russian.

8. Lebed' GK, Avdeenko AA. [On the approximation of periodic functions by Fejér sums]. *Izvestiya AN SSSR. Seriya matematicheskaya*. 1971;35(1):83–92. Russian.

9. Dzhrbashyan MM. [On the theory of Fourier series on rational functions]. *Izvestiya Akademii nauk Armyanskoi SSR. Seriya fiziko-matematicheskaya.* 1956;9(7):1–27. Russian.

10. Rusak VN. *Ratsional'nye funktsii kak apparat priblizheniya* [Rational functions as an apparatus of approximation]. Minsk: Belorusskii gosudarstvennyi universitet im. V. I. Lenina; 1979. 179 p. Russian.

11. Petrushev PP, Popov VA. *Rational approximation of real functions*. Cambridge: Cambridge University Press; 1987. 386 p.

12. Rusak VN. [Sharp order estimates for best rational approximations in classes of functions representable as convolutions]. *Doklady Akademii nauk SSSR.* 1984;279(4):810–812. Russian.

13. Rusak VN. [Exact order estimates for best rational approximations in classes of functions representable as convolution]. *Matematicheskii sbornik*. 1985;128(4):492–515. Russian.

14. Pekarskii AA. [Chebyshev rational approximations in a circle, on a circle, and on a segment]. *Matematicheskii sbornik*. 1987; 133(1):86–102. Russian.

15. Smotritskii KA. [Approximation by rational operators of Valle Poussin on a segment]. *Trudy Instituta matematiki NAN Belarusi.* 2001;9:136–139. Russian.

16. Rovba EA. Rational integral operators on a segment. *Vestnik BGU. Seriya 1. Fizika. Matematika. Informatika.* 1996;1:34–39. Russian.

17. Smotrytski KA. On the approximation of the convex functions by rational integral operators on the segment. *Vestnik BGU. Seriya 1. Fizika. Matematika. Informatika.* 2005;3:64–70. Russian.

18. Rovba EA. [Approximation of functions differentiable in the Riemann – Liouville sense by rational operators]. *Doklady of the National Academy of Sciences of Belarus.* 1996;40(6):18–22. Russian.

19. Rovba EA. [On the approximation by rational Fejer and Jackson operators of bounded variation functions]. *Doklady of the National Academy of Sciences of Belarus*. 1998;42(4):13–17. Russian.

20. Bernstein S. Sur la meilleure approximation de $|x|$ par des polynomes de degres donnes. *Acta Mathematica.* 1914;37:1–57. DOI: 10.1007/BF02401828.

21. Newman DJ. Rational approximation to $|x|$. *Michigan Mathematical Journal.* 1964;11(1):11–14. DOI: 10.1307/mmj/1028999029.

22. Bulanov AP. [Asymptotics for least deviation of $|x|$ by rational functions]. *Matematicheskii sbornik*. 1968;76(118-2):288–303. Russian.

23. Vyacheslavov NS. [On the approximation of the function $|x|$ by rational functions]. *Matematicheskie zametki.* 1974;16(1): 163–171. Russian.

24. Shtal' G. [Best uniform rational approximation of $|x|$ on $[-1, 1]$. *Matematicheskii sbornik*. 1992;183(8):85–118. Russian.

25. Bernstein SN. [Sur la meilleure approximation de $|x|^p$ par des polynomes de degres tres eleves]. *Izvestiya AN SSSR. Seriya matematicheskaya*. 1938;2(2):169–190. Russian.

26. Freud G, Szabados J. Rational approximation to x^a. *Acta Mathematica Academiae Scientiarum Hungaricae.* 1967;18(3–4): 393–399. DOI: 10.1007/BF02280298.

27. Gonchar AA. [On the rate of rational approximation of continuous functions with characteristic features]. *Matematicheskii sbornik*. 1967;73(4):630–638. Russian.

28. Vyacheslavov N. [On the approximation of x^a by rational functions]. *Izvestiya AN SSSR. Seriya matematicheskaya*. 1980;44(1): 92–109. Russian.

29. Stahl HR. Best uniform rational approximation of x^a on [0, 1]. *Bulletin of the American Mathematical Society.* 1993;28(1):116–122.

30. Revers M. On the asymptotics of polynomial interpolation to x^a at the Chebyshev nodes. *Journal of Approximation Theory.* 2013;65:70–82. DOI: 10.1016/j.jat.2012.09.005.

31. Raitsin RA. [Asymptotic properties of uniform approximations of functions with algebraic features by partial sums of the Fourier – Chebyshev series]. *Izvestiya vysshikh uchebnykh zavedenii. Matematika.* 1980;3:45–49. Russian.

32. Rouba Y, Patseika P, Smatrytski K. On one system of rational Chebyshev – Markov fractions. *Analysis Mathematica.* 2018; 44(1):115–140. DOI: 10.1007/s10476-018-0110-7.

33. Natanson IP. *Konstruktivnaya teoriya funktsii* [Constructive theory of functions]. Moscow: GITTL; 1949. 684 p. Russian.

34. Timan AF. *Teoriya priblizhenii funktsii deistvitel'nogo peremennogo* [Theory of approximations of functions of a real variable]. Moscow: GIFML; 1960. 624 p. Russian.

35. Titchmarsh E. *Teoriya funktsii* [Theory of functions]. Moscow: Nauka; 1980. 463 p. Russian.

36. Rovba EA, Potseiko PG. [Approximation of the function $|x|^s$ on the segment $[-1, 1]$ by partial sums of the rational Fourier – Chebyshev series]. *Vesnik Grodzenskaga dzjarzhawnaga wniversitjeta imja Janki Kupaly. Seryja 2. Matjematyka. Fizika. Infarmatyka, vylichal'naja tjehnika i kiravanne.* 2019;9(3):16–28. Russian.

37. Sidorov YuV, Fedoryuk MV, Shabunin MI. *Lektsii po teorii funktsii kompleksnogo peremennogo* [Lectures on the theory of functions of a complex variable]. Moscow: Nauka; 1989. 480 p. Russian.

38. Evgrafov MA. *Asimptoticheskie otsenki i tselye funktsii* [Asymptotic estimates and entire functions]. Moscow: Nauka; 1979. 320 p. Russian.

39. Fedoryuk MV. *Asimptotika. Integraly i ryady* [Asymptotics. Integrals and series]. Moscow: Nauka; 1987. 544 p. Russian.

40. Copson ET. *Asymptotic Expansions.* Cambridge: Cambridge University Press; 1965. 124 p. (Cambridge Tracts in Mathematics and Mathematical Physics; no. 55).

Статья поступила в редколлегию 14.10.2019.
Received by editorial board 14.10.2019.

Математическая логика, алгебра и теория чисел

Mathematical logic, algebra and number theory

УДК 512.542

О НЕКОТОРЫХ КЛАССАХ ПОДРЕШЕТОК РЕШЕТКИ ВСЕХ ПОДГРУПП

А. Н. СКИБА[1)]

[1)]*Гомельский государственный университет им. Франциска Скорины, ул. Советская, 104, 246019, г. Гомель, Беларусь*

В настоящей статье G всегда обозначает группу. Если K и H – подгруппы группы G, где K – нормальная подгруппа группы H, то фактор-группа группы H по K называется секцией группы G. Такая секция является нормальной, если K и H – нормальные подгруппы группы G, и тривиальной, если K и H равны. Назовем произвольное множество Σ нормальных секций группы G расслоением группы G, если оно содержит каждую тривиальную нормальную секцию группы G, и будем говорить, что расслоение Σ группы G является G-замкнутым, если Σ содержит каждую такую нормальную секцию группы G, которая G-изоморфна некоторой нормальной секции группы G, принадлежащей множеству Σ. Пусть теперь Σ – произвольное G-замкнутое расслоение группы G и пусть $\boldsymbol{L}$ – множество всех таких подгрупп A группы G, что фактор-группа группы V по W, где V – нормальное замыкание A в G, а W – нормальное ядро A в G, принадлежит Σ. Опишем условия на Σ, при которых множество $\boldsymbol{L}$ является подрешеткой решетки всех подгрупп группы G, а также обсудим некоторые применения этой подрешетки в теории обобщенных конечных T-групп.

Ключевые слова: группа; решетка подгрупп; модулярная решетка; формационное множество Фиттинга; формация Фиттинга.

Образец цитирования:
Скиба АН. О некоторых классах подрешеток решетки всех подгрупп. *Журнал Белорусского государственного университета. Математика. Информатика.* 2019;3:35–47 (на англ.).
https://doi.org/10.33581/2520-6508-2019-3-35-47

For citation:
Skiba AN. On some classes of sublattices of the subgroup lattice. *Journal of the Belarusian State University. Mathematics and Informatics.* 2019;3:35–47.
https://doi.org/10.33581/2520-6508-2019-3-35-47

Автор:
Александр Николаевич Скиба – доктор физико-математических наук, профессор; профессор кафедры алгебры и геометрии факультета математики и технологий программирования.

Author:
Alexander N. Skiba, doctor of science (physics and mathematics), full professor; professor at the department of algebra and geometry, faculty of mathematics and technologies of programming.
alexander.skiba49@gmail.com

Журнал Белорусского государственного университета. Математика. Информатика. 2019;3:35–47
Journal of the Belarusian State University. Mathematics and Informatics. 2019;3:35–47

ON SOME CLASSES OF SUBLATTICES OF THE SUBGROUP LATTICE

A. N. SKIBA[a]

[a]*Francisk Skorina Gomel State University,*
104 Saveckaja Street, Homiel 246019, Belarus

In this paper G always denotes a group. If K and H are subgroups of G, where K is a normal subgroup of H, then the factor group of H by K is called a section of G. Such a section is called normal, if K and H are normal subgroups of G, and trivial, if K and H are equal. We call any set Σ of normal sections of G a stratification of G, if Σ contains every trivial normal section of G, and we say that a stratification Σ of G is G-closed, if Σ contains every such a normal section of G, which is G-isomorphic to some normal section of G belonging Σ. Now let Σ be any G-closed stratification of G, and let $\boldsymbol{L}$ be the set of all subgroups A of G such that the factor group of V by W, where V is the normal closure of A in G and W is the normal core of A in G, belongs to Σ. In this paper we describe the conditions on Σ under which the set $\boldsymbol{L}$ is a sublattice of the lattice of all subgroups of G and we also discuss some applications of this sublattice in the theory of generalized finite T-groups.

Keywords: group; subgroup lattice; modular lattice; formation Fitting set; Fitting formation.

Introduction

In this paper G always denotes a group. Moreover, $\mathfrak{L}(G)$ denotes the set (the lattice) of all subgroups of G and $\mathfrak{L}_n(G)$ is the set (the lattice) of all normal subgroups of G. In this paper $\mathfrak{F}$ is a class of groups containing all identity groups, $\mathfrak{N}^*$ is the class of all finite quasinilpotent groups, $\mathfrak{N}$ is the class of all finite nilpotent groups and $\mathfrak{U}$ is the class of all finite supersoluble groups.

A class of groups $\mathfrak{F}$ is said to be a *Fitting formation* if the following conditions hold: (1) for every normal subgroup N of any group $G \in \mathfrak{F}$ both groups N and G/N belong to $\mathfrak{F}$; (2) $G \in \mathfrak{F}$ whenever G has normal subgroups A and B and either G/A, $G/B \in \mathfrak{F}$ and $A \cap B = 1$ or $G = AB$ and A, $B \in \mathfrak{F}$.

One of the organizing ideas of the group theory is the idea to study the group G depending on the presence in it a subgroup system $\mathfrak{L}$ having desired properties. Such an approach is the most effective in the case when $\mathfrak{L}$ forms a *sublattice* of $\mathfrak{L}(G)$, that is, $A \cap B \in \mathfrak{L}$ and $\langle A, B\rangle \in \mathfrak{L}$ for all A, $B \in \mathfrak{L}$. This circumstance makes the general problem of finding sublattices in $\mathfrak{L}(G)$ important and interesting.

One of the first results in this direction was obtained by Wielandt in his paper [1], where it was proved that the set $\mathfrak{L}_{sn}(G)$ of all subnormal subgroups of the group G having a composition series is a sublattice of $\mathfrak{L}(G)$. In the case when G is finite, an original generalization of the lattice $\mathfrak{L}_{sn}(G)$ was found by Kegel [2]. A subgroup A of G is called $\mathfrak{F}$-*subnormal in* G *in the sense of Kegel* [2] or K-$\mathfrak{F}$-*subnormal* in G [3, definition 6.1.4], if there is a subgroup chain $A = A_0 \leq A_1 \leq \ldots \leq A_t = G$ such that either $A_{i-1} \trianglelefteq A_i$ or $A_i/(A_{i-1})_{A_i} \in \mathfrak{F}$ for all $i = 1, \ldots, t$. Kegel proved [2] that if the class $\mathfrak{F}$ is closed under extensions, epimorphic images and subgroups, then the set $\mathfrak{L}_{\mathfrak{F}sn}(G)$ of all K-$\mathfrak{F}$-subnormal subgroups of a finite group G is a sublattice of the lattice $\mathfrak{L}(G)$. For every set π of primes, we may choose the class $\mathfrak{F}$ of all π-groups. In this way we obtain infinitely many functors $\mathfrak{L}_{\mathfrak{F}sn}$ assigning to every finite group G a sublattice of $\mathfrak{L}(G)$ containing $\mathfrak{L}_{sn}(G)$. Subsequently, this result was generalized (also in the universe of all finite groups) on the basis of methods of the formation theory (see, in particular, [4; 5] and chapter 6 in [3]).

In this paper, we develop a new approach for finding sublattices in $\mathfrak{L}(G)$, where G is an arbitrary group, and we also discuss some applications of such sublattices.

The main concepts and results

If $K \trianglelefteq H \leq G$, then H/K is called a *section* of G; such a section is called: *normal* if H and K are normal subgroups of G; *trivial* if $H = K$; a *chief factor* of G provided $K < H$ and for any normal subgroup L of G with $K \leq L \leq H$ we have either $K = L$ or $L = H$. We write $H/K \simeq_G T/L$ provided the normal sections H/K and T/L of G are G-isomorphic; $Ch_G(H/K)$ denotes the set of all chief factors T/L of G with $K \leq L < T \leq H$; A^G is the normal closure of the subgroup A in G and $A_G = \bigcap_{x \in G} A^x$. If Δ is any set of chief factors of G (not necessary non-empty),

then we write $\Sigma_G(\Delta)$ to denote the set of all normal sections H/K of G such that either $K = H$ or $K < H$ and the series $K < H$ can be refined to a chief series of G between K and H (of finite length) with $Ch_G(H/K) \subseteq \Delta$.

We call a set Σ of normal sections of G a *stratification* of G if Σ contains every trivial normal section of G and we say that a stratification Σ of G is *G-closed* provided $H/K \in \Sigma$ whenever H/K is a normal section of G with $H/K \simeq_G T/L \in \Sigma$.

Now let Σ be any stratification of G. Then write $\mathfrak{L}_\Sigma(G)$ to denote the set of all subgroups A of G with $A^G/A_G \in \Sigma$.

We will use $\Sigma_G(\mathfrak{F})$ to denote the set of normal sections H/K of G such that $H/K \in \mathfrak{F}$.

Definition. We say (by analogy with the definition of the *Fitting set* of a group [6, p. 537]) that a G-closed stratification Σ of G is a *formation Fitting set* of G if the following conditions hold:

(i) for every two normal sections H/K and T/K of G where $T/K \in \Sigma$ and $H \le T$, we have H/K, $T/H \in \Sigma$;

(ii) $H/(K \cap N) \in \Sigma$ for every two sections H/K, $H/N \in \Sigma$;

(iii) $HV/K \in \Sigma$ for every two sections H/K, $V/K \in \Sigma$.

The usefulness of this concept is primarily based on the following our three results.

Theorem 1. *If $\Sigma = \Sigma_G(\Delta)$ for some G-closed set Δ of chief factors of G or $\Sigma = \Sigma_G(\mathfrak{F})$ for some Fitting formation $\mathfrak{F}$, then Σ is a formation Fitting set of G.*

Theorem 2. *The set $\mathfrak{L}_\Sigma(G)$ forms a sublattice in $\mathfrak{L}(G)$ for each formation Fitting set Σ of G.*

Theorem 3. *The inclusion $\mathfrak{L}_n(G) \subseteq \mathfrak{L}_\Sigma(G)$ holds for every formation Fitting set Σ of G. Moreover, in the case when G satisfies the maximality condition the lattice $\mathfrak{L}_\Sigma(G)$ is distributive if and only if $\mathfrak{L}_\Sigma(G) = \mathfrak{L}_n(G)$ is distributive.*

From theorems 1 and 2 we get the following.

Corollary 1. *Let $\mathfrak{F}$ be either the class of all nilpotent groups, or the class of all soluble groups, or the class of all finite quasinilpotent groups. Then the set $\mathfrak{L}_{\Sigma_G(\mathfrak{F})}(G)$ forms a sublattice in $\mathfrak{L}(G)$.*

We say that a chief factor H/K of G is *$\mathfrak{F}$-central* in G [7] if

$$(H/K) \rtimes (G/C_G(H/K)) \in \mathfrak{F}.$$

Let $D = M \rtimes A$ and $R = N \rtimes B$. Then the pairs (M, A) and (R, B) are said to be *equivalent* provided there are isomorphisms $f : M \to N$ and $g : A \to B$ such that $f(a^{-1}ma) = g(a^{-1}) f(m) g(a)$ for all $m \in M$ and $a \in A$.

In fact, the following lemma is known (see, for example, lemma 3.27 in [7]) and it can be proved by the direct verification.

Lemma 1. *Let $D = M \rtimes A$ and $R = N \rtimes B$. If the pairs (M, A) and (R, B) are equivalent, then $D \simeq R$.*

Lemma 2. *Let N, M and $K < H \le G$ be normal subgroups of G, where H/K is a chief factor of G:*

(1) if $N \le K$, then $(H/K) \rtimes (G/C_G(H/K)) \simeq ((H/N)/(K/N)) \rtimes ((G/N)/C_{G/N}((H/N)/(K/N)))$;

(2) if T/L is a chief factor of G and H/K and T/L are G-isomorphic, then $C_G(H/K) = C_G(T/L)$ and $(H/K) \rtimes (G/C_G(H/K)) \simeq (T/L) \rtimes (G/C_G(T/L))$;

(3) $(MN/N) \rtimes (G/C_G(MN/N)) \simeq (M/M \cap N) \rtimes (G/C_G(M/M \cap N))$.

P r o o f. (1) In view of the G-isomorphisms $H/K \simeq (H/N)/(K/N)$ and

$$G/C_G(H/K) \simeq (G/N)/(C_G(H/K)/N),$$

the pairs

$$(H/K, G/C_G(H/K)), ((H/N)/(K/N), (G/N)/C_{G/N}((H/N)/(K/N)))$$

are equivalent. Hence statement (1) is a corollary of lemma 1.

(2) A direct check shows that $C = C_{G/N}(H/K) = C_G(T/L)$ and that the pairs $(H/K, G/C)$ and $(T/L, G/C)$ are equivalent. Hence statement (2) is also a corollary of lemma 1.

(3) This follows from the G-isomorphism $MN/N \simeq M/M \cap N$ and part (2).

The lemma is proved.

In view of lemma 2, we get from theorems 1 and 2 the following fact.

Corollary 2. *Let Δ be the set of all $\mathfrak{F}$-central chief factors of G. Then the set $\mathfrak{L}_{\Sigma(\Delta)}(G)$ forms a sublattice in $\mathfrak{L}(G)$.*

Remark 1. (i) Let $\Sigma(G)$ be the set of all formation Fitting sets of G. It is clear that $\Sigma(G)$ is partially ordered with respect to set inclusion and the formation Fitting set $\{H/K \mid H, K \in \mathfrak{L}_n(G)\}$ is the greatest element in $\Sigma(G)$. Moreover, for every set $\{\Sigma_i \mid i \in I\}$ of formation Fitting sets of G the intersection $\bigcap_{i \in I} \Sigma_i$ is also a formation Fitting set of G and so $\bigcap_{i \in I} \Sigma_i$ is the greatest lower bound for $\{\Sigma_i \mid i \in I\}$ in $\Sigma(G)$. Therefore $\Sigma(G)$ is a complete lattice. The set $\{H/H \mid H \trianglelefteq G\}$ is the smallest element in $\Sigma(G)$.

(ii) Let $\mathfrak{X}$ be any set of normal sections of G. Then the set $\{\Sigma_i \mid i \in I\}$ of all formation Fitting sets of G containing $\mathfrak{X}$ is non-empty and the intersection $\bigcap_{i \in I} \Sigma_i$ is a formation Fitting set of G by part (i). We say that $\bigcap_{i \in I} \Sigma_i$ is the *formation Fitting set of G generated by* $\mathfrak{X}$ and denote it by $\mathrm{formfit}(\mathfrak{X})$. If $\mathfrak{X} = \{T/L\}$ is a singleton set, we write $\mathrm{formfit}(T/L)$ instead of $\mathrm{formfit}(\{T/L\})$ and say that $\mathrm{formfit}(T/L)$ is a *one-generated formation Fitting set of G*.

(iii) Let E and N be subgroups of G, where N is normal in G. Then for any stratification Σ of G we use $\Sigma N/N$ and $\Sigma \cap E$ to denote the stratification $\{(NH/N)/(NK/N) \mid H/K \in \Sigma\}$ of G/N and the stratification $\{(T \cap E)/(L \cap E) \mid T/L \in \Sigma\}$ of E, respectively. If Σ is a formation Fitting set of G, then $\Sigma N/N$ is a formation Fitting set of G/N (see proposition (iv) below).

From theorem 1 we get the following useful result.

Corollary 3. *Let* $\mathfrak{X}$ *be a set of normal sections of G and* $T/L \in \Sigma = \mathrm{formfit}(\mathfrak{X})$. *Then the following statements hold:*

(i) $T/L \in \mathfrak{F}$ *for every Fitting formation* $\mathfrak{F}$ *containing* $\mathfrak{X}$;

(ii) if $H/K \in Ch(T/L)$, *then* $H/K \simeq_G F/S$ *for some* $F/S \in Ch(V/W)$ *and* $V/W \in \mathfrak{X}$.

For any two sections H/K and T/L of G we write $H/K \le T/L$ provided $K \le L$ and $H \le T$. Then the set of all sections of G is partially ordered with respect to $\le$.

The proofs of theorems 2 and 3 are based on the following useful observation.

Proposition. *Let* Σ *be a formation Fitting set of G and let E and N be subgroups of G, where* $N \trianglelefteq G$. *Then:*

(i) $\langle \Sigma, \le \rangle$ *is a lattice in which* HV/KW *is the least upper bound and* $(H \cap V)/(K \cap W)$ *is the greatest lower bound of* $\{H/K, V/W\}$ *for any two its sections* H/K, V/W;

(ii) if $T/L \in \Sigma$, *then* $\mathfrak{L}(T/L)$ *is isomorphic to the interval* $[T, L]$ *in* $\mathfrak{L}_\Sigma(G)$;

(iii) if $f : G \to G^*$ *is an isomorphism, then* $f(\Sigma) := \{T^f/L^f \mid T/L \in \Sigma\}$ *is a formation Fitting set of* G^*. *Moreover, if* Σ *is hereditary, then* $f(\Sigma)$ *is hereditary;*

(iv) $\Sigma N/N$ *is a formation Fitting set of* G/N *and* $\Sigma N/N = \{(H/N)/(K/N) \mid H/K \in \Sigma \text{ and } N \le K\}$.

P r o o f. (i) Since $H/K \in \Sigma$ and $K(V \cap H)/K \le H/K$, we have $K(V \cap H)/K \in \Sigma$. Hence from the G-isomorphism

$$(H \cap V)/(K \cap V) = (H \cap V)/(K \cap V \cap H) \simeq K(V \cap H)/K$$

we get that $(H \cap V)/(K \cap V) \in \Sigma$. Similarly, $(V \cap H)/(W \cap H) \in \Sigma$. But then we get that

$$(H \cap V)/((K \cap V) \cap (W \cap H)) = (H \cap V)/(K \cap W) \in \Sigma$$

since Σ is a formation Fitting set of G by hypothesis.

From the G-isomorphism

$$H(KW)/KW \simeq H/(H \cap KW) = H/K(H \cap W)$$

we get that $HKW/KW \in \Sigma$ since $(H \cap W)K/K \le H/K$. Similarly, one can get that $VKW/KW \in \Sigma$. Moreover,

$$HV/KW = (HKW/KW)(VKW/KW)$$

and so $HV/KW \in \Sigma$. Hence statement (i) holds.

(ii) This statement follows from the fact that for every subgroup H of G with $L \le H \le T$ we have $L \le H_G$ and $H^G \le T$.

(iii) This assertion can be proved by direct checking.

(iv) First note that, in view of part (i), $V/W \in \Sigma$ always implies that $VN/WN \in \Sigma$, so every normal section of G/N in $\Sigma N/N$ is of the form $(V/N)/(W/N)$ for some $V/W \in \Sigma$.

(1) $\Sigma N/N$ *is* (G/N)*-closed.*

Indeed, if

$$(H/N)/(K/N) \simeq_{G/N} (V/N)/(W/N) \in \Sigma N/N,$$

then $H/K \simeq_G (V/W) \in \Sigma$. Hence $H/K \in \Sigma$, so $(H/N)/(K/N) \in \Sigma N/N$.

(2) *For every two normal sections* $(H/N)/(K/N)$ *and* $(T/N)/(K/N)$ *of* G/N*, where* $H/N \le T/N$ *and* $(T/N)/(K/N) \in \Sigma N/N$ *both sections* $(H/N)/(K/N)$ *and* $(T/N)/(H/N)$ *belong to* $\Sigma N/N$. (This assertion is evident.)

(3) $(H/N)/((K/N) \cap (L/N)) \in \Sigma N/N$ *for every two normal sections* $(H/N)/(K/N)$, $(H/N)/(L/N) \in \Sigma N/N$.

From

$$(H/N)/(K/N),\ (H/N)/(L/N) \in \Sigma N/N$$

we get that H/K, $H/L \in \Sigma$ and so $H/(K \cap L) \in \Sigma$, which implies that

$$(H/N)/((K/N) \cap (L/N)) = (H/N)/((K \cap L)/N) \in \Sigma N/N.$$

(4) $(H/N)(V/N)/(K/N) \in \Sigma N/N$ *for every two normal sections* $(H/N)/(K/N)$, $(V/N)/(K/N) \in \Sigma N/N$.

From $(H/N)/(K/N)$, $(V/N)/(K/N) \in \Sigma N/N$ it follows that $HV/K \in \Sigma$, which implies that $(H/N)(V/N)/$ $/(K/N) \in \Sigma N/N$.

Hence statement (iv) holds.

The proposition is proved.

Before proceeding, consider some examples.

Example 1. (i) If $\mathfrak{X} = \{G/1\}$, then

$$\operatorname{formfit}(G/1) = \{H/K \mid H,\ K \trianglelefteq G\}$$

and so

$$\mathfrak{L}_{\operatorname{formfit}(G/1)}(G) = \mathfrak{L}(G).$$

(ii) If $\mathfrak{F}$ is the class of all identity groups, then $\mathfrak{L}_{\Sigma_G(\mathfrak{F})}(G) = \mathfrak{L}_n(G)$.

(iii) Let $p > q > 2$ be primes, where q divides $p - 1$. Let Q be a non-abelian group of order q^3. Then Q has a unique minimal normal subgroup, so there exists a simple $\mathbb{F}_p Q$-module P which is faithful for Q. Then $|P| > p$. Let $G = (P \rtimes Q) \times (C_p \rtimes C_q)$, where $C_p \rtimes C_q$ is a non-abelian group of order pq. Let Δ is the set of all those chief factors of G on which G induces an abelian group of automorphisms. Then

$$\mathfrak{L}(P) \not\subseteq \mathfrak{L}_{\Sigma_G(\Delta)}(G) = \mathfrak{L}_n(G) \cup \{AC_q^x \mid A \trianglelefteq G,\ x \in G\}.$$

Therefore for every Fitting formation $\mathfrak{F}$ we have $\mathfrak{L}_{\Sigma_G(\Delta)}(G) \ne \mathfrak{L}_{\Sigma_G(\mathfrak{F})}$ since otherwise $P \in \mathfrak{F}$ and so

$$\mathfrak{L}(P) \subseteq \mathfrak{L}_{\Sigma_G(\mathfrak{F})} = \mathfrak{L}_{\Sigma_G(\Delta)}(G).$$

(iv) Let A be a non-abelian simple group and $\mathfrak{F}$ the class of all groups B such that either $B = 1$ or B is the direct product of isomorphic copies of A. Let $G = A_0 \wr A = K \rtimes A$, where $A_0 \simeq A$ and $K = A_1 \times \cdots \times A_{|A|}$ is the base group of the regular wreath product G. Then K is the unique minimal normal subgroup of G by [6, chapter A, proposition 18.5]. Moreover,

$$\Sigma := \Sigma_G(\mathfrak{F}) = \{G/K,\ K/1,\ G/G,\ K/K,\ 1/1\}$$

is clearly a formation Fitting set of G, so $\mathfrak{L}_{\Sigma_G(\mathfrak{F})}(G)$ is a sublattice of $\mathfrak{L}(G)$. We show that $\mathfrak{L}_{\Sigma_G(\mathfrak{F})} \ne \mathfrak{L}_{\Sigma_G(\Delta)}(G)$ for every G-closed set Δ of chief factors of G. Indeed, assume that $\mathfrak{L}_{\Sigma_G(\mathfrak{F})}(G) = \mathfrak{L}_{\Sigma_G(\Delta)}(G)$. Then for all subgroups $L \le K$ and $K \le R \le G$ we have $L^G/L_G = K/1$ and $R^G/R_G = G/K$, so $L, R \in \mathfrak{L}_{\Sigma(\Delta)}(G)$. Therefore $R/1$, $G/K \in \Delta$ and

hence $G/1 \in \Sigma_G(\Delta)$. Thus $\mathfrak{L}_{\Sigma_G(\Delta)}(G) = \mathfrak{L}(G)$ and so $A \in \mathfrak{L}_{\Sigma_G(\mathfrak{F})}(G)$. But then $G/1 = A^G/A_G \in \mathfrak{F}$, which means that G is the direct product of isomorphic copies of A. This contradiction shows that

$$\mathfrak{L}_{\Sigma_G(\mathfrak{F})} \neq \mathfrak{L}_{\Sigma_G(\Delta)}(G)$$

for every G-closed set Δ of chief factors of G.

(v) The class of groups $\mathfrak{F}$ is called a *saturated* if $\mathfrak{F}$ contains every finite group G with $G/\Phi(G) \in \mathfrak{F}$.

Now let A be a maximal subgroup of a finite group G and let $\mathfrak{F}$ be a saturated Fitting formation. Let Δ be the set of all $\mathfrak{F}$-central chief factors of G. Then $G/A_G = A^G/A_G \in \mathfrak{F}$ if and only if $A^G/A_G \in \Sigma_G(\Delta)$ (see lemma 5 below). Therefore $A \in \mathfrak{L}_{\Sigma_G(\mathfrak{F})}(G)$ if and only if $A \in \mathfrak{L}_{\Sigma_G(\Delta)}(G)$.

In conclusion of this section note that some special versions of theorems 2 and 3 were proved in the papers [8–10]. In particular, in the paper [9], the following results were proved.

Corollary 4 (see theorem 1.4(ii) in [9]). *Let G be a finite group and $\Sigma = \Sigma(\Delta)$, where Δ is the set of all central chief factors of G. Then the lattice $\mathfrak{L}_\Sigma(G)$ is distributive if and only if $\mathfrak{L}_\Sigma(G) = \mathfrak{L}_n(G)$ is distributive.*

Corollary 5 (see theorem 1.2 in [9]). *Let G be a finite group and either $\Sigma = \Sigma(\Delta)$, where Δ is the set of all $\mathfrak{F}$-central chief factors of G for some class of groups containing all identity groups $\mathfrak{F}$, or $\Sigma = \Sigma_G(\mathfrak{F})$ for some Fitting formation $\mathfrak{F}$, then $\mathfrak{L}_\Sigma(G)$ is a sublattice in $\mathfrak{L}(G)$.*

Some further applications

A group is called *primary* if it is a finite p-group for some prime p. If $\sigma = \{\sigma_i \mid i \in I\}$ is any partition of the set of all primes $\mathbb{P}$, that is, $\mathbb{P} = \bigcup_{i \in I} \sigma_i$ and $\sigma_i \cap \sigma_j = \varnothing$ for all $i \neq j$, then we say, following [11], that the group G is: σ-*primary* if it is a finite σ_i-group for some i; σ-*soluble* if G is finite and every its chief factor is σ-primary; σ-*nilpotent* or σ-*decomposable* [12] if $G = G_1 \times \cdots \times G_n$ for some σ-primary groups $G_1, \ldots, G_n$. Observe that a finite group is primary (respectively soluble, nilpotent) if and only if it is σ-primary (respectively σ-soluble, σ-nilpotent), where $\sigma = \{\{2\}, \{3\}, \ldots\}$.

In this section we discuss some applications of the lattice $\mathfrak{L}_\Sigma(G)$ in the theory of finite groups. And we start with one application of the lattices $\mathfrak{L}_{\Sigma_G(\mathfrak{N}_\sigma)}(G)$ and $\mathfrak{L}_{\Sigma_G(\Delta)}(G)$, where $\mathfrak{N}_\sigma$ is the class of all σ-nilpotent groups and Δ is the set of all σ-central, that is, $\mathfrak{N}_\sigma$-central chief factors of G, in the theory of generalized T-groups.

Lattice characterizations of finite σ-soluble $P\sigma T$-groups. We say, following [11], that the subgroup A of G is σ-*subnormal in* G if it is $\mathfrak{N}_\sigma$-*subnormal in* G in the sense of Kegel. Note that a subgroup A of G is subnormal in G if and only if A is σ-subnormal in G, where $\sigma = \{\{2\}, \{3\}, \ldots\}$.

A subgroup A of a finite group G is said to be: *quasinormal* (respectively *S-quasinormal* or *S-permutable* [13]) in G if A permutes with all subgroups (respectively with all Sylow subgroups) H of G, that is, $AH = HA$; σ-*permutable* in G [11] if A permutes with all Hall σ_i-subgroups of G for all i.

Recall that a finite group G is said to be a *T-group* (respectively *PT-group*, *PST-group*) if every subnormal subgroup of G is normal (respectively permutable, S-permutable) in G; G is said to be a $P\sigma T$-*group* if every σ-subnormal subgroup of G is σ-permutable in G.

The description of *PST*-groups, that are groups, in which every subnormal subgroup is S-permutable, was first obtained by Agrawal [14], for the soluble case, and by Robinson in [15], for the general case. In the further publications, authors (see, for example, the recent papers [16–25]) have found out and described many other interesting characterizations of soluble *PST*-groups. Some characterizations of $P\sigma T$-groups were obtained in the papers [11; 26]. Theorem 2.4 allows to prove the following result in this line research.

Theorem 4. *Suppose that G is a finite σ-soluble group. Then G is a $P\sigma T$-group if and only if $\mathfrak{L}_{\Sigma_G(\mathfrak{N}_\sigma)}(G) = \mathfrak{L}_{\Sigma_G(\Delta)}(G)$, where Δ is the set of all σ-central chief factors of G.*

The proof of theorem 4 consists of many steps and it uses theorems 1 and 2 and also the following lemmas.

Lemma 3. *Let $\mathfrak{F}$ be a class of groups, N be a normal subgroup of G and Σ be a formation Fitting set of G.*

(1) If $\Sigma = \Sigma_G(\Delta)$, where Δ is the set of all $\mathfrak{F}$-central chief factors of G, then $\Sigma N/N = \Sigma_{G/N}(\Delta^*)$, where Δ^* is the set of all $\mathfrak{F}$-central chief factors of G/N.

(2) $\Sigma_G(\mathfrak{F})N/N = \Sigma_{G/N}(\mathfrak{F})$.

P r o o f. (1) This follows from proposition (iv) and the fact that a chief factor $(H/N)/(K/N)$ is $\mathfrak{F}$-central in G/N if and only if the chief factor H/K is $\mathfrak{F}$-central in G (see lemma 2(1)).

(2) This follows from proposition (iv).

The lemma is proved.

Lemma 4. *Let Σ be a formation Fitting set of G and let $A \in \mathfrak{L}_\Sigma(G)$ and $N \le H \le G$, where $N \trianglelefteq G$:*

(1) $AN/N \in \mathfrak{L}_{\Sigma N/N}(G/N)$;

(2) if $H/N \in \mathfrak{L}_{\Sigma N/N}(G/N)$, then $H \in \mathfrak{L}_\Sigma(G)$;

(3) $A \cap E \in \mathfrak{L}_{\mathrm{formfit}(\Sigma \cap E)}(E)$ for every subgroup E of G.

P r o o f. (1) Since $A \in \mathfrak{L}_\Sigma(G)$, $A^G/A_G \in \Sigma$ and so

$$\left(A^G N/N\right)/\left(A_G N/N\right) \in \Sigma N/N.$$

On the other hand, we have that

$$\left(AN/N\right)^{G/N} = \left(AN\right)^G/N = A^G N/N,$$

where $A_G N/N \le \left(AN/N\right)_{G/N}$. Hence

$$\left(AN/N\right)^{G/N}/\left(AN/N\right)_{G/N} \in \Sigma N/N$$

since $\Sigma N/N$ is a formation Fitting set of G/N by proposition (iv), so $AN/N \in \mathfrak{L}_{\Sigma N/N}(G/N)$.

(2) Since $H/N \in \mathfrak{L}_{\Sigma N/N}(G/N)$, we have

$$\left(H^G/N\right)/\left(H_G/N\right) = \left(H/N\right)^{G/N}/\left(H/N\right)_{G/N} \in \Sigma N/N$$

and so $H^G/H_G \in \Sigma$ by proposition (i). Hence $H \in \mathfrak{L}_\Sigma(G)$.

(3) Let $\Sigma_0 = \mathrm{formfit}(\Sigma \cap E)$. It is clear that

$$\left(A^G \cap E\right)/\left(A_G \cap E\right) \in \Sigma \cap E \subseteq \Sigma_0.$$

On the other hand, we have

$$A_G \cap E \le \left(A \cap E\right)_E \le A \cap E \le \left(A \cap E\right)^E \le A^G \cap E$$

and so $\left(A \cap E\right)^E / \left(A \cap E\right)_E \in \Sigma_0$ since Σ_0 is a formation Fitting set of E. Hence $A \cap E \in \mathfrak{L}_{\Sigma_0}(E)$.

The lemma is proved.

Lemma 5. *Let $\mathfrak{F}$ be a saturated formation and G be a finite group:*

(1) if $G \in \mathfrak{F}$, then every chief factor of G is $\mathfrak{F}$-central in G;

(2) if G has a normal subgroup N with $G/N \in \mathfrak{F}$ such that every chief factor of G below N is $\mathfrak{F}$-central in G, then $G \in \mathfrak{F}$.

P r o o f. (1) This part directly follows from the Barnes – Kegel result [6, chapter IV, proposition 1.5].

(2) In fact, in view of part (1) and the Jordan – Hölder's theorem for the chief series, it is enough to show that if every chief factor of G is $\mathfrak{F}$-central in G, then $G \in \mathfrak{F}$. Assume that this is false and let G be a counter-example of minimal order. Then G has a unique minimal normal subgroup, R say, and $R \not\le \Phi(G)$. Moreover, R is abelian since otherwise we have $G \simeq G/C_G(R) = G/1 \in \mathfrak{F}$. Hence $R = C_G(R)$ by [6, chapter A, theorem 15.6] and for some maximal subgroup M of G we have $G = R \rtimes M$. Therefore the map

$$f : G \to R \rtimes \left(G/C_G(R)\right) = R \rtimes \left(G/R\right)$$

with $f(rm) = (r, mR)$ for all $r \in R$ and $m \in M$ is isomorphism, so $G \in \mathfrak{F}$ since the factor $R/1$ is $\mathfrak{F}$-central in G by hypothesis.

The lemma is proved.

Recall that the σ-*nilpotent residual* $G^{\mathfrak{N}_\sigma}$ of a finite groups G is the intersection of all normal subgroups N of G with σ-nilpotent quotient G/N.

Lemma 6 (see theorem A in [26]). *Let $D = G^{\mathfrak{N}_\sigma}$ be the σ-nilpotent residual of a finite group G. If G is σ-soluble $P\sigma T$-group, then the following conditions hold:*

(1) $G = D \rtimes M$, where D is an abelian Hall subgroup of G of odd order, M is σ-nilpotent and every element of G induces a power automorphism in D;

(2) $O_{\sigma_i}(D)$ has a normal complement in a Hall σ_i-subgroup of G for all i.

Conversely, if conditions (1) and (2) hold for some subgroups D and M of G, then G is a $P\sigma T$-group.

Lemma 7. *Let N be a normal subgroup of a finite group G such that every chief factor of G below N is G-central in G. Then N is σ-nilpotent, and if N is a σ_i-group, then $O^{\sigma_i}(G) \le C_G(N)$.*

P r o o f. Let $1 = Z_0 < Z_1 < \ldots < Z_t = N$ be a chief series of G below N and $C_i = C_G\left(Z_i/Z_{i-1}\right)$. First we show that N is σ-nilpotent. By hypothesis, Z_1 and G/G_1 are σ_j-groups for some j. Now let H/K be any chief factor of N such that $H \le Z_1$. From the isomorphism $C_1N/N \simeq N/\left(C_1 \cap N\right)$ it follows that H/K and $N/C_N\left(H/K\right)$ are σ_j-groups. Therefore every chief factor of N below Z_1 is N_σ-central in N. On the other hand, N/Z_1 is σ-nilpotent by induction and so N is σ-nilpotent by lemma 5, condition (2).

Finally, assume that N is a σ_i-group and let $C = C_1 \cap \ldots \cap C_t$. Then G/C is a σ_i-group. On the other hand, $C/C_G\left(N\right) \simeq A \le Aut\left(N\right)$ stabilizes the series $1 = Z_0 < Z_1 < \ldots < Z_t = N$, so $C/C_G\left(N\right)$ is a $\pi\left(N\right)$-group by [6, chapter A, corollary 12.4]. Hence $C/C_G\left(N\right)$ is a σ_i-group, so $O^{\sigma_i}\left(G\right) \le C_G\left(N\right)$. The lemma is proved.

Now consider some applications of theorem 4.

Recall that $Z_\sigma\left(G\right)$ denotes the σ-*hypercentre* of G [11], that is, the largest normal subgroup of G such that every chief factor of G below $Z_\sigma\left(G\right)$ is σ-central in G. We say, following [13, p. 20], that a subgroup H of a finite group G is σ-*hypercentrally embedded in* G if $H/H_G \le Z_\sigma\left(G/H_G\right)$ and *hypercentrally embedded in* G if $H/H_G \le Z_\infty\left(G/H_G\right)$.

Corollary 6 (see theorem 4.1 in [11]). *Let G be a finite σ-soluble group. If every σ-subnormal subgroup of G is σ-hypercentrally embedded in G, then G is a $P\sigma T$-group.*

In the case where $\sigma = \left\{\left\{2\right\}, \left\{3\right\}, \ldots\right\}$ we get from theorem 3.1 the following known characterization of finite soluble *PST*-groups.

Corollary 7 (see theorem 1.3 in [10]). *Suppose that G is a finite soluble group. Then G is a PST-group if and only if $\mathfrak{L}_{\Sigma_G(\mathfrak{N})}\left(G\right) = \mathfrak{L}_{\Sigma(\Delta)}\left(G\right)$, where Δ is the set of all central chief factors H/K of G, that is, $C_G\left(H/K\right) = G$.*

Corollary 8 (see theorem 2.4.4 in [13]). *Let G be a finite group. G is a soluble PST-group if and only if every subnormal subgroup H of G is hypercentrally embedded in G (that is $H/H_G \le Z_\infty\left(G/H_G\right)$).*

Groups with Σ-normal and Σ-abnormal subgroups. Let Σ be a formation Fitting set of G. Then we say that a subgroup A of G is: (i) Σ-*normal* in G if $A \in \mathfrak{L}_\Sigma\left(G\right)$; (ii) Σ-*abnormal* in G provided $H \notin \mathfrak{L}_{\mathrm{formfit}(\Sigma \cap E)}\left(E\right)$ for all subgroups $H < E$ of G, where $A \le H$.

Example 2. (i) A subgroup A of G is normal in G if and only if it is Σ-normal in G, where $\Sigma = \left\{H/H \mid H \trianglelefteq G\right\}$.

(ii) A subgroup A of G is called *abnormal* in G if $g \in \left\langle A, A^g \right\rangle$ for all $g \in G$. If G is a soluble finite group, then A is abnormal in G if and only if A is Σ-abnormal in G, where $\Sigma = \Sigma_G\left(\mathfrak{N}\right)$, by [12, chapter IV, theorem 1.7.1].

(iii) Let Δ be the set of all $\mathfrak{F}$-central chief factors of G and $\Sigma = \Sigma_G\left(\Delta\right)$. If G is finite, then a subgroup A of G is called: (a) $\mathfrak{F}$-*normal* in G [8] if $A^G/A_G \in \Sigma$, (b) $\mathfrak{F}$-*abnormal* in G [8] if H is not $\mathfrak{F}$-normal in E for every two subgroups $H < E$ of G such that $A \le H$. Therefore a subgroup A of G is $\mathfrak{F}$-normal ($\mathfrak{F}$-abnormal) in G if and only if it is Σ-normal (respectively Σ-abnormal) in G, where $\Sigma = \Sigma_G\left(\Delta\right)$.

(iv) Let G be finite. If A is σ-hypercentrally embedded in G, that is, $A/A_G \le Z_\sigma\left(G/A_G\right)$, then $A^G/A_G \le Z_\sigma\left(G/A_G\right)$. In particular, if A is hypercentrally embedded in G, then $A^G/A_G \le Z_\infty\left(G/A_G\right)$. Therefore A is σ-hypercentrally (hypercentrally) embedded in G if and only if it is Σ-abnormal in G, where $\Sigma = \Sigma_G\left(\Delta\right)$ and Δ is the set of all σ-central (respectively central) chief factors of G.

Recall that a finite group G is a *DM-group* [8] if $G = D \rtimes M$ and the following conditions hold: (1) $D = G' \ne 1$ is abelian; (2) $M = \left\langle x \right\rangle$ is a cyclic abnormal Sylow p-subgroup of G, where p is the smallest prime dividing $\left|G\right|$; (3) $M_G = \left\langle x^p \right\rangle = Z\left(G\right)$; (4) x induces a fixed-point-free power automorphism on D.

In the paper [27], Fattahi defined B-groups to be a finite groups in which every subgroup is either normal or abnormal and he showed that a non-nilpotent finite group G is a B-group if and only if G is a *DM*-group. As a generalization of this result, Ebert and Bauman classified the group in which every subgroup is either subnormal or abnormal [28]. In further, the results in [27] have been developed in many other directions (see, for example, the recent papers [8; 29–33]).

We say that G is a ΣNA-*group* if every subgroup of G is either Σ-normal or Σ-abnormal in G for some formation Fitting set Σ of G.

The results in [8; 27–33] and also many other known results of this type are the motivation for the following question.

Question 1. Let Σ be a formation Fitting set of a finite group G. What we can say about the structure of G in the case when at least one of the following conditions holds: *(i) every subgroup of G is Σ-normal in G; (ii) G is a ΣNA-group, where $\Sigma = \Sigma_G(\Delta)$ for some G-closed set Δ of chief factors of G or $\Sigma = \Sigma_G(\mathfrak{F})$ for some hereditary (in the sense of Mal'cev* [34]*) Fitting formation $\mathfrak{F}$?*

Note that the answer to question 1 for some special Σ is known. Let, for example, $\Sigma = \{H/H \mid H \trianglelefteq G\}$. Then: (i) every subgroup of G is Σ-normal in G if and only if G is a Dedekind group; (ii) G is a ΣNA-group if and only if G is a P-group by example 2(i) and 2(ii) since every P-group is clearly soluble.

Now let Δ be the set of all $\mathfrak{F}$-central chief factors of a finite group G and $\Sigma = \Sigma_G(\Delta)$, where $\mathfrak{F}$ is a hereditary saturated formation containing all nilpotent groups. Then G is a ΣNA-group if and only if every subgroup of G is either $\mathfrak{F}$-normal or $\mathfrak{F}$-abnormal in G by example 2(iii). Such a class of finite groups is also known.

Theorem 5 (see theorem 1.4 in [8]). *Let $\mathfrak{F}$ be a hereditary saturated formation containing all nilpotent groups. If every subgroup of a finite group G is either $\mathfrak{F}$-normal or $\mathfrak{F}$-abnormal in G, then G is of either of the following types:*

(I) $G \in \mathfrak{F}$;

(II) $G = D \rtimes M$ is a DM-group, where $D = G^{\mathfrak{F}}$, and M is an $\mathfrak{F}$-abnormal subgroup of G with $M_G = Z_{\mathfrak{F}}(G)$.

Conversely, in a group G of type (I) or (II) every subgroup is either $\mathfrak{F}$-normal or $\mathfrak{F}$-abnormal.

In this theorem $Z_{\mathfrak{F}}(G)$ denotes the $\mathfrak{F}$-*hypercentre of* G, that is the product of all normal subgroups N of G such that either $N = 1$ or $N \neq 1$ and every chief factor of G below N is $\mathfrak{F}$-central in G.

Finite groups G with modular lattices $\mathfrak{L}_\Sigma(G)$ and $\mathfrak{L}_{sn}(G)$. A subgroup A of G is called: *subnormal* in G if there exists a subgroup series $A = A_0 \trianglelefteq A_1 \trianglelefteq \cdots \trianglelefteq A_{t-1} \trianglelefteq A_t = G$ $(*)$; *composition* in G if every factor A_i/A_{i-1} of the series $(*)$ is a simple group. Note that a subgroup A of a finite group G is subnormal in G if and only if it is composition in G.

Now let Σ be a formation Fitting set of G. We say a subgroup A of G is Σ-*subnormal* in G if there exists a subgroup series $A = A_0 \trianglelefteq A_1 \trianglelefteq \cdots \trianglelefteq A_{t-1} \trianglelefteq A_t = G$ of G such that A_{i-1} is Σ_i-normal in A_i, where $\Sigma_i = \mathrm{formfit}(\Sigma \cap A_i)$, for all $i = 1, \ldots, t$.

By classical Wielandt's result [35, theorem 1.1.5], the set $\mathfrak{L}_{sn}(G)$ of all composition subgroups of G forms a sublattice of $\mathfrak{L}(G)$.

Question 2. Let G be finite. For which conditions on the formation Fitting set Σ of G the set of all Σ-subnormal subgroups of G forms a sublattice of $\mathfrak{L}(G)$?

In some special cases the answer to question 2 is known. Indeed, $\mathfrak{L}_n(G) = \mathfrak{L}_\Sigma(G)$, where $\Sigma = \{H/H \mid H \trianglelefteq G\}$, is modular. In the paper [9] the following result in this direction was obtained.

Theorem 6 (see theorem 1.4 in [9]). *Let G be finite and $\Sigma = \Sigma_G(\Delta)$, where Δ is the set of all central chief factors of G. Then the lattice $\mathfrak{L}_\Sigma(G)$ is modular if and only if every two subgroups $A, B \in \mathfrak{L}_\Sigma(G)$ are permutable, that is $AB = BA$.*

Zappa, in his paper [36], described conditions under which the lattice $\mathfrak{L}_{sn}(G)$, where G is finite, is modular.

Theorem 7 (see theorem 9.2.3 in [35]). *The following properties of the finite group G are equivalent:*

(a) the lattice $\mathfrak{L}_{sn}(G)$ is modular;

(b) if $T \trianglelefteq S$, where S is subnormal in G and S/T is a p-group, p a prime, then $\mathfrak{L}(S/T)$ is modular;

(c) if $T \trianglelefteq S$, where S is subnormal in G and $|S/T| = p^3$, p a prime, then $\mathfrak{L}(S/T)$ is modular.

A new characterization of finite groups with modular lattice of the subnormal subgroups was given in the paper [9].

Theorem 8 (see theorem 1.3 in [9]). *Let G be a finite group. Then the lattice $\mathfrak{L}_{sn}(G)$ is modular if and only if for every two subnormal subgroups $L \leq T$ of G, where $L \in \mathfrak{L}_\Sigma(T)$ and $\Sigma = \Sigma_T(\mathfrak{N}^*)$, L permutes with every subnormal subgroup M of T.*

Finite groups factorized by Σ-normal subgroups. It is well-known that the product $G = AB$ of two normal finite supersoluble groups A and B is not supersoluble in general. Nevertheless, such a product is supersoluble if the indices $|G:A|$ and $|G:B|$ are coprime [37, chapter 4, theorem 3.4]. Moreover, by Doerk's result [38], the finite group G is supersoluble if it has four supersoluble subgroups A_1, A_2, A_3, A_4 whose indices $|G:A_1|, |G:A_2|, |G:A_3|, |G:A_4|$ are pairwise coprime. In this paper, we prove the following result in this line research.

Theorem 9. *Suppose that G is finite and let* Δ *is the set of all cyclic chief factors of G and* $\Sigma = \Sigma_G(\Delta)$. *Then G is supersoluble if and only if G has three* Σ*-normal supersoluble subgroups* A_1, A_2, A_3 *whose indices* $|G:A_1|$, $|G:A_2|$, $|G:A_3|$ *are pair coprime.*

Lemma 8 (see lemma 4.5 in [6, chapter IV]). *Let G be a finite group in* $\mathfrak{F}$, *where* $\mathfrak{F}$ *is a saturated Fitting formation and let* $p \in \pi(G)$. *If* $X = G/O_{p',p}(G)$ *and R is an irreducible* $\mathbb{F}_p X$*-module, then* $R \rtimes X \in \mathfrak{F}$.

Proof of theorem 9. We need only to show that the sufficiency of the condition of the theorem holds. Assume that this is false and let G be a counterexample of minimal order. Then $G \neq A_i \neq 1$ for all i and G is soluble by Wielandt's theorem [6, chapter I, theorem 3.4]. Moreover, from $(|G:A_i|, |G:A_j|) = 1$ for $i \neq j$ it follows that $G = A_1A_2 = A_1A_3 = A_2A_3$.

Let R be a minimal normal subgroup of G. Then R is a p-group for some prime p. Note also that $\Sigma R/R = \Sigma_{G/R}(\Delta^*)$, where Δ^* is the set of all cyclic chief factors of G/R by lemma 3(1). On the other hand, the subgroup A_iR/R belongs the lattice $\mathfrak{L}_{\Sigma R/R}(G)$ by lemma 4(1), so $A_iR/R \in \mathfrak{L}_{\Sigma_{G/R}(\Delta^*)}(G/R)$. Note also that $A_iR/R \simeq A_i/(A_i \cap R)$ is supersoluble. Therefore the hypothesis hods for G/R. Hence G/R is supersoluble, so R is the unique minimal normal subgroup of G and $R \not\leq \Phi(G)$. Thus $R = C_G(R) = O_p(G)$ for some prime p by [6, chapter A, theorem 15.6]. Let G_p be a Sylow p-subgroup of G.

From the hypothesis it follows that for some $i \neq j$ and some $x, y \in G$ we have $R \leq G_p^x \leq A_i$ and $R \leq G_p^y \leq A_j$. Since $R = C_G(R)$, $F(A_i) = O_p(A_i)$. On the other hand, A_i is supersoluble and so $A_i/F(A_i) = A_i/O_p(A_i)$ is abelian. Hence $A_i \leq N_G(G_p^x)$. It follows that $A_i^{x^{-1}} \leq N_G(G_p)$. Similarly, $A_j^{y^{-1}} \leq N_G(G_p)$. Then

$$G = A_iA_j = A_i^{x^{-1}}A_j^{y^{-1}} \leq N_G(G_p)$$

and so

$$R = O_p(G) = G_p = O_p(A_i) = O_p(A_j).$$

Now we show that $R \leq A_k$, where $j \neq k \neq i$. Assume that $R \not\leq A_k$. Then $(A_k)_G = 1$ and $A_k^G \neq 1$ since $A_k \neq 1$. Hence $R \leq A_k^G$, which implies that $R/1$ is cyclic and so G is supersoluble. This contradiction shows that $R \leq A_3$, so $R = G_p = O_p(A_k) = F(A_k)$.

Therefore A_1R/R, A_2R/R, A_3R/R are abelian subgroup of G/R whose indices

$$|G/R : A_1R/R|, \ |G/R : A_2R/R|, \ |G/R : A_3R/R|$$

are pair coprime, so G/R is nilpotent by Kegel's theorem [39]. Moreover, for every Sylow subgroup Q/R of G/R we have that $Q/R \leq A_i/R$ or $Q/R \leq A_j/R$. Hence for some subgroups $A/R \trianglelefteq A_i/R$ and $B/R \trianglelefteq A_j/R$ we have $G/R = (A/R) \times (B/R)$. It is clear that the subgroups A and B are supersoluble and so the group $A \times B$ is supersoluble. It is clear also that $O_{p',p}(A) = R = O_{p',p}(B)$. Hence

$$X = (A \times B)/O_{p',p}(A \times B) \simeq (A/R) \times (B/R) \simeq G/R.$$

But then G is supersoluble by lemma 8. This contradiction completes the proof of the result.

A subgroup M of G is called *modular* in G if M is a modular element (in the sense of Kurosh [35, p. 43]) of the lattice $\mathfrak{L}(G)$. It is known that [35, theorem 5.2.3] for every modular subgroup A of G all chief factors of G between A_G and A^G are cyclic. Therefore we get from theorem 9 the following result.

Corollary 9. *If G is finite and G has three modular supersoluble subgroups* A_1, A_2, A_3 *whose indices* $|G:A_1|$, $|G:A_2|$, $|G:A_3|$ *are pair coprime, then G is supersoluble.*

Библиографические ссылки

1. Wielandt H. Eine Verallgemenerung der invarianten Untergruppen. *Mathematische Zeitschrift.* 1939;45(1):209–244. DOI: 10.1007/BF01580283.

2. Kegel OH. Untergruppenverbände endlicher Gruppen, die Subnormalteilorverband echt enthalten. *Archiv der Mathematik.* 1978; 30(1):225–228. DOI: 10.1007/BF01226043.

3. Ballester-Bolinches A, Ezquerro LM. *Classes of Finite Groups.* Dordrecht: Springer; 2006. 381 p. (Mathematics and its applications; volume 584). DOI: 10.1007/1-4020-4719-3.

4. Ballester-Bolinches A, Doerk K, Pèrez-Ramos MD. On the lattice of $\mathfrak{F}$-subnormal subgroups. *Journal of Algebra.* 1992;148(1): 42–52. DOI: 10.1016/0021-8693(92)90235-E.

5. Васильев АФ, Каморников СФ, Семенчук ВН. О решетках подгрупп конечных групп. В: *Бесконечные группы и примыкающие алгебраические структуры.* Киев: Институт математики АН Украины; 1993. с. 27–54.

6. Doerk K, Hawkes T. *Finite soluble groups.* Berlin: Walter de Gruyter; 1992. 910 p. (de Gruyter expositions in mathematics; book 4).

7. Шеметков ЛА, Скиба АН. *Формации алгебраических систем.* Москва: Наука; 1989. 256 с.

8. Hu B, Huang J, Skiba AN. Finite groups with only $\mathfrak{F}$-normal and $\mathfrak{F}$-abnormal subgroups. *Journal of Group Theory.* 2019;22: 915–926. DOI: 10.1515/jgth-2018-0199.

9. Chi Z, Skiba AN. On two sublattices of the subgroup lattice of a finite group. *Journal of Group Theory.* 2019;22(6):1035–1047. DOI: 10.1515/jgth-2019-0039.

10. Chi Z, Skiba AN. On a lattice characterization of finite soluble *PST*-groups. *Bulletin of the Australian Mathematical Society.* 2019;99(3):1–8. DOI: 10.1017/S0004972719000741.

11. Skiba AN. On σ-subnormal and σ-permutable subgroups of finite groups. *Journal of Algebra.* 2015;436:1–16. DOI: 10.1016/j.jalgebra.2015.04.010.

12. Шеметков ЛА. *Формации конечных групп.* Москва: Наука; 1978. 272 с.

13. Ballester-Bolinches A, Esteban-Romero R, Asaad M. *Products of Finite Groups.* Berlin: Walter de Gruyter; 2010. (de Gruyter expositions in mathematics; volume 53). DOI: 10.1515/9783110220612.

14. Agrawal RK. Finite groups whose subnormal subgroups permute with all Sylow subgroups. *Proceedings of the American Mathematical Society.* 1975;47:77–83. DOI: 10.1090/S0002-9939-1975-0364444-4.

15. Robinson DJS. The structure of finite groups in which permutability is a transitive relation. *Journal of the Australian Mathematical Society.* 2001;70(2):143–160. DOI: 10.1017/S1446788700002573.

16. Brice RA, Cossey J. The Wielandt subgroup of a finite soluble groups. *Journal of the London Mathematical Society.* 1989; 40(2):244–256. DOI: 10.1112/jlms/s2-40.2.244.

17. Beidleman JC, Brewster B, Robinson DJS. Criteria for permutability to be transitive in finite groups. *Journal of Algebra.* 1999; 222(2):400–412. DOI: 10.1006/jabr.1998.7964.

18. Ballester-Bolinches A, Esteban-Romero R. Sylow permutable subnormal subgroups of finite groups. *Journal of Algebra.* 2002; 251(2):727–738. DOI: 10.1006/jabr.2001.9138.

19. Ballester-Bolinches A, Beidleman JC, Heineken H. Groups in which Sylow subgroups and subnormal subgroups permute. *Illinois Journal of Mathematics.* 2003;47(1–2):63–69. DOI: 10.1215/ijm/1258488138.

20. Ballester-Bolinches A, Beidleman JC, Heineken H. A local approach to certain classes of finite groups. *Communications in Algebra.* 2003;31(12):5931–5942. DOI: 10.1081/AGB-120024860.

21. Asaad M. Finite groups in which normality or quasinormality is transitive. *Archiv der Mathematik.* 2004;83(4):289–296. DOI: 10.1007/s00013-004-1065-4.

22. Ballester-Bolinches A, Cossey J. Totally permutable products of finite groups satisfying *SC* or *PST*. *Monatshefte für Mathematik.* 2005;145(2):89–94. DOI: 10.1007/s00605-004-0263-9.

23. Al-Sharo KA, Beidleman JC, Heineken H, Ragland MF. Some characterizations of finite groups in which semipermutability is a transitive relation. *Forum Mathematicum.* 2010;22(5):855–862. DOI: 10.1515/forum.2010.045.

24. Beidleman JC, Ragland MF. Subnormal, permutable, and embedded subgroups in finite groups. *Central European Journal of Mathematics.* 2011;9(4):915–921. DOI: 10.2478/s11533-011-0098-8.

25. Yi X, Skiba AN. Some new characterizations of *PST*-groups. *Journal of Algebra.* 2014;399:39–54. DOI: 10.1016/j.jalgebra.2013.10.001.

26. Skiba AN. Some characterizations of finite σ-soluble $P\sigma T$-groups. *Journal of Algebra.* 2018;495:114–129. DOI: 10.1016/j.jalgebra.2017.11.009.

27. Fattahi A. Groups with only normal and abnormal subgroups. *Journal of Algebra.* 1974;28(1):15–19. DOI: 10.1016/0021-8693(74)90019-2.

28. Ebert G, Bauman S. A note on subnormal and abnormal chains. *Journal of Algebra.* 1975;36(2):287–293. DOI: 10.1016/0021-8693(75)90103-9.

29. Semenchuk VN, Skiba AN. On one generalization of finite $\mathfrak{U}$-critical groups. *Journal of Algebra and its Applications.* 2016; 15(4):1650063. DOI: 10.1142/S0219498816500638.

30. Монахов ВС. Конечные группы с абнормальными и $\mathfrak{U}$-субнормальными подгруппами. *Сибирский математический журнал.* 2016;57(2):447–462. DOI: 10.17377/smzh.2016.57.217.

31. Монахов ВС, Сохор ИЛ. Конечные группы с формационно субнормальными примарными подгруппами. *Сибирский математический журнал.* 2017;58(4):851–863. DOI: 10.17377/smzh.2017.58.412.

32. Monakhov VS, Sokhor IL. On groups with formational subnormal Sylow subgroups. *Journal of Group Theory.* 2018;21:273–287. DOI: 10.1515/jgth-2017-0039.

33. Monakhov VS, Sokhor IL. Finite groups with abnormal or formational subnormal primary subgroups. *Communications in Algebra.* 2019;47(10):3941–3949. DOI: 10.1080/00927872.2019.1572174.

34. Мальцев АИ. *Алгебраические системы.* Москва: Наука; 1970. 392 с.

35. Schmidt R. *Subgroup lattices of groups*. Berlin: Walter de Gruyter; 1994. 572 p. (de Gruyter expositions of mathematics; volume 14).

36. Zappa G. Sui gruppi finiti per cui il reticolo dei sottogruppi di composizione è modulare. *Bollettino dell'Unione Matematica Italiana. Serie 3.* 1956;11(3):315–318.

37. Bray HB. *Between nilpotent and solvable*. Weinstein M, editor. Passaic: Polygonal Publishing House; 1982. 231 p.

38. Doerk K. Minimal nicht überauflösbare, endlicher Gruppen. *Mathematische Zeitschrift.* 1966;91(3):198–205. DOI: 10.1007/BF01312426.

39. Kegel OH. Zur Struktur mehrafach faktorisierbarer endlicher Gruppen. *Mathematische Zeitschrift.* 1965;87(1):42–48. DOI: 10.1007/BF01109929.

References

1. Wielandt H. Eine Verallgemenerung der invarianten Untergruppen. *Mathematische Zeitschrift.* 1939;45(1):209–244. DOI: 10.1007/BF01580283.

2. Kegel OH. Untergruppenverbände endlicher Gruppen, die Subnormalteilorverband echt enthalten. *Archiv der Mathematik.* 1978; 30(1):225–228. DOI: 10.1007/BF01226043.

3. Ballester-Bolinches A, Ezquerro LM. *Classes of Finite Groups.* Dordrecht: Springer; 2006. 381 p. (Mathematics and its applications; volume 584). DOI: 10.1007/1-4020-4719-3.

4. Ballester-Bolinches A, Doerk K, Pèrez-Ramos MD. On the lattice of $\mathfrak{F}$-subnormal subgroups. *Journal of Algebra.* 1992;148(1): 42–52. DOI: 10.1016/0021-8693(92)90235-E.

5. Vasil'ev AF, Kamornikov SF, Semenchuk VN. [On lattices of subgroups of finite groups]. In: *Beskonechnye gruppy i primykayushchie algebraicheskie struktury* [Infinite groups and related algebraic structures]. Kiev: Institute of Mathematics of National Academy of Sciences of Ukraine; 1993. p. 27–54. Russian.

6. Doerk K, Hawkes T. *Finite soluble groups.* Berlin: Walter de Gruyter; 1992. 910 p. (de Gruyter expositions in mathematics; book 4).

7. Shemetkov LA, Skiba AN. *Formatsii algebraicheskikh sistem* [Formations of algebraic systems]. Moscow: Nauka; 1989. 256 p. Russian.

8. Hu B, Huang J, Skiba AN. Finite groups with only $\mathfrak{F}$-normal and $\mathfrak{F}$-abnormal subgroups. *Journal of Group Theory.* 2019;22: 915–926. DOI: 10.1515/jgth-2018-0199.

9. Chi Z, Skiba AN. On two sublattices of the subgroup lattice of a finite group. *Journal of Group Theory.* 2019;22(6):1035–1047. DOI: 10.1515/jgth-2019-0039.

10. Chi Z, Skiba AN. On a lattice characterization of finite soluble *PST*-groups. *Bulletin of the Australian Mathematical Society.* 2019;99(3):1–8. DOI: 10.1017/S0004972719000741.

11. Skiba AN. On σ-subnormal and σ-permutable subgroups of finite groups. *Journal of Algebra.* 2015;436:1–16. DOI: 10.1016/j.jalgebra.2015.04.010.

12. Shemetkov LA. *Formatsii konechnykh grupp* [Formations of finite groups]. Moscow: Nauka; 1978. 272 p. Russian.

13. Ballester-Bolinches A, Esteban-Romero R, Asaad M. *Products of Finite Groups.* Berlin: Walter de Gruyter; 2010. (de Gruyter expositions in mathematics; volume 53). DOI: 10.1515/9783110220612.

14. Agrawal RK. Finite groups whose subnormal subgroups permute with all Sylow subgroups. *Proceedings of the American Mathematical Society.* 1975;47:77–83. DOI: 10.1090/S0002-9939-1975-0364444-4.

15. Robinson DJS. The structure of finite groups in which permutability is a transitive relation. *Journal of the Australian Mathematical Society.* 2001;70(2):143–160. DOI: 10.1017/S1446788700002573.

16. Brice RA, Cossey J. The Wielandt subgroup of a finite soluble groups. *Journal of the London Mathematical Society.* 1989; 40(2):244–256. DOI: 10.1112/jlms/s2-40.2.244.

17. Beidleman JC, Brewster B, Robinson DJS. Criteria for permutability to be transitive in finite groups. *Journal of Algebra.* 1999; 222(2):400–412. DOI: 10.1006/jabr.1998.7964.

18. Ballester-Bolinches A, Esteban-Romero R. Sylow permutable subnormal subgroups of finite groups. *Journal of Algebra.* 2002; 251(2):727–738. DOI: 10.1006/jabr.2001.9138.

19. Ballester-Bolinches A, Beidleman JC, Heineken H. Groups in which Sylow subgroups and subnormal subgroups permute. *Illinois Journal of Mathematics.* 2003;47(1–2):63–69. DOI: 10.1215/ijm/1258488138.

20. Ballester-Bolinches A, Beidleman JC, Heineken H. A local approach to certain classes of finite groups. *Communications in Algebra.* 2003;31(12):5931–5942. DOI: 10.1081/AGB-120024860.

21. Asaad M. Finite groups in which normality or quasinormality is transitive. *Archiv der Mathematik.* 2004;83(4):289–296. DOI: 10.1007/s00013-004-1065-4.

22. Ballester-Bolinches A, Cossey J. Totally permutable products of finite groups satisfying *SC* or *PST*. *Monatshefte für Mathematik.* 2005;145(2):89–94. DOI: 10.1007/s00605-004-0263-9.

23. Al-Sharo KA, Beidleman JC, Heineken H, Ragland MF. Some characterizations of finite groups in which semipermutability is a transitive relation. *Forum Mathematicum.* 2010;22(5):855–862. DOI: 10.1515/forum.2010.045.

24. Beidleman JC, Ragland MF. Subnormal, permutable, and embedded subgroups in finite groups. *Central European Journal of Mathematics.* 2011;9(4):915–921. DOI: 10.2478/s11533-011-0098-8.

25. Yi X, Skiba AN. Some new characterizations of *PST*-groups. *Journal of Algebra.* 2014;399:39–54. DOI: 10.1016/j.jalgebra.2013.10.001.

26. Skiba AN. Some characterizations of finite σ-soluble $P\sigma T$-groups. *Journal of Algebra.* 2018;495:114–129. DOI: 10.1016/j.jalgebra.2017.11.009.

27. Fattahi A. Groups with only normal and abnormal subgroups. *Journal of Algebra.* 1974;28(1):15–19. DOI: 10.1016/0021-8693(74)90019-2.

28. Ebert G, Bauman S. A note on subnormal and abnormal chains. *Journal of Algebra.* 1975;36(2):287–293. DOI: 10.1016/0021-8693(75)90103-9.

29. Semenchuk VN, Skiba AN. On one generalization of finite $\mathfrak{U}$-critical groups. *Journal of Algebra and its Applications.* 2016; 15(4):1650063. DOI: 10.1142/S0219498816500638.

30. Monakhov VS. [Finite groups with abnormal and $\mathfrak{U}$-subnormal subgroups]. *Sibirskii matematicheskii zhurnal.* 2016;57(2): 447–462. Russian. DOI: 10.17377/smzh.2016.57.217.

31. Monakhov VS, Sokhor IL. [Finite groups with formation subnormal primary subgroups]. *Sibirskii matematicheskii zhurnal.* 2017;58(4):851–863. Russian. DOI: 10.17377/smzh.2017.58.412.

32. Monakhov VS, Sokhor IL. On groups with formational subnormal Sylow subgroups. *Journal of Group Theory.* 2018;21:273–287. DOI: 10.1515/jgth-2017-0039.

33. Monakhov VS, Sokhor IL. Finite groups with abnormal or formational subnormal primary subgroups. *Communications in Algebra.* 2019;47(10):3941–3949. DOI: 10.1080/00927872.2019.1572174.

34. Mal'tsev AI. *Algebraicheskie sistemy* [Algebraic systems]. Moscow: Nauka; 1970. 392 p. Russian.

35. Schmidt R. *Subgroup lattices of groups*. Berlin: Walter de Gruyter; 1994. 572 p. (de Gruyter expositions of mathematics; volume 14).

36. Zappa G. Sui gruppi finiti per cui il reticolo dei sottogruppi di composizione è modulare. *Bollettino dell'Unione Matematica Italiana. Serie 3*. 1956;11(3):315–318.

37. Bray HB. *Between nilpotent and solvable*. Weinstein M, editor. Passaic: Polygonal Publishing House; 1982. 231 p.

38. Doerk K. Minimal nicht überauflösbare, endlicher Gruppen. *Mathematische Zeitschrift.* 1966;91(3):198–205. DOI: 10.1007/BF01312426.

39. Kegel OH. Zur Struktur mehrafach faktorisierbarer endlicher Gruppen. *Mathematische Zeitschrift.* 1965;87(1):42–48. DOI: 10.1007/BF01109929.

Received by editorial board 18.04.2019.

Дифференциальные уравнения и оптимальное управление

Differential equations and optimal control

УДК 517.948.32:517.544

ГИПЕРСИНГУЛЯРНЫЕ ИНТЕГРО-ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ СО СТЕПЕННЫМИ МНОЖИТЕЛЯМИ В КОЭФФИЦИЕНТАХ

А. П. ШИЛИН[1)]

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*

Предложена схема исследования линейного гиперсингулярного интегро-дифференциального уравнения произвольного порядка на замкнутой кривой, расположенной в комплексной плоскости, в случае, когда его коэффициенты имеют некоторую частную структуру. Схема предусматривает использование обобщенных формул Сохоцкого, решение краевой задачи Римана и решение в классе аналитических функций линейных дифференциальных уравнений. По этой схеме решены явно два уравнения, коэффициенты которых содержат степенные множители, вследствие чего наряду с задачей Римана конструктивно решены возникающие дифференциальные уравнения. Приведены условия разрешимости, формулы решения, рассмотрены примеры.

Ключевые слова: интегро-дифференциальные уравнения; гиперсингулярные интегралы; обобщенные формулы Сохоцкого; краевая задача Римана; линейные дифференциальные уравнения.

Образец цитирования:
Шилин АП. Гиперсингулярные интегро-дифференциальные уравнения со степенными множителями в коэффициентах. *Журнал Белорусского государственного университета. Математика. Информатика.* 2019;3:48–56.
https://doi.org/10.33581/2520-6508-2019-3-48-56

For citation:
Shilin AP. Hypersingular integro-differential equations with power factors in coefficients. *Journal of the Belarusian State University. Mathematics and Informatics.* 2019;3:48–56. Russian.
https://doi.org/10.33581/2520-6508-2019-3-48-56

Автор:
Андрей Петрович Шилин – кандидат физико-математических наук, доцент; доцент кафедры высшей математики и математической физики физического факультета.

Author:
Andrei P. Shilin, PhD (physics and mathematics), docent; associate professor at the department of higher mathematics and mathematical physics, faculty of physics.
a.p.shilin@gmail.com

HYPERSINGULAR INTEGRO-DIFFERENTIAL EQUATIONS WITH POWER FACTORS IN COEFFICIENTS

A. P. SHILIN [a]

[a]*Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*

The linear hypersingular integro-differential equation of arbitrary order on a closed curve located on the complex plane is considered. A scheme is proposed to study this equation in the case when its coefficients have some particular structure. This scheme providers for the use of generalized Sokhotsky formulas, the solution of the Riemann boundary value problem and the solution in the class of analytical functions of linear differential equations. According to this scheme, the equations are explicitly solved, the coefficients of which contain power factors, so that along with the Riemann problem the arising differential equations are constructively solved. Solvability conditions, solution formulas, examples are given.

Keywords: integro-differential equations; hypersingular integrals; generalized Sokhotsky formulas; Riemann boundary problem; linear differential equations.

Введение

Пусть L – простая замкнутая гладкая кривая на расширенной комплексной плоскости, D_+ и D_- – области с границей L, $0 \in D_+$, $\infty \in D_-$. Выберем на кривой L ту ориентацию, которая оставляет область D_+ слева. В работе [1] решено интегро-дифференциальное уравнение

$$\sum_{k=0}^{n}\left(a_k \varphi^{(k)}(t) + \frac{k! b_k}{\pi i}\int_L \frac{\varphi(\tau)d\tau}{(\tau - t)^{k+1}}\right) = f(t),\ t \in L. \tag{1}$$

В этом уравнении искомая функция $\varphi(t)$ вместе со своими производными до порядка n включительно и заданная функция $f(t)$ предполагаются H-непрерывными (т. е. удовлетворяющими условию Гельдера) на кривой L. Коэффициенты a_k и b_k – заданные (комплексные) числа, $k = \overline{0, n}$, $n \in \mathbb{N}$. Интегралы в уравнении (1) понимаются в смысле конечной части по Адамару [2], такие интегралы называются также гиперсингулярными.

Гиперсингулярные интегральные уравнения возникают в задачах аэродинамики, гидродинамики, квантовой физики, трещиноустойчивости. Основные методы их решения численные (например, [3; 4]). Как сказано в работе [5], «отсутствует общая теория гиперсингулярных интегральных уравнений», и в этой же работе создана основа подобной теории для некоторого класса таких уравнений.

Уравнение (1) является, по-видимому, первым исследованным в математической литературе интегро-дифференциальным уравнением с гиперсингулярными интегралами. Частные случаи переменных коэффициентов в этом уравнении изучались затем в [6; 7]. В настоящей работе решено уравнение (1) с переменными коэффициентами, содержащими степенные множители. Наличие степенных множителей приводит к более «благополучной», чем в [1], картине разрешимости из-за отсутствия бесконечного числа условий разрешимости.

Общая схема исследования. Два уравнения для последующего применения этой схемы

Пусть в уравнении (1) коэффициенты имеют вид

$$a_k = a(t)A_k(t) + b(t)B_k(t),\ b_k = a(t)A_k(t) - b(t)B_k(t),\ t \in L, \tag{2}$$

где $a(t) \neq 0$, $b(t) \neq 0$, $A_k(t)$, $B_k(t)$ – H-непрерывные заданные функции, $k = \overline{0, n}$. При этом все функции $A_k(t)$ и $B_k(t)$ аналитически продолжимы в области D_+ и D_- соответственно, лишь у функций $B_k(t)$ допускаются полюсы в точке $z = \infty$.

Введем интеграл типа Коши

$$\Phi_{\pm}(z) = \frac{1}{2\pi i}\int_L \frac{\varphi(\tau)d\tau}{\tau - z},\ z \in D_{\pm}.$$

Используя для предельных значений функций $\Phi_{\pm}(z)$ и их производных обобщенные формулы Сохоцкого [8]

$$\begin{cases} \Phi_{+}^{(k)}(t) - \Phi_{-}^{(k)}(t) = \varphi^{(k)}(t), \\ \Phi_{+}^{(k)}(t) + \Phi_{-}^{(k)}(t) = \dfrac{k!}{\pi i}\displaystyle\int_{L} \frac{\varphi(\tau)d\tau}{(\tau - t)^{k+1}}, \ t \in L, \ k = \overline{0, n}, \end{cases}$$

сведем уравнение (1) к задаче линейного сопряжения

$$2a(t)\sum_{k=0}^{n} A_k(t)\Phi_{+}^{(k)}(t) = 2b(t)\sum_{k=0}^{n} B_k(t)\Phi_{-}^{(k)}(t) + f(t), \ t \in L. \tag{3}$$

Введем аналитические функции

$$F_{+}(z) = \sum_{k=0}^{n} A_k(z)\Phi_{+}^{(k)}(z), \ z \in D_{+}, \tag{4}$$

$$F_{-}(z) = \sum_{k=0}^{n} B_k(z)\Phi_{-}^{(k)}(z), \ z \in D_{-} \tag{5}$$

с H-непрерывными предельными значениями $F_{\pm}(t)$ на L и из (3) получим краевую задачу Римана

$$F_{+}(t) = \frac{b(t)}{a(t)}F_{-}(t) + \frac{f(t)}{2a(t)}, \ t \in L. \tag{6}$$

Эта задача должна решаться в классе функций, имеющих на бесконечности поведение, вытекающее из формулы (5).

Если задача Римана (6) окажется разрешимой, то соотношения (4), (5) станут линейными дифференциальными уравнениями для нахождения функций $\Phi_{\pm}(z)$. Решение уравнения (5) следует находить с учетом условия $\Phi_{-}(\infty) = 0$, выражающего известное свойство интеграла типа Коши. Решив эти дифференциальные уравнения, получим решение исходного уравнения в виде

$$\varphi(t) = \Phi_{+}(t) - \Phi_{-}(t), \ t \in L. \tag{7}$$

Далее будем решать следующие два уравнения:

$$\sum_{k=0}^{n}\left(\frac{a(t)k!\,\alpha_k + b(t)(-1)^k n!\,t^k}{k!}\varphi^{(k)}(t) + \frac{a(t)k!\,\alpha_k - b(t)(-1)^k n!\,t^k}{\pi i}\int_{L}\frac{\varphi(\tau)d\tau}{(\tau - t)^{k+1}}\right) = f(t), \ t \in L, \tag{8}$$

$$a(t)\sum_{k=0}^{n}\alpha_k\left(\varphi^{(k)}(t) + \frac{k!}{\pi i}\int_{L}\frac{\varphi(\tau)d\tau}{(\tau - t)^{k+1}}\right) +$$

$$+ b(t)\left(\varphi(\tau) - t^{2n}\varphi^{(n)}(t) - \frac{1}{\pi i}\int_{L}\frac{\varphi(\tau)d\tau}{\tau - t} + \frac{n!\,t^{2n}}{\pi i}\int_{L}\frac{\varphi(\tau)d\tau}{(\tau - t)^{n+1}}\right) = f(t), \ t \in L, \tag{9}$$

где α_k – заданные числа, $k = \overline{0, n}$. Оба уравнения (8) и (9) являются частными случаями (1) с коэффициентами вида (2), в которых $A_k(t) = \alpha_k$, $k = \overline{0, n}$, а функции $B_k(t)$ различны для каждого из уравнений:

• в случае (8): $B_k(t) = \dfrac{(-1)^k n!\,t^k}{k!}$, $k = \overline{0, n}$;

• в случае (9): $B_0(t) = 1$, $B_n(t) = -t^{2n}$, $B_k(t) = 0$, $k = \overline{1, n-1}$.

В (9) считаем для удобства $n \geq 2$. При $n = 1$ решение этого уравнения также осуществляется по общей схеме (и притом особенно просто), однако выкладки, проводимые с произвольным n, требуют для $n = 1$ частых оговорок.

Вспомогательные факты

1. Задача Римана (6) хорошо изучена [9]. Условия ее разрешимости (если они возникают) записываются в виде

$$\int_L \frac{f(\tau)\tau^k d\tau}{a(\tau)X_+(\tau)} = 0, \tag{10}$$

а сами решения (если они существуют) – в виде

$$F_\pm(z) = X_\pm(z)\left(\frac{1}{4\pi i}\int_L \frac{f(\tau)d\tau}{a(\tau)X_+(\tau)(\tau - z)} + P(z)\right),\ z \in D_\pm. \tag{11}$$

В этих формулах $X_\pm(z)$ – канонические функции задачи (6). Значения k в (10) и выражение для функции $P(z)$ в (11) зависят от индекса $\alpha = \mathrm{Ind}_L \dfrac{b(t)}{a(t)}$ и поведения искомой функции $F_-(z)$ на бесконечности. Это поведение будет разным для уравнений (8), (9), поэтому остальные необходимые пояснения к формулам (10), (11) удобно сделать в дальнейшем.

2. Для обоих уравнений (8), (9) соответствующее дифференциальное уравнение (4) может быть записано в виде

$$\sum_{k=0}^{n} \alpha_k \Phi_+^{(k)}(z) = F_+(z),\ z \in D_+. \tag{12}$$

Пусть $\lambda_1, \lambda_2, \ldots, \lambda_n$ – корни характеристического уравнения $\sum\limits_{k=0}^{n} \alpha_k \lambda^k = 0,$ которые мы для простоты далее считаем попарно различными. Решение уравнения (12) представимо формулой

$$\Phi_+(z) = \sum_{j=1}^{n}\left(C_j + N_j \int_0^z e^{-\lambda_j \zeta} F_+(\zeta)d\zeta\right)e^{\lambda_j z}, \tag{13}$$

где C_j – произвольные постоянные; $N_j = \dfrac{(-1)^{n+1}}{\alpha_n \prod\limits_{\substack{m=1,\\ m\neq j}}^{n}(\lambda_m - \lambda_j)},\ j = \overline{1,\ n}.$ (При $n = 1$ произведение $\prod\limits_{\substack{m=1,\\ m\neq j}}^{n}(\lambda_m - \lambda_j)$ следует заменить на 1.)

Формула (13) взята из работы [1], причем здесь приведен более подробный ее вариант. Отметим еще, что если среди корней характеристического уравнения будут кратные, то (13) может быть надлежащим образом видоизменена.

Решение уравнения (8)

Для уравнения (8) соответствующее уравнение (5) принимает вид

$$\sum_{k=0}^{n} \frac{(-1)^k n!}{k!} z^k \Phi_-^{(k)}(z) = F_-(z),\ z \in D_-. \tag{14}$$

Поскольку $\Phi_-(\infty) = 0,$ то все слагаемые в левой части (14) также, очевидно, на бесконечности равны нулю, поэтому $F_-(\infty) = 0.$ Следовательно, задачу Римана (6) следует решать в классе функций, исчезающих на бесконечности.

Уравнение (14) есть линейное уравнение Эйлера. Известно [10, с. 562], что фундаментальную систему решений соответствующего однородного уравнения образуют функции $z, z^2, \ldots, z^n$. Решать это уравнение удобнее, не используя известные методы, а делая замену $\tilde{\Phi}_-(z) = \dfrac{\Phi_-(z)}{z}$, после которой получим

$$\Phi_-(z) = \tilde{\Phi}_-(z)z,\ \ \Phi_-^{(k)}(z) = \tilde{\Phi}_-^{(k)}(z)z + k\tilde{\Phi}_-^{(k-1)}(z),\ \ k = \overline{1,\ n},$$

а (14) примет вид

$$(-1)^{n+1}\tilde{\Phi}_-^{(n)}z^{n+1}z^{n+1} = F_-(z),$$

откуда

$$\tilde{\Phi}_-(z) = Q(z) + (-1)^{n+1}\int_\infty^z d\zeta_1\int_\infty^{\zeta_1} d\zeta_2 \ldots \int_\infty^{\zeta_{n-1}} \frac{F_-(\zeta_n)}{\zeta_n^{n+1}}d\zeta_n,$$

где $Q(z)$ – многочлен степени $n-1$ с произвольными коэффициентами. Следовательно,

$$\Phi_-(z) = zQ(z) + (-1)^{n+1} z\int_\infty^z d\zeta_1\int_\infty^{\zeta_1} d\zeta_2 \ldots \int_\infty^{\zeta_{n-1}} \frac{F_-(\zeta_n)}{\zeta_n^{n+1}}d\zeta_n. \tag{15}$$

Поскольку $\frac{F_-(\zeta_n)}{\zeta_n^{n+1}} = O\left(\frac{1}{\zeta_n^{n+2}}\right)$ при $\zeta_n \to \infty$, то n-кратное интегрирование в (15) приведет к тому, что $(-1)^{n+1} z\int_\infty^z d\zeta_1\int_\infty^{\zeta_1} d\zeta_2 \ldots \int_\infty^{\zeta_{n-1}} \frac{F_-(\zeta_n)}{\zeta_n^{n+1}}d\zeta_n = O\left(\frac{1}{z}\right)$ при $z \to \infty$, и тогда для выполнения условия $\Phi_-(\infty) = 0$ следует положить $Q(z) \equiv 0$.

Итак, если соответствующая задача Римана (6) разрешима, то всегда можно с помощью формул (13), (15) найти функции $\Phi_\pm(z)$, а затем по формуле (7) записать решение исходного уравнения. Таким образом, справедливо следующее утверждение.

Теорема 1. *При* $\alpha \geq 0$ *уравнение (8) безусловно разрешимо. При* $\alpha < 0$ *для его разрешимости необходимо и достаточно выполнения условий (10), в которых* $k = \overline{0, -\alpha - 1}$.

В случае разрешимости уравнения (8) его решение содержит $n + \max(0, \alpha)$ *произвольных постоянных и дается формулой*

$$\varphi(t) = \sum_{j=1}^{n}\left(C_j + N_j\int_0^t e^{-\lambda_j\zeta}F_+(\zeta)d\zeta\right)e^{\lambda_j t} + (-1)^n t\int_\infty^t d\zeta_1\int_\infty^{\zeta_1} d\zeta_2 \ldots \int_\infty^{\zeta_{n-1}} \frac{F_-(\zeta_n)}{\zeta_n^{n+1}}d\zeta_n,$$

где C_j *– произвольные постоянные,* $j = \overline{1, n}$*; функции* $F_\pm(z)$ *выражаются по формуле (11), в которой* $P(z)$ *– многочлен степени* $\alpha - 1$ *с произвольными коэффициентами при* $\alpha \geq 1$, $P(z) \equiv 0$ *при* $\alpha < 1$.

Пример 1. Рассмотрим уравнение (8), полагая в нем $n = 2$, $a(t) = e^{-t}$, $b(t) = t$, $\alpha_0 = \alpha_2 = 1$, $\alpha_1 = 2$, $f(t) = 2e^{-t}$. Получим

$$\left(e^{-t} + 2t\right)\varphi(t) + 2\left(e^{-t} - t^2\right)\varphi'(t) + \left(e^{-t} + t^3\right)\varphi''(t) +$$

$$+ \frac{e^{-t} - 2t}{\pi i}\int_L \frac{\varphi(\tau)d\tau}{\tau - t} + \frac{2\left(e^{-t} + t^2\right)}{\pi i}\int_L \frac{\varphi(\tau)d\tau}{(\tau - t)^2} + \frac{2\left(e^{-t} - t^3\right)}{\pi i}\int_L \frac{\varphi(\tau)d\tau}{(\tau - t)^3} = 2e^{-t},\ t \in L.$$

В этом случае задача Римана (6) записывается в виде

$$F_+(t) = te^t F_-(t) + 1,\ t \in L,$$

и имеет в классе исчезающих на бесконечности функций решение $F_+(z) = 1 + \beta e^z$, $F_-(z) = \frac{\beta}{z}$, где β – произвольная постоянная.

Дифференциальные уравнения (4), (5) соответственно приобретают вид

$$\Phi_+(z) + 2\Phi_+'(z) + \Phi_+''(z) = 1 + \beta e^z,\ \ 2\Phi_-(z) - 2z\Phi_-'(z) + z^2\Phi_-''(z) = \frac{\beta}{z}$$

и имеют решения

$$\Phi_+(z) = C_1e^{-z} + C_2ze^{-z} + \frac{\beta}{4}e^z + 1,\ \Phi_-(z) = \frac{\beta}{6z},$$

где C_1, C_2 – произвольные постоянные.

Решение примера 1 согласно формуле (7) есть

$$\varphi(t) = e^{-t}(C_1 + C_2t) + \beta\left(\frac{e^t}{4} - \frac{1}{6t}\right) + 1.$$

Решение уравнения (9)

Для уравнения (9) соответствующее уравнение (5) принимает вид

$$z^{2n}\Phi_-^{(n)}(z) - \Phi_-(z) = -F_-(z), \tag{16}$$

из которого понятно, что $F_-(z) = O\left(z^{n-1}\right)$ при $z \to \infty$, т. е. задачу Римана (6) надо решать в классе функций, допускающих на бесконечности полюс порядка не выше $n - 1$.

Пусть $\varepsilon_k = e^{\frac{2\pi ki}{n}}$ – комплексные корни n-й степени из единицы, $k = \overline{1, n}$. Известно [10, с. 484], что фундаментальную систему решений однородного уравнения (16) образуют функции $z^{n-1}e^{-\frac{\varepsilon_k}{z}}$, $k = \overline{1, n}$.

Лемма 1. *Для* $k = \overline{1, n}$, $m = \overline{1, n-1}$ *справедлива формула*

$$\left(z^{n-1}e^{-\frac{\varepsilon_k}{z}}\right)^{(m)} = e^{-\frac{\varepsilon_k}{z}}\sum_{j=0}^{m}\varepsilon_k^j l_{mj} z^{n-m-j-1}, \tag{17}$$

где l_{mj} *– некоторые постоянные (не зависящие от k), причем* $l_{mm} = 1$.

Доказательство. Проведем доказательство методом математической индукции по m. Для $m = 1$ формула (17), очевидно, верна. Пусть она верна для некоторого $m \geq 1$. Тогда

$$\left(z^{n-1}e^{-\frac{\varepsilon_k}{z}}\right)^{(m+1)} = \left(e^{-\frac{\varepsilon_k}{z}}\sum_{j=0}^{m}\varepsilon_k^j l_{mj} z^{n-m-j-1}\right)' =$$

$$= e^{-\frac{\varepsilon_k}{z}}\frac{\varepsilon_k}{z^2}\sum_{j=0}^{m}\varepsilon_k^j l_{mj} z^{n-m-j-1} + e^{-\frac{\varepsilon_k}{z}}\sum_{j=0}^{m}\varepsilon_k^j l_{mj}\left(n-m-j-1\right)z^{n-m-j-2} = e^{-\frac{\varepsilon_k}{z}}\sum_{j=0}^{m+1}\varepsilon_k^j l_{m+1,\,j} z^{n-m-j-2},$$

где $l_{m+1,0} = l_{m,0}(n-m-1)$; $l_{m+1,\,j} = l_{m,\,j-1} + l_{m,\,j}(n-m-1-j)$, $j = \overline{1, m}$; $l_{m+1,\,m+1} = l_{m,\,m} = 1$. Полученное выражение соответствует замене m на $m + 1$ в формуле (17). Лемма 1 доказана.

Лемма 2. *Определитель W Вронского функций* $z^{n-1}e^{-\frac{\varepsilon_k}{z}}$ *равен определителю Вандермонда чисел* ε_k, $k = \overline{1, n}$.

Доказательство. Запишем определитель W, используя лемму 1:

$$\begin{vmatrix} e^{-\frac{\varepsilon_1}{z}}z^{n-1} & e^{-\frac{\varepsilon_2}{z}}z^{n-1} & \dots & e^{-\frac{\varepsilon_n}{z}}z^{n-1} \\ e^{-\frac{\varepsilon_1}{z}}\left(l_{10}z^{n-2} + \varepsilon_1 z^{n-3}\right) & e^{-\frac{\varepsilon_2}{z}}\left(l_{10}z^{n-2} + \varepsilon_2 z^{n-3}\right) & \dots & e^{-\frac{\varepsilon_n}{z}}\left(l_{10}z^{n-2} + \varepsilon_n z^{n-3}\right) \\ \dots & \dots & \dots & \dots \\ e^{-\frac{\varepsilon_1}{z}}\sum_{j=0}^{n-1}\varepsilon_1^j l_{n-1,\,j}z^{-j} & e^{-\frac{\varepsilon_2}{z}}\sum_{j=0}^{n-1}\varepsilon_2^j l_{n-1,\,j}z^{-j} & \dots & e^{-\frac{\varepsilon_n}{z}}\sum_{j=0}^{n-1}\varepsilon_n^j l_{n-1,\,j}z^{-j} \end{vmatrix}.$$

Вынесем общие множители столбцов $e^{-\frac{\varepsilon_k}{z}}$, $k = \overline{1, n}$, за знак определителя. Получим перед определителем коэффициент $\prod_{k=1}^{n}e^{-\frac{\varepsilon_k}{z}} = e^{-\frac{\varepsilon_1+\varepsilon_2+\dots+\varepsilon_n}{z}} = 1$, поскольку $\varepsilon_1 + \varepsilon_2 + \dots + \varepsilon_n = 0$.

Оставшийся определитель представим в виде надлежащей суммы определителей, элементами которых служат отдельные слагаемые строк определителя. Все указанные определители из-за пропорциональности элементов каких-либо строк обратятся в нуль, ненулевым будет лишь определитель из последних слагаемых всех строк:

$$\begin{vmatrix} z^{n-1} & z^{n-1} & \dots & z^{n-1} \\ \varepsilon_1 z^{n-3} & \varepsilon_2 z^{n-3} & \dots & \varepsilon_n z^{n-3} \\ \dots & \dots & \dots & \dots \\ \varepsilon_1^{n-1}z^{-n+1} & \varepsilon_2^{n-1}z^{-n+1} & \dots & \varepsilon_n^{n-1}z^{-n+1} \end{vmatrix}.$$

Вынесем общие множители строк последнего определителя за знак определителя, тогда перед определителем будет коэффициент $z^{n-1}z^{n-3}\cdots z^{-n+1}=1$ и в результате останется, очевидно, определитель Вандермонда чисел $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$. Лемма 2 доказана.

Лемма 3. *Пусть* $k=\overline{1,\, n}$, $k\neq j$. *Тогда для каждого* $j=\overline{1,\, n}$ *определитель Вронского функций* $z^{n-1}e^{-\frac{\varepsilon_k}{z}}$ *равен* $W_j z^{n-1}e^{-\frac{\varepsilon_j}{z}}$, *где* W_j – *определитель Вандермонда чисел* ε_k.

Доказательство леммы 3 не приводится, поскольку оно вполне аналогично доказательству леммы 2.

Будем решать уравнение (16) методом вариации произвольных постоянных. Для этого вначале следует решить систему линейных алгебраических уравнений

$$\sum_{j=1}^{n}\left(z^{n-1}e^{-\frac{\varepsilon_j}{z}}\right)^{(k)}\tilde{D}_j(z)=\tilde{F}_k(z),\ k=\overline{0,\, n-1},$$

где $\tilde{F}_k(z)=0$ при $k=\overline{0,\, n-2}$; $\tilde{F}_{n-1}(z)=-\dfrac{F_-(z)}{z^{2n}}$. Решая ее по правилу Крамера и используя леммы 2, 3 для записи необходимых определителей, получим

$$\tilde{D}_j(z)=-\frac{(-1)^{n+j}W_j z^{n-1}e^{-\frac{\varepsilon_j}{z}}F_-(z)}{Wz^{2n}}=\frac{M_j e^{\frac{\varepsilon_j}{z}}F_-(z)}{z^{n+1}},$$

где $M_j=\dfrac{(-1)^{n+j-1}W_j}{W}=\dfrac{(-1)^{n+j-1}\prod\limits_{\substack{m,k=1,\,m>k,\\ m\neq j,\,k\neq j}}^{n}(\varepsilon_m-\varepsilon_k)}{\prod\limits_{\substack{m,k=1,\\ m>k}}^{n}(\varepsilon_m-\varepsilon_k)}=\dfrac{(-1)^n}{\prod\limits_{\substack{m=1,\\ m\neq j}}^{n}(\varepsilon_m-\varepsilon_j)},\ j=\overline{1,\, n}.$

Метод вариации произвольных постоянных позволяет записать общее решение уравнения (16) в виде

$$\Phi_-(z)=z^{n-1}\sum_{j=1}^{n}\left(D_j+M_j\int_{\infty}^{z}\frac{e^{\frac{\varepsilon_j}{\zeta}}F_-(\zeta)}{\zeta^{n+1}}d\zeta\right)e^{-\frac{\varepsilon_j}{z}},\tag{18}$$

где D_j – произвольные постоянные.

Заметим, что $\dfrac{e^{\frac{\varepsilon_j}{\zeta}}F_-(\zeta)}{\zeta^{n+1}}=O\left(\dfrac{1}{\zeta^2}\right)$ при $\zeta\to\infty$, поэтому интегралы из (18) сходятся и дают исчезающие на бесконечности функции.

Функция $\Phi_-(z)$ в формуле (18), вообще говоря, имеет полюс порядка $n-1$ на бесконечности. Покажем, что постоянные D_j можно подобрать, и притом единственным образом, так, чтобы добиться равенства $\Phi_-(\infty)=0$.

Пусть $\dfrac{\gamma_1}{z}+\dfrac{\gamma_2}{z^2}+\ldots+\dfrac{\gamma_{n-1}}{z^{n-1}}+\ldots$ есть разложение в ряд Тейлора в окрестности бесконечности функции $\sum\limits_{j=1}^{n}M_j\left(\int\limits_{\infty}^{z}\dfrac{e^{\frac{\varepsilon_j}{\zeta}}F_-(\zeta)}{\zeta^{n+1}}d\zeta\right)e^{-\frac{\varepsilon_j}{z}}$. Делая аналогичные разложения для функций $e^{-\frac{\varepsilon_j}{z}}$, в окрестности бесконечности получим

$$\Phi_-(z)=z^{n-1}\left(\sum_{j=1}^{n}\left(1-\frac{\varepsilon_j}{1!z}+\frac{\varepsilon_j^2}{2!z^2}-\ldots+(-1)^{n-1}\frac{\varepsilon_j^{n-1}}{(n-1)!z^{n-1}}+\ldots\right)D_j+\frac{\gamma_1}{z}+\frac{\gamma_2}{z^2}+\ldots+\frac{\gamma_{n-1}}{z^{n-1}}+\ldots\right).$$

Устраним полюс на бесконечности, подчинив постоянные D_j требованиям

$$\begin{cases} \sum\limits_{j=1}^{n} D_j = 0, \\ \sum\limits_{j=1}^{n} \varepsilon_j^k D_j = (-1)^{k+1} \gamma_k k!, \; k = \overline{1,\, n-1}. \end{cases} \tag{19}$$

Эти требования представляют собой систему линейных алгебраических уравнений для нахождения постоянных D_j, причем определитель системы есть определитель Вандермонда $W \neq 0$, поэтому она всегда имеет единственное решение.

Теперь установлены все факты, позволяющие сформулировать результат в отношении уравнения (9).

Теорема 2. *При $n + \alpha \geq 0$ уравнение (9) безусловно разрешимо. При $n + \alpha < 0$ для разрешимости уравнения (9) необходимо и достаточно выполнения условий (10), в которых $k = \overline{0,\, -n-\alpha-1}$. В случае разрешимости уравнения (9) его решение содержит $n + \max(0,\, n+\alpha)$ произвольных постоянных и дается формулой*

$$\varphi(t) = \sum_{j=1}^{n} \left(\left(C_j + N_j \int_0^t e^{-\lambda_j \zeta} F_+(\zeta)\, d\zeta \right) e^{\lambda_j t} - t^{n-1} \left(D_j + M_j \int_\infty^t \frac{e^{\frac{\varepsilon_j}{\zeta}} F_-(\zeta)}{\zeta^{n+1}}\, d\zeta \right) e^{-\frac{\varepsilon_j}{t}} \right),$$

где C_j – произвольные постоянные; D_j – вполне определенные постоянные, являющиеся решением системы (19), $j = \overline{1,\, n}$.

Функции $F_\pm(z)$ выражаются формулой (11), в которой $P(z)$ – многочлен степени $n + \alpha - 1$ с произвольными коэффициентами при $n + \alpha - 1 \geq 0$, $P(z) \equiv 0$ при $n + \alpha - 1 < 0$.

Пример 2. Рассмотрим уравнение

$$3\varphi(t) - (2t^4 + 1)\varphi''(t) - \frac{1}{\pi i} \int_L \frac{\varphi(\tau)\, d\tau}{\tau - t} + \frac{2(2t^4 - 1)}{\pi i} \int_L \frac{\varphi(\tau)\, d\tau}{(\tau - t)^3} = e^{2t}, \; t \in L.$$

Такой вид можно придать уравнению (9) при $n = 2$, $a(t) = 1$, $b(t) = 2$, $\alpha_0 = 1$, $\alpha_1 = 0$, $\alpha_2 = -1$, $f(t) = e^{2t}$.

Тогда краевая задача Римана (6) запишется как

$$F_+(t) = 2F_-(t) + \frac{1}{2} e^{2t}, \; t \in L.$$

Решать ее следует в классе функций, допускающих полюс 1-го порядка на бесконечности. В итоге получим

$$F_+(z) = \frac{1}{2} e^{2z} + \beta_0 + \beta_1 z, \; F_-(z) = \frac{\beta_0}{2} + \frac{\beta_1 z}{2},$$

где β_0 и β_1 – произвольные постоянные.

Далее следует решать дифференциальные уравнения, которые для рассматриваемого примера имеют вид

$$\Phi_+(z) - \Phi_+''(z) = \frac{1}{2} e^{2z} + \beta_0 + \beta_1 z, \; \Phi_-(z) - z^4 \Phi_-''(z) = \frac{\beta_0}{2} + \frac{\beta_1 z}{2}.$$

Решением являются функции

$$\Phi_+(z) = \beta_0 + \beta_1 z - \frac{1}{6} e^{2z} + C_1 e^z + C_2 e^{-z}, \; \Phi_-(z) = \frac{1}{2} \left(\beta_0 + \beta_1 z - \beta_0 z \operatorname{sh} \frac{1}{z} - \beta_1 z \operatorname{ch} \frac{1}{z} \right),$$

здесь C_1, C_2 – произвольные постоянные, а решение $\Phi_-(z)$ найдено с учетом условия $\Phi_-(\infty) = 0$.

Наконец, решение примера получим по формуле (7)

$$\varphi(t) = \frac{1}{2} \left(\beta_0 + \beta_1 t + \beta_0 t \operatorname{sh} \frac{1}{t} + \beta_1 t \operatorname{ch} \frac{1}{t} \right) - \frac{1}{6} e^{2t} + C_1 e^t + C_2 e^{-t}.$$

Заключение

Формулы решений, указанные в формулировках теорем 1 и 2, верны как для однородных, так и неоднородных уравнений (8), (9). Отметим как очевидное следствие этих теорем безусловную разрешимость однородных уравнений и наличие у них нетривиальных решений.

Наряду с уравнениями (8), (9) по указанной в настоящей статье схеме можно решить многие уравнения (1), коэффициенты которых имеют вид (2). Важно при этом уметь решать в классе аналитических функций возникающие дифференциальные уравнения (4), (5).

Библиографические ссылки

1. Зверович ЭИ. Решение гиперсингулярного интегро-дифференциального уравнения с постоянными коэффициентами. *Доклады Национальной академии наук Беларуси*. 2010;54(6):5–8.

2. Адамар Ж. *Задача Коши для линейных уравнений с частными производными гиперболического типа*. Москва: Наука; 1978. 352 с.

3. Boykov IV, Ventsel ES, Boykova AI. An approximate solution of hypersingular integral equations. *Applied Numerical Mathematics*. 2010;60(6):607–628. DOI: 10.1016/j.apnum.2010.03.003.

4. Chan Y-S, Fannjiang AC, Paulino GH. Integral equations with hypersingular kernels – theory and application to fracture mechanics. *International Journal of Engineering Science.* 2003;41(7):683–720. DOI: 10.1016/S0020-7225(02)00134-9.

5. Бойков ИВ. О разрешимости гиперсингулярных интегральных уравнений. *Известия высших учебных заведений. Поволжский регион. Физико-математические науки*. 2016;3(39):86–102. DOI: 10.21685/2072-3040-2016-3-6.

6. Зверович ЭИ, Шилин АП. Решение интегро-дифференциальных уравнений с сингулярными и гиперсингулярными интегралами специального вида. *Известия Национальной академии наук Беларуси. Серия физико-математических наук*. 2018; 54(4):404–407. DOI: 10.29235/1561-2430-2018-54-4-404-407.

7. Шилин АП. Явное решение одного гиперсингулярного интегро-дифференциального уравнения второго порядка. *Журнал Белорусского государственного университета. Математика. Информатика.* 2019;2:67–72. DOI: 10.33581/2520-6508-2019-2-67-72.

8. Зверович ЭИ. Обобщение формул Сохоцкого. *Известия Национальной академии наук Беларуси. Серия физико-математических наук.* 2012;2:24–28.

9. Гахов ФД. *Краевые задачи*. Москва: Наука; 1977. 640 с.

10. Камке Э. *Справочник по обыкновенным дифференциальным уравнениям. 6-е издание*. Фомин СВ, переводчик. Санкт-Петербург: Лань; 2003. 576 с.

References

1. Zverovich EI. Solution of the hypersingular integro-differential equation with constant coefficients. *Doklady of the National Academy of Sciences of Belarus*. 2010;54(6):5–8. Russian.

2. Adamar Zh. *Zadacha Koshi dlya lineinykh uravnenii s chastnymi proizvodnymi giperbolicheskogo tipa* [The Cauchy problem of linear equations with partial derivatives of hyperbolic type]. Moscow: Nauka; 1978. 352 p. Russian.

3. Boykov IV, Ventsel ES, Boykova AI. An approximate solution of hypersingular integral equations. *Applied Numerical Mathematics*. 2010;60(6):607–628. DOI: 10.1016/j.apnum.2010.03.003.

4. Chan Y-S, Fannjiang AC, Paulino GH. Integral equations with hypersingular kernels – theory and application to fracture mechanics. *International Journal of Engineering Science.* 2003;41(7):683–720. DOI: 10.1016/S0020-7225(02)00134-9.

5. Boykov IV. On solubility of hypersingular integral equations. *Izvestiya vysshikh uchebnykh zavedenii. Povolzhskii region. Fiziko-matematicheskie nauki.* 2016:3(39):86–102. Russian. DOI: 10.21685/2072-3040-2016-3-6.

6. Zverovich EI, Shilin AP. [Solution of the integro-differential equations with a singular and hypersingular integrals]. *Proceedings of the National Academy of Sciences of Belarus. Physics and Mathematics Series*. 2018;54(4):404–407. Russian. DOI: 10.29235/1561-2430-2018-54-4-404-407.

7. Shilin AP. Explicit solution of one hypersingular integro-differential equation of the second order. *Journal of the Belarusian State University. Mathematics and Informatics*. 2019;2:67–72. Russian. DOI: 10.33581/2520-6508-2019-2-67-72.

8. Zverovich EI. [Generalization of Sohotsky formulas]. *Proceedings of the National Academy of Sciences of Belarus. Physics and Mathematics Series.* 2012;2:24–28. Russian.

9. Gakhov FD. *Kraevye zadachi* [Boundary value problems]. Moscow: Nauka; 1977. 640 p. Russian.

10. Kamke E. *Differentialgleichungen. Lösungsmethoden und Lösungen. I. Gewöhnliche Differentialgleichungen.* Leipzig: B. G. Teubner; 1977. DOI: 10.1007/978-3-663-05925-7.

Russian edition: Kamke E. *Spravochnik po obyknovennym differentsial'nym uravneniyam. 6-e izdanie*. Fomin SV, translator. Saint Petersburg: Lan'; 2003. 576 p.

Статья поступила в редколлегию 08.09.2019.
Received by editorial board 08.09.2019.

Теория вероятностей и математическая статистика

Probability theory and mathematical statistics

УДК 519.872

МНОГОЛИНЕЙНАЯ СИСТЕМА МАССОВОГО ОБСЛУЖИВАНИЯ С РЕЗЕРВНЫМИ ПРИБОРАМИ

В. И. КЛИМЕНОК[1)]

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*

Рассматривается многолинейная система с неограниченным буфером, которая может использоваться при проектировании экономных схем энергопотребления и как математическая модель ненадежных реальных стохастических систем. Запросы поступают в систему в групповом марковском потоке, времена обслуживания распределены по фазовому закону. Если время обслуживания запроса прибором превышает некоторую случайную величину, распределенную по фазовому закону, этот прибор получает помощь от резервного прибора из конечного множества резервных приборов. В статье найдены стационарное распределение вероятностей состояний и основные характеристики производительности системы.

Ключевые слова: система массового обслуживания; резервные приборы; групповой марковский поток; фазовое распределение времени обслуживания; стационарное распределение; характеристики производительности.

Благодарность. Исследование выполнено в рамках совместного гранта Белорусского республиканского фонда фундаментальных исследований (грант № Ф18Р-136) и Российского фонда фундаментальных исследований (грант № 18-57-00002).

Образец цитирования:
Клименок ВИ. Многолинейная система массового обслуживания с резервными приборами. *Журнал Белорусского государственного университета. Математика. Информатика.* 2019;3:57–70.
https://doi.org/10.33581/2520-6508-2019-3-57-70

For citation:
Klimenok VI. Multi-server queueing system with reserve servers. *Journal of the Belarusian State University. Mathematics and Informatics.* 2019;3:57–70. Russian.
https://doi.org/10.33581/2520-6508-2019-3-57-70

Автор:
Валентина Ивановна Клименок – доктор физико-математических наук, профессор; главный научный сотрудник научно-исследовательской лаборатории прикладного вероятностного анализа факультета прикладной математики и информатики.

Author:
Valentina I. Klimenok, doctor of science (physics and mathematics), full professor; chief researcher at the laboratory of applied probabilistic analysis, faculty of applied mathematics and computer science.
vklimenok@yandex.ru

Журнал Белорусского государственного университета. Математика. Информатика. 2019;3:57–70
Journal of the Belarusian State University. Mathematics and Informatics. 2019;3:57–70

MULTI-SERVER QUEUEING SYSTEM WITH RESERVE SERVERS

V. I. KLIMENOK[a]

[a]*Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*

In this paper, we investigate a multi-server queueing system with an unlimited buffer, which can be used in the design of energy consumption schemes and as a mathematical model of unreliable real stochastic systems. Customers arrive to the system in a batch Markovian arrival process, the service times are distributed according to the phase law. If the service time of the customer by the server exceeds a certain random value distributed according to the phase law, this server receives assistance from the reserve server from a finite set of reserve servers. In the paper, we calculate the stationary distribution and performance characteristics of the system.

Keywords: queueing system; reserve servers; batch Markovian arrival process; phase type distribution; stationary distribution; performance characteristics.

Acknowledgements. This work has been financially supported by the joint grant of Belarusian Republican Foundation for Fundamental Research (No. Ф18Р-136) and Russian Foundation for Fundamental Research (No. 18-57-00002).

Введение

Проблемы, связанные с энергосбережением во многих реальных системах, в частности в центрах обработки данных при облачных вычислениях, могут решаться путем резервирования (с дальнейшим адаптивным подключением) обслуживающих ресурсов. В системах обслуживания разнородной информации, например в кол-центрах, резервирование устройств для приоритетной информации может существенно повысить качество работы. В ненадежных системах передачи данных наличие резервных каналов позволяет улучшить качество передачи. Вследствие стохастического характера обработки и передачи информации актуальным является математическое моделирование систем с резервированием в рамках теории массового обслуживания. Упомянем лишь некоторые публикации в данной области.

В [1; 2] исследуются двухфазные системы массового обслуживания с резервированием каналов для приоритетных заявок на второй фазе, которые моделируют кол-центры, обрабатывающие запросы различных категорий важности. В публикациях [3–6] рассмотрены математические модели гибридных систем связи, состоящих из ненадежного канала FSO (free space optics) и резервного абсолютно надежного радиоканала либо из FSO-канала и резервного канала миллиметрового диапазона, которые могут выходить из строя в непересекающихся интервалах времени. В статье [7] анализируется однолинейная система с бесконечным буфером и резервным прибором, в которой предполагается, что резервный прибор активируется, когда обслуживание заявки становится слишком длинным. После активации основной и резервный приборы обслуживают эту заявку одновременно. В [7] показано, что процесс функционирования системы является векторным процессом гибели и размножения. Доказывается условие эргодичности, вычисляются стационарное распределение и характеристики производительности системы, выводится преобразование Лапласа – Стилтьеса распределения времени ожидания.

В настоящей статье результаты [7] обобщаются на случай многолинейной системы с конечным пулом резервных приборов. Полученные результаты могут быть использованы для решения проблемы нахождения компромисса между энергопотреблением и качеством обслуживания посредством использования резервных приборов, которые подключаются к обслуживанию заявки в случае превышения времени пребывания заявки на основном приборе (будем говорить также «в случае истечения таймера»). Такой сценарий обслуживания при разумной экономии энергии позволяет избежать слишком больших задержек в системе. Последняя также может рассматриваться как модель ненадежной системы, где при отказе основного ненадежного прибора резервный прибор завершает обслуживание текущей заявки. В этом случае время пребывания на основном приборе интерпретируется как время до поломки этого прибора.

Описание системы

Мы рассматриваем N-линейную систему массового обслуживания с потоком заявок, представляющим собой BMAP-поток (batch markovian arrival process). Последний задается управляющим процессом ν_t, $t \geq 0$, который является неприводимой цепью Маркова с непрерывным временем, конечным пространством состояний $\{0, \ldots, W\}$ и $(W+1)\times(W+1)$-матрицами D_k, $k \geq 0$, где (ν, ν')-й элемент матрицы

D_k, $k \geq 1$, есть интенсивность перехода управляющего процесса из состояния ν в состояние ν', сопровождающегося генерацией группы, состоящей из k запросов. Недиагональные элементы матрицы D_0 есть интенсивности переходов управляющего процесса из состояния ν в состояние ν', не сопровождающихся генерацией запросов. Диагональные элементы матрицы D_0 есть взятые с противоположным знаком интенсивности выхода управляющего процесса из своих состояний. Отметим, что матрицы D_k, $k \geq 0$, полностью определяются их производящей функцией $D(z) = \sum_{k=0}^{\infty} D_k z^k$, $|z| \leq 1$. При этом матрица $D(1)$ является инфинитезимальным генератором управляющего процесса ν_t, $t \geq 0$.

Приведем некоторые важные характеристики BMAP. Интенсивность поступления заявок в BMAP (fundamental rate) определяется как $\lambda = \boldsymbol{\theta} D'(1)\boldsymbol{e}$, где $\boldsymbol{\theta}$ – единственное решение системы $\boldsymbol{\theta} D(1) = \mathbf{0}$, $\boldsymbol{\theta}\boldsymbol{e} = 1$, $\boldsymbol{e}$ – вектор-столбец, состоящий из единиц. Интенсивность λ_b поступления групп заявок определяется как $\lambda_b = \boldsymbol{\theta}(-D_0)\boldsymbol{e}$. Коэффициент вариации интервалов между моментами поступления групп заявок находится по формуле $c_{\text{var}}^2 = 2\lambda_b \boldsymbol{\theta}(-D_0)^{-1}\boldsymbol{e} - 1$. Коэффициент корреляции соседних интервалов между моментами поступления групп заявок определяется как $c_{\text{corr}} = \dfrac{\lambda_b \boldsymbol{\theta}(-D_0)^{-1}\left(D(1) - D_0\right)(-D_0)^{-1}\boldsymbol{e} - 1}{c_{\text{var}}^2}$.

(Дополнительную информацию о BMAP см. в [8].)

Полагаем, что все приборы одинаковы и независимы друг от друга. Время обслуживания заявки прибором имеет распределение *PH*-типа с неприводимым представлением $(\boldsymbol{\beta}, S)$, т. е. указанное время интерпретируется как время, за которое цепь Маркова m_t, $t \geq 0$, с пространством состояний $\{1, \ldots, M+1\}$ достигнет поглощающего состояния $M + 1$. Переходы цепи m_t, $t \geq 0$, с пространством состояний $\{1, \ldots, M\}$ задаются субгенератором S, а интенсивности переходов в поглощающее состояние – вектором $\boldsymbol{S}_0 = -S\boldsymbol{e}$. В момент начала обслуживания состояние процесса m_t, $t \geq 0$, выбирается из пространства состояний $\{1, \ldots, M\}$ в соответствии с вероятностным вектором-строкой $\boldsymbol{\beta}$. Полагаем, что матрица $S + \boldsymbol{S}_0\beta$ неприводима. Интенсивность обслуживания задается как $\mu = -\left(\boldsymbol{\beta} S^{-1}\boldsymbol{e}\right)^{-1}$. Более подробно о *PH*-распределении и его свойствах можно узнать из [9].

Если в момент поступления группы заявок необходимое количество приборов свободно, то заявки занимают соответствующие приборы. Если свободных приборов недостаточно (или все приборы заняты), часть заявок (или все заявки) помещаются в конец очереди в бесконечном буфере в случайном порядке.

Кроме рабочих обслуживающих приборов, в системе имеется R резервных приборов. Считаем, что $R \leq N$. Свободный резервный прибор подключается к обслуживанию текущей заявки, если время обслуживания этой заявки превышает некоторое предельное время нахождения на приборе. Это предельное время определяется как случайная величина (таймер), имеющая *PH*-распределение с неприводимым представлением $(\boldsymbol{\tau}, T)$ и пространством состояний управляющего процесса $(1, 2, \ldots, L)$. Интенсивности переходов в абсорбирующее состояние задаются вектором $\boldsymbol{T}_0 = -T\boldsymbol{e}$. Интенсивность таймера вычисляется как $\tau = -\left(\boldsymbol{\gamma} T^{-1}\boldsymbol{e}\right)^{-1}$. При подключении к обслуживанию резервного прибора распределение времени до конца обслуживания заявки задается как *PH*-распределение с неприводимым представлением $(\tilde{\boldsymbol{\beta}}, \tilde{S})$ и пространством состояний управляющего процесса $(1, 2, \ldots, \tilde{M})$. Интенсивности переходов в абсорбирующее состояние определяются вектором $\tilde{\boldsymbol{S}}_0 = -\tilde{S}\boldsymbol{e}$. Интенсивность такого обслуживания задается как $\tilde{\mu} = -\left(\tilde{\boldsymbol{\beta}}\, \tilde{S}^{-1}\boldsymbol{e}\right)^{-1}$.

Если в момент окончания предельного времени обслуживание заявки еще не закончилось, а свободных резервных приборов нет, то с вероятностью p заявка покидает систему недообслуженной и с дополнительной вероятностью $1 - p$ обслуживание на этом приборе начинается заново.

Процесс изменения состояний системы

Пусть в момент времени t:

- i_t – число заявок в системе, $i_t \geq 0$;
- r_t – число занятых резервных приборов, $r_t = \overline{0, \min\{i_t, R\}}$;
- $m_t^{(j)}$ – состояние управляющего процесса обслуживания на j-м приборе, работающем без поддержки, $m_t^{(j)} = \overline{1, M}$ (полагаем, что работающие без поддержки приборы нумеруются в порядке их занятия,

т. е. прибор, который начинает обслуживание, нумеруется максимальным числом среди всех занятых приборов. Когда прибор заканчивает работу, происходит перенумерация);

• $\eta_t^{(j)}$ – состояние управляющего процесса таймера на j-м приборе, работающем без поддержки, $\eta_t^{(j)} = \overline{1, L}$;

• $\tilde{m}_t^{(j)}$ – состояние управляющего процесса на j-м приборе, работающем с поддержкой, $\tilde{m}_t^{(j)} = \overline{1, \tilde{M}}$ (полагаем, что прибор, на котором только что закончился таймер, получает первый номер среди всех приборов, работающих с поддержкой, а номера остальных таких приборов увеличиваются на единицу. Когда на каком-либо из этих приборов заканчивается обслуживание, остальные приборы перенумеровываются);

• ν_t – состояние управляющего процесса BMAP, $\nu_t = \overline{0, W}$, $t \geq 0$.

Процесс изменения состояний системы описывается регулярной неприводимой цепью Маркова с непрерывным временем и пространством состояний

$$\Omega = \left\{(i, \nu),\ i = 0, \nu = \overline{0, W}\right\} \cup$$

$$\cup \left\{\left(i, r, \nu, m^{(1)}, l^{(1)}, m^{(2)}, l^{(2)}, \ldots, m^{(\min\{i, N\} - r)}, l^{(\min\{i, N\} - r)}, \tilde{m}^{(1)}, \ldots, \tilde{m}^{(r)}\right),\ i > 0,\right.$$

$$r = \overline{0, \min\{i, R\}},\ \nu = \overline{0, W},\ m^{(1)}, \ldots, m^{(\min\{i, N\} - r)} = \overline{1, M},$$

$$\left. l^{(1)}, \ldots, l^{(\min\{i, N\} - r)} = \overline{1, L},\ \tilde{m}^{(1)}, \ldots, \tilde{m}^{(r)} = \overline{1, \tilde{M}}\right\}.$$

Число состояний пространства состояний при $i = 0$ вычисляется как

$$K_0 = (W+1)\sum_{i=0}^{N}\sum_{r=0}^{\min\{i, R\}} (ML)^{i-r}\tilde{M}^r,$$

и для любого фиксированного $i > 0$ число состояний равно

$$K = (W+1)\sum_{r=0}^{R} (ML)^{N-r}\tilde{M}^r.$$

В дальнейшем будем использовать следующие обозначения:

• $\otimes$ ($\oplus$) – кронекерово произведение (сумма) матриц;

• $A^{\otimes l} = \underbrace{A \otimes \ldots \otimes A}_{l}$, $l \geq 1$, $A^{\otimes 0} = 1$;

• $A^{\oplus l} = \sum_{m=0}^{l-1} I_{n^m} \otimes A \otimes I_{n^{l-m-1}}$, $l \geq 1$, для матрицы A, имеющей n строк;

• $\bar{W} = W + 1$; $a = ML$;

• $\mathcal{B}^{(N, r)} = I_{\bar{W}} \otimes \left(\boldsymbol{S}_0 \boldsymbol{\beta} \otimes \boldsymbol{e}\boldsymbol{\tau}\right)^{\oplus N-r} \otimes I_{\tilde{M}^r}$, $r = \overline{0, R}$;

• $\tilde{\mathcal{B}}^{(N, r)} = I_{\bar{W}} \otimes I_{a^{N-r}} \otimes (\boldsymbol{\beta} \otimes \boldsymbol{\tau}) \otimes \tilde{\boldsymbol{S}}_0^{\oplus r}$, $r = \overline{1, R}$;

• $\mathcal{T}_0^{(i, r)} = I_{\bar{W}} \otimes \left[\left(\boldsymbol{e}_M \otimes \boldsymbol{T}_0\right)^{\oplus \min\{i, N\} - r} \otimes I_{\tilde{M}^r}\right]\left(I_{a^{\min\{i, N\} - r - 1}} \otimes \tilde{\boldsymbol{\beta}} \otimes I_{\tilde{M}^r}\right)$, $i > r$, $r = \overline{0, R}$;

• $\mathcal{S}_0^{(i, r)} = I_{\bar{W}} \otimes \left(\boldsymbol{S}_0 \otimes \boldsymbol{e}_L\right)^{\oplus i - r} \otimes I_{\tilde{M}^r}$, $r = \overline{0, i}$, $i = \overline{1, N}$;

• $\tilde{\mathcal{S}}_0^{(i, r)} = I_{\bar{W}} \otimes I_{a^{i-r}} \otimes \tilde{\boldsymbol{S}}_0^{\oplus r}$, $r = \overline{0, i}$, $i = \overline{1, N}$;

• $\mathcal{C}^{(i, r)} = D_0 \oplus \left(S \oplus T\right)^{\oplus \min\{i, N\} - r} \oplus \tilde{S}^{\oplus r}$, $r = \overline{0, \min\{i, R\}}$, $i \geq 0$;

• $\mathcal{D}_k^{(i, r)} = D_k \otimes I_{a^{i-r}} \otimes (\boldsymbol{\beta} \otimes \boldsymbol{\tau})^{\otimes \min\{k, N-i\}} \otimes I_{\tilde{M}^r}$, $k \geq 1$, $r = \overline{0, \min\{i, R\}}$, $i = \overline{0, N}$.

Полагаем, что состояния цепи Маркова ξ_t, $t \geq 0$, перенумерованы в лексикографическом порядке.

Лемма. *Инфинитезимальный генератор Q цепи Маркова ξ_t, $t \geq 0$, имеет следующую блочную структуру:*

$$Q = \begin{pmatrix} F_0 & F_1 & F_2 & F_3 & \dots \\ \tilde{Q}_{-1} & Q_0 & Q_1 & Q_2 & \dots \\ O & Q_{-1} & Q_0 & Q_1 & \dots \\ O & O & Q_{-1} & Q_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Матрица F_0 порядка $K_0 \times K_0$ имеет блочный вид $F_0 = \left(F_0^{(i,\,j)}\right)_{i,\,j=\overline{0,\,N}}$, где

$$F_0^{(i,\,i-1)} = \begin{pmatrix} \mathcal{S}_0^{(i,\,0)} & 0 & \dots & 0 & 0 \\ \tilde{\mathcal{S}}_0^{(i,\,1)} & \mathcal{S}_0^{(i,\,1)} & \dots & 0 & 0 \\ 0 & \tilde{\mathcal{S}}_0^{(i,\,2)} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \tilde{\mathcal{S}}_0^{(i,\,i-1)} & \mathcal{S}_0^{(i,\,i-1)} \\ 0 & 0 & \dots & 0 & \tilde{\mathcal{S}}_0^{(i,\,i)} \end{pmatrix}, \tag{1}$$

$$F_0^{(i,\,i)} = \begin{pmatrix} \mathcal{C}^{(i,\,0)} & \mathcal{T}_0^{(i,\,0)} & \dots & 0 & 0 \\ 0 & \mathcal{C}^{(i,\,1)} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \mathcal{C}^{(i,\,i-1)} & \mathcal{T}_0^{(i,\,i-1)} \\ 0 & 0 & \dots & 0 & \mathcal{C}^{(i,\,i)} \end{pmatrix}, \; 0 \le i \le R,$$

$$F_0^{(i,\,i+k)} = \begin{pmatrix} \mathcal{D}_k^{(i,\,0)} & O & O & \dots & O & O & \dots & O \\ O & \mathcal{D}_k^{(i,\,1)} & O & \dots & O & O & \dots & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ O & O & O & \dots & \mathcal{D}_k^{(i,\,i)} & O_{\tilde{M}^i \times a^{k-1}\tilde{M}^{i+1}} & \dots & O_{\tilde{M}^i \times a^{i+k-\min\{i+k,\,R\}}\tilde{M}^{\min\{i+k,\,R\}}} \end{pmatrix}, \; 1 \le k \le N - i,$$

$$F_0^{(i,\,i-1)} = \begin{pmatrix} \mathcal{S}_0^{(i,\,0)} & 0 & \dots & 0 & 0 \\ \tilde{\mathcal{S}}_0^{(i,\,1)} & \mathcal{S}_0^{(i,\,1)} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \mathcal{S}_0^{(i,\,R-1)} & 0 \\ 0 & 0 & \dots & \tilde{\mathcal{S}}_0^{(i,\,R)} & \mathcal{S}_0^{(i,\,R)} + \left(1 - \delta_{N,\,R}\right) p I_{\bar{W}} \otimes \\ & & & & \otimes \left(\boldsymbol{e}_M \otimes \boldsymbol{T}_0\right)^{\oplus i-R} \otimes I_{\bar{M}^R} \end{pmatrix}, \tag{2}$$

$$F_0^{(i,\,i)} = \begin{pmatrix} \mathcal{C}^{(i,\,0)} & \mathcal{T}_0^{(i,\,0)} & \dots & 0 & 0 \\ 0 & \mathcal{C}^{(i,\,1)} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \mathcal{C}^{(i,\,R-1)} & \mathcal{T}_0^{(i,\,R-1)} \\ 0 & 0 & \dots & 0 & \mathcal{C}^{(i,\,R)} + (1-p) I_{\bar{W}} \otimes \\ & & & & \otimes \left(I_M \otimes \boldsymbol{T}_0 \boldsymbol{\tau}\right)^{\oplus i-R} \otimes I_{\bar{M}^R} \end{pmatrix}, \; R < i < N, \tag{3}$$

$$F_0^{(i,\,i+k)} = \begin{pmatrix} \mathcal{D}_k^{(i,\,0)} & 0 & \dots & 0 \\ 0 & \mathcal{D}_k^{(i,\,1)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathcal{D}_k^{(i,\,R)} \end{pmatrix}, 1 \le k \le N - i.$$

Если $R < N$, то матрицы $F_0^{(N,\,N-1)}$ и $F_0^{(N,\,N)}$ вычисляются по формулам (2) и (3) соответственно, где полагается $i = N$. Если $R = N$, то матрица $F_0^{(N,\,N-1)}$ вычисляется по формуле (1), а матрица $F_0^{(N,\,N)}$ – по формуле (3).

Матрицы F_k, $k \ge 1$, порядка $K_0 \times K$ имеют следующий вид:

$$F_k = \begin{pmatrix} \operatorname{diag}\left\{\mathcal{D}_{N+k}^{(0,\,r)},\ r = \overline{0,\ \min\{0,\,R\}}\right\} & O_{\bar{W} \times \bar{W} \sum\limits_{r=1}^{R} a^{N-r} \bar{M}^r} \\ \operatorname{diag}\left\{\mathcal{D}_{N+k-1}^{(1,\,r)},\ r = \overline{0,\ \min\{1,\,R\}}\right\} & O_{\bar{W} \sum\limits_{r=0}^{1} a^{1-r} \bar{M}^r \times \bar{W} \sum\limits_{r=2}^{R} a^{N-r} \bar{M}^r} \\ \vdots & \vdots \\ \operatorname{diag}\left\{\mathcal{D}_{k+1}^{(N-1,\,r)},\ r = \overline{0,\ \min\{N-1,\,R\}}\right\} & O_{\bar{W} \sum\limits_{r=0}^{N-1} a^{N-1-r} \bar{M}^r \times \bar{W} \sum\limits_{r=N-1}^{R} a^{N-r} \bar{M}^r} \\ \operatorname{diag}\left\{\mathcal{D}_k^{(N,\,r)},\ r = \overline{0,\,R}\right\} & \end{pmatrix}, k \ge 1.$$

Матрица $\tilde{Q}_{-1}$ имеет порядок $K \times K_0$ и задается как

$$\tilde{Q}_{-1} = \left(O_{K \times (K_0 - K)} \middle| Q_{-1}\right).$$

Матрицы Q_k, $k = -1, 0, 1, \dots$, имеют порядок $K \times K$ и задаются как

$$Q_{-1} = \begin{pmatrix} \mathcal{B}^{(N,\,0)} & 0 & \dots & 0 & 0 \\ \tilde{\mathcal{B}}^{(N,\,1)} & \mathcal{B}^{(N,\,1)} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \mathcal{B}^{(N,\,R-1)} & 0 \\ 0 & 0 & \dots & \tilde{\mathcal{B}}^{(N,\,R)} & \mathcal{B}^{(N,\,R)} + pI_{\bar{W}} \otimes \left(\boldsymbol{e}_M \boldsymbol{\beta} \otimes \boldsymbol{T}_0 \boldsymbol{\tau}\right)^{\oplus N-R} \otimes I_{\bar{M}^R} \end{pmatrix},$$

$$Q_0 = \begin{pmatrix} \mathcal{C}^{(N,\,0)} & \mathcal{T}_0^{(N,\,0)} & \dots & 0 & 0 \\ 0 & \mathcal{C}^{(N,\,1)} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \mathcal{C}^{(N,\,R-1)} & \mathcal{T}_0^{(N,\,R-1)} \\ 0 & 0 & \dots & 0 & \mathcal{C}^{(N,\,R)} + (1-p) I_{\bar{W}} \otimes \left(I_M \otimes \boldsymbol{T}_0 \boldsymbol{\tau}\right)^{\oplus N-R} \otimes I_{\bar{M}^R} \end{pmatrix},$$

$$Q_k = \begin{pmatrix} \mathcal{D}_k^{(N,\,0)} & 0 & \dots & 0 \\ 0 & \mathcal{D}_k^{(N,\,1)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathcal{D}_k^{(N,\,R)} \end{pmatrix}, k \ge 1.$$

Следствие. *Цепь Маркова ξ_t, $t \ge 0$, принадлежит классу квазитеплицевых цепей Маркова с непрерывным временем* [10].

Доказательство. Инфинитезимальный генератор Q рассматриваемой цепи имеет блочную верхнюю хессенбергову структуру, и блоки, образованные интенсивностями переходов из состояний со значением i счетной компоненты, $i \geq N+1$, в состояния со значением j этой компоненты, зависят от значений i, j только через их разность $i - j$. Согласно определению квазитеплицевых цепей Маркова, приведенному в [10], это означает, что рассматриваемая цепь принадлежит классу квазитеплицевых цепей Маркова с $N+1$ граничным состоянием.

Следствие доказано.

Условие эргодичности. Стационарное распределение

Пусть $\mathcal{B}^{(r)}$, $\tilde{\mathcal{B}}^{(r)}$, $\mathcal{T}_0^{(r)}$ – матрицы, полученные из матриц $\mathcal{B}^{(N,\,r)}$, $\tilde{\mathcal{B}}^{(N,\,r)}$, $\mathcal{T}_0^{(N,\,r)}$ соответственно путем формального удаления выражений $I_{\bar{W}} \otimes$, $\mathcal{C}^{(r)}$ – матрица, полученная из матрицы $\mathcal{C}^{(N,\,r)}$ удалением выражения $D_0 \oplus$. Пусть также $\mathcal{Q}_{-1}$, $\mathcal{Q}_0$ – матрицы, полученные из Q_{-1}, Q_0 формальной заменой $\mathcal{B}^{(N,\,r)}$, $\mathcal{T}_0^{(N,\,r)}$, $\mathcal{S}_0^{(N,\,r)}$, $\mathcal{C}^{(N,\,r)}$, $r = \overline{0, R}$, на матрицы $\mathcal{B}^{(r)}$, $\mathcal{T}_0^{(r)}$, $\mathcal{S}_0^{(r)}$, $\mathcal{C}^{(r)}$ соответственно. Выражение $pI_{\bar{W}} \otimes \left(\boldsymbol{e}_M \boldsymbol{\beta} \otimes \boldsymbol{T}_0 \boldsymbol{\tau}\right)^{\oplus N-R} \otimes I_{\bar{M}^R}$ заменим на $p\left(\boldsymbol{e}_M \boldsymbol{\beta} \otimes \boldsymbol{T}_0 \boldsymbol{\tau}\right)^{\oplus N-R} \otimes I_{\bar{M}^R}$ и выражение $(1-p) I_{\bar{W}} \otimes \left(I_M \otimes \boldsymbol{T}_0 \boldsymbol{\tau}\right)^{\oplus N-R} \otimes I_{\bar{M}^R}$ – на $(1-p)\left(I_M \otimes \right.$ $\left. \otimes \boldsymbol{T}_0 \boldsymbol{\tau}\right)^{\oplus N-R} \otimes I_{\bar{M}^R}$.

Теорема. *Необходимым и достаточным условием эргодичности цепи Маркова ξ_t, $t \geq 0$, является выполнение неравенства*

$$\lambda < \sum_{r=0}^{R} \left(\boldsymbol{x}_r^{(1)} \boldsymbol{S}_0 \otimes \boldsymbol{e}_L\right)^{\oplus N-r} \boldsymbol{e} + p \boldsymbol{x}_R^{(1)} \left(\boldsymbol{e}_M \otimes \boldsymbol{T}_0\right)^{\oplus N-R} \boldsymbol{e} + \sum_{r=1}^{R} \boldsymbol{x}_r^{(2)} \tilde{\boldsymbol{S}}_0^{\oplus r} \boldsymbol{e}, \tag{4}$$

где

$$\boldsymbol{x}_r^{(1)} = \boldsymbol{x}_r \left(I_{a^{N-r}} \otimes \boldsymbol{e}_{\tilde{M}^r}\right), \; \boldsymbol{x}_r^{(2)} = \boldsymbol{x}_r \left(\boldsymbol{e}_{a^{N-r}} \otimes I_{\tilde{M}^r}\right),$$

а вектор $\boldsymbol{x} = \left(\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_R\right)$ – единственное решение системы линейных алгебраических уравнений

$$\boldsymbol{x}\left(\mathcal{Q}_{-1} + \mathcal{Q}_0\right) = \boldsymbol{0}, \; \boldsymbol{x}\boldsymbol{e} = 1. \tag{5}$$

Доказательство. Как следует из [10], необходимое и достаточное условие эргодичности квазитеплицевой цепи Маркова ξ_t, $t \geq 0$, может быть сформулировано в терминах блоков генератора Q следующим образом:

$$\boldsymbol{y} \sum_{k=0}^{\infty} (k+1) Q_k \boldsymbol{e} < 0, \tag{6}$$

где вектор $\boldsymbol{y}$ – единственное решение системы линейных алгебраических уравнений

$$\boldsymbol{y} \sum_{k=-1}^{\infty} Q_k = \boldsymbol{0}, \; \boldsymbol{y}\boldsymbol{e} = 1. \tag{7}$$

Вектор $\boldsymbol{y}$ представим в виде

$$\boldsymbol{y} = \left(\boldsymbol{\theta} \otimes \boldsymbol{x}_0, \boldsymbol{\theta} \otimes \boldsymbol{x}_1, \ldots, \boldsymbol{\theta} \otimes \boldsymbol{x}_R\right).$$

Принимая во внимание соотношение $\sum_{k=0}^{\infty} D_k \boldsymbol{e} = 0$, легко проверить, что такой вектор является единственным решением системы (7) тогда и только тогда, когда $\boldsymbol{x} = \left(\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_R\right)$ является единственным решением системы (5). Заметим, что система (5) имеет единственное решение, так как матрица $\mathcal{Q}_{-1} + \mathcal{Q}_0$ есть неприводимый генератор. Таким образом, мы перешли от системы (7) для вектора $\boldsymbol{y}$ к системе (5) для вектора $\boldsymbol{x}$.

Теперь перейдем от неравенства (6) к неравенству (4). Перепишем (6) в виде

$$\boldsymbol{y} \sum_{k=1}^{\infty} k Q_k \boldsymbol{e} < -\boldsymbol{y} \sum_{k=0}^{\infty} Q_k \boldsymbol{e}.$$

Используя равенство $\sum_{k=0}^{\infty} Q_k \boldsymbol{e} = -Q_1 \boldsymbol{e}$, имеем

$$\boldsymbol{y} \sum_{k=1}^{\infty} k Q_k \boldsymbol{e} < \boldsymbol{y} Q_{-1} \boldsymbol{e}. \tag{8}$$

Подставляя в левую часть неравенства (8) выражения для Q_k, $k \geq 1$, и принимая во внимание, что $\boldsymbol{\theta}\sum_{k=0}^{\infty} kD_k \boldsymbol{e} = \lambda,$ получим

$$\boldsymbol{y}\sum_{k=1}^{\infty} kQ_k \boldsymbol{e} = \lambda. \tag{9}$$

Теперь рассмотрим правую часть неравенства (8). Имеет место следующая цепочка соотношений:

$$\boldsymbol{y}Q_{-1}\boldsymbol{e} = \left(\boldsymbol{\theta}\otimes \boldsymbol{x}_0, \boldsymbol{\theta}\otimes \boldsymbol{x}_1, \ldots, \boldsymbol{\theta}\otimes \boldsymbol{x}_R\right)\begin{pmatrix} \mathcal{B}^{(N,0)}\boldsymbol{e} \\ \mathcal{B}^{(N,1)}\boldsymbol{e} + \tilde{\mathcal{B}}^{(N,1)}\boldsymbol{e} \\ \vdots \\ \mathcal{B}^{(N,R-1)}\boldsymbol{e} + \tilde{\mathcal{B}}^{(N,R-1)}\boldsymbol{e} \\ \mathcal{B}^{(N,R)}\boldsymbol{e} + \tilde{\mathcal{B}}^{(N,R)}\boldsymbol{e} + p\left[I_{\bar{W}}\otimes \right. \\ \left. \otimes \left(\boldsymbol{e}_M\boldsymbol{\beta}\otimes \boldsymbol{T}_0\boldsymbol{\tau}\right)^{\oplus N-R}\otimes I_{\bar{M}^R}\right]\boldsymbol{e} \end{pmatrix} =$$

$$= \boldsymbol{x}_0\left(\boldsymbol{S}_0\boldsymbol{\beta}\otimes \boldsymbol{e}\boldsymbol{\tau}\right)^{\oplus N}\boldsymbol{e} + \sum_{r=1}^{R}\boldsymbol{x}_r\left[\left(\boldsymbol{S}_0\boldsymbol{\beta}\otimes \boldsymbol{e}\boldsymbol{\tau}\right)^{\oplus N-r}\otimes I_{\tilde{M}^r}\right]\boldsymbol{e} +$$

$$+\sum_{r=1}^{R}\boldsymbol{x}_r\left[I_{a^{N-r}}\otimes\left(\boldsymbol{\beta}\otimes\boldsymbol{\tau}\right)\otimes\tilde{\boldsymbol{S}}_0^{\oplus r}\right]\boldsymbol{e} + p\boldsymbol{x}_R\left[\left(\boldsymbol{e}_M\boldsymbol{\beta}\otimes\boldsymbol{T}_0\boldsymbol{\tau}\right)^{\oplus N-R}\otimes I_{\bar{M}^R}\right]\boldsymbol{e} =$$

$$= \boldsymbol{x}_0\left(\boldsymbol{S}_0\boldsymbol{\beta}\otimes\boldsymbol{e}\boldsymbol{\tau}\right)^{\oplus N}\boldsymbol{e} + \sum_{r=1}^{R}\boldsymbol{x}_r\left(I_{a^{N-r}}\otimes \boldsymbol{e}_{\tilde{M}^r}\right)\left(\boldsymbol{S}_0\otimes\boldsymbol{e}_L\right)^{\oplus N-r}\boldsymbol{e} +$$

$$+\sum_{r=1}^{R}\boldsymbol{x}_r\left(\boldsymbol{e}_{a^{N-r}}\otimes I_{\tilde{M}^r}\right)\tilde{\boldsymbol{S}}_0^{\oplus r}\boldsymbol{e} + p\boldsymbol{x}_R\left(I_{a^{N-r}}\otimes \boldsymbol{e}_{\tilde{M}^r}\right)\left(\boldsymbol{e}_M\otimes\boldsymbol{T}_0\right)^{\oplus N-R}\boldsymbol{e} =$$

$$= \sum_{r=0}^{R}\boldsymbol{x}_r^{(1)}\left(\boldsymbol{S}_0\otimes\boldsymbol{e}_L\right)^{\oplus N-r}\boldsymbol{e} + \sum_{r=1}^{R}\boldsymbol{x}_r^{(2)}\tilde{\boldsymbol{S}}_0^{\oplus r}\boldsymbol{e} + p\boldsymbol{x}_R^{(1)}\left(\boldsymbol{e}_M\otimes\boldsymbol{T}_0\right)^{\oplus N-R}\boldsymbol{e}. \tag{10}$$

Используя (9), (10) в (8), получим искомое неравенство (4). Теорема доказана.

В дальнейшем будем считать, что условие эргодичности (4) выполняется.

Перенумеруем стационарные вероятности цепи ξ_t, $t \geq 0$, в лексикографическом порядке и образуем векторы-строки $\boldsymbol{p}_i$ вероятностей, соответствующих значению i первой компоненты цепи, $i \geq 0$. Чтобы вычислить векторы $\boldsymbol{p}_i$, $i \geq 0$, применим адаптированный к нашему случаю численно устойчивый алгоритм, который был разработан в [10] для многомерных квазитеплицевых цепей Маркова. При разработке этого алгоритма использовались элементы теории матриц и техника сенсорных цепей Маркова (см., например, [11; 12]). Для удобства читателя приводим здесь основные шаги адаптированного алгоритма.

Алгоритм. 1. Находим матрицу G как минимальное неотрицательное решение матричного уравнения

$$\sum_{n=-1}^{\infty} Q_n G^{n+1} = O.$$

2. Вычисляем матрицу G_1, используя уравнение

$$Q_{-1} + \sum_{n=0}^{\infty} Q_n G^n G_1 = O,$$

из которого следует, что

$$G_1 = -\left(\sum_{n=0}^{\infty} Q_n G^n\right)^{-1} Q_{-1}.$$

3. Находим матрицу G_0 из уравнения

$$\tilde{Q}_{-1} + \left(Q_0 + \sum_{n=1}^{\infty} Q_n G^{n-1} G_1 \right) G_0 = O,$$

откуда

$$G_0 = -\left(Q_0 + \sum_{n=1}^{\infty} Q_n G^{n-1} G_1 \right)^{-1} \tilde{Q}_{-1}.$$

4. Вычисляем матрицы

$$\bar{Q}_{i,\,l} = \begin{cases} F_l + \sum\limits_{n=l+1}^{\infty} F_n G_{n-1} G_{n-2} \cdots G_l, \; i = 0, \; l \geq 0, \\ Q_{l-i} + \sum\limits_{n=l+1}^{\infty} Q_{n-i} G_{n-1} G_{n-2} \cdots G_l, \; i \geq 1, \; l \geq i, \end{cases}$$

где $G_i = G$, $i \geq 2$.

5. Находим матрицы Φ_l по рекуррентной формуле

$$\Phi_l = \left(\bar{Q}_{0,\,l} + \sum_{i=1}^{l-1} \Phi_i \bar{Q}_{i,\,l} \right) \left(-\bar{Q}_{l,\,l} \right)^{-1}, \; l \geq 1.$$

6. Вычисляем вектор π как единственное решение системы

$$\begin{cases} \boldsymbol{\pi} \bar{Q}_{0,\,0} = \mathbf{0}, \\ \boldsymbol{\pi} \left(\boldsymbol{e}_{K_0} + \sum\limits_{l=1}^{\infty} \Phi_l \boldsymbol{e}_K \right) = 1. \end{cases}$$

7. Находим векторы $\boldsymbol{p}_{N+l}$ следующим образом: $\boldsymbol{p}_{N+l} = \boldsymbol{\pi} \Phi_l$, $l \geq 1$.

8. Вычисляем векторы $\boldsymbol{p}_l$:

$$\boldsymbol{p}_l = \boldsymbol{\pi} \begin{pmatrix} O & O & O \\ \bar{W} \sum\limits_{i=0}^{l-1} \sum\limits_{r=0}^{\min\{i,\,R\}} a^{i-r} \tilde{M}^r & & \\ O & I & O \\ & \bar{W} \sum\limits_{r=0}^{\min\{l,\,R\}} a^{i-r} \tilde{M}^r & \\ O & O & O \\ & & \bar{W} \sum\limits_{i=l+1}^{N} \sum\limits_{r=0}^{\min\{i,\,R\}} a^{i-r} \tilde{M}^r \end{pmatrix}, \; l = \overline{0,\,N}.$$

Характеристики производительности

Определив стационарное распределение $\boldsymbol{p}_i$, $i \geq 0$, можно найти ряд стационарных характеристик производительности системы. Приведем наиболее важные из них:

- среднее число заявок в системе $L = \sum\limits_{i=0}^{\infty} i \boldsymbol{p}_i \boldsymbol{e}$;
- вероятность того, что система пуста, $p_0 = \boldsymbol{p}_0 \boldsymbol{e}$;
- стационарное распределение числа занятых приборов

$$p_n = \boldsymbol{p}_n \boldsymbol{e}, \; n = \overline{0,\,N-1}, \; p_N = \sum_{i=N}^{\infty} \boldsymbol{p}_i \boldsymbol{e};$$

- среднее число занятых приборов $N_{\text{busy}} = \sum\limits_{n=1}^{N} n p_n$;
- совместная вероятность в произвольный момент застать r занятых резервных приборов, $\min\{i, N\} - r$ приборов, работающих без поддержки, и i заявок в системе

$$p_i(r) = \boldsymbol{p}_i J^{(i,r)} \boldsymbol{e},\ i \geq 0,\ r = \overline{0,\ \min\{i, R\}},$$

где

$$J^{(i,r)} = \begin{pmatrix} O & & \\ & \bar{W} \sum_{n=0}^{r-1} a^{\min\{i,N\}-n} \tilde{M}^n \times \bar{W} a^{\min\{i,N\}-r} \tilde{M}^r & \\ & & I_{\bar{W} a^{\min\{i,N\}-r} \tilde{M}^r} \\ O & & \\ & \bar{W} \sum_{n=r+1}^{\min\{i,R\}} a^{\min\{i,N\}-n} \tilde{M}^n \times \bar{W} a^{\min\{i,N\}-r} \tilde{M}^r & \end{pmatrix}. \quad (11)$$

Дадим краткое пояснение формулы (11). Умножая вектор $\boldsymbol{p}_i$ на матрицу $J^{(i,r)}$, мы выделяем часть этого вектора, соответствующую r занятым резервным приборам. Суммируя все элементы этого вектора, получаем искомую вероятность $p_i(r)$;

• стационарное распределение числа занятых резервных приборов

$$q_r = \sum_{i=r}^{\infty} p_i(r),\ r = \overline{0, R};$$

• среднее число занятых резервных приборов

$$N_{\text{busy}}^{(\text{reserve})} = \sum_{r=1}^{R} r q_r;$$

• стационарное распределение числа занятых приборов, работающих без поддержки,

$$g_l = \sum_{i=0}^{\infty} p_i q_{\min\{i,N\}-l},\ l = \overline{0, N};$$

• среднее число занятых приборов, работающих без поддержки,

$$N_{\text{busy}}^{(\text{non-support})} = \sum_{l=1}^{N} l g_l;$$

• вероятность потери произвольной заявки

$$P_{\text{loss}} = 1 - \frac{\left(\sum_{i=1}^{\infty} \boldsymbol{p}_i \sum_{r=1}^{\min\{i,R\}} J^{(i,r)} \right) \left(S_0^{(\min\{i,N\},\, r)} \boldsymbol{e} + \tilde{S}_0^{(\min\{i,N\},\, r)} \boldsymbol{e} \right)}{\boldsymbol{\lambda}}. \quad (12)$$

В формуле (12) выражение $\boldsymbol{p}_i J^{(i,r)} \cdot S_0^{(\min\{i,N\},\, r)} \boldsymbol{e}$ есть интенсивность выходного потока заявок, обслуженных на основных приборах, работающих без поддержки резервных приборов, когда в системе находится i заявок и r занятых резервных приборов. Выражение $\boldsymbol{p}_i J^{(i,r)} \cdot \tilde{S}_0^{(\min\{i,N\},\, r)} \boldsymbol{e}$ есть интенсивность выходного потока заявок, обслуженных на основных приборах, работающих с поддержкой резервных приборов, когда в системе находится i заявок и r занятых резервных приборов. Тогда числитель дроби в (12) – это суммарная интенсивность выходного потока, в то время как знаменатель – интенсивность входного потока. Значит, вычитаемое в (12) есть вероятность того, что произвольная заявка не будет потеряна, а искомое значение P_{loss} вычисляется как дополнительная вероятность.

Численные примеры

Цель этого раздела – продемонстрировать выполнимость представленных алгоритмов, привести графики зависимости характеристик производительности системы от корреляции и интенсивности входного потока и пример численного решения задачи оптимизации. Для этого ниже представлены результаты двух численных экспериментов.

Эксперимент 1. В данном эксперименте исследуется влияние коэффициента корреляции во входном потоке на среднее число заявок в системе и вероятность потери заявок. Рассмотрим три потока MAP (markovian arrival process) с разными коэффициентами корреляции: $MAP^{(0)}$, $MAP^{(0,2)}$ и $MAP^{(0,4)}$.

$MAP^{(0)}$ является стационарным пуассоновским потоком. Он имеет коэффициент корреляции $c_{\text{corr}} = 0$. Задающие его матрицы вырождаются в скаляры: $D_0 = -1{,}0$; $D = 1{,}0$.

$MAP^{(0,2)}$ имеет коэффициент корреляции $c_{\text{corr}} = 0{,}2$ и задается матрицами

$$D_0 = \begin{pmatrix} -1,352\,6 & 0 \\ 0 & -0,043\,91 \end{pmatrix}, \; D = \begin{pmatrix} 1,343\,6 & 0,009 \\ 0,024\,46 & 0,019\,45 \end{pmatrix}.$$

$MAP^{(0,4)}$ имеет коэффициент корреляции $c_{\text{corr}} = 0{,}4$ и задается матрицами

$$D_0 = \begin{pmatrix} -3,398\,23 & 0 \\ 0,001\,01 & -0,110\,24 \end{pmatrix}, \; D = \begin{pmatrix} 3,362\,83 & 0,035\,4 \\ 0,012\,14 & 0,097\,09 \end{pmatrix}.$$

На основе этих MAP-потоков построим три BMAP-потока: $MAP^{(0)}$, $MAP^{(0,2)}$ и $MAP^{(0,4)}$, каждый из которых определяется матрицами D_k, $k = 0, \ldots, 4$. Матрица D_0 такая же, как матрица D_0 в соответствующем MAP-потоке, а остальные матрицы вычисляются как $D_k = \dfrac{Dq^{k-1}(1-q)}{(1-q^4)}$, $k = \overline{0,4}$, где $q = 0{,}8$.

Время обслуживания на основном приборе, работающем без поддержки резервного прибора, имеет распределение Эрланга порядка 2 и задается вектором $\boldsymbol{\beta} = (1 \quad 0)$ и матрицей $S = \begin{pmatrix} -20 & 20 \\ 0 & -20 \end{pmatrix}$.

Время обслуживания на основном приборе, работающем с поддержкой резервного прибора, имеет *PH*-распределение, которое задается вектором $\tilde{\boldsymbol{\beta}} = (0{,}05 \quad 0{,}95)$ и матрицей $\tilde{S} = \begin{pmatrix} -3,72 & 1,0 \\ 10,0 & -292,0 \end{pmatrix}$. Время до срабатывания таймера имеет гиперэкспоненциальное распределение, которое задается вектором $\boldsymbol{\tau} = (0{,}4 \quad 0{,}6)$ и матрицей $T = \begin{pmatrix} -5,971\,281 & 0,0 \\ 0,0 & -0,507\,18 \end{pmatrix}$.

Рассматриваемая система имеет $N = 5$ основных обслуживающих приборов, $R = 5$ резервных приборов, и с вероятностью $p = 0{,}5$ заявка покидает систему, если в момент срабатывания таймера нет свободных резервных приборов.

На рис. 1 и 2 представлены графики зависимости среднего числа заявок в системе от интенсивности входного потока для BMAP с различными коэффициентами корреляции.

Как видно из рис. 1 и 2, с увеличением интенсивности входного потока ожидаемо возрастает среднее число L заявок в системе. Более интересно заметить тот факт, что при одинаковой интенсивности потока величина L существенно зависит от коэффициента корреляции – с увеличением последнего растет и среднее число заявок. Можно также заметить, что при приближении к точке $\lambda = 105$, после которой условие эргодичности нарушается, происходит повышение скорости возрастания данной характеристики.

Далее представим графики зависимости вероятности потерь P_{loss} от интенсивности входного потока для BMAP с различными коэффициентами корреляции (рис. 3). При этом частично изменим приведенные ранее данные эксперимента, взяв $R \neq N$, а именно $N = 5$, $R = 1$. Такое изменение вызвано тем, что при любых N, R таких, что $N = R$, вероятность P_{loss} равна нулю.

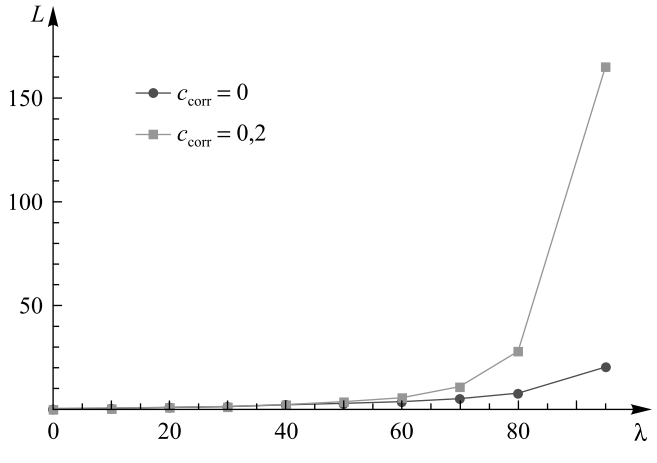


Рис. 1. Зависимость среднего числа L заявок в системе от λ при различных значениях c_{corr}

Fig. 1. Mean number of customers in the system, L; as a function of λ for BMAPs at various values of coefficients of correlation c_{corr}

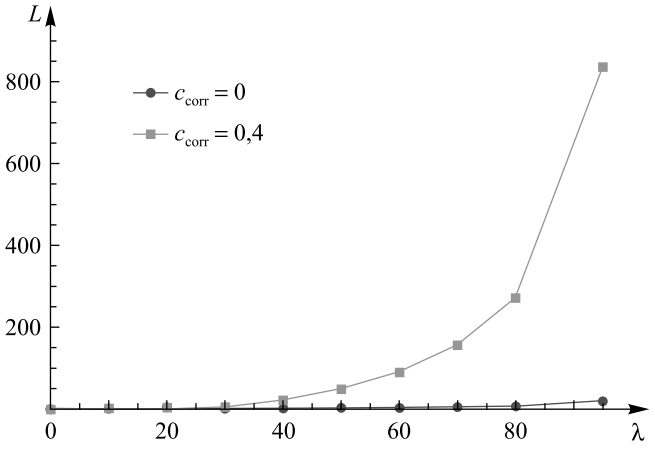


Рис. 2. Зависимость среднего числа L заявок в системе от λ при различных значениях c_{corr}

Fig. 2. Mean number of customers in the system, L; as a function of λ for BMAPs at various values of coefficients of correlation c_{corr}

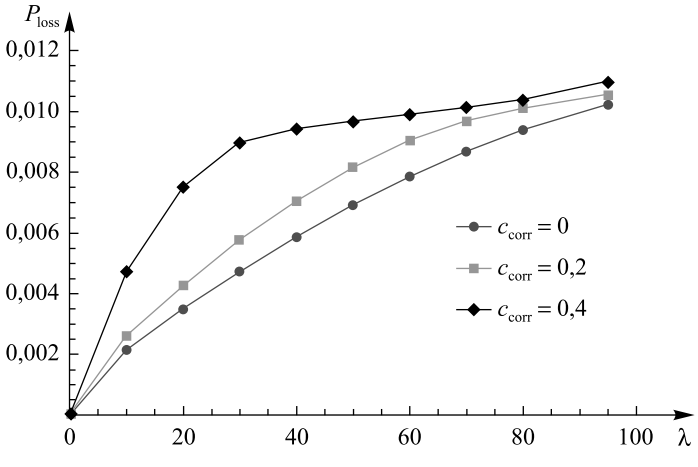


Рис. 3. Зависимость вероятности потерь P_{loss} от λ при различных значениях c_{corr}

Fig. 3. Loss probability P_{loss} as a function of λ for BMAP at various values of coefficients of correlation c_{corr}

Как видно из рис. 3, вероятность потери произвольной заявки зависит как от интенсивности входного потока, так и от коэффициента корреляции. Представленные графики показывают, что с увеличением коэффициента корреляции вероятность потерь также возрастает.

Результаты данного эксперимента свидетельствуют о том, что факт наличия корреляции во входном потоке нельзя игнорировать, иначе можно получить слишком оптимистические оценки характеристик производительности системы.

Эксперимент 2. В этом эксперименте решается задача численной оптимизации параметров системы: числа резервных приборов R и интенсивности таймера τ. Критерием качества служит стоимостный критерий – средний штраф в единицу времени

$$J = aL + cN_{busy}^{(reserve)},$$

где a – штраф за пребывание одной заявки в системе в единицу времени; c – стоимость единицы времени использования резервного прибора.

Задача оптимизации состоит в нахождении величин R и τ, доставляющих минимум критерию J. Минимум ищется путем перебора значений R и с определенным шагом τ (отсчет τ ведется от точки, где начинает выполняться условие существования стационарного режима).

В данном случае положим $N = R = 5$ и $p = 0$. В качестве входного потока возьмем $BMAP^{(0,2)}$, нормализовав матрицы D_k, $k = \overline{0, 4}$, так, чтобы получить интенсивность $\lambda = 125$. Вид распределений

времен обслуживания и таймера такой же, как в эксперименте 1, но матрицы S и $\tilde{S}$ нормализуются таким образом, чтобы получить интенсивности обслуживания $\mu = 10$ и $\tilde{\mu} = 60$. Для стоимостных коэффициентов полагаем $a = 1$ и $c = 200$.

Зависимость критерия качества J от R и τ изображена на рис. 4. Соответствующие значения приведены также в таблице.

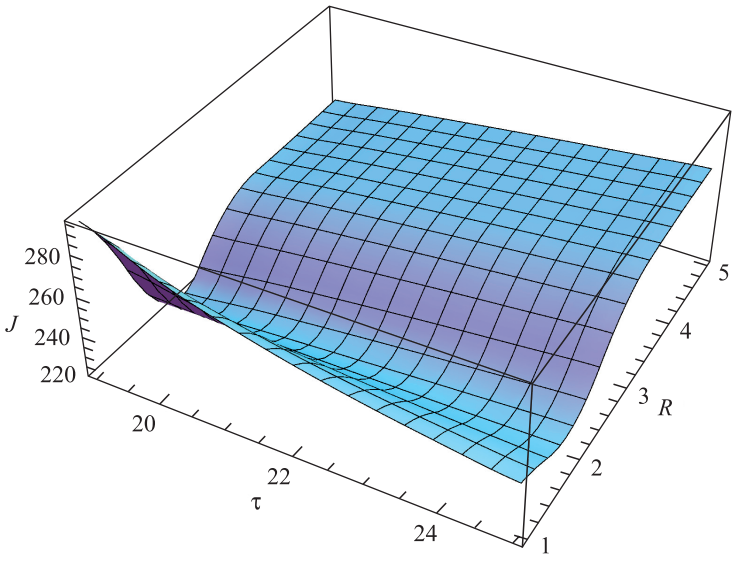


Рис. 4. Значения критерия J при различных величинах R и τ ($a = 1$, $c = 200$)

Fig. 4. The cost criterion J as a function of the number R of backup servers and the rate τ ($a = 1$, $c = 200$)

Значения критерия качества при различных величинах R и τ ($a = 1$, $c = 200$)

Dependence of the cost criterion on R and τ ($a = 1$, $c = 200$)

R	τ						
	19	20	21	22	23	24	25
1	294,80	280,78	270,25	262,08	255,60	250,35	246,05
2	**224,19**	226,31	228,32	230,24	232,08	233,84	235,52
3	244,11	247,24	250,23	253,07	255,79	258,39	260,87
4	249,26	252,77	256,12	259,33	262,40	265,35	268,17
5	249,80	253,37	256,79	260,06	263,20	266,21	269,10

Примечание. Полужирным шрифтом выделено минимальное значение.

Заключение

Рассмотрена модель многолинейной системы массового обслуживания, потенциально полезная для оптимизации работы реальных систем, где требуется удовлетворить противоречивые цели высокой производительности и низкого энергопотребления. В данной системе предполагается интенсификация процесса обслуживания, если последнее длится слишком долго, за счет подключения дополнительного ресурса (резервных приборов). Система исследована при довольно общих предположениях о процессе поступления заявок и процессе обслуживания. Получено нетривиальное условие эргодичности, найдены стационарное распределение вероятностей состояний системы и формулы для основных характеристик ее производительности, выполнены численные эксперименты по изучению поведения характеристик производительности системы в зависимости от интенсивности входного потока и от корреляции в этом потоке.

Эксперименты показали, что качество обслуживания в системе значительно зависит от корреляции во входном потоке. В частности, с увеличением корреляции растет и вероятность потерь. Это означает, что наличие корреляции должно быть учтено при проектировании и оценке производительности реальных систем. Также в статье численно решается задача оптимизации по нахождению числа резервных приборов и параметра таймера, которые доставляют минимум экономическому критерию качества функционирования системы.

Полученные результаты могут использоваться для поддержки экспертных решений при проектировании и оценке производительности реальных систем в целях уменьшения энергозатрат при сохранении их высокой производительности.

Библиографические ссылки

1. Klimenok V, Savko R. A retrial tandem queue with two types of customers and reservation of channels. In: Dudin A, Klimenok V, Tsarenkov G, Dudin S, editors. *Modern probabilistic methods for analysis of telecommunication networks. Proceedings on the Belarusian winter workshops in queueing theory. BWWQT-2013; 2013 January 28–31; Minsk, Belarus.* Berlin: Springer; 2013. p. 105–114. (Communications in Computer and Information Science; volume 356). DOI: 10.1007/978-3-642-35980-4_12.

2. Kim CS, Klimenok V, Taramin O. A tandem retrial queueing system with two Markovian flows and reservation of channels. *Computers and Operations Research*. 2010;37(7):1238–1246. DOI: 10.1016/j.cor.2009.03.030.

3. Arnon S, Barry J, Karagiannidis G, Schober R, Uysal M, editors. *Advanced optical wireless communication systems*. Cambridge: Cambridge University Press; 2012.

4. Vishnevsky V, Kozyrev D, Semenova OV. Redundant queueing system with unreliable servers. In: *Proceedings of the 6th International congress on ultra modern telecommunications and control systems and workshops (ICUMT); 2014 October 6–8; Saint Petersburg, Russia.* [S. l.]: IEEE; 2014. p. 383–386. DOI: 10.1109/ICUMT.2014.7002116.

5. Vishnevsky VM, Semenova OV, Sharov SYu. Modeling and analysis of a hybrid communication channel based on free-space optical and radio-frequency technologies. *Automation and Remote Control*. 2013;74(3):521–528. DOI: 10.1134/S0005117913030144.

6. Шаров СЮ, Семёнова ОВ. Имитационная модель беспроводного канала связи на основе лазерной и радиотехнологий. В: *Distributed computer and communication networks. Theory and applications (DCCN-2010). Proceedings of the 14th International Conference; 26–28 октября 2010 г.; Москва, Россия.* Москва: Информационные и сетевые технологии; 2010. с. 368–374.

7. Klimenok V, Dudin A, Vishnevsky V, Shumchenya V, Krishnamoorthy A. Performance measures and optimization of queueing system with reserve server. In: Vishnevskiy V, Samouylov K, Kozyrev D, editors. *Distributed Computer and Communication Networks. DCCN-2016. 19th International Conference; 2016 November 21–25; Moscow, Russia.* Cham: Springer; 2017. p. 74–88. (Communications in Computer and Information Science; volume 678). DOI: 10.1007/978-3-319-51917-3_8.

8. Lucantoni DM. New results on the single server queue with a batch Markovian arrival process. *Stochastic Models*. 1991;7(1): 1–46. DOI: 10.1080/15326349108807174.

9. Neuts MF. *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Baltimore: The Johns Hopkins University Press; 1981. 332 p.

10. Klimenok VI, Dudin AN. Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. *Queueing Systems*. 2006;54(4):245–259. DOI: 10.1007/s11134-006-0300-z.

11. Гантмахер ФР. *Теория матриц*. Москва: Наука; 1967. 576 с.

12. Kemeni JG, Snell JL, Knapp AW. *Denumerable Markov chains*. New York: Springer-Verlag; 1976. 484 p. (Graduate texts in mathematics; volume 40). DOI: 10.1007/978-1-4684-9455-6.

References

1. Klimenok V, Savko R. A retrial tandem queue with two types of customers and reservation of channels. In: Dudin A, Klimenok V, Tsarenkov G, Dudin S, editors. *Modern probabilistic methods for analysis of telecommunication networks. Proceedings on the Belarusian winter workshops in queueing theory. BWWQT-2013; 2013 January 28–31; Minsk, Belarus.* Berlin: Springer; 2013. p. 105–114. (Communications in Computer and Information Science; volume 356). DOI: 10.1007/978-3-642-35980-4_12.

2. Kim CS, Klimenok V, Taramin O. A tandem retrial queueing system with two Markovian flows and reservation of channels. *Computers and Operations Research*. 2010;37(7):1238–1246. DOI: 10.1016/j.cor.2009.03.030.

3. Arnon S, Barry J, Karagiannidis G, Schober R, Uysal M, editors. *Advanced optical wireless communication systems*. Cambridge: Cambridge University Press; 2012.

4. Vishnevsky V, Kozyrev D, Semenova OV. Redundant queueing system with unreliable servers. In: *Proceedings of the 6th International congress on ultra modern telecommunications and control systems and workshops (ICUMT); 2014 October 6–8; Saint Petersburg, Russia.* [S. l.]: IEEE; 2014. p. 383–386. DOI: 10.1109/ICUMT.2014.7002116.

5. Vishnevsky VM, Semenova OV, Sharov SYu. Modeling and analysis of a hybrid communication channel based on free-space optical and radio-frequency technologies. *Automation and Remote Control*. 2013;74(3):521–528. DOI: 10.1134/S0005117913030144.

6. Sharov SYu, Semenova OV. [Simulation model of wireless channel based on FSO and RF technologies]. In: *Distributed computer and communication networks. Theory and applications (DCCN-2010). Proceedings of the 14th International Conference; 2010 October 26–28; Moscow, Russia.* Moscow: Informatsionnye i setevye tekhnologii; 2010. p. 368–374. Russian.

7. Klimenok V, Dudin A, Vishnevsky V, Shumchenya V, Krishnamoorthy A. Performance measures and optimization of queueing system with reserve server. In: Vishnevskiy V, Samouylov K, Kozyrev D, editors. *Distributed Computer and Communication Networks. DCCN-2016. 19th International Conference; 2016 November 21–25; Moscow, Russia.* Cham: Springer; 2017. p. 74–88. (Communications in Computer and Information Science; volume 678). DOI: 10.1007/978-3-319-51917-3_8.

8. Lucantoni DM. New results on the single server queue with a batch Markovian arrival process. *Stochastic Models*. 1991;7(1): 1–46. DOI: 10.1080/15326349108807174.

9. Neuts MF. *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Baltimore: The Johns Hopkins University Press; 1981. 332 p.

10. Klimenok VI, Dudin AN. Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. *Queueing Systems*. 2006;54(4):245–259. DOI: 10.1007/s11134-006-0300-z.

11. Gantmakher FR. *Teoriya matrits.* Moscow: Nauka; 1967. 576 p. Russian.

12. Kemeni JG, Snell JL, Knapp AW. *Denumerable Markov chains*. New York: Springer-Verlag; 1976. 484 p. (Graduate texts in mathematics; volume 40). DOI: 10.1007/978-1-4684-9455-6.

Статья поступила в редколлегию 01.11.2019.
Received by editorial board 01.11.2019.

Вычислительная математика

Computational mathematics

УДК 519.632:[537.84+536.252]

МОНОТОННАЯ РАЗНОСТНАЯ СХЕМА ПОВЫШЕННОГО ПОРЯДКА ТОЧНОСТИ ДЛЯ ДВУМЕРНЫХ УРАВНЕНИЙ КОНВЕКЦИИ – ДИФФУЗИИ

В. К. ПОЛЕВИКОВ[1)]

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*

Для двумерного стационарного уравнения конвекции – диффузии общего вида построена, теоретически обоснована и испытана на тестовой задаче устойчивая конечно-разностная схема, определенная на минимальном шаблоне равномерной сетки, удовлетворяющая принципу максимума и обладающая четвертым порядком аппроксимации. Монотонность схемы контролируется двумя параметрами регуляризации, введенными в разностный оператор. Схема ориентирована на решение прикладных задач конвекции – диффузии в условиях развитого пограничного слоя, включая гравитационную и термомагнитную конвекцию, диффузию частиц в магнитной жидкости. Схема апробирована на известной задаче высокоинтенсивной гравитационной конвекции в горизонтальном канале квадратного сечения при однородном нагреве сбоку. Проведено детальное сравнение с монотонной схемой Самарского второго порядка аппроксимации на последовательности квадратных сеток с числом разбиений от 10 до 1000 на каждой стороне квадрата во всем диапазоне чисел Рэлея, соответствующих режиму ламинарной конвекции. Показано значительное преимущество схемы четвертого порядка в скорости сходимости при уменьшении шага сетки.

Ключевые слова: гравитационная конвекция; термомагнитная конвекция; диффузия частиц; уравнение конвекции – диффузии; разностная схема повышенного порядка аппроксимации; принцип максимума; параметры регуляризации.

Благодарность. Работа выполнена в рамках государственной программы научных исследований «Конвергенция» (задание 1.5.03.2).

Образец цитирования:
Полевиков ВК. Монотонная разностная схема повышенного порядка точности для двумерных уравнений конвекции – диффузии. *Журнал Белорусского государственного университета. Математика. Информатика.* 2019;3:71–83 (на англ.).
https://doi.org/10.33581/2520-6508-2019-3-71-83

For citation:
Polevikov VK. A monotone finite-difference high order accuracy scheme for the 2D convection – diffusion equations. *Journal of the Belarusian State University. Mathematics and Informatics.* 2019;3:71–83.
https://doi.org/10.33581/2520-6508-2019-3-71-83

Автор:
Виктор Кузьмич Полевиков – кандидат физико-математических наук, доцент; доцент кафедры вычислительной математики факультета прикладной математики и информатики.

Author:
Viktor K. Polevikov, PhD (physics and mathematics), docent; associate professor at the department of computational mathematics, faculty of applied mathematics and computer science.
polevikov@bsu.by

Журнал Белорусского государственного университета. Математика. Информатика. 2019;3:71–83
Journal of the Belarusian State University. Mathematics and Informatics. 2019;3:71–83

A MONOTONE FINITE-DIFFERENCE HIGH ORDER ACCURACY SCHEME FOR THE 2D CONVECTION − DIFFUSION EQUATIONS

V. K. POLEVIKOV[a]

[a]*Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*

A stable finite-difference scheme is constructed on a minimum stencil of a uniform mesh for a two-dimensional steady-state convection – diffusion equation of a general form; the scheme is theoretically studied and tested. It satisfies the maximum principle and has the fourth order of approximation. The scheme monotonicity is controlled by two regularization parameters introduced into the difference operator. The scheme is focused on solving applied convection – diffusion problems with a developed boundary layer, including gravitational convection, thermomagnetic convection, and diffusion of particles in a magnetic fluid. The scheme is tested on the well-known problem of a high-intensive gravitational convection in a horizontal channel of a square cross-section with a uniform heating from the side. A detailed comparison is performed with the monotone Samarskii scheme of the second order approximation on the sequences of square meshes with the number of partitions from 10 to 1000 on each side of the square domain and over the entire range of the Rayleigh numbers, corresponding to the laminar convection mode. A significant advantage of the fourth order scheme in the convergence rate is shown for the decreasing mesh step.

Keywords: gravitational convection; thermomagnetic convection; diffusion of particles; convection – diffusion equation; finite-difference high order approximation scheme; maximum principle; parameters of regularization.

Acknowledgements. The study was supported by the state program of scientific research «Convergence» (project 1.5.03.2).

Introduction

A solution of the applied convective heat transfer problems requires a transition to the region of high values of the Rayleigh numbers, which is characterized by a formation of boundary layers with large velocity and temperature gradients and small-scale convective motions. Similarly, the concentration of solid suspended particles in colloidal systems is redistributed because of their diffusion under the action of mass forces. For example, the ferromagnetic particles in a magnetic fluid diffuse in the direction of the magnetic-field gradient, creating zones near the boundary with large gradients of the particle concentration [1; 2]. This imposes strong requirements on stabilization and approximation properties of a difference scheme. The problem is particularly crucial in a three-dimensional case. An increase of the approximation order of the difference scheme is one of the way to solve the problem, although it is very difficult to fulfill contradictory requirements of stability and accuracy.

A standard way to increase an approximation order of a difference scheme consists in a replacement of the high order derivatives in the main part of the approximation error by the lower order derivatives, which are suitable for a difference approximation on a minimum stencil, with the help of the original differential equation under assumption of sufficiently smooth functions of the equation. The stable schemes of fourth order approximation were constructed in this way in [3] for the two-dimensional Poisson equation with steps $\frac{1}{\sqrt{5}} \le \frac{h_1}{h_2} \le \sqrt{5}$ on a uniform mesh. In principle, it is not difficult to get the fourth order scheme for the convection – diffusion equation with variable coefficients, but a serious problem is ensuring the scheme monotonicity, i. e. fulfilling conditions of the maximum principle. The practice of numerical solution of convection and diffusion problems has shown that the property of monotonicity is an important factor of a scheme applicability in conditions of a developed boundary layer.

A lot of current publications in computational mathematics are devoted to the development of numerical methods for convection – diffusion problems including two-dimensional ones (see, e. g., [4–7]). To solve them, effective finite-difference and finite-element algorithms of the first or second order of accuracy are developed.

In this work, a monotone finite-difference scheme of the fourth order of approximation is constructed for the two-dimensional steady-state convection – diffusion equations in magnetic and non-magnetic fluids. The scheme is defined on a minimum nine-point stencil of a uniform mesh. Its monotonicity is provided by two regularization parameters introduced into the difference operator. The scheme is tested on the well-known problem of natural convection.

Equations of gravitational and thermomagnetic convection

One has to deal with the problem of controlling convective heat exchange in closed cavities in design of many technological devices (e. g., cooling systems for high-voltage electric cables, power transformers, electric generators and electric motors, nuclear reactors, etc.). There are two mechanisms for convection in a non-isothermal magnetic fluid located in gravitational and non-uniform magnetic fields: gravitational and magnetic one. The first mechanism is due to the dependence of density on temperature, the second one is due to the dependence of magnetization on temperature. The presence of the magnetic mechanism opens up real possibilities in controlling the structure and the intensity of convective process by applied magnetic field. This is especially important under zero-gravity conditions, when the gravitational mechanism is absent.

The most common and investigated model of a thermomagnetic convection is a model for homogeneous, non-conducting and incompressible magnetic fluid without heat sources in the temperature equation and with the linear state equations [8–12]. The system of the steady-state convective equations for this model under the Boussinesq approximation for the density and the non-inductive approximation for the magnetic field takes the form

$$(\mathbf{v}\cdot\nabla)\mathbf{v} = \nu\nabla^2\mathbf{v} + \frac{1}{\rho_0}\left(-\nabla p + \rho\mathbf{g} + \mu_0 M\nabla H\right), \tag{1}$$

$$\nabla\cdot\mathbf{v} = 0,\ \mathbf{v}\cdot\nabla T = a\nabla^2 T; \tag{2}$$

$$\rho = \rho_0\left[1-\beta\left(T-T_0\right)\right],\ M = M\left(T_0, H_0\right) - K\left(T-T_0\right) + \frac{\partial M\left(T_0, H_0\right)}{\partial H}\left(H-H_0\right),$$

$$\rho_0 = \rho\left(T_0\right),\ \beta = -\frac{1}{\rho_0}\frac{\partial\rho\left(T_0\right)}{\partial T},\ K = -\frac{\partial M\left(T_0, H_0\right)}{\partial T},$$

where $\mathbf{v}$ is the velocity vector of the convective motion; T is the absolute temperature of the fluid; p is the pressure; H is the given value of the magnetic-field intensity; ρ is the fluid density; $\mathbf{g}$ is the gravitational acceleration vector; $M = M\left(T, H\right)$ is the magnetization of the fluid for the uniform distribution of magnetic particles; $\mu_0 = 4\pi\cdot 10^{-7}$ H/m is the magnetic constant; T_0 and H_0 are the characteristic values of the temperature and the field intensity in the fluid bulk; ν, a and β are the coefficients of the kinematic viscosity, the thermal conductivity and the volumetric thermal expansion of the fluid; K is the pyromagnetic coefficient. The last two terms in equation (1) define the gravitational and magnetic mechanisms of convection, respectively.

The idea of the non-inductive approximation consists in neglecting the influence of the fluid on the external magnetic field. The validity of the non-inductive approximation is shown in [8–11] for a wide class of thermomagnetic convection problems.

A Cartesian coordinate system x_1, x_2, x_3 with the coordinate orts $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ is introduced. We set in equations (1), (2) that $\mathbf{v} = \mathbf{v}\left(\upsilon_1, \upsilon_2, 0\right)$, $\upsilon_1 = \upsilon_1\left(x_1, x_2\right)$, $\upsilon_2 = \upsilon_2\left(x_1, x_2\right)$, $T = T\left(x_1, x_2\right)$, $\rho = \rho\left(x_1, x_2\right)$, $H = H\left(x_1, x_2\right)$, $\mathbf{g} = \mathbf{g}\left(g_1, g_2, 0\right)$ assuming that the convective problem is two-dimensional. Let us define a stream function $\psi\left(x_1, x_2\right)$ and a vorticity $\omega\left(x_1, x_2\right)$ associated with the velocity components by relations

$$\upsilon_1 = \frac{\partial\psi}{\partial x_2},\ \upsilon_2 = -\frac{\partial\psi}{\partial x_1},\ \omega = \frac{\partial\upsilon_2}{\partial x_1} - \frac{\partial\upsilon_1}{\partial x_2}. \tag{3}$$

The continuity equation $\nabla\cdot\mathbf{v} = 0$ is automatically satisfied in these variables. We obtain the vector equation for the vorticity by applying the rotor operator to motion equation (1) and taking into account (3):

$$\nabla\times\left[\mathbf{v}\times\left(\omega\mathbf{k}\right)\right] = \nu\nabla\times\left[\nabla\times\left(\omega\mathbf{k}\right)\right] + \beta\nabla\times\left(T\mathbf{g}\right) + \left(\frac{\mu_0 K}{\rho_0}\right)\nabla T\times\nabla H.$$

Thus, equations (1), (2) in 2D case are transformed into a system of three scalar equations for the temperature T, the stream function ψ and the vorticity ω:

$$\begin{gathered}\frac{\partial\psi}{\partial x_2}\frac{\partial T}{\partial x_1} - \frac{\partial\psi}{\partial x_1}\frac{\partial T}{\partial x_2} = a\left(\frac{\partial^2 T}{\partial x_1^2} + \frac{\partial^2 T}{\partial x_2^2}\right),\ \frac{\partial^2\psi}{\partial x_1^2} + \frac{\partial^2\psi}{\partial x_2^2} + \omega = 0,\\ \frac{\partial\psi}{\partial x_2}\frac{\partial\omega}{\partial x_1} - \frac{\partial\psi}{\partial x_1}\frac{\partial\omega}{\partial x_2} = \nu\left(\frac{\partial^2\omega}{\partial x_1^2} + \frac{\partial^2\omega}{\partial x_2^2}\right) - \beta\left(\frac{\partial T}{\partial x_1}g_2 - \frac{\partial T}{\partial x_2}g_1\right) + \frac{\mu_0 K}{\rho_0}\left(\frac{\partial T}{\partial x_2}\frac{\partial H}{\partial x_1} - \frac{\partial T}{\partial x_1}\frac{\partial H}{\partial x_2}\right).\end{gathered} \tag{4}$$

Let Ox_2-axis be the vertical axis in the Cartesian coordinate system x_1, x_2, in which case $g_1 = 0$, $g_2 = -g$. Let T_0 and $T_1 = T_0 + \Delta T$ define the given minimum and maximum values of the temperature on the walls. We introduce dimensionless variables by choosing the characteristic size of the computational domain l as the length scale, the kinematic viscosity ν as the scale for the stream function, the relation νl^2 as the scale for the vorticity, the relation νl as the scale for the velocity, the temperature difference ΔT as the scale for the temperature, and the value γl as the scale for the magnetic field intensity where γ is a characteristic value of the field gradient. For convenience, we denote the dimensionless variables in the same way as the dimensional ones, and write system (4) in new variables (see [12]):

$$\begin{gathered}\mathbf{v}\cdot\nabla T = \frac{1}{\mathrm{Pr}}\nabla^2 T,\ \nabla^2\psi + \omega = 0,\ \mathbf{v}\cdot\nabla\omega = \nabla^2\omega + f,\ f = \mathrm{Gr}\frac{\partial T}{\partial x_1} + \mathrm{Gr}_m\frac{\partial(H,T)}{\partial(x_1,x_2)};\\ \upsilon_1 = \frac{\partial\psi}{\partial x_2},\ \upsilon_2 = -\frac{\partial\psi}{\partial x_1};\ \mathrm{Pr} = \frac{\nu}{a},\ \mathrm{Gr} = \frac{\beta g l^3\Delta T}{\nu^2},\ \mathrm{Gr}_m = \frac{\mu_0 K l^3\Delta T\gamma}{\rho_0\nu^2},\end{gathered} \tag{5}$$

where Pr is the Prandtl number; Gr is the Grashof number and Gr_m is the magnetic Grashof number. Equations (5) at $\mathrm{Gr}_m = 0$ describe the process of natural (gravitational) convection.

Equation of particle diffusion in magnetic fluid

The magnetic fluid is a stable colloidal suspension of ferromagnetic nanoparticles in a nonmagnetic carrier liquid. A particle size is of the order of 10 nm = 10^{-8} m and they are in a Brownian motion in the carrier liquid. Due to the magnetic properties of particles, not only the Brownian motion but also the diffusion of particles under the action of a non-uniform magnetic field (magnetophoresis) occurs in the magnetic fluid. The particles are distributed in the fluid bulk as a result of the competition between these two mechanisms.

The steady-state diffusion equation for magnetic particles in a magnetic fluid in the presence of a convective motion takes the form [1; 2; 13]:

$$\begin{gathered}\nabla^2 C - \left(\frac{1}{D}\mathbf{v} + \alpha\right)\cdot\nabla C - qC = 0,\\ q = \nabla\cdot\alpha,\ \alpha = \mathcal{L}(\xi)\nabla\xi,\ \xi = \frac{\mu_0 m}{kT}H,\ \mathcal{L}(\xi) = \coth(\xi) - \frac{1}{\xi} > 0,\end{gathered} \tag{6}$$

where C is the volume particle concentration in the colloid; D is the diffusion coefficient; $\mathcal{L}(\xi)$ is the Langevin function; m is the magnetic moment of a particle; $k = 1.380\,656\,8\cdot 10^{-23}$ J/K is the Boltzmann constant; T is the particle temperature.

The magnetization M is a function of the field intensity and the particles concentration, i. e. $M = M(H, C)$, for isothermal magnetic fluids. Under the condition $M(H, C) \ll H$, the Maxwell equations are of the form $\nabla\times\mathbf{H} = 0$, $\nabla\cdot\mathbf{H} = 0$. In 2D case of Cartesian coordinates, it follows that

$$\frac{\partial H_2}{\partial x_1} - \frac{\partial H_1}{\partial x_2} = 0,\ \frac{\partial H_1}{\partial x_1} + \frac{\partial H_2}{\partial x_2} = 0, \tag{7}$$

where H_1, H_2 are the components of the intensity vector $\mathbf{H}$.

From the point of view of stability of the difference scheme, it is important to show that the coefficient q in equation (6) takes only positive values. We prove this taking into account (7). Consider first

$$\begin{aligned}|\nabla H|^2 &= \left(\frac{\partial H}{\partial x_1}\right)^2 + \left(\frac{\partial H}{\partial x_2}\right)^2 = \frac{1}{H^2}\left[\left(H_1\frac{\partial H_1}{\partial x_1} + H_2\frac{\partial H_2}{\partial x_1}\right)^2 + \left(H_1\frac{\partial H_1}{\partial x_2} + H_2\frac{\partial H_2}{\partial x_2}\right)^2\right] =\\ &= \frac{1}{H^2}\left[H_1^2\left(\frac{\partial H_1}{\partial x_1}\right)^2 + H_2^2\left(\frac{\partial H_2}{\partial x_1}\right)^2 + H_1^2\left(\frac{\partial H_1}{\partial x_2}\right)^2 + H_2^2\left(\frac{\partial H_2}{\partial x_2}\right)^2\right] +\\ &+ \frac{2}{H^2}H_1H_2\underbrace{\left(\frac{\partial H_1}{\partial x_1}\frac{\partial H_2}{\partial x_1} + \frac{\partial H_1}{\partial x_2}\frac{\partial H_2}{\partial x_2}\right)}_{\text{equal 0 by virtue of (7)}} = \left(\frac{\partial H_1}{\partial x_1}\right)^2 + \left(\frac{\partial H_2}{\partial x_1}\right)^2 \overset{(7)}{=} \left(\frac{\partial H_2}{\partial x_2}\right)^2 + \left(\frac{\partial H_1}{\partial x_2}\right)^2.\end{aligned} \tag{8}$$

Then

$$\nabla^2 H = \frac{\partial}{\partial x_1}\left(\frac{1}{2H}\frac{\partial\left(H_1^2 + H_2^2\right)}{\partial x_1}\right) + \frac{\partial}{\partial x_2}\left(\frac{1}{2H}\frac{\partial\left(H_1^2 + H_2^2\right)}{\partial x_2}\right) =$$

$$= \frac{1}{H}\left[\left(\frac{\partial H_1}{\partial x_1}\right)^2 + \left(\frac{\partial H_2}{\partial x_1}\right)^2 + \left(\frac{\partial H_1}{\partial x_2}\right)^2 + \left(\frac{\partial H_2}{\partial x_2}\right)^2 + H_1\nabla^2 H_1 + H_2\nabla^2 H_2\right] -$$

$$- \frac{1}{2H^2}\left(\frac{\partial H}{\partial x_1}\frac{\partial\left(H^2\right)}{\partial x_1} + \frac{\partial H}{\partial x_2}\frac{\partial\left(H^2\right)}{\partial x_2}\right) \overset{(7),\,(8)}{=} \frac{1}{H}\left|\nabla H\right|^2 > 0 \text{ if } \nabla H \neq 0.$$

Hence

$$\nabla^2 \xi = \frac{1}{\xi}\left|\nabla\xi\right|^2 > 0 \text{ if } \nabla\xi \neq 0. \tag{9}$$

Taking into account (9), we obtain

$$q = \nabla\cdot\alpha = \nabla\cdot\left(\mathcal{L}(\xi)\nabla\xi\right) = \nabla\mathcal{L}(\xi)\cdot\nabla\xi + \mathcal{L}(\xi)\nabla^2\xi = \frac{d}{d\xi}\left(\xi\mathcal{L}(\xi)\right)\frac{1}{\xi}\left|\nabla\xi\right|^2 \geq 0.$$

Thus, we get that $q \geq 0$ in concentration equation (6). Moreover, we have $q \equiv 0$ if and only if $\nabla H \equiv 0$, i. e. when the magnetic field is uniform or absent.

Difference scheme of high order accuracy

Let us consider a two-dimensional steady-state convection – diffusion equation

$$\sum_{\alpha=1}^{2}\mathcal{L}_\alpha^{(k,\,\upsilon)}u - q(x)u = -f(x),\ x = (x_1,\,x_2) \in G, \tag{10}$$

where $\mathcal{L}_\alpha^{(k,\,\upsilon)}u = \mathcal{L}_\alpha u - \upsilon_\alpha(x)k(x)\dfrac{\partial u}{\partial x_\alpha}$, $\mathcal{L}_\alpha u = \dfrac{\partial}{\partial x_\alpha}\left(k(x)\dfrac{\partial u}{\partial x_\alpha}\right)$, $k(x) > 0$, $q(x) \geq 0$, $u = u(x)$, is the unknown function satisfying equation (10); k, q, υ_1, υ_2, and f are the given functions; x_1, x_2 are the space coordinates. All functions are assumed to be sufficiently smooth. The first term in the differential operator $\mathcal{L}_\alpha^{(k,\,\upsilon)}u$ is the diffusion term, the second one is the convective term. Note that each of equations (4)–(6) can be written in form (10).

Scheme construction. We construct the finite-difference scheme for equation (10) which has the fourth order of approximation on the minimal nine-point stencil of a uniform mesh and satisfies the maximum principle. Note, that for $q \equiv 0$ the high order scheme is presented in [12].

We approximate the differential operators $\mathcal{L}_\alpha^{(k,\,\upsilon)}$, $\alpha = 1,\,2$, by monotone difference operators $\Lambda_\alpha^{(a,\,b)}$ of the form

$$\Lambda_\alpha^{(a,\,b)}u = æ_\alpha\left(a_\alpha u_{\bar{x}_\alpha}\right)_{x_\alpha} - b_\alpha^+ a_\alpha u_{\bar{x}_\alpha} - b_\alpha^- a_\alpha^{(+1_\alpha)}u_{x_\alpha} =$$

$$= \left(1 + æ_\alpha R_\alpha^4\right)\left(a_\alpha u_{\bar{x}_\alpha}\right)_{x_\alpha} - \frac{1}{2}b_\alpha\left(a_\alpha u_{\bar{x}_\alpha} + a_\alpha^{(+1_\alpha)}u_{x_\alpha}\right) \tag{11}$$

with the coefficients

$$a_\alpha = \frac{6}{\left(\dfrac{1}{k^{(-1_\alpha)}} + \dfrac{4}{k^{(-0.5_\alpha)}} + \dfrac{1}{k}\right)} > 0,\ b_\alpha = \upsilon_\alpha + O\left(h^4\right),\ h = \sqrt{h_1^2 + h_2^2},$$

$$æ_\alpha = \frac{1}{1 + R_\alpha + R_\alpha^2 + R_\alpha^3} > 0,\ R_\alpha = \frac{1}{2}h_\alpha\left|b_\alpha\right| > 0, \tag{12}$$

$$b_\alpha^+ = \frac{1}{2}\left(b_\alpha + \left|b_\alpha\right|\right) \geq 0,\ b_\alpha^- = \frac{1}{2}\left(b_\alpha - \left|b_\alpha\right|\right) \leq 0.$$

Here h_1 and h_2 are steps of a uniform mesh relative to variables x_1 and x_2, respectively. The standard non-index notations are used for the left and right difference derivatives and for the function values at the peripheral points of the stencil:

$$u_{\bar{x}_\alpha} = \frac{u - u^{(-1_\alpha)}}{h_\alpha},\ u_{x_\alpha} = \frac{u^{(+1_\alpha)} - u}{h_\alpha},\ \alpha = 1, 2,$$

$$u = u(x),\ u^{(\pm 1_1)} = u(x_1 \pm h_1,\ x_2),\ u^{(\pm 1_2)} = u(x_1,\ x_2 \pm h_2),$$

$$k = k(x),\ k^{(\pm 0.5_1)} = k(x_1 \pm 0.5h_1,\ x_2),\ k^{(\pm 0.5_2)} = k(x_1,\ x_2 \pm 0.5h_2),$$

where $x = (x_1, x_2)$ is the central node of the stencil.

The finite-difference operators $\Lambda_\alpha^{(a,b)}u$ approximate the corresponding differential operators $L_\alpha^{(k,\upsilon)}u$ with the second order. We note that operators (11) are analogous to the operators of the well-known monotone scheme of the second order described in the book of A. A. Samarskii [3], but we define the scheme coefficients a_α, b_α and $æ_\alpha$ in a different way.

Under assumptions (12) for the coefficients a_α, the following asymptotic expansions at the center node of the mesh stencil are valid:

$$\begin{aligned} a_\alpha u_{\bar{x}_\alpha} &= k\frac{\partial u}{\partial x_\alpha} - \frac{1}{2}h_\alpha L_\alpha u + \frac{1}{6}h_\alpha^2\sqrt{k}\,\frac{\partial}{\partial x_\alpha}\left(\frac{1}{\sqrt{k}}L_\alpha u\right) - \frac{1}{24}h_\alpha^3 L_\alpha\left(\frac{1}{k}L_\alpha u\right) + O\left(h_\alpha^4\right), \\ a_\alpha^{(+1_\alpha)}u_{x_\alpha} &= k\frac{\partial u}{\partial x_\alpha} + \frac{1}{2}h_\alpha L_\alpha u + \frac{1}{6}h_\alpha^2\sqrt{k}\,\frac{\partial}{\partial x_\alpha}\left(\frac{1}{\sqrt{k}}L_\alpha u\right) + \frac{1}{24}h_\alpha^3 L_\alpha\left(\frac{1}{k}L_\alpha u\right) + O\left(h_\alpha^4\right),\ \alpha = 1, 2. \end{aligned} \tag{13}$$

Taking into account (13) we get the following relation

$$\sum_{\alpha=1}^{2} L_\alpha^{(k,\upsilon)}u = \sum_{\alpha=1}^{2}\Lambda_\alpha^{(a,b)}u + \frac{h^2}{12}Eu + O\left(h^4\right), \tag{14}$$

connecting the differential and difference operators for any sufficiently smooth function $u(x_1, x_2)$, where

$$Eu = \sum_{\alpha=1}^{2}\left\{\delta_\alpha^2\left[-L_\alpha^{(k,\upsilon)}\left(\frac{1}{k}L_\alpha^{(k,\upsilon)}u\right) + p_\alpha L_\alpha u - r_\alpha k\frac{\partial u}{\partial x_\alpha}\right]\right\},$$

$$p_\alpha = \upsilon_\alpha^2 + \frac{\upsilon_\alpha}{k}\frac{\partial k}{\partial x_\alpha} - 2\frac{\partial \upsilon_\alpha}{\partial x_\alpha},\ r_\alpha = L_\alpha^{(k,\upsilon)}\left(\frac{\upsilon_\alpha}{k}\right),\ \delta_\alpha = \frac{h_\alpha}{h}.$$

Following the conventional methodology of increasing the approximation order on the minimal stencil, we modify the operator Eu, by expressing $L_1^{(k,\upsilon)}u = -L_2^{(k,\upsilon)}u + qu - f$, $L_2^{(k,\upsilon)}u = -L_1^{(k,\upsilon)}u + qu - f$ from equation (10) and substituting them into a term with the derivatives of the order 3–4. We exclude in this way the derivatives of a high order, which are not suitable for difference approximation on the minimum stencil. In addition we introduce in the operator Eu some regularization parameters $\sigma_0 = \sigma_0(x) \geq 0$ and $\sigma_1 = \sigma_1(x)$ by adding a term, which is identically zero on the solution of equation (10) $u = u(x)$.

Due to these changes we get

$$\begin{aligned} Eu = \sum_{\alpha=1}^{2}\left\{\delta_\alpha^2\left[L_\alpha^{(k,\upsilon)}\left(\frac{1}{k}\left(L_\beta^{(k,\upsilon)}u - qu + f\right)\right) + p_\alpha L_\alpha u - r_\alpha k\frac{\partial u}{\partial x_\alpha}\right]\right\} + \\ + (\sigma_0 + \sigma_1)\underbrace{\left(\sum_{\alpha=1}^{2}L_\alpha^{(k,\upsilon)}u - qu + f\right)}_{\text{equal } 0},\ \beta \neq \alpha. \end{aligned}$$

The introduced regularization parameters allow regulating the basic properties of the difference scheme providing the maximum-principle conditions and keeping the fourth order of approximation.

By simple manipulations, the operator Eu is reduced to the final form

$$Eu = \sum_{\alpha=1}^{2}\left\{\delta_\alpha^2\left[L_\alpha^{(k,\upsilon)}\left(\frac{1}{k}L_\beta^{(k,\upsilon)}u\right) + p_\alpha L_\alpha u - r_\alpha k\frac{\partial u}{\partial x_\alpha} - \frac{q}{k}L_\alpha^{(k,\upsilon)}u - \right.\right.$$

$$\left.\left. - 2k\frac{\partial\left(\frac{q}{k}\right)}{\partial x_\alpha}\frac{\partial u}{\partial x_\alpha} - L_\alpha^{(k,\upsilon)}\left(\frac{q}{k}\right)u\right] + \left(\sigma_0+\sigma_1\right)L_\alpha^{(k,\upsilon)}u\right\} +$$

$$+ \sum_{\alpha=1}^{2}\delta_\alpha^2 L_\alpha^{(k,\upsilon)}\left(\frac{f}{k}\right) + \left(\sigma_0+\sigma_1\right)\left(f-qu\right) =$$

$$= \sum_{\alpha=1}^{2}\left\{\delta_\alpha^2\left[L_\alpha^{(k,\upsilon)}\left(\frac{1}{k}L_\beta^{(k,\upsilon)}u\right) + p_\alpha L_\alpha u - \left(r_\alpha + 2\frac{\partial\left(\frac{q}{k}\right)}{\partial x_\alpha}\right)k\frac{\partial u}{\partial x_\alpha} + \right.\right.$$

$$\left.\left. + \frac{q}{k}\left(L_\beta^{(k,\upsilon)}u - qu + f\right)\right] + \left(\sigma_0+\sigma_1\right)L_\alpha^{(k,\upsilon)}u\right\} + \sum_{\alpha=1}^{2}\delta_\alpha^2 L_\alpha^{(k,\upsilon)}\left(\frac{f}{k}\right) -$$

$$- \sum_{\alpha=1}^{2}\delta_\alpha^2 L_\alpha^{(k,\upsilon)}\left(\frac{q}{k}\right)u + \left(\sigma_0+\sigma_1\right)\left(f-qu\right).$$

Thus, we get for the main part of the approximation error

$$\frac{h^2}{12}Eu = \sum_{\alpha=1}^{2}\left(\tilde{k}_\alpha L_\alpha^{(k,\tilde{\upsilon})}u + \frac{h^2}{12}\sigma_0 L_\alpha^{(k,\upsilon)}u\right) + \frac{h_1^2}{12}L_1^{(k,\upsilon)}\left(\frac{1}{k}L_2^{(k,\upsilon)}u\right) + \frac{h_2^2}{12}L_2^{(k,\upsilon)}\left(\frac{1}{k}L_1^{(k,\upsilon)}u\right) - \tilde{q}u + \tilde{f}, \tag{15}$$

where

$$\tilde{k}_\alpha = 1 + \frac{h^2}{12}\tilde{p}_\alpha,\ \ \tilde{p}_\alpha = \delta_\alpha^2 p_\alpha + \delta_\beta^2\frac{q}{k} + \sigma_1,\ \ p_\alpha = \upsilon_\alpha^2 + \frac{\upsilon_\alpha}{k}\frac{\partial k}{\partial x_\alpha} - 2\frac{\partial\upsilon_\alpha}{\partial x_\alpha},$$

$$\tilde{\upsilon}_\alpha = \frac{1}{\tilde{k}_\alpha}\left(\upsilon_\alpha + \frac{h^2}{12}\tilde{r}_\alpha\right),\ \ \tilde{r}_\alpha = \delta_\alpha^2\left(r_\alpha + 2\frac{\partial\left(\frac{q}{k}\right)}{\partial x_\alpha}\right) + \left(\delta_\beta^2\frac{q}{k} + \sigma_1\right)\upsilon_\alpha,\ \ r_\alpha = L_\alpha^{(k,\upsilon)}\left(\frac{\upsilon_\alpha}{k}\right), \tag{16}$$

$$\tilde{q} = q + \frac{h^2}{12}\left[\sum_{\alpha=1}^{2}\delta_\alpha^2 L_\alpha^{(k,\upsilon)}\left(\frac{q}{k}\right) + \left(\frac{q}{k} + \sigma_0 + \sigma_1\right)q\right],$$

$$\tilde{f} = f + \frac{h^2}{12}\left[\sum_{\alpha=1}^{2}\delta_\alpha^2 L_\alpha^{(k,\upsilon)}\left(\frac{f}{k}\right) + \left(\frac{q}{k} + \sigma_0 + \sigma_1\right)f\right].$$

We choose the regularization parameter σ_1 in expression (15) from the conditions $\tilde{k}_\alpha \geq 1$ and $\tilde{q} \geq 0$ for $\sigma_0 \geq 0$. A feasible value of the parameter σ_1 is determined from these inequalities:

$$\sigma_1 = -\min\left[\min_\alpha\left(\delta_\alpha^2 p_\alpha\right),\ \frac{1}{q}\sum_{\alpha=1}^{2}\delta_\alpha^2 L_\alpha^{(k,\upsilon)}\left(\frac{q}{k}\right)\right] = O(1). \tag{17}$$

Taking into account representation (15) for the main part of the approximation error, the scheme of high approximation order can be written in the following form

$$\sum_{\alpha=1}^{2}\left(\tilde{a}_\alpha\Lambda_\alpha^{(a,\tilde{b})}y+\frac{h^2}{12}\sigma_0\Lambda_\alpha^{(a,b)}y\right)+\frac{h_1^2}{12}\Lambda_1^{(a,b)}\left(\frac{1}{k}\Lambda_2^{(a,b)}y\right)+\frac{h_2^2}{12}\Lambda_2^{(a,b)}\left(\frac{1}{k}\Lambda_1^{(a,b)}y\right)-\tilde{d}y+\tilde{\varphi}=0, \tag{18}$$

where $y=y(x)$ is the solution of the difference problem; $x=(x_1,x_2)$ is the internal mesh node,

$$\Lambda_\alpha^{(a,\tilde{b})}y=\tilde{æ}_\alpha\left(a_\alpha y_{\bar{x}_\alpha}\right)_{x_\alpha}-\tilde{b}_\alpha^{+}a_\alpha y_{\bar{x}_\alpha}-\tilde{b}_\alpha^{-}a_\alpha^{(+1_\alpha)}y_{x_\alpha},\ \tilde{æ}_\alpha=\frac{1}{1+\tilde{R}_\alpha+\tilde{R}_\alpha^2+\tilde{R}_\alpha^3},\ \tilde{R}_\alpha=\frac{1}{2}h_\alpha\left|\tilde{b}_\alpha\right|,$$
$$\tilde{a}_\alpha=\tilde{k}_\alpha+O\left(h^4\right)\geq 1,\ \tilde{b}_\alpha=\tilde{\upsilon}_\alpha+O\left(h^4\right),\ \tilde{d}=\tilde{q}+O\left(h^4\right)\geq 0,\ \tilde{\varphi}=\tilde{f}+O\left(h^4\right). \tag{19}$$

Obviously, scheme (18) is defined on the minimum nine-point stencil.

Approximation order. Let us consider the approximation error for scheme (18):

$$\nu=\sum_{\alpha=1}^{2}\left(\tilde{a}_\alpha\Lambda_\alpha^{(a,\tilde{b})}u-L_\alpha^{(k,\upsilon)}u+\frac{h^2}{12}\sigma_0\Lambda_\alpha^{(a,b)}u\right)+$$
$$+\frac{h_1^2}{12}\Lambda_1^{(a,b)}\left(\frac{1}{k}\Lambda_2^{(a,b)}u\right)+\frac{h_2^2}{12}\Lambda_2^{(a,b)}\left(\frac{1}{k}\Lambda_1^{(a,b)}u\right)-\tilde{d}u+\tilde{\varphi}+qu-f,$$

where $u=u(x)$ is the solution of differential equation (10). Taking into account (12), (14) and (19), we have

$$\nu=\sum_{\alpha=1}^{2}\left[\tilde{k}_\alpha\left(\Lambda_\alpha^{(k,\tilde{\upsilon})}u-L_\alpha^{(k,\tilde{\upsilon})}u\right)+\frac{h^2}{12}\tilde{p}_\alpha L_\alpha u-\frac{h^2}{12}\tilde{r}_\alpha k\frac{\partial u}{\partial x_\alpha}+\frac{h^2}{12}\sigma_0\Lambda_\alpha^{(k,\upsilon)}u\right]+$$
$$+\frac{h_1^2}{12}L_1^{(k,\upsilon)}\left(\frac{1}{k}L_2^{(k,\upsilon)}u\right)+\frac{h_2^2}{12}L_2^{(k,\upsilon)}\left(\frac{1}{k}L_1^{(k,\upsilon)}u\right)-(\tilde{q}-q)u+\tilde{f}-f+$$
$$+\sigma_0O\left(h^4\right)+O\left(h^4\right)=\sum_{\alpha=1}^{2}\left[\tilde{k}_\alpha\left(\Lambda_\alpha^{(k,\tilde{\upsilon})}u-L_\alpha^{(k,\tilde{\upsilon})}u\right)\right]+\frac{h^2}{12}Eu+\sigma_0O\left(h^4\right)+O\left(h^4\right)=$$
$$=\frac{h^2}{12}\sum_{\alpha=1}^{2}\left[\tilde{p}_\alpha\left(\Lambda_\alpha^{(k,\tilde{\upsilon})}u-L_\alpha^{(k,\tilde{\upsilon})}u\right)\right]+\sigma_0O\left(h^4\right)+O\left(h^4\right)=\sigma_0O\left(h^4\right)+O\left(h^4\right). \tag{20}$$

It follows that scheme (18) has the fourth order approximation for $\sigma_0=O(1)$. A concrete value of the parameter σ_0 is determined from the monotonicity conditions of the difference scheme.

Stability and convergence. We investigate the stability of scheme (18) using the maximum principle [3]. For this purpose, scheme (18) is rewritten in the canonical form of the maximum principle:

$$Cy=\sum_{\alpha=1}^{2}\left(A_\alpha y^{(-1_\alpha)}+B_\alpha y^{(+1_\alpha)}\right)+A_{12}y^{(-1_1,-1_2)}+B_{12}y^{(+1_1,+1_2)}+D_{12}y^{(-1_1,+1_2)}+D_{21}y^{(+1_1,-1_2)}+\tilde{\varphi},$$

where

$$A_\alpha=\frac{1}{12h_\alpha^2}\eta_\alpha\left(\sigma_0h^2-A_\alpha^*\right),\ B_\alpha=\frac{1}{12h_\alpha^2}\xi_\alpha\left(\sigma_0h^2-B_\alpha^*\right),$$
$$A_{12}=\frac{1}{12}\left[\frac{1}{h_2^2}\eta_1\left(\frac{\eta_2}{k}\right)^{(-1_1)}+\frac{1}{h_1^2}\eta_2\left(\frac{\eta_1}{k}\right)^{(-1_2)}\right]>0, \tag{21}$$
$$B_{12}=\frac{1}{12}\left[\frac{1}{h_2^2}\xi_1\left(\frac{\xi_2}{k}\right)^{(+1_1)}+\frac{1}{h_1^2}\xi_2\left(\frac{\xi_1}{k}\right)^{(+1_2)}\right]>0,$$

$$D_{\alpha\beta} = \frac{1}{12}\left[\frac{1}{h_\beta^2}\eta_\alpha\left(\frac{\xi_\beta}{k}\right)^{(-1_\alpha)} + \frac{1}{h_\alpha^2}\xi_\beta\left(\frac{\eta_\alpha}{k}\right)^{(+1_\beta)}\right] > 0,$$

$$C = \sum_{\alpha=1}^{2}\left(A_\alpha + B_\alpha\right) + A_{12} + B_{12} + D_{12} + D_{21} + \tilde{d};$$

$$A_\alpha^* = -12\frac{\tilde{\eta}_\alpha\tilde{a}_\alpha}{\eta_\alpha} + \frac{h_\alpha^2}{h_\beta^2}\left(\frac{\eta_\beta + \xi_\beta}{k}\right)^{(-1_\alpha)} + \frac{\eta_\beta + \xi_\beta}{k},\ B_\alpha^* = -12\frac{\tilde{\xi}_\alpha\tilde{a}_\alpha}{\xi_\alpha} + \frac{h_\alpha^2}{h_\beta^2}\left(\frac{\eta_\beta + \xi_\beta}{k}\right)^{(+1_\alpha)} + \frac{\eta_\beta + \xi_\beta}{k},$$

$$\eta_\alpha = a_\alpha\left(æ_\alpha + h_\alpha b_\alpha^+\right) > 0,\ \xi_\alpha = a_\alpha^{(+1_\alpha)}\left(æ_\alpha - h_\alpha b_\alpha^-\right) > 0,$$

$$\tilde{\eta}_\alpha = a_\alpha\left(\tilde{æ}_\alpha + h_\alpha\tilde{b}_\alpha^+\right) > 0,\ \tilde{\xi}_\alpha = a_\alpha^{(+1_\alpha)}\left(\tilde{æ}_\alpha - h_\alpha\tilde{b}_\alpha^-\right) > 0.$$

The coefficients A_α and B_α correspond to left, right, lower and upper peripheral nodes of the stencil relative to the central node. Their signs depend on the choice of the regularization parameter σ_0. The angular coefficients A_{12}, B_{12}, D_{12} и D_{21} are positive for any mesh steps regardless of the regularization parameters.

From the requirement that the coefficients A_α and B_α for $\alpha = 1, 2$ are non-negative, we get the sufficient condition under which scheme (18) satisfies the maximum principle:

$$\sigma_0 = \begin{cases} \frac{1}{h^2}\max\left(A_\alpha^*, B_\alpha^*\right), \text{ if } \max\limits_\alpha\left(A_\alpha^*, B_\alpha^*\right) > 0, \\ 0, \text{ if } \max\limits_\alpha\left(A_\alpha^*, B_\alpha^*\right) \le 0. \end{cases} \tag{22}$$

Analysis of the coefficients A_α^*, B_α^* shows that they can be positive on coarse meshes. In this case we have $\sigma_0 = O\left(h^{-2}\right)$ and $\nu = \sigma_0 O\left(h^4\right) + O\left(h^4\right) = O\left(h^2\right)$ according to (20), (22). The use of formula (22) may seem unreasonable due to the threat of a decrease in the order of approximation. However, it follows from formulas (21) that $A_\alpha^*, B_\alpha^* \underset{h\to 0}{\to} -10 + 2\left(\frac{h_\alpha}{h_\beta}\right)^2 < 0$ if $\frac{h_\alpha}{h_\beta} < \sqrt{5}$, $\beta \ne \alpha$. Consequently, if the mesh steps are related by the condition

$$\frac{1}{\sqrt{5}} < \frac{h_1}{h_2} < \sqrt{5}, \tag{23}$$

all coefficients A_α^*, B_α^* should become negative for sufficiently small steps h_1, h_2, thereby providing $\sigma_0 = 0$ and therefore the approximation error $\nu = O\left(h^4\right)$. For instance, for the test problem in the following section all coefficients A_α^*, B_α^* become negative on meshes with the step $h \le \frac{1}{53}$ at the Grashof number $\mathrm{Gr} = 10^6$ and on meshes with the step $h \le \frac{1}{135}$ at the Grashof number $\mathrm{Gr} = 10^7$.

Thus, scheme (18) with coefficients (16), (17), (19), (22) subject to constraint (23) satisfies the maximum principle and has the fourth order of approximation. This means that scheme (18), supplemented by difference boundary conditions with the same approximation and stabilization properties, converges with the rate of $O\left(h^4\right)$ as $h \to 0$, i. e. is of fourth order of accuracy.

It should be noted that condition (23) relates the mesh steps but does not limit their values. It agrees with the convergence condition of the high order accuracy scheme for the two-dimensional Poisson equation [3] which corresponds to $k \equiv 1$, $q \equiv 0$, $\upsilon_1 \equiv 0$, $\upsilon_2 \equiv 0$.

Scheme testing

Scheme (18) has been tested on the well-known problem of a natural convection in a horizontal channel of a square cross-section with a uniform heating of the right vertical wall [12; 14; 15]. The problem geometry and the boundary conditions for the temperature are shown in fig. 1.

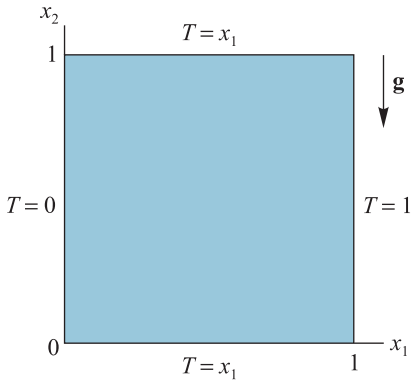


Fig. 1. Illustration of the test problem statement

The dimensionless mathematical model of the test problem is defined by equations (5) with respect to the temperature $T(x_1, x_2)$, the stream function $\psi(x_1, x_2)$ and the vorticity $\omega(x_1, x_2)$ at $\mathrm{Gr}_m = 0$ with the boundary temperature conditions: $T(0, x_2) = 0$, $T(1, x_2) = 1$, $T(x_1, 0) = T(x_1, 1) = x_1$.

The test computations were carried out for the Prandtl number Pr = 1 and the Grashof number in the range $\mathrm{Gr} \le 5 \cdot 10^7$ corresponding to the laminar mode of convection. A square mesh was used with a step $h = h_1 = h_2$ and the number of partitions $10 \le N = \dfrac{1}{h} \le 1000$ on each side of the square domain. Note that the numerical solution for $N = 1000$ requires to solve a system with more than 3 million of nonlinear difference equations. An approximate condition of the fourth order was applied for the vorticity on the boundary [12; 16]. The realization of the difference scheme was carried out by a relaxation method described in [12; 17].

Figures 2 and 3 illustrate the temperature distribution (left) and the flow pattern (right) obtained with scheme (18) on the square mesh with the step $h = \dfrac{1}{500}$ for the Grashof numbers $\mathrm{Gr} = 10^6$ and $\mathrm{Gr} = 5 \cdot 10^7$. The last of them is close to the critical value at which a turbulization of a laminar flow begins. The resulting thermoconvective structures are characterized by a formed boundary layer, in which the dominant velocity and temperature gradients are concentrated. Due to this, an extensive stagnation zone is formed in the central part of the domain with a constant vertical gradient of the temperature $|\nabla T| = \left|\dfrac{\partial T}{\partial x_2}\right| = 0.656$.

Figure 4 and table below show the dependences of the maximum values of the stream function and vorticity on the number of the mesh partition, which are obtained by applying fourth order scheme (18) and the second order monotone Samarskii scheme [3] to the test problem. The upper numbers in the cells of table correspond to the second order scheme, the lower numbers – to the fourth order scheme. The comparison of the simulations results shows that the fourth order scheme has significant advantages in the rate of convergence as $N \to \infty$. For example, the solution, obtained by the fourth order scheme for $N = 100$, is not inferior in accuracy to the solution,

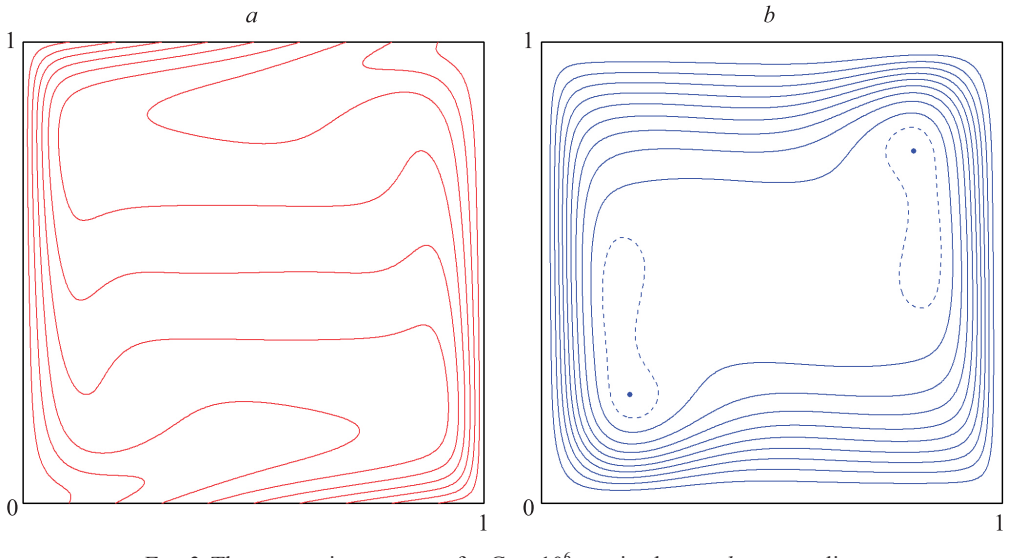


Fig. 2. The convection structure for $\mathrm{Gr} = 10^6$: *a* – isotherms; *b* – streamlines

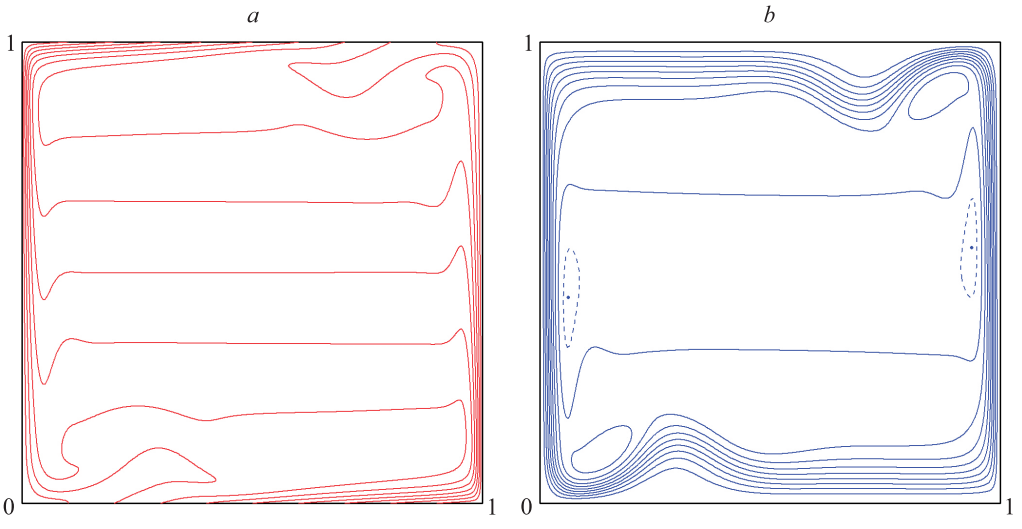


Fig. 3. The convection structure for $\mathrm{Gr} = 5 \cdot 10^7$: *a* – isotherms; *b* – streamlines

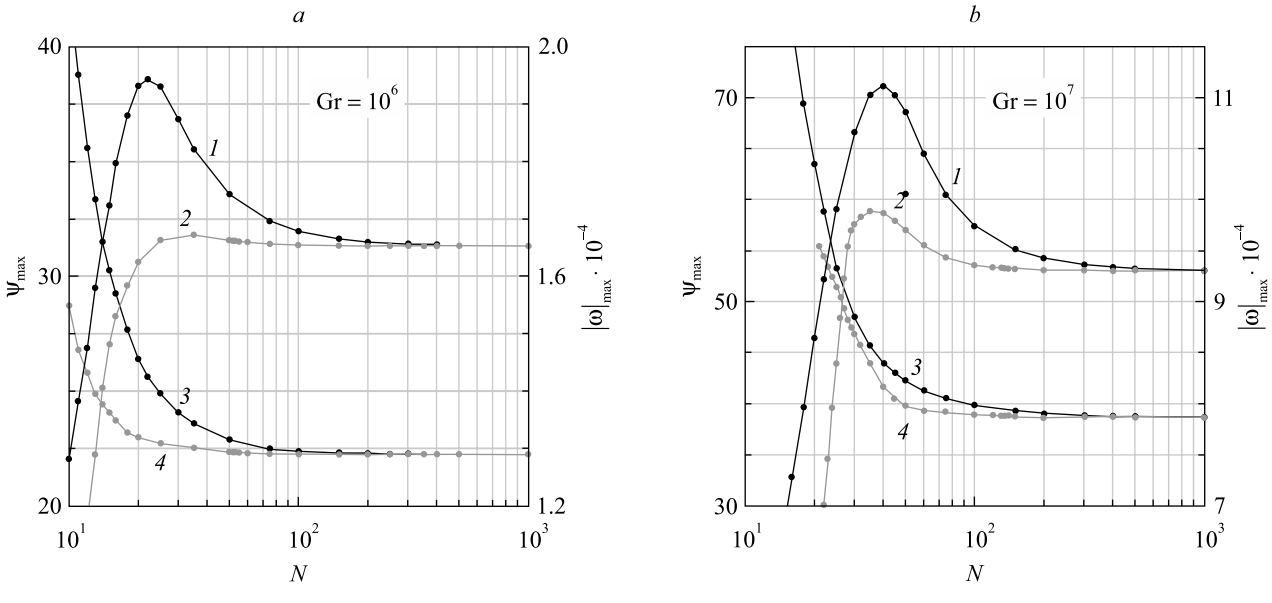


Fig. 4. The dependence of the maximum values of vorticity $|\omega|_{\max}$ (lines *1*, *2*) and stream function $\psi_{\max}$ (lines *3*, *4*) on the mesh number $N = \dfrac{1}{h}$ for $\mathrm{Gr} = 10^6$ (*a*) and $\mathrm{Gr} = 10^7$ (*b*):
1, *3* – the monotone second-order scheme (Samarskii); *2*, *4* – fourth order scheme (18)

obtained by the second order scheme for $N = 300$. It means a nine-fold decrease in size of the system of nonlinear difference equations as well as decrease in the number of iterations to solve this system with the same accuracy. However, the gain in time, expected due to the nine-fold decrease in the number of nodes as well as due to higher convergence rate of iteration process for larger mesh steps, is somewhat compensated by the time difference for the one iteration, which is a 4–5 times higher for scheme (18) than for the scheme of the second order.

Maximum values of the stream function ($\psi_{\max}$) and vorticity $|\omega|_{\max}$ depending on the mesh step (h) for $\mathrm{Gr} = 10^7$

h	$\frac{1}{20}$	$\frac{1}{50}$	$\frac{1}{100}$	$\frac{1}{200}$	$\frac{1}{300}$	$\frac{1}{400}$	$\frac{1}{500}$	$\frac{1}{1000}$
$\psi_{\max}$	63.505 –	42.241 39.840	39.874 38.946	39.072 38.734	38.879 38.701	38.800 38.689	38.760 38.685	38.704 38.683
$\lvert\omega\rvert_{\max}$	86 445.7 –	108 591.1 97 033.3	97 442.5 93 580.4	94 287.3 93 132.4	93 624.6 93 065.3	93 377.2 93 044.7	93 257.3 93 036.5	93 088.9 93 028.7

The test computations show that the constructed scheme of the higher approximation order becomes effective at the Rayleigh numbers Ra = GrPr corresponding to the developed laminar convection. Although the high order scheme significantly complicates a computational algorithm, it could have significant advantages over monotone schemes of the first and the second order [3; 12; 14; 18] for the Rayleigh numbers close to the beginning of a convective flow turbulization because it allows to get numerical solutions with a high accuracy on relatively coarse meshes.

Conclusion

The finite-difference scheme of high order accuracy for the two-dimensional steady-state convection – diffusion equation is constructed. The scheme defined on the minimal stencil of a uniform mesh, has the fourth order of approximation and satisfies the maximum principle for any mesh steps satisfied the condition $\frac{1}{\sqrt{5}} < \frac{h_1}{h_2} < \sqrt{5}$. The scheme is focused on solving a wide range of applied problems of convection – diffusion such as the gravitational convection, a thermomagnetic convection and a diffusion of particles in magnetic fluids. The high approximation and stabilization properties, compared with other methods, provide a higher accuracy with less calculation cost. It is especially important for modeling of convection and diffusion processes in developed boundary layers with the large gradients of velocity, temperature and particle concentration. The proposed scheme is tested on the well-known problem of the high-intensity gravitational convection in the horizontal channel of a square cross-section with the uniform heating from the side. A detailed comparison with the monotone Samarskii scheme of the second order [3] is performed on the sequences of square meshes with the number of partitions from 10 to 1000 on each side of the square domain in the whole range of the Rayleigh numbers $\mathrm{Ra} \le 5 \cdot 10^7$, corresponding to the laminar convection mode. A significant advantage of the fourth order scheme in the convergence rate is shown for the decreasing mesh step.

Библиографические ссылки

1. Beresnev S, Polevikov V, Tobiska L. Numerical study of the influence of diffusion of magnetic particles on equilibrium shapes of a free magnetic fluid surface. *Communications in Nonlinear Science and Simulation.* 2009;14(4):1403–1409. DOI: 10.1016/j.cnsns.2008.04.005.

2. Polevikov V, Tobiska L. Influence of diffusion of magnetic particles on stability of a static magnetic fluid seal under the action of external pressure drop. *Communications in Nonlinear Science and Numerical Simulation.* 2011;16(10):4021–4027. DOI: 10.1016/j.cnsns.2011.02.025.

3. Samarskii AA. *The theory of difference schemes.* New York: Marcel Dekker; 2001. 73 p.

4. Roos H-G, Stynes M, Tobiska L. *Robust numerical methods for singularly perturbed differential equations: convection-diffusion-reaction and flow problems.* Berlin: Springer; 2008. 604 p. (Springer Series in Computational Mathematics; volume 24). DOI: 10.1007/978-3-540-34467-4.

5. Stynes M, Stynes D. *Convection-diffusion problems: an introduction to their analysis and numerical solution.* USA: American Mathematical Society; 2018. 156 p. (Graduate studies in mathematics; volume 196). Co-published by Atlantic Association for Research in the Mathematical Sciences.

6. Самарский АА, Вабищевич ПН. *Численные методы решения задач конвекции-диффузии.* 4-е издание. Москва: Либроком; 2009. 248 с.

7. Лемешевский СВ, Матус ПП, Якубук РМ. Двухслойные разностные схемы повышенного порядка точности для уравнения конвекции-диффузии. *Доклады НАН Беларуси.* 2012;56(2):15–18.

8. Berkovski B, Bashtovoi V, editors. *Magnetic fluids and applications handbook.* New York: Begell House; 1996. 831 p. (UNESCO series of learning materials).

9. Berkovsky BM, Medvedev VF, Krakov MS. *Magnetic fluids: engineering applications.* New York: Oxford University Press; 1993. 243 p.

10. Фертман ВЕ. *Магнитные жидкости – естественная конвекция и теплообмен.* Минск: Наука и техника; 1978. 205 с.

11. Bashtovoi VG, Berkovsky BM, Vislovich AN. *Introduction to thermomechanics of magnetic fluids.* Washington: Hemisphere Publishing; 1988. 228 p.

12. Берковский БМ, Полевиков ВК. *Вычислительный эксперимент в конвекции.* Минск: Университетское; 1988. 167 с.

13. Polevikov V, Tobiska L. On the solution of the steady-state diffusion problem for ferromagnetic particles in a magnetic fluid. *Mathematical Modeling and Analysis.* 2008;13(2):233–240. DOI: 10.3846/1392-6292.2008.13.233-240.

14. Berkovskii BM, Polevikov VK. Effect of the Prandtl number on the convection field and the heat transfer during natural convection. *Journal of Engineering Physics.* 1973;24(5):598–603. DOI: 10.1007/BF00838619.

15. Gershuni GZ, Zhukhovitskii EM, Tarunin EL. Numerical investigation of convective motion in a closed cavity. *Fluid Dynamics.* 1966;1(5):38–42. DOI: 10.1007/BF01022148.

16. Полевиков ВК. Разностная схема четвертого порядка точности для расчета функции вихря на границе в задачах динамики жидкости. *Доклады Академии наук Белорусской ССР.* 1979;23(10):872–875.

17. Polevikov VK. Application of the relaxation method to solve steady difference problems of convection. *USSR Computational Mathematics and Mathematical Physics.* 1981;21(1):126–137. DOI: 10.1016/0041-5553(81)90138-5.

18. Gosman AD, Pun WM, Runchal AK, Spalding DB, Wolfshtein M. *Heat and mass transfer in recirculating flows.* London: Academic Press; 1969. 338 p.

References

1. Beresnev S, Polevikov V, Tobiska L. Numerical study of the influence of diffusion of magnetic particles on equilibrium shapes of a free magnetic fluid surface. *Communications in Nonlinear Science and Simulation.* 2009;14(4):1403–1409. DOI: 10.1016/j.cnsns.2008.04.005.

2. Polevikov V, Tobiska L. Influence of diffusion of magnetic particles on stability of a static magnetic fluid seal under the action of external pressure drop. *Communications in Nonlinear Science and Numerical Simulation.* 2011;16(10):4021–4027. DOI: 10.1016/j.cnsns.2011.02.025.

3. Samarskii AA. *The theory of difference schemes.* New York: Marcel Dekker; 2001. 73 p.

4. Roos H-G, Stynes M, Tobiska L. *Robust numerical methods for singularly perturbed differential equations: convection-diffusion-reaction and flow problems.* Berlin: Springer; 2008. 604 p. (Springer Series in Computational Mathematics; volume 24). DOI: 10.1007/978-3-540-34467-4.

5. Stynes M, Stynes D. *Convection-diffusion problems: an introduction to their analysis and numerical solution.* USA: American Mathematical Society; 2018. 156 p. (Graduate studies in mathematics; volume 196). Co-published by Atlantic Association for Research in the Mathematical Sciences.

6. Samarskii AA, Vabishchevich PN. *Chislennye metody resheniya zadach konvektsii-diffuzii* [Numerical methods for solving convection-diffusion problems]. 4th edition. Moscow: Librokom; 2009. 248 p. Russian.

7. Lemeshevsky SV, Matus PP, Yakubuk RM. Two-layered higher-order difference schemes for the convection-diffusion equation. *Doklady of the National Academy of Sciences of Belarus.* 2012;56(2):15–18. Russian.

8. Berkovski B, Bashtovoi V, editors. *Magnetic fluids and applications handbook.* New York: Begell House; 1996. 831 p. (UNESCO series of learning materials).

9. Berkovsky BM, Medvedev VF, Krakov MS. *Magnetic fluids: engineering applications.* New York: Oxford University Press; 1993. 243 p.

10. Fertman VE. *Magnitnye zhidkosti – estestvennaya konvektsiya i teploobmen* [Magnetic fluids: natural convection and heat transfer]. Minsk: Nauka i tekhnika; 1978. 205 p. Russian.

11. Bashtovoi VG, Berkovsky BM, Vislovich AN. *Introduction to thermomechanics of magnetic fluids.* Washington: Hemisphere Publishing; 1988. 228 p.

12. Berkovsky BM, Polevikov VK. *Vychislitel'nyi eksperiment v konvektsii* [Computational experiment in convection]. Minsk: Universitetskoe; 1988. 167 p. Russian.

13. Polevikov V, Tobiska L. On the solution of the steady-state diffusion problem for ferromagnetic particles in a magnetic fluid. *Mathematical Modeling and Analysis.* 2008;13(2):233–240. DOI: 10.3846/1392-6292.2008.13.233-240.

14. Berkovskii BM, Polevikov VK. Effect of the Prandtl number on the convection field and the heat transfer during natural convection. *Journal of Engineering Physics.* 1973;24(5):598–603. DOI: 10.1007/BF00838619.

15. Gershuni GZ, Zhukhovitskii EM, Tarunin EL. Numerical investigation of convective motion in a closed cavity. *Fluid Dynamics.* 1966;1(5):38–42. DOI: 10.1007/BF01022148.

16. Polevikov VK. [The difference scheme of fourth order accuracy to compute the boundary vorticity in hydrodynamics problems]. *Doklady Akademii nauk Belorusskoi SSR.* 1979;23(10):872–875. Russian.

17. Polevikov VK. Application of the relaxation method to solve steady difference problems of convection. *USSR Computational Mathematics and Mathematical Physics.* 1981;21(1):126–137. DOI: 10.1016/0041-5553(81)90138-5.

18. Gosman AD, Pun WM, Runchal AK, Spalding DB, Wolfshtein M. *Heat and mass transfer in recirculating flows.* London: Academic Press; 1969. 338 p.

Received by editorial board 03.07.2019.

Дискретная математика и математическая кибернетика

Discrete mathematics and mathematical cybernetics

УДК 519.67

ОБОБЩЕННЫЙ БЛОЧНЫЙ АЛГОРИТМ ФЛОЙДА – УОРШЕЛЛА

Н. А. ЛИХОДЕД[1)]***, Д. С. СИПЕЙКО***[1)]

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*

Одним из наиболее используемых на практике алгоритмов для поиска кратчайших путей между всеми парами вершин во взвешенных графах является алгоритм Флойда – Уоршелла. Блочная версия алгоритма служит основой для получения эффективных параллельных алгоритмов при реализации на многоядерных центральных процессорах, компьютерах с распределенной памятью, графических процессорах. Увеличение зернистости вычислений в блочных версиях алгоритмов приводит к более эффективному использованию кешей и более эффективной организации параллельных вычислений. В этой работе предложено обобщение блочного алгоритма Флойда – Уоршелла. Порядок выполнения блоков вычислений реорганизован таким образом, чтобы элементы массива, участвующие в коммуникационных операциях как чтения, так и записи, реже вытеснялись из памяти с быстрым доступом. Тогда при реализации алгоритма на графическом процессоре реже, по сравнению с исходным блочным алгоритмом, используется медленная глобальная память.

Ключевые слова: параллельные алгоритмы; поиск кратчайших путей; графы; алгоритм Флойда – Уоршелла; блочный алгоритм; графический процессор; GPU.

Благодарность. Работа выполнена в рамках государственной программы научных исследований Республики Беларусь «Конвергенция-2020» (подпрограмма «Методы математического моделирования сложных систем»).

Образец цитирования:
Лиходед НА, Сипейко ДС. Обобщенный блочный алгоритм Флойда – Уоршелла. *Журнал Белорусского государственного университета. Математика. Информатика.* 2019;3:84–92.
https://doi.org/10.33581/2520-6508-2019-3-84-92

For citation:
Likhoded NA, Sipeyko DS. Generalized blocked Floyd – Warshall algorithm. *Journal of the Belarusian State University. Mathematics and Informatics.* 2019;3:84–92. Russian.
https://doi.org/10.33581/2520-6508-2019-3-84-92

Авторы:
Николай Александрович Лиходед – доктор физико-математических наук, профессор; профессор кафедры вычислительной математики факультета прикладной математики и информатики.
Дмитрий Сергеевич Сипейко – магистрант факультета прикладной математики и информатики. Научный руководитель – Н. А. Лиходед.

Authors:
Nikolai A. Likhoded, doctor of science (physics and mathematics), full professor; professor at the department of computational mathematics, faculty of applied mathematics and computer science.
likhoded@bsu.by
Dmitry S. Sipeyko, master's degree student at the faculty of applied mathematics and computer science.
mintaids@ya.ru

GENERALIZED BLOCKED FLOYD – WARSHALL ALGORITHM

***N. A. LIKHODED*[a], *D. S. SIPEYKO*[a]**

[a]*Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*
Corresponding author: N. A. Likhoded (likhoded@bsu.by)

One of the most commonly used on practice all-pairs shortest paths algorithms on weighted graphs is Floyd – Warshall algorithm. Blocked version serves as a basis for obtaining effective parallel algorithms to be implemented on multi-core central processing units, on computers with distributed memory, on graphics processing units (GPU). Increasing computation granularity in blocked versions of algorithm leads to a more efficient usage of caches and more efficient organization of parallel computations. In this paper we introduce generalization of blocked Floyd – Warshall algorithm. Computing blocks execution order was reorganized in such a way that array elements which participate in communication operations, both reading and writing, are pushed out of fast-access memory less often. This means that in GPU implementation slow global memory is used less often, compared with the original blocked algorithm.

Keywords: parallel algorithms; shortest paths; graphs; Floyd – Warshall algorithm; block algorithm; GPU.

Acknowledgements. The prepared report was sponsored by the government program of scientific research of the Republic of Belarus «Convergence-2020» (subprogram «Methods of mathematical modeling of complex systems»).

Введение

Поиск в графах играет важную роль при анализе больших наборов данных. Широкое применение в задачах маршрутизации и логистики имеют алгоритмы поиска кратчайших путей. Одним из наиболее используемых на практике алгоритмов для поиска кратчайших путей между всеми парами вершин во взвешенных графах является алгоритм Флойда – Уоршелла.

Все алгоритмы решения задачи поиска кратчайших путей между всеми парами вершин графа имеют нелинейную трудоемкость, поэтому для решения задач большого размера требуются методы, хорошо приспособленные для реализации на современных вычислительных устройствах. Среди них – блочный алгоритм Флойда – Уоршелла [1; 2]. Он служит основой для получения эффективных параллельных версий алгоритма для реализации на многоядерных центральных процессорах, компьютерах с распределенной памятью, графических процессорах (GPU) [3–6]. Как и во многих других случаях применения блочных алгоритмов, увеличение зерна вычислений приводит к более эффективному использованию кешей и более эффективной организации параллельных вычислений.

Цель работы – построение модификации блочного алгоритма Флойда – Уоршелла, в которой порядок выполнения блоков вычислений реорганизован таким образом, чтобы реже вытеснялись из памяти с быстрым доступом (кеши, регистры) элементы массива, участвующие не только в коммуникационных операциях чтения, но и в операциях записи. Можно ожидать, что тогда при реализации алгоритма будет более эффективно, по сравнению с исходным блочным алгоритмом, использоваться память с быстрым доступом. Предлагаемая модификация задает параметрическое семейство блочных алгоритмов, которое включает в себя и исходный алгоритм.

Степень использования памяти с быстрым доступом отражает вычислительное свойство метода, называемое локальностью, которая при реализации алгоритмов на многопроцессорных вычислительных устройствах играет важнейшую роль в достижении высокой производительности [7–9]. В данной работе построенный обобщенный алгоритм Флойда – Уоршелла реализован на графическом процессоре, при вычислениях на котором быстрым является процесс обращения к регистрам, разделяемой памяти мультипроцессора и кешам, но не обращение к глобальной памяти GPU. Реализация на основе обобщенного алгоритма приводит к сокращению числа обращений к глобальной памяти и, как показали вычислительные эксперименты, к уменьшению времени выполнения.

Точечный алгоритм. Зависимости алгоритма

Пусть $G(V, E)$ – некоторый граф, V – множество вершин, E – множество ребер. Будем считать, что вершины графа занумерованы последовательными целыми числами от 1 до n, а граф задан матрицей смежности A размером $n \times n$.

Приведем основную часть последовательного точечного (т. е. не блочного) алгоритма Флойда – Уоршелла:

do $k = 1, n$
  do $i = 1, n$
    do $j = 1, n$
      $S_1(k, i, j)$: $a(i, j) = \min\left(a(i, j), a(i, k) + a(k, j)\right)$
    enddo
  enddo
enddo

Перед началом выполнения алгоритма матрица расстояний A заполняется длинами ребер графа (или заведомо большим числом, если ребра нет). На каждом шаге k матрица A обновляется. По завершении работы алгоритма матрица A содержит длины кратчайших путей между всеми вершинами графа.

В гнезде циклов имеется один выполняемый оператор S_1 и используется один массив a размерности 2. Область изменения параметров циклов (область итераций) для оператора S_1 имеет размерность 3:

$$V_1 = \left\{(k, i, j) \in \mathbb{Z}^3 \,|\, 1 \le k \le n, 1 \le i \le n, 1 \le j \le n\right\}.$$

Формально между операциями одного слоя k существуют информационные зависимости, которые связаны с обновлением или использованием элементов массивов $a(i, j)$ при $i = k$ или $j = k$. Устранить указанные зависимости можно введением дополнительных массивов [5]. Но непосредственно из записи алгоритма Флойда – Уоршелла следует, что данные $a(i, j)$ не обновляются, если $i = k$ или $j = k$. Поэтому для фиксированного k все операции фактически используют данные, вычисленные на предыдущем $(k-1)$-м шаге. Можно допустить, что все операции при фиксированном k не зависят друг от друга и возможен произвольный порядок их выполнения.

Рассмотрим зависимости и векторы зависимостей алгоритма с учетом сделанного допущения. Векторы зависимостей будем для наглядности помечать элементами матрицы, фигурирующими на порождающих зависимости вхождениях. Например, вектор $d^{a(i,j),a(i,k)}$ порождается зависимостью между данными $a(i, j)$ в левой части оператора S_1 и данными $a(i, k)$ в его правой части. Укажем итерации, порождающие зависимости, и векторы зависимостей:

• $S_1(k-1, i, j) \to S_1(k, i, j)$: данное $a(i, j)$, вычисленное на итерации $(k-1, i, j)$, является аргументом $a(i, j)$ для вычислений на итерации (k, i, j); $d^{a(i,j),a(i,j)} = (1, 0, 0)$;

• $S_1(k-1, i, k) \to S_1(k, i, j)$: $a(i, k)$, вычисленное на итерации $(k-1, i, k)$, является аргументом для вычислений на итерациях (k, i, j); $d^{a(i,j),a(i,k)} = (1, 0, j-k)$;

• $S_1(k-1, k, j) \to S_1(k, i, j)$: $a(k, j)$, вычисленное на итерации $(k-1, k, j)$, является аргументом для вычислений на итерациях (k, i, j); $d^{a(i,j),a(k,j)} = (1, i-k, 0)$.

Блочный алгоритм

Блочный алгоритм Флойда – Уоршелла с трехмерными (3D) блоками впервые предложен в работе [1]. Выделим $Q \times Q \times Q$ блоков размером $r \times r \times r$, где $Q = \left\lceil \frac{n}{r} \right\rceil$, r – параметр, задающий размер. Отметим, что одинаковые размеры блока имеют существенное значение, а не выбраны для простоты.

Пусть k^{gl}, i^{gl}, j^{gl} – номера частей, на которые при формировании блоков разбиваются области значений параметров k, i, j циклов, $0 \le k^{gl}, i^{gl}, j^{gl} \le Q - 1$. Блоки вычислений называют также тайлами. Блок $\mathrm{Tile}\left(k^{gl}, i^{gl}, j^{gl}\right)$ имеет следующий вид:

do $k = 1 + k^{gl}r, \min\left(\left(k^{gl}+1\right)r, n\right)$
  do $i = 1 + i^{gl}r, \min\left(\left(i^{gl}+1\right)r, n\right)$
    do $j = 1 + j^{gl}r, \min\left(\left(j^{gl}+1\right)r, n\right)$
      $a(i, j) = \min\left(a(i, j), a(i, k) + a(k, j)\right)$
    enddo
  enddo
enddo

Установим корректный порядок выполнения блоков и обоснуем корректный порядок выполнения вычислений в блоке.

Рассмотрим блоки некоторой блочной итерации k^{gl} (некоторого блочного слоя k^{gl}), т. е. блоки $\text{Tile}\left(k^{gl},\ i^{gl},\ j^{gl}\right)$ при фиксированном k^{gl}. Назовем:

• $\text{Tile}\left(k^{gl},\ k^{gl},\ k^{gl}\right)$ – ведущим блоком;

• $\text{Tile}\left(k^{gl},\ k^{gl},\ j^{gl}\right)$, $0 \le j^{gl} \le Q-1$, $j^{gl} \ne k^{gl}$, – блоком ведущей строки;

• $\text{Tile}\left(k^{gl},\ i^{gl},\ k^{gl}\right)$, $0 \le i^{gl} \le Q-1$, $i^{gl} \ne k^{gl}$, – блоком ведущего столбца.

Анализ зависимостей показывает следующее.

1. Вычисления любого ведущего блока $\text{Tile}\left(k^{gl},\ k^{gl},\ k^{gl}\right)$ не зависят от вычислений других блоков блочного слоя, для вычисления элементов ведущего блока нужны только его элементы. Ведущий блок на блочном слое следует вычислять первым. Ведущие блоки назовем I-блоками (independent blocks [4]).

2. Вычисления блоков ведущей строки и ведущего столбца зависят от вычислений ведущего блока блочного слоя. Для вычислений этих блоков необходимы их собственные элементы и уже подсчитанные элементы ведущего блока. Между собой эти блоки не конкурируют, поэтому последовательность их вычислений на блочном слое может быть произвольной. Назовем блоки ведущих строк и столбцов SD-блоками (singly dependent blocks).

3. Остальные блоки $\text{Tile}\left(k^{gl},\ i^{gl},\ j^{gl}\right)$, $0 \le i^{gl}, j^{gl} \le Q-1$, $i^{gl} \ne k^{gl}$, $j^{gl} \ne k^{gl}$, зависят от вычислений блоков ведущих строк и столбцов. Для вычислений этих блоков нужны их собственные элементы, а также элементы соответствующих блоков ведущей строки и ведущего столбца. Указанные блоки вычисляются на блочном слое в произвольном порядке после вычисления блоков ведущей строки и столбца. Блоки вне ведущих строк и столбцов назовем DD-блоками (doubly dependent blocks).

Опишем шаги блочной итерации k^{gl}:

1) производятся вычисления ведущего блока (I-блока). Фактически выполняется обычный точечный алгоритм Флойда – Уоршелла, в итоге сохраняется версия элементов подматрицы ведущего блока на итерации $\left(k^{gl}+1\right)r$ (или $\min\left(\left(k^{gl}+1\right)r,\ n\right)$ для последнего слоя);

2) производятся вычисления блоков ведущей строки и ведущего столбца (SD-блоков). При обращении к элементам ведущего блока происходит обращение к их последней версии. Блоки могут вычисляться независимо, в произвольном порядке;

3) производятся вычисления оставшихся блоков (DD-блоков). При обращении к элементам блоков ведущей строки и ведущего столбца происходит обращение к их последней версии. Блоки могут вычисляться независимо, в произвольном порядке.

Замечание 1. Результаты промежуточных вычислений точечного и блочного алгоритмов Флойда – Уоршелла могут не совпадать. Тем не менее блочный алгоритм Флойда – Уоршелла приводит к корректному результату [2].

Основная часть алгоритма Флойда – Уоршелла с выделенными 3D-блоками имеет следующий вид [4] (циклы, итерации которых заведомо можно выполнять независимо, запишем как dopar):

```
do k^gl = 0, Q – 1
   Tile(k^gl, k^gl, k^gl) // вычисления I-блока
   dopar j^gl = 0, Q – 1 (j^gl ≠ k^gl)
      Tile(k^gl, k^gl, j^gl) // вычисления SD-блоков ведущей строки
   enddopar
   dopar i^gl = 0, Q – 1 (i^gl ≠ k^gl)
      Tile(k^gl, i^gl, k^gl) // вычисления SD-блоков ведущего столбца
   enddopar
   dopar i^gl = 0, Q – 1 (i^gl ≠ k^gl)
      dopar j^gl = 0, Q – 1 (j^gl ≠ k^gl)
         Tile(k^gl, i^gl, j^gl) // вычисления DD-блоков
      enddopar
   enddopar
enddo (k^gl)
```

Замечание 2. В DD-блоках вычисления всех $a(i, j)$ происходят независимо друг от друга: $a(i, k)$ и $a(k, j)$ вычисляются вне DD-блока, между операциями блока остаются только зависимости, задаваемые вектором $d^{a(i,j),\,a(i,j)} = (1, 0, 0)$. На практике в DD-блоках производят перестановку циклов так, чтобы цикл с параметром k стал самым внутренним. Обозначим такой блок через $\mathrm{Tile}_{\mathrm{DD}}\left(k^{gl}, i^{gl}, j^{gl}\right)$ и запишем его явный вид:

dopar $i = 1 + i^{gl}r,\ \min\left(\left(i^{gl}+1\right)r,\ n\right)$
  dopar $j = 1 + j^{gl}r,\ \min\left(\left(j^{gl}+1\right)r,\ n\right)$
    do $k = 1 + k^{gl}r,\ \min\left(\left(k^{gl}+1\right)r,\ n\right)$
      $a(i, j) = \min\left(a(i, j),\ a(i, k) + a(k, j)\right)$
    enddo
  enddopar
enddopar

Модифицированный блочный алгоритм

В рассмотренном блочном алгоритме последовательно выполняются блочные итерации k^{gl}. Для каждого фиксированного k^{gl} сначала производятся вычисления I-блока $\mathrm{Tile}\left(k^{gl}, k^{gl}, k^{gl}\right)$, затем (в произвольном порядке) – вычисления $2(Q-1)$ SD-блоков $\mathrm{Tile}\left(k^{gl}, k^{gl}, j^{gl}\right)$ и $\mathrm{Tile}\left(k^{gl}, i^{gl}, k^{gl}\right)$, далее (в произвольном порядке) – вычисления $(Q-1)\times(Q-1)$ DD-блоков $\mathrm{Tile}\left(k^{gl}, i^{gl}, j^{gl}\right)$. Как уже отмечалось, размеры блока ($r \times r \times r$ итераций) могут быть только одинаковые.

Реорганизуем порядок выполнения блоков вычислений таким образом, чтобы атомарно, как одна макрооперация, выполнялось некоторое количество тайлов с идущими подряд номерами первой блочной координаты и фиксированными номерами второй и третьей координаты. Такие объединенные тайлы будем называть мультитайлами (вообще говоря, мультитайл может содержать только один тайл). В макрооперациях-мультитайлах выполнение большего числа идущих подряд итераций k приводит к более редкому вытеснению из памяти с быстрым доступом (кеши, регистры) элементов массива на вхождениях $a(i, j)$, требующих как операций чтения, так и операций записи, в то время как вхождения $a(i, k)$ и $a(k, j)$ требуют только операций чтения. Можно ожидать, что при реализации алгоритма будет более эффективно использоваться память с быстрым доступом.

Пусть к – некоторое число в пределах от 1 до Q, l – некоторое число в пределах от 0 до к – 1. Параметр к задает количество блочных итераций k^{gl} (число блочных слоев), используемых для образования мультитайлов. От параметра l зависит число блочных итераций k^{gl} (число блочных слоев), которые объединяются для получения мультитайлов.

Введем в рассмотрение процедуры выполнения мультитайлов.

$\mathrm{CalcLeadBlock}\left(k^{gl}, l\right)$ – процедура вычисления (при фиксированных параметрах k^{gl}, l процедуры) мультитайла

$$\mathrm{Tile}\left(k^{gl}, k^{gl}, k^{gl}\right), \text{ если } l = 0,$$

$$\bigcup_{m=0}^{l-1} \mathrm{Tile}_{\mathrm{DD}}\left(k^{gl}+m, k^{gl}+l, k^{gl}+l\right) \bigcup \mathrm{Tile}\left(k^{gl}+l, k^{gl}+l, k^{gl}+l\right), \text{ если } l \neq 0,$$

объединяющего l DD-блоков и один I-блок исходного блочного алгоритма. Мультитайл включает I-блок и невычисленные DD-блоки (если таковые имеются, т. е. если $l > 0$) с такими же, как у I-блока, номерами второй и третьей блочной координаты и с меньшими номерами первой блочной координаты.

$\mathrm{CalcLeadRowAndColumn}\left(k^{gl}, l\right)$ – процедура вычисления $Q - 1$ мультитайлов, каждый из которых включает один SD-блок $(k^{gl} + l)$-й блочной строки исходного блочного алгоритма, и $Q - 1$ мультитайлов, каждый из которых включает один SD-блок $\left(k^{gl} + l\right)$-го блочного столбца исходного блочного алгоритма. Кроме того, каждый мультитайл включает DD-блоки (если таковые имеются) с такими же, как у SD-блока, номерами второй и третьей блочной координаты и с меньшими номерами первой блочной координаты.

Мультитайл с SD-блоком $\left(k^{gl}+l\right)$-й блочной строки имеет вид

$$\mathrm{Tile}\left(k^{gl},\ k^{gl},\ j^{gl}\right),$$

$$\text{если } l=0,\ j^{gl}=0,\ 1,\ \ldots,\ Q-1\ \left(j^{gl}\neq k^{gl}\right),$$

$$\bigcup_{m=0}^{l-1}\mathrm{Tile}_{\mathrm{DD}}\left(k^{gl}+m,\ k^{gl}+l,\ j^{gl}\right)\bigcup\mathrm{Tile}\left(k^{gl}+l,\ k^{gl}+l,\ j^{gl}\right),$$

$$\text{если } l\neq 0,\ j^{gl}=0,\ 1,\ \ldots,\ Q-1\ \left(j^{gl}\neq k^{gl},\ \ldots,\ k^{gl}+l\right),$$

$$\bigcup_{m=j^{gl}-k^{gl}+1}^{l-1}\mathrm{Tile}_{\mathrm{DD}}\left(k^{gl}+m,\ k^{gl}+l,\ j^{gl}\right)\bigcup\mathrm{Tile}\left(k^{gl}+l,\ k^{gl}+l,\ j^{gl}\right),$$

$$\text{если } l\neq 0,\ j^{gl}=k^{gl},\ \ldots,\ k^{gl}+l-1.$$

Мультитайл с SD-блоком $\left(k^{gl}+l\right)$-го блочного столбца представляется в виде

$$\mathrm{Tile}\left(k^{gl},\ i^{gl},\ k^{gl}\right),$$

$$\text{если } l=0,\ i^{gl}=0,\ 1,\ \ldots,\ Q-1\ \left(i^{gl}\neq k^{gl}\right),$$

$$\bigcup_{m=0}^{l-1}\mathrm{Tile}_{\mathrm{DD}}\left(k^{gl}+m,\ i^{gl},\ k^{gl}+l\right)\bigcup\mathrm{Tile}\left(k^{gl}+l,\ i^{gl},\ k^{gl}+l\right),$$

$$\text{если } l\neq 0,\ i^{gl}=0,\ 1,\ \ldots,\ Q-1\ \left(i^{gl}\neq k^{gl},\ \ldots,\ k^{gl}+l\right),$$

$$\bigcup_{m=i^{gl}-k^{gl}+1}^{l-1}\mathrm{Tile}_{\mathrm{DD}}\left(k^{gl}+m,\ i^{gl},\ k^{gl}+l\right)\bigcup\mathrm{Tile}\left(k^{gl}+l,\ i^{gl},\ k^{gl}+l\right),$$

$$\text{если } l\neq 0,\ i^{gl}=k^{gl},\ \ldots,\ k^{gl}+l-1.$$

$\mathrm{CalcLeadRowAndColumnReverse}\left(k^{gl},\ \kappa,\ l\right)$ – процедура вычисления (при фиксированных параметрах k^{gl}, κ, l, здесь $l<\kappa-1$) мультитайлов $\left(k^{gl}+l\right)$-й блочной строки исходного блочного алгоритма и $\left(k^{gl}+l\right)$-го блочного столбца исходного блочного алгоритма. Каждый мультитайл объединяет $\kappa-l-1$ DD-блоков с фиксированными номерами второй и третьей блочной координаты и с большими, чем $k^{gl}+l$, номерами первой блочной координаты:

$$\bigcup_{m=l+1}^{\kappa-1}\mathrm{Tile}_{\mathrm{DD}}\left(k^{gl}+m,\ k^{gl}+l,\ j^{gl}\right),$$

$$j^{gl}=0,\ 1,\ \ldots,\ k^{gl}+l,\ j^{gl}=k^{gl}+\kappa,\ \ldots,\ Q-1,$$

$$\bigcup_{m=l+1}^{\kappa-1}\mathrm{Tile}_{\mathrm{DD}}\left(k^{gl}+m,\ i^{gl},\ k^{gl}+l\right),$$

$$i^{gl}=0,\ 1,\ \ldots,\ k^{gl}+l-1,\ i^{gl}=k^{gl}+\kappa,\ \ldots,\ Q-1.$$

$\mathrm{CalcRestBlocks}\left(k^{gl},\ \kappa\right)$, $\kappa\neq Q$, есть процедура вычисления $\left(Q-\kappa\right)\times\left(Q-\kappa\right)$ мультитайлов

$$\bigcup_{m=0}^{\kappa-1}\mathrm{Tile}_{\mathrm{DD}}\left(k^{gl}+m,\ i^{gl},\ j^{gl}\right),$$

$$i^{gl}=0,\ 1,\ \ldots,\ Q-1\ \left(i^{gl}\neq k^{gl},\ k^{gl}+1,\ \ldots,\ k^{gl}+\kappa-1\right),$$

$$j^{gl}=0,\ 1,\ \ldots,\ Q-1\ \left(j^{gl}\neq k^{gl},\ k^{gl}+1,\ \ldots,\ k^{gl}+\kappa-1\right),$$

каждый из которых объединяет κ DD-блоков (вне ведущих блочных строк и столбцов) исходного блочного алгоритма.

Отметим, что во всех процедурах вычисление мультитайлов при фиксированных k^{gl}, κ, l можно выполнять независимо (кроме процедуры $\mathrm{CalcLeadBlock}\left(k^{gl},\ l\right)$, вычисляющей только один мультитайл).

Основную часть обобщенного блочного алгоритма Флойда – Уоршелла можно представить следующим образом:

do $k^{gl} = 0, Q - 1, \kappa$ // вычисления с шагом κ
do $l = 0, \kappa - 1$
CalcLeadBlock$\left(k^{gl}, l\right)$
CalcLeadRowAndColumn$\left(k^{gl}, l\right)$
enddo
do $l = \kappa - 2, 0, -1$ // вычисления с шагом –1
CalcLeadRowAndColumnReverse$\left(k^{gl}, \kappa, l\right)$
enddo
CalcRestBlocks$\left(k^{gl}, \kappa\right)$
enddo $\left(k^{gl}\right)$

Если $\kappa = 1$, то получим известный блочный алгоритм Флойда – Уоршелла (цикл do $l = \kappa - 2, 0, -1$ отсутствует).

Если $\kappa = 2$, то имеем случай, рассмотренный в магистерской диссертации О. И. Сычевой[1].

Если $\kappa = Q$, то внешний цикл do $k^{gl} = 0, Q - 1, \kappa$ вырождается в одну итерацию $k^{gl} = 0$, циклы с параметром l охватывают вычисление всех блоков, функция CalcRestBlocks$\left(k^{gl}, \kappa\right)$ отсутствует.

Реализация на графическом процессоре

Графический процессор осуществляет множество параллельных потоков вычислений. Потоки объединяются в блоки вычислений, каждый блок потоков выполняется атомарно на одном из мультипроцессоров графического процессора. При этом должны быть указаны блоки, которые могут выполняться мультипроцессорами одновременно и независимо друг от друга.

Точечный алгоритм Флойда – Уоршелла обладает естественным параллелизмом в пределах одной итерации. Поэтому на каждой итерации k ($k = 1, 2, \ldots, n$) можно выделить двумерные (2D) блоки вычислений, которые могут выполняться независимо друг от друга. GPU-реализация алгоритма Флойда – Уоршелла, основанная на 2D-блоках вычислений, предложена в работе [10]. Матрица A хранится в глобальной памяти GPU, поэтому на каждой итерации k необходима запись всех обновленных элементов матрицы в глобальную память, из которой считываются все нужные данные, подсчитанные на предыдущей итерации.

При вычислениях на GPU быстрым является процесс обращения к разделяемой памяти мультипроцессора и к кешам, но не обращение к глобальной памяти GPU. В работе [11] реализован на GPU алгоритм Флойда – Уоршелла с 3D-блоками. Использование блочного алгоритма с 3D-блоками позволило существенно уменьшить время выполнения алгоритма. Оно сократилось главным образом за счет того, что запись обновленных элементов матрицы в глобальную память производится не на каждой итерации k, а на каждой r-й итерации k. На каждой блочной итерации k^{gl} требуются три так называемых запуска ядра (т. е. три процедуры выполнения блоков вычислений):

• запускаются вычисления ведущего блока (I-блока). Используется $r \times r$ потоков – один поток вычисляет один элемент матрицы. Все потоки в одном блоке можно запустить, если $r \le 32$ (в одном блоке может быть до 1024 потоков). Для каждого потока нужно один элемент скопировать из глобальной памяти в разделяемую, обновить его на r слоях и вернуть новое значение в глобальную память;

• запускаются вычисления блоков ведущей строки и ведущего столбца (SD-блоков), в каждом блоке используется $r \times r$ ($r \le 32$) потоков. Напомним, что для вычислений этих блоков необходимы их собственные элементы и уже подсчитанные элементы ведущего блока. Для каждого из запускаемых блоков в разделяемой памяти хранятся две матрицы размером $r \times r$: помимо своих элементов потокам суммарно требуются $r \times r$ элементов I-блока;

• запускаются вычисления DD-блоков, в каждом блоке используется $r \times r$ потоков ($r \le 32$). Для каждого из запускаемых блоков в разделяемой памяти хранятся три матрицы размером $r \times r$: помимо своих элементов потокам суммарно требуются $r \times r$ элементов блока ведущей строки и $r \times r$ элементов блока ведущего столбца.

[1] *Сычева О. И.* Разработка и программная реализация новых параллельных версий алгоритма Флойда – Уоршелла : магист. дис. Минск : БГУ, 2016. 58 с.

Замечание 3. При вычислении DD-блоков в разделяемой памяти достаточно хранить только матрицы, связанные с вхождениями $a(i, k)$ и $a(k, j)$: потокам не понадобится обращаться к элементам $a(i, j)$, которые пересчитывают другие потоки, – нужен будет только свой элемент и элементы двух SD-блоков. Каждый поток может хранить элемент, который он пересчитывает, в своей регистровой памяти.

Дальнейшая оптимизация алгоритма путем уменьшения объема занимаемой разделяемой памяти в блочном алгоритме Флойда – Уоршелла предложена в работе [5]. Главная идея подхода – многостадийное чтение SD-блоков при вычислении DD-блоков. Сокращение размеров разделяемой памяти, необходимой блоку в каждый момент времени, обусловливает заметный выигрыш в производительности, так как уменьшение объема разделяемой памяти, используемой в блоке, позволяет мультипроцессору выполнять большее количество блоков одновременно.

С помощью построенного обобщенного алгоритма Флойда – Уоршелла можно реализовать на графическом процессоре 3D-блоки размером $(l+1)r \times r \times r$, где, напомним, l изменяется от 0 до к – 1. Запись обновленных элементов матрицы в глобальную память производится на $(l+1)r$-й итерации k, а не r-й итерации, как в случаях блоков размером $r \times r \times r$.

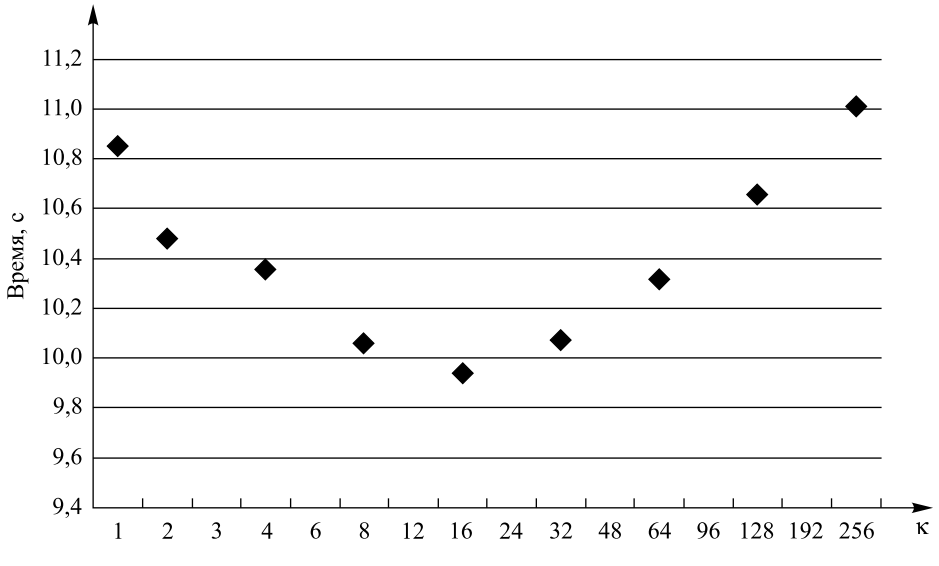
Опишем вычислительные эксперименты, в которых использовалось многостадийное чтение SD-блоков, число потоков в одном блоке 1024 ($r \times r = 32 \times 32$).

Эксперименты проводились на графическом процессоре NVIDIA GeForce GTX 670. Некоторые характеристики этого GPU:

Число мультипроцессоров	7
Количество ядер в устройстве	1344
Объем глобальной памяти	2 Гб
Объем разделяемой памяти	48 Кб на каждый мультипроцессор
Количество 32-битных регистров в мультипроцессоре	65 536
Используемая архитектура	Kepler

На рисунке представлен график зависимости времени вычислений реализации алгоритма от параметра к, влияющего на число записей в глобальную память.

Рисунок показывает, что при нескольких значениях параметра к обобщенного блочного алгоритма достигается меньшее время вычислений по сравнению со случаем к = 1 (классический блочный алгоритм). В этом примере число вершин графа равно 8192. Аналогичная картина наблюдается и в примерах с другим числом вершин.



Зависимость времени вычислений реализации обобщенного блочного алгоритма ($n = 8192$, $r = 32$) от параметра к

Dependency of computation time of generalized blocked algorithm implementation ($n = 8192$, $r = 32$) on the parameter к

Замечание 4. При обращении к функции CalcLeadBlock выполняется только один мультитайл, поэтому активен только один (из нескольких) мультипроцессор GPU. Можно организовать вычисление функции CalcLeadBlock одновременно на всех мультипроцессорах. Эксперименты показали, что общее время реализации алгоритма при этом уменьшается очень незначительно, так как почти все временные затраты приходятся на вычисление других функций.

Таким образом, в работе построено параметрическое семейство блочных алгоритмов Флойда – Уоршелла, которое включает в себя и классический блочный алгоритм, рассмотрена реализация предложенного алгоритма на графическом процессоре.

Библиографические ссылки

1. Venkataraman G, Sahni S, Mukhopadhyaya S. A blocked all-pairs shortest-paths algorithm. *Journal of Experimental Algorithmics*. 2003;8:857–874. DOI: 10.1145/996546.996553.

2. Park J, Penner M, Prasanna VK. Optimizing graph algorithms for improved cache performance. *IEEE Transactions on Parallel and Distributed Systems*. 2004;15(9):769–782. DOI: 10.1109/TPDS.2004.44.

3. Srinivasan T, Balakrishnan R, Gangadharan SA, Hayawardh V. A scalable parallelization of all-pairs shortest path algorithm for a high performance cluster environment. In: *Proceedings of the 13th International Conference on Parallel and Distributed Systems; 2007 December 5–7; Hsinchu, Taiwan.* Washington: IEEE Computer Society; 2007. p. 1–8. DOI: 10.1109/ICPADS.2007.4447721.

4. Lund BD, Smith JW. A multi-stage CUDA kernel for Floyd – Warshall. arXiv:1001.4108. 2010 [Preprint]. 2010 [cited 2019 June 3]. Available from: https://arxiv.org/abs/1001.4108.

5. Mullapudi RT, Bondhugula U. Tiling for dynamic scheduling. In: *IMPACT 2014. Proceedings of the 4th International Workshop on Polyhedral Compilation Techniques; 2014 January 20–22; Vienna, Austria.* [S. l.]: [s. n.]; 2014.

6. Прихожий АА, Карасик ОН. Разнородный блочный алгоритм поиска кратчайших путей между всеми парами вершин графа. *Системный анализ и прикладная информатика*. 2017;3:68–75. DOI: 10.21122/2309-4923-2017-3-68-75.

7. Воеводин ВлВ, Воеводин ВадВ. Спасительная локальность суперкомпьютеров. *Открытые системы. СУБД*. 2013; 9:12–15.

8. Buluc A, Gilberta JR, Budak C. Solving path problems on the GPU. *Parallel Computing*. 2010;36(5–6):241–253. DOI: 10.1016/j.parco.2009.12.002.

9. Лиходед НА, Полещук МА. Условия приватизации элементов массива потоками вычислений. *Журнал Белорусского государственного университета. Математика. Информатика.* 2018;3:59–67.

10. Harish P, Narayanan P. Accelerating large graph algorithms on the GPU using CUDA. In: *High Performance Computing – HiPC 2007. Proceedings of the 14th International Conference; 2007 December 18–21; Goa, India.* Berlin: Springer-Verlag; 2007. p. 197–208. DOI: 10.1007/978-3-540-77220-0_21.

11. Katz GJ, Kider JT. All-pairs shortest-paths for large graphs on the GPU. In: *Proceedings of the 23rd ACM SIGGRAPH/EUROGRAPHICS symposium on Graphics hardware; 2008 June 20–21; Sarajevo, Bosnia and Herzegovina.* Aire-la-Ville: Eurographics Association; 2008. p. 47–55.

References

1. Venkataraman G, Sahni S, Mukhopadhyaya S. A blocked all-pairs shortest-paths algorithm. *Journal of Experimental Algorithmics*. 2003;8:857–874. DOI: 10.1145/996546.996553.

2. Park J, Penner M, Prasanna VK. Optimizing graph algorithms for improved cache performance. *IEEE Transactions on Parallel and Distributed Systems*. 2004;15(9):769–782. DOI: 10.1109/TPDS.2004.44.

3. Srinivasan T, Balakrishnan R, Gangadharan SA, Hayawardh V. A scalable parallelization of all-pairs shortest path algorithm for a high performance cluster environment. In: *Proceedings of the 13th International Conference on Parallel and Distributed Systems; 2007 December 5–7; Hsinchu, Taiwan.* Washington: IEEE Computer Society; 2007. p. 1–8. DOI: 10.1109/ICPADS.2007.4447721.

4. Lund BD, Smith JW. A multi-stage CUDA kernel for Floyd – Warshall. arXiv:1001.4108. 2010 [Preprint]. 2010 [cited 2019 June 3]. Available from: https://arxiv.org/abs/1001.4108.

5. Mullapudi RT, Bondhugula U. Tiling for dynamic scheduling. In: *IMPACT 2014. Proceedings of the 4th International Workshop on Polyhedral Compilation Techniques; 2014 January 20–22; Vienna, Austria.* [S. l.]: [s. n.]; 2014.

6. Prihozhy AA, Karasik ON. Heterogenious blocked all-pairs shortest paths algorithm. *Sistemnyi analiz i prikladnaya informatika.* 2017;3:68–75. Russian. DOI: 10.21122/2309-4923-2017-3-68-75.

7. Voevodin VlV, Voevodin VadV. [The fortunate locality of supercomputers]. *Otkrytye sistemy. SUBD*. 2013;9:12–15. Russian.

8. Buluc A, Gilberta JR, Budak C. Solving path problems on the GPU. *Parallel Computing*. 2010;36(5–6):241–253. DOI: 10.1016/j.parco.2009.12.002.

9. Likhoded NA, Paliashchuk MA. Conditions for privatizing the elements of arrays by computing threads. *Journal of the Belarusian State University. Mathematics and Informatics.* 2018;3:59–67. Russian.

10. Harish P, Narayanan P. Accelerating large graph algorithms on the GPU using CUDA. In: *High Performance Computing – HiPC 2007. Proceedings of the 14th International Conference; 2007 December 18–21; Goa, India.* Berlin: Springer-Verlag; 2007. p. 197–208. DOI: 10.1007/978-3-540-77220-0_21.

11. Katz GJ, Kider JT. All-pairs shortest-paths for large graphs on the GPU. In: *Proceedings of the 23rd ACM SIGGRAPH/EUROGRAPHICS Symposium on Graphics Hardware; 2008 June 20–21; Sarajevo, Bosnia and Herzegovina.* Aire-la-Ville: Eurographics Association; 2008. p. 47–55.

Статья поступила в редколлегию 22.09.2019.
Received by editorial board 22.09.2019.

Теоретические основы информатики

Theoretical foundations of computer science

УДК 519.17;519.85;004.02

УЛУЧШЕННЫЕ ВЕРХНИЕ ОЦЕНКИ В ЗАДАЧЕ ОПТИМАЛЬНОГО РАЗБИЕНИЯ ГРАФА НА КЛИКИ

А. Б. БЕЛЫЙ [1)], ***С. Л. СОБОЛЕВСКИЙ*** [2), 3), 4)], ***А. Н. КУРБАЦКИЙ*** [5)], ***К. РАТТИ*** [4)]

[1)]*СМАРТ-центр, проезд Криэйт, 1, 138602, г. Сингапур, Сингапур*
[2)]*Национальный исследовательский университет ИТМО, пр. Кронверкский, 49, 197101, г. Санкт-Петербург, Россия*
[3)]*Нью-Йоркский университет, ул. Джэй, 370, 11201, г. Нью-Йорк, США*
[4)]*Массачусетский технологический институт, пр. Массачусетс, 77, 02139, г. Кембридж, США*
[5)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*

Рассматривается задача нахождения разбиения полного взвешенного графа на клики так, что сумма весов ребер между вершинами, принадлежащими одной клике, максимальна. Данная задача, известная как задача разбиения графа на клики (clique partitioning problem), возникает во многих приложениях и представляет собой вариант

Образец цитирования:
Белый АБ, Соболевский СЛ, Курбацкий АН, Ратти К. Улучшенные верхние оценки в задаче оптимального разбиения графа на клики. *Журнал Белорусского государственного университета. Математика. Информатика.* 2019;3:93–104.
https://doi.org/10.33581/2520-6508-2019-3-93-104

For citation:
Belyi AB, Sobolevsky SL, Kurbatski AN, Ratti C. Improved upper bounds in clique partitioning problem. *Journal of the Belarusian State University. Mathematics and Informatics.* 2019; 3:93–104. Russian.
https://doi.org/10.33581/2520-6508-2019-3-93-104

Авторы:
Александр Борисович Белый – инженер-программист.
Станислав Леонидович Соболевский – доктор физико-математических наук; профессор Института дизайна и урбанистики[2)], ассоциированный профессор практики Центра исследований и развития городов[3)], исследователь[4)].
Александр Николаевич Курбацкий – доктор технических наук, профессор; заведующий кафедрой технологий программирования факультета прикладной математики и информатики.
Карло Ратти – кандидат наук; профессор практики кафедры урбанистических исследований и планирования, директор лаборатории MIT Senseable City Lab.

Authors:
Alexander B. Belyi, software engineer.
alex.bely@smart.mit.edu
http://orcid.org/0000-0001-5650-3182
Stanislav L. Sobolevsky, doctor of science (physics and mathematics); professor at the Institute Design and Urban Science[b], associate professor of practice at the Center for Urban Science and Progress[c], researcher[d].
sobolevsky@nyu.edu
Alexander N. Kurbatski, doctor of science (engineering), full professor; head of the department of software engineering, faculty of applied mathematics and computer science.
kurb@unibel.by
Carlo Ratti, PhD; professor of the practice at the department of urban studies and planning, director of MIT Senseable City Lab.
ratti@mit.edu

классической задачи кластеризации. Она, как и многие другие задачи комбинаторной оптимизации, является NP-трудной, поэтому нахождение ее точного решения зачастую оказывается трудоемким. В данной работе предлагается новый метод построения верхней оценки для функции качества разбиения и показывается, как полученная оценка применяется в методе ветвей и границ при нахождении точного решения. Предлагаемый подход накладывает ограничения на максимально возможное качество разбиения. Новизна метода заключается в возможности использования треугольников, пересекающихся по ребрам, что позволяет находить гораздо более точные оценки, чем при рассмотрении только непересекающихся подграфов. Помимо построения начальной оценки в статье описывается способ ее пересчета при фиксировании ребер на каждом шаге метода ветвей и границ. Приводятся результаты тестирования предлагаемого алгоритма на сгенерированных наборах случайных графов. Показывается, что версия, использующая новые оценки, работает в несколько раз быстрее ранее известных методов.

Ключевые слова: разбиение графа на клики; точное решение; метод ветвей и границ; верхние оценки.

Благодарность. Исследование выполнено при поддержке Национального исследовательского фонда (офис премьер-министра Сингапура) в рамках программы CREATE, Singapore-MIT Alliance for Research and Technology (SMART) Future Urban Mobility (FM) IRG.

IMPROVED UPPER BOUNDS IN CLIQUE PARTITIONING PROBLEM

A. B. BELYI [a], ***S. L. SOBOLEVSKY*** [b, c, d], ***A. N. KURBATSKI*** [e], ***C. RATTI*** [d]

[a] *SMART Centre, 1 Create Way, Singapore 138602, Singapore*
[b] *ITMO University, 49 Kronverksky Avenue, Saint Petersburg 197101, Russia*
[c] *New York University, 370 Jay Street, New York 11201, USA*
[d] *Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge 02139, USA*
[e] *Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*

Corresponding author: A. B. Belyi (alex.bely@smart.mit.edu)

In this work, a problem of partitioning a complete weighted graph into cliques in such a way that sum of edge weights between vertices belonging to the same clique is maximal is considered. This problem is known as a clique partitioning problem. It arises in many applications and is a varian of classical clustering problem. However, since the problem, as well as many other combinatorial optimization problems, is NP-hard, finding its exact solution often appears hard. In this work, a new method for constructing upper bounds of partition quality function values is proposed, and it is shown how to use these upper bounds in branch and bound technique for finding an exact solution. Proposed method is based on the usage of triangles constraining maximal possible quality of partition. Novelty of the method lies in possibility of using triangles overlapping by edges, which allows to find much tighter bounds than when using only non-overlapping subgraphs. Apart from constructing initial estimate, a method of its recalculation, when fixing edges on each step of branch and bound method, is described. Test results of proposed algorithm on generated sets of random graphs are provided. It is shown, that version that uses new bounds works several times faster than previously known methods.

Keywords: clique partitioning; branch and bound method; exact solution; upper bounds.

Acknowledgements. This research is supported by the National Research Foundation (prime minister's office, Singapore), under its CREATE programme, Singapore-MIT Alliance for Research and Technology (SMART) Future Urban Mobility (FM) IRG.

Введение

Задача оптимального разбиения полного взвешенного графа на клики возникает во многих приложениях, особенно часто там, где нужно кластеризовать объекты, учитывая только отношения между ними [1]. Ярким примером целого класса таких задач может служить поиск сообществ в комплексных сетях [2–5], который сводится к задаче оптимального разбиения на клики, если осуществлять поиск путем максимизации целевой функции, такой как, например, модулярность [6; 7] или длина кода [8]. Практическая значимость задачи в последнее время вызывает все больший интерес ученых. Однако, поскольку она является NP-трудной [9], большинство исследований сконцентрированы на предложении эвристик, способных относительно быстро находить решения, близкие к оптимальным [10–15]. В то же время алгоритмы, приводящие к точному решению, предлагаются редко и в основном представляют собой вариации метода ветвей и границ [11; 16; 17]. Один из наиболее новых и эффективных

алгоритмов был представлен в работе [17], основной вклад авторов которой заключается в получении нового метода оценки верхней границы значений функции качества разбиения.

В настоящей статье предлагается улучшенный метод, позволяющий находить более эффективные оценки. Его использование приводит к существенному увеличению скорости схождения метода ветвей и границ. Показывается, что новый метод работает в несколько раз быстрее ранее известных подходов.

Задача разбиения графа на клики может быть сформулирована следующим образом [1]. Пусть задан взвешенный полный граф $G = (V, E)$, в котором веса ребер есть вещественные числа как положительные, так и отрицательные, V – множество вершин, $E = \left\{ (i, j, e_{ij}) \middle| i, j \in V, e_{ij} \in \mathbb{R} \right\}$ – множество ребер с весами. Если изначально граф не полный, то отсутствующие ребра могут быть представлены ребрами с весом 0. Далее ребра с положительным либо отрицательным весом будем называть положительными либо отрицательными соответственно. На данном графе вводится функция качества разбиения $Q(C)$, которая для каждого разбиения C множества вершин на кластеры равна сумме весов ребер, соединяющих вершины из одного кластера:

$$Q(C) = \sum_{C_i = C_j} e_{ij},$$

где C_i – номер кластера, в который попадает вершина i; e_{ij} – вес ребра (i, j).

Требуется найти такое разбиение множества вершин V на кластеры, при котором значение Q максимально.

Данная задача часто возникает на практике, особенно там, где нужно разделить объекты на заранее неизвестное число групп. Тогда объектам сопоставляются вершины графа, а некоторой мере схожести объектов соответствуют веса ребер. Примеры таких задач можно найти в биологии [1], в сферах планирования [18] и групповой технологии [16; 19]. При исследовании комплексных сетей важным аспектом является обнаружение в них структуры сообществ [2–5]. Некоторые наиболее популярные подходы основываются на сведении задачи поиска сообществ к задаче разбиения на клики путем замены изначальной сети графом, веса ребер которого задаются специальной функцией, такой как модулярность [6; 7] или длина кода [8]. Таким образом, нахождение точного решения задачи разбиения на клики может быть применено при определении оптимальной с точки зрения конкретной функции качества структуры сообществ [20].

Построение верхней оценки

Очевидной тривиальной верхней оценкой значений функции Q является сумма всех положительных весов ребер:

$$Q_{\text{trivial\_max}} = \sum_{e_{ij} > 0} e_{ij}.$$

Однако на практике данная оценка обычно очень далека от реально достижимого максимума. Для получения более точной оценки можно учесть следующее наблюдение. Рассмотрим тройку вершин (a, b, c), соединенных ребрами с весами e_{ab}, e_{bc}, e_{ac} такими, что $e_{ab} > 0$, $e_{ac} > 0$, $e_{bc} < 0$. Несложно заметить, что при любом разбиении вершин на кластеры либо хотя бы одно положительное ребро не попадет в сумму (т. е. или вершины a и b, или вершины a и c окажутся в разных кластерах), либо отрицательное ребро будет включено в сумму (т. е. вершины b и c окажутся в одном кластере). Такую упорядоченную тройку вершин (a, b, c) с весами ребер $e_{ab} > 0$, $e_{ac} > 0$, $e_{bc} < 0$ будем называть штрафующим треугольником, а величину $p_{abc} = \min(e_{ab}, e_{ac}, -e_{bc})$ – штрафом, поскольку из рассуждений выше следует, что

$$Q(C) \le Q_{\text{trivial\_max}} - p_{abc}$$

для любого разбиения C, или, что то же самое,

$$Q^* \le Q_{\text{trivial\_max}} - p_{abc},$$

где Q^* – максимально возможное значение $Q(C)$, достигаемое при оптимальном C.

Если штрафующие треугольники не пересекаются по ребрам, то каждый из них накладывает свой штраф на максимально возможное значение Q. То есть если S – множество непересекающихся по ребрам штрафующих треугольников, то

$$Q^* \le Q_{\text{trivial\_max}} - \sum_{\{a,b,c\} \in S} p_{abc}.$$

Обозначив $P_G = Q_{\text{trivial\_max}} - Q^*$ минимальный штраф разбиения графа G, предыдущее утверждение можно записать как $\sum_{\{a,b,c\} \in S} p_{abc} \le P_G$. Данное наблюдение было доказано и использовано в работе [17]. Ниже показывается, как, применяя идею из [5], построить более эффективную оценку, учитывая штрафующие треугольники, пересекающиеся по ребрам.

Теорема. *Пусть* (a, b, c) *– штрафующий треугольник в графе* $G = (V, E)$. *Построим новый граф* $G' = (V, E')$ *на вершинах V, вычтя из положительных весов и добавив к отрицательному весу ребер* e_{ab}, e_{ac} *и* e_{bc} *штраф* p_{abc}, *т. е.* $E' = E \setminus \{(a, b, e_{ab}), (a, c, e_{ac}), (b, c, e_{bc})\} \cup \{(a, b, e_{ab} - p_{abc}), (a, c, e_{ac} - p_{abc}), (b, c, e_{bc} + p_{abc})\}$. *Тогда* $Q^* \le Q_{\text{trivial\_max}} - p_{abc} - P_{G'}$ *или* $P_{G'} + p_{abc} \le P_G$.

Доказательство. По определению $P_{G'} = Q'_{\text{trivial\_max}} - Q'^*$, где $Q'_{\text{trivial\_max}}$ – сумма положительных ребер в графе G'; Q'^* – максимальное значение Q при оптимальном разбиении графа G'. Тогда $Q_{\text{trivial\_max}} - p_{abc} - P_{G'} = Q_{\text{trivial\_max}} - p_{abc} - Q'_{\text{trivial\_max}} + Q'^* = Q_{\text{trivial\_max}} - Q'_{\text{trivial\_max}} - p_{abc} + Q'^* = p_{abc} + Q'^* =$ (поскольку $Q_{\text{trivial\_max}}$ равно сумме положительных ребер в графе G, а $Q'_{\text{trivial\_max}}$ – сумме положительных ребер в G' и G и G' отличаются весами только двух положительных ребер) $= e_{ab} + e_{ac} - (e_{ab} - p_{abc}) - (e_{ac} - p_{abc}) - p_{abc} + Q'^* = p_{abc} + Q'^*$. Остается показать, что $Q^* - Q'^* \le p_{abc}$.

Пусть C – оптимальное разбиение, при котором достигается $Q^* = Q(C)$. Рассмотрим то же самое разбиение графа G' и положим для него $Q = Q'(C)$. По определению $Q'(C) \le Q'^*$, т. е. достаточно показать, что $Q(C) - Q'(C) \le p_{abc}$. Поскольку все ребра в G и G', кроме трех, имеют одинаковые веса, то разность в левой части неравенства зависит только от того, какие из весов ребер e_{ab}, e_{ac} и e_{bc} включены в суммы. Возможны три случая:

1) все три вершины a, b и c принадлежат разным кластерам. Тогда ни одно ребро не включено в суммы и выполняется

$$Q(C) - Q'(C) = 0 \le p_{abc};$$

2) вершины a, b и c принадлежат одному кластеру. Тогда все три ребра включены в суммы и выполняется

$$Q(C) - Q'(C) = e_{ab} + e_{ac} + e_{bc} - (e_{ab} - p_{abc}) - (e_{ac} - p_{abc}) - (e_{bc} + p_{abc}) = p_{abc};$$

3) две вершины из a, b и c принадлежат одному кластеру. Тогда ровно одно ребро включено в суммы и либо $Q(C) - Q'(C) = e_{ab} - (e_{ab} - p_{abc}) = p_{abc}$, либо $Q(C) - Q'(C) = e_{ac} - (e_{ac} - p_{abc}) = p_{abc}$, либо $Q(C) - Q'(C) = e_{bc} - (e_{bc} + p_{abc}) = -p_{abc}$. Следовательно, в любом случае утверждение теоремы выполнено и она доказана.

Таким образом, чтобы построить оценку $Q_{\max} \ge Q^*$, можно последовательно находить штрафующие треугольники и вычитать их штраф из весов ребер, пока в графе не останется штрафующих треугольников. Сумма полученных штрафов $\sum p_{abc}$ будет оценкой снизу для P_G, и тогда $\sum p_{abc} \le P_G = Q_{\text{trivial\_max}} - Q^*$, $Q^* \le Q_{\text{trivial\_max}} - \sum p_{abc} = Q_{\max}$.

Нахождение точного решения

Применяя полученную оценку, для определения точного решения можно использовать метод ветвей и границ. На первом шаге найдем какое-нибудь разбиение, желательно со значением Q, близким к максимуму. Эта величина будет начальной нижней оценкой $Q_{\min}$ достижимого значения Q. Далее построим набор штрафующих треугольников и начальную оценку $Q_{\max}$. Затем на каждой итерации будем рассматривать очередное ребро и два возможных разбиения: разбиение, в котором данное ребро зафиксировано как внутрикластерное, т. е. концы ребра попадают в один кластер, и разбиение, в котором данное ребро зафиксировано как межкластерное, т. е. концы ребра лежат в разных кластерах. Для каждого разбиения будем пересчитывать оценку $Q_{\max}$ с учетом нового ограничения. Если на очередном шаге оценка $Q_{\max}$ меньше уже достигнутого значения $Q_{\min}$, то при текущих ограничениях получить решение, лучшее уже найденного, невозможно, и данная ветвь отсекается. Иначе рекурсивно рассматривается очередное ребро. Ниже опишем каждый шаг более детально.

Начальное разбиение и оценку $Q_{\min}$ можно найти с помощью какой-либо из множества эвристик, описанных в литературе. В данной работе используется алгоритм Combo [15], как один из наиболее точных и быстрых методов со свободно доступным исходным кодом, позволяющим достичь наилучших (по сравнению с другими подходами) значений целевой функции.

Для построения начального набора штрафующих треугольников применяется жадный алгоритм. На каждой итерации выбирается треугольник с наименьшим (наибольшим по модулю) весом отрицательного ребра. Веса ребер соответствующего треугольника корректируются с учетом полученного штрафа, и данный процесс повторяется, пока в графе есть штрафующие треугольники. Описанный процесс может быть представлен в виде нижеприведенного алгоритма 1. Далее при описании всех алгоритмов подразумевается, что они имеют доступ к графу $G = (V, E)$.

Алгоритм 1. Построение набора штрафующих треугольников *S*

Выход: набор штрафующих треугольников *S*.

1. Изначально набор *S* пуст.
2. Найти все отрицательные ребра.
3. Упорядочить их по возрастанию веса.
4. Для каждого ребра (b, c, e_{bc}) из полученного списка:
   для каждой вершины *a*, неинцидентной ребру:
   если $e_{ab} > 0$ и $e_{ac} > 0$:
   добавить треугольник (a, b, c) в *S*;
   // обновить веса ребер треугольника (a, b, c) с учетом штрафа:
   $p_{abc} = \min(e_{ab}, e_{ac}, -e_{bc})$;
   $e_{ab} = e_{ab} - p_{abc}$;
   $e_{ac} = e_{ac} - p_{abc}$;
   $e_{bc} = e_{bc} + p_{abc}$.
5. Вернуть набор *S*.

Отрицательные ребра можно найти путем простого перебора всех ребер за $O(n^2)$ итераций, где $n = |V|$ – число вершин. На шаге 4 алгоритма 1 рассматриваются некоторые тройки вершин ровно один раз. Так как всего троек вершин порядка n^3, то алгоритм 1 выполняется за время $O(n^3)$.

На каждой итерации метода ветвей и границ очередное незафиксированное ребро объявляется либо внутрикластерным, либо межкластерным. В каждом случае нужно обновить статус остальных ребер, чтобы выполнялось условие транзитивности, т. е. если ребра (a, b) и (b, c) внутрикластерные, то таковым должно быть и ребро (a, c), и если ребро (a, b) внутрикластерное, а ребро (b, c) межкластерное, то ребро (a, c) также должно быть межкластерным. Это может быть сделано с помощью следующего алгоритма 2.

Алгоритм 2. Обновление зафиксированных ребер

Вход: зафиксированное ребро (a, b).

Выход: набор зафиксированных ребер.

1. Найти вершины, принадлежащие кластеру вершины *a* (множество $A = \{i \mid C_i = C_a\}$), кластеру вершины *b* (множество $B = \{i \mid C_i = C_b\}$); вершины, про которые уже известно, что они принадлежат кластеру, отличному от кластера вершины *a* (множество $X = \{i \mid C_i \neq C_a\}$); вершины, про которые уже известно, что они принадлежат кластеру, отличному от кластера вершины *b* (множество $Y = \{i \mid C_i \neq C_b\}$).
2. Если ребро (a, b) зафиксировано как внутрикластерное, то таковыми же объявить все ребра, соединяющие вершины множеств *A* и *B*, а все ребра, соединяющие *A* и *Y*, а также *B* и *X*, объявить межкластерными;
   иначе все ребра, соединяющие *A* и *B*, объявить межкластерными.
3. Возвратить все вновь зафиксированные ребра.

Первый шаг может быть выполнен путем рассмотрения всех вершин графа, т. е. за $O(n)$ итераций. На втором шаге обновляется статус некоторых ребер, которых всего не более n^2. Таким образом, сложность алгоритма 2 есть $O(n^2)$.

После обновления статуса ребер нужно обновить текущее значение оценки сверху $Q_{\max}$. Пусть (a, b, c) – штрафующий треугольник. Если на текущем шаге положительное ребро фиксируется как

межкластерное или внутрикластерным объявляется отрицательное ребро, то абсолютное значение его веса добавляется к штрафу и все штрафующие треугольники, включающие данное ребро, перестают учитываться. Если же внутрикластерным объявляется положительное ребро или межкластерным объявляется отрицательное, то штраф треугольника обновляется: если (a, b) фиксируется внутрикластерным, то $p_{abc} = \min(e_{ac}, -e_{bc})$, если (a, c) фиксируется внутрикластерным, то $p_{abc} = \min(e_{ab}, -e_{bc})$, если (b, c) фиксируется межкластерным, то $p_{abc} = \min(e_{ab}, e_{ac})$. Таким образом, для обновления множества штрафующих треугольников и пересчета $Q_{\max}$ можно использовать алгоритм 3.

Алгоритм 3. Обновление $Q_{\max}$ и S

Вход: набор штрафующих треугольников S.

Выход: обновленный S и новая $Q_{\max}$.

1. $Q_{\max} = Q_{\text{trivial\_max}}$.
2. Для каждого зафиксированного ребра (a, b):

   если $e_{ab} > 0$ и (a, b) является межкластерным, то

   $Q_{\max} = Q_{\max} - e_{ab}$;

   если $e_{ab} < 0$ и (a, b) является внутрикластерным, то

   $Q_{\max} = Q_{\max} + e_{ab}$.
3. Для каждого треугольника (a, b, c) из S:

   если межкластерным является (a, b) или (a, c) либо (b, c) есть внутрикластерное, то

   исключить (a, b, c) из S;

   $p_{abc} = 0$;

   иначе, если (a, b) является внутрикластерным, то

   $p_{abc} = \min(e_{ac}, -e_{bc})$;

   $e_{ac} = e_{ac} - p_{abc}$;

   $e_{bc} = e_{bc} + p_{abc}$;

   иначе, если (a, c) является внутрикластерным, то

   $p_{abc} = \min(e_{ab}, -e_{bc})$;

   $e_{ab} = e_{ab} - p_{abc}$;

   $e_{bc} = e_{bc} + p_{abc}$;

   иначе, если (b, c) является межкластерным, то

   $p_{abc} = \min(e_{ab}, e_{ac})$;

   $e_{ab} = e_{ab} - p_{abc}$;

   $e_{ac} = e_{ac} - p_{abc}$;

   иначе:

   $p_{abc} = \min(e_{ab}, e_{ac}, -e_{bc})$;

   $e_{ab} = e_{ab} - p_{abc}$;

   $e_{ac} = e_{ac} - p_{abc}$;

   $e_{bc} = e_{bc} + p_{abc}$;

   $Q_{\max} = Q_{\max} - p_{abc}$.
4. Пока в графе есть штрафующие треугольники (a, b, c):

   $p_{abc} = \min(e_{ab}, e_{ac}, -e_{bc})$;

   $e_{ab} = e_{ab} - p_{abc}$;

   $e_{ac} = e_{ac} - p_{abc}$;

   $e_{bc} = e_{bc} + p_{abc}$;

   $Q_{\max} = Q_{\max} - p_{abc}$.
5. Возвратить обновленный S и $Q_{\max}$.

Добавление в S каждого штрафующего треугольника обнуляет как минимум одно ребро в графе. Таким образом, штрафующих треугольников не более n^2, и шаги 2 и 3 выполняются за $O(n^2)$ итераций. Шаг 4 может быть выполнен рассмотрением всех треугольников, т. е. за $O(n^3)$ итераций. Таким образом, сложность алгоритма 3 есть $O(n^3)$.

В методе ветвей и границ важным аспектом является очередность, в которой происходит ветвление. В работе [17] показано, что выбирать очередное ребро нужно таким образом, чтобы при объявлении его межкластерным значение $Q_{\max}$ максимально уменьшалось. В нашем случае получение точных значений изменения $Q_{\max}$ на каждом шаге требует вызова алгоритма 3, что слишком трудоемко. Чтобы найти оценки изменения $Q_{\max}$, для каждого положительного ребра рассмотрим, как $Q_{\max}$ меняется при объявлении межкластерным только этого ребра. И в дальнейшем ребра рассматриваются в порядке убывания величины вычисленного изменения.

Алгоритм 4. Упорядочивание ребер

Вход: набор штрафующих треугольников S, дающий оценку $Q_{\max}$.

Выход: упорядоченный список ребер.

1. Для каждого положительного ребра (a, b):
   объявить ребро (a, b) межкластерным;
   вычислить новую $Q'_{\max}$ с помощью алгоритма 3;
   ребру (a, b) поставить в соответствие значение $Q_{\max} - Q'_{\max}$.
2. Отсортировать ребра в порядке убывания соответствующих им значений.
3. Вернуть упорядоченный список.

Алгоритм 4 будет вызываться только один раз в начале работы метода ветвей и границ. Так как положительных ребер может быть порядка n^2 и производится их сортировка, то алгоритм 4 выполняется за время $O\left(n^2 \log(n)\right)$.

Далее представлен рекурсивный алгоритм обхода в глубину дерева поиска метода ветвей и границ.

Алгоритм 5. Рекурсивный обход в глубину дерева поиска

Вход: список ребер для рассмотрения, индекс текущего ребра, набор штрафующих треугольников S, значение $Q_{\min}$.

Выход: значение $Q_{\min}$ и сохраненное оптимальное разбиение.

1. Если индекс выходит за рамки списка ребер:
   // значит, все положительные ребра зафиксированы
   // и полученное разбиение является новым наилучшим разбиением.
   Незафиксированные отрицательные ребра объявить межкластерными.
   Запомнить разбиение.
   Обновить $Q_{\min} = Q$.
   Возвратить $Q_{\min}$.
2. Выбрать из списка ребро с заданным индексом.
3. Пока выбранное ребро зафиксировано:
   выбирать следующее ребро и увеличивать индекс.
4. Повторить для {выбранное ребро объявляется внутрикластерным, выбранное ребро объявляется межкластерным}:
   с помощью алгоритма 2 найти и зафиксировать остальные ребра, которые нужно зафиксировать;
   с помощью алгоритма 3 найти новый набор штрафующих треугольников S' и новое значение $Q_{\max}$.
   Если $Q_{\max} > Q_{\min}$, то
      обновить значение $Q_{\min}$ значением, полученным рекурсивным вызовом алгоритма 5 со следующим индексом, с набором S' и текущим значением $Q_{\min}$.
   Возвратить в исходное состояние ребра, зафиксированные в начале шага 4.
5. Возвратить $Q_{\min}$.

Окончательный алгоритм нахождения оптимального разбиения графа на клики может быть представлен следующим алгоритмом 6.

Алгоритм 6. Нахождение точного решения задачи разбиения графа на клики

1. Используя алгоритм Combo, построить начальное разбиение и оценку $Q_{\min}$.
2. По алгоритму 1 найти начальный набор штрафующих треугольников S и оценку $Q_{\max}$.
3. С помощью алгоритма 4 упорядочить ребра.
4. Вызвать алгоритм 5 со списком, полученным на шаге 3, индексом 1, набором штрафующих треугольников S и значением $Q_{\min}$.
5. Возвратить оптимальное разбиение и значение $Q_{\min}$.

Данный алгоритм рассматривает все возможные комбинации, запоминает оптимальное разбиение и наибольшее достижимое значение $Q_{\min}$.

Вычислительный эксперимент и анализ результатов

Предлагаемый алгоритм реализован на языке C++. Для его тестирования использовались наборы искусственных графов, предложенные в [17]. Первый набор состоит из графов на n вершинах, веса ребер которых выбирались случайным образом из равномерного распределения на отрезке $[-q, q]$. Для возможности сравнения с результатами из работы [17] применялась аналогичная процедура: для каждого n от 10 до 20 и каждого q из набора $\{1, 2, 3, 5, 10, 50, 100\}$ генерировалось 5 случайных графов.

В табл. 1 приведены результаты работы предлагаемого алгоритма, а также результаты из [17] для сравнения. Каждое значение в табл. 1 равно сумме соответствующих значений для 35 случайных графов. *Nodes* обозначает количество рассмотренных вершин в дереве поиска решения, t – время выполнения в секундах, $Q_{\text{init\_min}}$ и $Q_{\text{init\_max}}$ – начальные оценки $Q_{\min}$ и $Q_{\max}$, полученные на шагах 1 и 2 алгоритма 6 соответственно. Значения $Q_{\text{trivial\_max}}$, $Q_{\text{init\_max}}$, $Q_{\text{init\_min}}$ нормализованы относительно оптимального решения Q^*.

Таблица 1

Результаты на первом наборе случайных графов

Table 1

Results for the first set of random graphs

n	$Q^a_{\text{trivial\_max}}$	$Q^a_{\text{init\_max}}$	$Q^a_{\text{init\_min}}$	$Q^b_{\text{trivial\_max}}$	$Q^b_{\text{init\_max}}$	$Q^b_{\text{init\_min}}$	$Nodes^a$	$Nodes^b$	t^a	t^b
10	1,720	**1,099**	**1,000**	1,764	1,226	0,994	**362**	1205	**0,05**	0,05
11	1,819	**1,119**	**0,998**	1,831	1,272	0,988	**877**	4236	**0,10**	0,13
12	1,758	**1,120**	**0,994**	1,932	1,305	0,993	**1401**	7577	**0,16**	0,18
13	2,014	**1,191**	**0,990**	1,867	1,287	0,986	**3723**	20 005	**0,42**	0,47
14	2,006	**1,199**	**0,991**	1,971	1,355	0,983	**6590**	50 101	**0,77**	1,28
15	2,065	**1,223**	0,995	2,071	1,367	**0,996**	**10 776**	185 336	**1,58**	5,26
16	2,119	**1,261**	0,991	2,043	1,341	**0,999**	**43 150**	499 569	**6,35**	16,3
17	2,154	**1,229**	0,994	2,189	1,419	**0,997**	**63 916**	4 186 427	**10,26**	155
18	2,226	**1,305**	0,988	2,230	1,433	**0,993**	**199 302**	9 811 533	**39,84**	466
19	2,183	**1,264**	0,986	2,236	1,439	**0,994**	**479 192**	37 572 347	**101,08**	1849
20	2,313	**1,314**	**0,989**	2,251	1,440	0,988	**918 177**	185 321 420	**259,61**	11 299

Примечание. Здесь и в табл. 2–4 обозначены результаты: ^{a} – предлагаемого алгоритма; ^{b} – алгоритма из [17]. Полужирным шрифтом выделены лучшие результаты.

Графы второго набора получены в итоге следующей процедуры. Для каждого графа из n вершин фиксировался параметр p. Затем для каждой вершины строился бинарный вектор длины p, у которого значения 0 или 1 выбирались равновероятно. Вес ребра между вершинами i и j полагался равным p минус удвоенное количество позиций, в которых векторы i и j отличаются. Как и в первом наборе, в целях удобного сравнения с результатами [17] для каждого значения n от 10 до 24 и p из набора $\{1, 2, 3, 5, 10, 50, 100\}$ генерировалось 5 случайных графов (табл. 2). Каждое значение в табл. 2 равно сумме соответствующих значений для 35 реализаций случайных графов.

Таблица 2

Результаты на втором наборе случайных графов

Table 2

Results for the second set of random graphs

n	$Q^a_{\text{trivial\_max}}$	$Q^a_{\text{init\_max}}$	$Q^a_{\text{init\_min}}$	$Q^b_{\text{trivial\_max}}$	$Q^b_{\text{init\_max}}$	$Q^b_{\text{init\_min}}$	$Nodes^a$	$Nodes^b$	t^a	t^b
10	1,370	**1,057**	**0,998**	1,387	1,122	0,985	**167**	488	**0,03**	0,04
11	1,427	**1,089**	**0,998**	1,476	1,177	0,995	**376**	962	**0,05**	0,05

Окончание табл. 2
Ending table 2

n	$Q^a_{\text{trivial\_max}}$	$Q^a_{\text{init\_max}}$	$Q^a_{\text{init\_min}}$	$Q^b_{\text{trivial\_max}}$	$Q^b_{\text{init\_max}}$	$Q^b_{\text{init\_min}}$	$Nodes^a$	$Nodes^b$	t^a	t^b
12	1,466	**1,064**	0,994	1,421	1,149	**1,000**	**388**	972	**0,07**	0,05
13	1,480	**1,092**	0,992	1,437	1,144	**0,997**	**791**	2178	**0,11**	0,08
14	1,522	**1,102**	**0,994**	1,516	1,173	0,991	**1597**	6158	**0,25**	0,19
15	1,523	**1,116**	0,993	1,546	1,178	**0,995**	**2134**	7819	**0,38**	0,22
16	1,560	**1,126**	0,988	1,541	1,181	**0,992**	**4215**	21 752	**0,78**	0,71
17	1,560	**1,113**	**0,996**	1,569	1,188	0,988	**5277**	138 305	**1,08**	5,08
18	1,553	**1,115**	**0,996**	1,575	1,195	0,992	**8532**	160 195	**2,17**	6,52
19	1,617	**1,148**	**0,997**	1,592	1,214	0,987	**34 968**	1 389 759	**11,14**	66,4
20	1,602	**1,131**	**0,991**	1,630	1,228	0,986	**38 270**	2 598 775	**12,56**	136
21	1,640	**1,148**	**0,997**	1,631	1,229	0,983	**38 993**	11 977 231	**14,87**	741
22	1,708	**1,175**	**0,994**	1,639	1,232	0,990	**160 856**	14 413 288	**71,73**	962
23	1,691	**1,172**	**0,993**	1,632	1,218	0,992	**295 452**	25 313 750	**134,63**	1805
24	1,724	**1,192**	**0,995**	1,728	1,269	0,984	**1 615 930**	778 958 420	**782,56**	67 034

Третий набор состоит из графов, первый шаг построения которых совпадал с процедурой генерации графов первого набора, но на втором шаге вес каждого ребра мог быть обнулен с заданной вероятностью. Было сгенерировано два поднабора, в первом вероятность обнуления веса ребра составляла 40 %, во втором – 80 % (табл. 3 и 4).

Таблица 3

Результаты на третьем наборе случайных графов с вероятностью обнуления веса ребра 40 %

Table 3

Results for the third set of random graphs with 40 % probability of changing an edge weight to zero

n	$Q^a_{\text{trivial\_max}}$	$Q^a_{\text{init\_max}}$	$Q^a_{\text{init\_min}}$	$Q^b_{\text{trivial\_max}}$	$Q^b_{\text{init\_max}}$	$Q^b_{\text{init\_min}}$	$Nodes^a$	$Nodes^b$	t^a	t^b
10	1,327	**1,076**	**0,998**	1,468	1,153	0,987	**254**	388	0,03	**0,01**
11	1,374	**1,085**	**0,988**	1,494	1,173	0,985	**451**	810	0,04	**0,01**
12	1,489	**1,117**	**0,996**	1,498	1,167	0,983	**596**	2442	0,07	**0,04**
13	1,517	**1,142**	**0,990**	1,513	1,192	0,988	**1269**	5128	0,12	**0,09**
14	1,566	**1,158**	**0,993**	1,492	1,184	0,990	**2422**	4836	0,21	**0,08**
15	1,638	**1,182**	**0,994**	1,616	1,243	0,983	**3783**	25 647	**0,41**	0,54
16	1,703	**1,180**	**0,988**	1,696	1,267	0,986	**10 789**	54 728	**1,21**	1,38
17	1,699	**1,181**	**0,990**	1,750	1,307	0,975	**13 918**	140 765	**1,79**	3,93
18	1,736	**1,201**	**0,988**	1,699	1,263	0,974	**46 884**	382 507	**7,00**	11,59
19	1,736	**1,199**	**0,988**	1,800	1,315	0,984	**56 757**	1 469 527	**8,85**	55,69
20	1,864	**1,231**	**0,986**	1,850	1,326	0,985	**122 830**	3 195 924	**25,30**	114

Таблица 4

Результаты на третьем наборе случайных графов с вероятностью обнуления веса ребра 80 %

Table 4

Results for the third set of random graphs with 80 % probability of changing an edge weight to zero

n	$Q^a_{\text{trivial\_max}}$	$Q^a_{\text{init\_max}}$	$Q^a_{\text{init\_min}}$	$Q^b_{\text{trivial\_max}}$	$Q^b_{\text{init\_max}}$	$Q^b_{\text{init\_min}}$	$Nodes^a$	$Nodes^b$	t^a	t^b
10	1,064	1,040	**1,000**	1,060	**1,037**	0,992	51	**30**	0,01	**0**
11	1,061	1,013	0,987	1,067	1,013	**1,000**	47	**14**	0,01	**0**
12	1,079	1,015	0,984	1,050	**1,006**	**0,995**	82	**55**	0,01	**0**
13	1,093	1,043	**0,990**	1,088	**1,023**	0,977	**149**	167	0,01	0,01
14	1,074	1,029	0,977	1,048	**1,019**	**0,999**	176	**71**	0,02	**0**
15	1,101	**1,042**	0,974	1,110	1,059	**0,982**	**228**	472	0,02	**0,01**
16	1,133	**1,041**	0,978	1,135	1,062	**0,979**	**243**	720	0,02	**0,01**
17	1,174	**1,055**	**0,996**	1,114	1,080	0,989	**642**	789	0,04	**0,02**
18	1,159	**1,078**	0,981	1,143	1,082	**0,985**	**650**	979	0,05	**0,02**
19	1,148	**1,068**	**0,987**	1,195	1,108	0,985	**891**	2601	0,08	**0,06**
20	1,164	1,098	0,982	1,147	**1,072**	**0,986**	**2366**	2423	0,20	**0,05**

Тестирование выполнялось на компьютере с оперативной памятью 8 Гб и процессором Intel Core i7 частотой 2,3 ГГц под управлением MacOS 10.13.6. Как видно из табл. 1–3, предложенный алгоритм работает быстрее, чем алгоритм из [17], который, в свою очередь, как показали его авторы, выполняется быстрее, чем другие известные алгоритмы. Для графов до 20 вершин, в которых большая часть ребер имеет вес 0, описанный алгоритм работает медленнее, чем алгоритм из [17], однако оба алгоритма осуществляются за доли секунды, и с ростом размера графа предлагаемый алгоритм рассматривает меньшее число вариантов (см. табл. 4). Отметим, что представленные для сравнения результаты из [17] получены на компьютере с частотой процессора 1,86 ГГц, в то время как тестирование предлагаемого алгоритма проводилось на несколько более быстром компьютере. Кроме того, данные сравниваются на разных реализациях случайных графов, сгенерированных по одним и тем же правилам. Однако увеличение скорости работы явно не пропорционально увеличению производительности компьютера и в большей степени вызвано улучшением эффективности алгоритма. Убедиться в этом можно, например учитывая тот факт, что пересчет значения $Q_{\max}$ на каждой итерации в обоих алгоритмах выполняется за время $O\left(n^3\right)$, при этом по предлагаемому алгоритму рассматривается гораздо меньше вершин дерева поиска, чем по алгоритму из [17].

Заключение

Предлагаемый алгоритм может быть применен в целях более быстрого нахождения точного решения задачи оптимального разбиения графа на клики. Как и в ранее известном методе, рассмотрение каждой вершины дерева поиска метода ветвей и границ выполняется за время $O\left(n^3\right)$, однако константа в сложности пересчета в описанном алгоритме оказывается несколько выше, что иногда ведет к незначительному проигрыванию по времени. Но данный эффект заметен только на очень простых графах, с которыми оба алгоритма справляются за доли секунды. Для более сложных графов ускорение за счет предложенных в данной работе более точных верхних оценок качества разбиения оказывается существенным. По сравнению с ранее известным методом описанный алгоритм позволяет получить ускорение по времени более чем в 85 раз на некоторых типах графов (с 18,5 ч до 13 мин). При этом сокращение числа рассматриваемых вершин дерева поиска метода ветвей и границ может быть более чем в 480 раз.

Библиографические ссылки

1. Grötschel M, Wakabayashi Y. A cutting plane algorithm for a clustering problem. *Mathematical Programming. Series B*. 1989; 45(1–3):59–96. DOI: 10.1007/BF01589097.

2. Fortunato S. Community detection in graphs. *Physics reports*. 2010;486(3–5):75–174. DOI: 10.1016/j.physrep.2009.11.002.

3. Belyi A, Bojic I, Sobolevsky S, Sitko I, Hawelka B, Rudikova L, et al. Global multi-layer network of human mobility. *International Journal of Geographical Information Science*. 2017;31(7):1381–1402. DOI: 10.1080/13658816.2017.1301455.

4. Belyi A, Bojic I, Sobolevsky S, Rudikova L, Kurbatski A, Ratti C. Community structure of the world revealed by Flickr data. В: *Технологии информатизации и управления. ТИМ-2016. Материалы III Международной научно-практической конференции; 14–15 апреля 2016 г.; Гродно, Беларусь*. Гродно: ГрГУ; 2016. с. 1–9.

5. Sobolevsky S, Belyi A, Ratti C. Optimality of community structure in complex networks. arXiv:1712.05110 [Preprint]. 2017 [cited 2019 March 22]: [17 p.]. Available from: https://arxiv.org/abs/1712.05110.

6. Newman ME, Girvan M. Finding and evaluating community structure in networks. *Physical Review E*. 2004;69(2):026113. DOI: 10.1103/PhysRevE.69.026113.

7. Newman ME. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*. 2006;103(23):8577–8582. DOI: 10.1073/pnas.0601602103.

8. Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. *Sciences of the United States of America*. 2008;105(4):1118–1123. DOI: 10.1073/pnas.0706851105.

9. Wakabayashi Y. *Aggregation of binary relations: algorithmic and polyhedral investigations*. Augsburg: University of Augsburg; 1986. 191 p.

10. De Amorim SG, Barthélemy JP, Ribeiro CC. Clustering and clique partitioning: simulated annealing and tabu search approaches. *Journal of Classification*. 1992;9(1):17–41. DOI: 10.1007/BF02618466.

11. Dorndorf U, Pesch E. Fast clustering algorithms. *ORSA Journal on Computing*. 1994;6(2):141–153. DOI: 10.1287/ijoc.6.2.141.

12. Charon I, Hudry O. Noising methods for a clique partitioning problem. *Discrete Applied Mathematics*. 2006;154(5):754–769. DOI: 10.1016/j.dam.2005.05.029.

13. Zhou Y, Hao JK, Goëffon A. A three-phased local search approach for the clique partitioning problem. *Journal of Combinatorial Optimization*. 2016;32(2):469–491. DOI: 10.1007/s10878-015-9964-9.

14. Brimberg J, Janićijević S, Mladenović N, Urošević D. Solving the clique partitioning problem as a maximally diverse grouping problem. *Optimization Letters*. 2017;11(6):1123–1135. DOI: 10.1007/s11590-015-0869-4.

15. Sobolevsky S, Campari R, Belyi A, Ratti C. General optimization technique for high-quality community detection in complex networks. *Physical Review E*. 2014;90(1):012811. DOI: 10.1103/PhysRevE.90.012811.

16. Oosten M, Rutten JHGC, Spieksma FCR. The clique partitioning problem: facets and patching facets. *Networks: An International Journal*. 2001;38(4):209–226. DOI: 10.1002/net.10004.

17. Jaehn F, Pesch E. New bounds and constraint propagation techniques for the clique partitioning problem. *Discrete Applied Mathematics*. 2013;161(13–14):2025–2037. DOI: 10.1016/j.dam.2013.02.011.

18. Dorndorf U, Jaehn F, Pesch E. Modelling robust flight-gate scheduling as a clique partitioning problem. *Transportation Science*. 2008;42(3):292–301. DOI: 10.1287/trsc.1070.0211.

19. Wang H, Alidaee B, Glover F, Kochenberger G. Solving group technology problems via clique partitioning. *International Journal of Flexible Manufacturing Systems*. 2006;18(2):77–97. DOI: 10.1007/s10696-006-9011-3.

20. Aloise D, Cafieri S, Caporossi G, Hansen P, Perron S, Liberti L. Column generation algorithms for exact modularity maximization in networks. *Physical Review E*. 2010;82(4):046112. DOI: 10.1103/PhysRevE.82.046112.

References

1. Grötschel M, Wakabayashi Y. A cutting plane algorithm for a clustering problem. *Mathematical Programming. Series B*. 1989; 45(1–3):59–96. DOI: 10.1007/BF01589097.

2. Fortunato S. Community detection in graphs. *Physics reports*. 2010;486(3–5):75–174. DOI: 10.1016/j.physrep.2009.11.002.

3. Belyi A, Bojic I, Sobolevsky S, Sitko I, Hawelka B, Rudikova L, et al. Global multi-layer network of human mobility. *International Journal of Geographical Information Science*. 2017;31(7):1381–1402. DOI: 10.1080/13658816.2017.1301455.

4. Belyi A, Bojic I, Sobolevsky S, Rudikova L, Kurbatski A, Ratti C. Community structure of the world revealed by Flickr data. In: *Tekhnologii informatizatsii i upravleniya. TIM-2016. Materialy III Mezhdunarodnoi nauchno-prakticheskoi konferentsii; 14–15 aprelya 2016 g.; Grodno, Belarus'* [Technologies of Information and Management TIM-2016. Materials of the 3rd International science and training conference; 2016 April 14–15; Grodno, Belarus]. Grodno: Yanka Kupala State University of Grodno; 2016. p. 1–9.

5. Sobolevsky S, Belyi A, Ratti C. Optimality of community structure in complex networks. arXiv:1712.05110 [Preprint]. 2017 [cited 2019 March 22]: [17 p.]. Available from: https://arxiv.org/abs/1712.05110.

6. Newman ME, Girvan M. Finding and evaluating community structure in networks. *Physical Review E*. 2004;69(2):026113. DOI: 10.1103/PhysRevE.69.026113.

7. Newman ME. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*. 2006;103(23):8577–8582. DOI: 10.1073/pnas.0601602103.

8. Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. *Sciences of the United States of America*. 2008;105(4):1118–1123. DOI: 10.1073/pnas.0706851105.

9. Wakabayashi Y. *Aggregation of binary relations: algorithmic and polyhedral investigations*. Augsburg: University of Augsburg; 1986. 191 p.

10. De Amorim SG, Barthélemy JP, Ribeiro CC. Clustering and clique partitioning: simulated annealing and tabu search approaches. *Journal of Classification*. 1992;9(1):17–41. DOI: 10.1007/BF02618466.

11. Dorndorf U, Pesch E. Fast clustering algorithms. *ORSA Journal on Computing*. 1994;6(2):141–153. DOI: 10.1287/ijoc.6.2.141.

12. Charon I, Hudry O. Noising methods for a clique partitioning problem. *Discrete Applied Mathematics*. 2006;154(5):754–769. DOI: 10.1016/j.dam.2005.05.029.

13. Zhou Y, Hao JK, Goëffon A. A three-phased local search approach for the clique partitioning problem. *Journal of Combinatorial Optimization.* 2016;32(2):469–491. DOI: 10.1007/s10878-015-9964-9.

14. Brimberg J, Janićijević S, Mladenović N, Urošević D. Solving the clique partitioning problem as a maximally diverse grouping problem. *Optimization Letters.* 2017;11(6):1123–1135. DOI: 10.1007/s11590-015-0869-4.

15. Sobolevsky S, Campari R, Belyi A, Ratti C. General optimization technique for high-quality community detection in complex networks. *Physical Review E.* 2014;90(1):012811. DOI: 10.1103/PhysRevE.90.012811.

16. Oosten M, Rutten JHGC, Spieksma FCR. The clique partitioning problem: facets and patching facets. *Networks: An International Journal.* 2001;38(4):209–226. DOI: 10.1002/net.10004.

17. Jaehn F, Pesch E. New bounds and constraint propagation techniques for the clique partitioning problem. *Discrete Applied Mathematics.* 2013;161(13–14):2025–2037. DOI: 10.1016/j.dam.2013.02.011.

18. Dorndorf U, Jaehn F, Pesch E. Modelling robust flight-gate scheduling as a clique partitioning problem. *Transportation Science.* 2008;42(3):292–301. DOI: 10.1287/trsc.1070.0211.

19. Wang H, Alidaee B, Glover F, Kochenberger G. Solving group technology problems via clique partitioning. *International Journal of Flexible Manufacturing Systems*. 2006;18(2):77–97. DOI: 10.1007/s10696-006-9011-3.

20. Aloise D, Cafieri S, Caporossi G, Hansen P, Perron S, Liberti L. Column generation algorithms for exact modularity maximization in networks. *Physical Review E.* 2010;82(4):046112. DOI: 10.1103/PhysRevE.82.046112.

Статья поступила в редколлегию 26.08.2019.
Received by editorial board 26.08.2019.

УДК 681.3

СИНТЕЗ РЕЧИ ТОНАЛЬНЫХ ЯЗЫКОВ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ НЕПРЯМЫХ МАРКЕРОВ И КОЛИЧЕСТВЕННОГО ПРИБЛИЖЕНИЯ ЦЕЛИ

***Т. Й. ТХАЙ*[1)], *Х. Н. ХУИ*[2)], *Д. В. ТУИЕТ*[3), 4)],**
***С. В. АБЛАМЕЙКО*[3)], *Н. В. ХУНГ*[5)], *Д. В. ХОА*[5)]**

[1)]*Ханойский университет бизнеса и технологий,*
ул. Вин Туи, 29А, Вин Туи, Хай Ба Трунг, г. Ханой, Вьетнам
[2)]*Университет электроэнергии Министерства промышленности и торговли Вьетнама,*
ул. Хоанг Куок Вьет, 235, 129823, Ко Нуэ, Ту Лием, г. Ханой, Вьетнам
[3)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*
[4)]*Университет Бинь Зыонг, пр. Бин Дуонг, 504, 820000, г. Тху Дау Мот, район Бин Дуонг, Вьетнам*
[5)]*Военный институт науки и техники, ул. Хоанг Сэм, 17, Нгхиа Ду, Кау Гиау, г. Ханой, Вьетнам*

Синтезирующие тоны играют важную роль в системах преобразования текста в речь тональных языков. Для этого необходимо выполнить два важных шага: определить маркеры высоты тона голосовых высказываний и синтезировать траектории F_0 для лексических тонов. В этой статье мы предлагаем два эффективных алгоритма, один из которых заключается в расположении маркеров высоты тона на пиках кумулятивного сигнала каждой озвученной части входного высказывания, а другой – в генерации F_0-траекторий тонов с количественными параметрами приближения цели (qTA). Эксперимент показал, что предложенные алгоритмы представляют маркеры высоты звука с высокой точностью, что позволило нам генерировать тоны со сложной формой.

Ключевые слова: маркеры основного тона; кумулятивный сигнал; модель Сюй; qTA; полиномиальное приближение.

Образец цитирования:
Тхай ТЙ, Хуи ХН, Туиет ДВ, Абламейко СВ, Хунг НВ, Хоа ДВ. Синтез речи тональных языков с использованием методов непрямых маркеров и количественного приближения цели. *Журнал Белорусского государственного университета. Математика. Информатика.* 2019;3:105–121 (на англ.).
https://doi.org/10.33581/2520-6508-2019-3-105-121

For citation:
Thai TY, Huy HN, Tuyet DV, Ablameyko SV, Hung NV, Hoa DV. Tonal languages speech synthesis using an indirect pitch markers and the quantitative target approximation methods. *Journal of the Belarusian State University. Mathematics and Informatics.* 2019;3:105–121.
https://doi.org/10.33581/2520-6508-2019-3-105-121

Авторы:
Та Йен Тхай – лектор на факультете информатики.
Хоан Нго Хуи – кандидат наук (информатика); заместитель декана факультета информатики.
Дао Ван Туиет – старший исследователь Центра биомедицинской информатики[4)]; аспирант кафедры веб-технологий и компьютерного регулирования механико-математического факультета[3)]. Научный руководитель – С. В. Абламейко.
Сергей Владимирович Абламейко – академик НАН Беларуси, доктор технических наук, профессор; профессор кафедры веб-технологий и компьютерного регулирования механико-математического факультета.
Нгуен Ван Хунг – кандидат наук (информатика); лектор на факультете информатики.
Доан Ван Хоа – кандидат наук (информатика); лектор на факультете информатики.

Authors:
Ta Yen Thai, lecturer at the faculty of informatics.
thaity@hubt.edu.vn
Hoang Ngo Huy, PhD (informatics); vice dean of the faculty informatics.
huynh@epu.edu.vn
Dao Van Tuyet, senior researcher at the Biomedical Informatics Center[d] and postgraduate student at the department of web-technologies and computer simulation, faculty of mechanics and mathematics[c].
daovt@bsu.by
Sergey V. Ablameyko, academician of the National Academy of Sciences of Belarus, doctor of science (engineering), full professor; professor at the department of web-technologies and computer simulation, faculty of mechanics and mathematics.
ablameyko@bsu.by
Nguyen Van Hung, PhD (informatics); lecturer at the faculty of informatics.
nvhnt73@gmail.com
Doan Van Hoa, PhD (informatics); lecturer at the faculty of informatics.
doanvanhoa@gmail.com

Журнал Белорусского государственного университета. Математика. Информатика. 2019;3:105–121
Journal of the Belarusian State University. Mathematics and Informatics. 2019;3:105–121

TONAL LANGUAGES SPEECH SYNTHESIS USING AN INDIRECT PITCH MARKERS AND THE QUANTITATIVE TARGET APPROXIMATION METHODS

***T. Y. THAI*[a], *H. N. HUY*[b], *D. V. TUYET*[c, d],**
***S. V. ABLAMEYKO*[c], *N. V. HUNG*[e], *D. V. HOA*[e]**

[a]*Hanoi University of Business and Technology, 29A Vinh Tuy Street, Vinh Tuy Ward, Hai Ba Trung Dist, Hanoi, Vietnam*
[b]*Electric Power University, Vietnam Ministry of Industry and Trade, 235 Hoang Quoc Viet Street, Co Nhue, Tu Liem, Hanoi 129823, Vietnam*
[c]*Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*
[d]*Binh Duong University, 504 Binh Duong Avenue, Thu Dau Mot Town 820000, Binh Duong Province, Vietnam*
[e]*Military Institute of Science and Technology, 17 Hoang Sam Street, Nghia Do Ward, Cau Giay District, Hanoi, Vietnam*
Corresponding author: D. V. Tuyet (daovt@bsu.by)

Synthesizing tones plays an important role in text-to-speech systems of tonal languages. To accomplish this, the two important steps are to determine the pitch markers of voice utterances and synthesize F_0 trajectories for lexical tones. In this paper, we propose two efficient algorithms, one of them is to locate the pitch markers at the peaks of the cumulative signal of each voiced part of the input utterance and the other is to generate F_0 trajectories of tones with quantitative target approximation (qTA) parameters of Xu model. The experimentation has shown that the proposed algorithms present pitch markers with high accuracy which has enabled us to generate tones with complex shapes.

Keywords: pitch markers; cumulative signal; Xu model; qTA; polynomial approximation.

Introduction

Nowadays, text to speech (TTS) systems and speech to text (STT) systems are increasingly used by the radiologist to create radiology study reports. Besides, TTS systems and STT systems can help people with disabilities integrate into the community by using computer easier. With the integration of the laboratory information system (LIS) and radiological information systems (RIS) patient identifiers and examination information can automatically map into examination reports. There are many potential benefits of report automation to radiologists including improvements in efficiency, accuracy, and fatigue [1].

Besides, TTS systems can help people with disabilities integrate into the community by using computer easier. For example the JAWS (job access with speech) software, is the world's most popular screen reader, developed for computer users whose vision loss prevents them from seeing screen content or navigating with a mouse. JAWS provides speech as an output for the most popular computer applications on your PC such as Microsoft Office, Web browsers etc. JAWS users around the world sent us videos about the impact JAWS has made on their lives [2].

Due to the application needs, the research on speech representation has been increasingly developed, the issues of research on estimation and modeling of fundamental frequency trajectories is still open research issues until now.

Frequency relates to the individual pulsations produced by vocal cord vibrations for a unit of time. The rate of vibration depends on the length, thickness, and tension of the vocal cords, and thus is different for child, adult male and female speech. A speech sound contains an important type of frequencies namely fundamental frequency (F_0) which relates to vocal cord function and reflects the rate of vocal cord vibration during phonation (pitch).

Pitch markers (PM) play a central role in phonetics signal analytic because pitch is a big part of hearing music, we can be tricky sounds without clear F_0. In addition PM is also useful for coding or representation for extracting information of speech for telephony and communication.

The fundamental frequency, pitch markers and hearing quality. The fundamental frequency is the primary element of speech signal and because the pitch marker indicates the beginning of each cycle of the waveform, PM plays a very important role in generating and recognizing speech sentences. However, pitch is an inherently subjective quantity and cannot be directly measured from the speech signal. It is a nonlinear function of the signal's spectral and temporal energy distribution. Therefore, PM estimation is an unsolved and challenging problem. It is one of the key technologies that determines the performance of speech processing.

Autocorrelation method or average magnitude difference function (AMDF) is commonly used. In addition, modified autocorrelation method [3] is also commonly used to compute the auto correlation instead of speech signal [2]. However, these methods suffer from error estimation in noisy environment. Robust algorithm for pitch tracking (RAPT) is well-known and widely used F_0 estimation method since it does offer low delay, low computational amount and robust against noise [4].

The YIN [5] algorithm uses a difference function based on the autocorrelation function as the candidate generator in conjunction with a number of optimization steps. Named after the oriental yin-yang principle of duality, it aims to balance between the autocorrelation and the cancelation that it involves.

The dynamic programming projected phase-slope algorithms (DYPSA) [6] was originally designed for automatic estimation of glottal closure instants (GCIs) in voiced speech but as a consequence also gives pitch information. The algorithm is based on an enhancement of the group delay algorithm [7] by R. Smits and B. Yegnanarayana, which is used as the primary candidate generator. DYPSA uses dynamic programming to identify the best GCI candidates by minimizing some cost functions. The DYPSA algorithm operates on the speech signal alone and does not require an electroglottography reference signal. The pitch estimate is derived from the inter GCI duration and mapped into frames.

F_0 trajectory representation and analysis-by-synthesis. From a modeling perspective, a model is of little use if it is not *predictive*. To make a model predictive, however, it is critical to first determine what the predictors should be. If, as suggested above, communicative functions like tone, focus and sentence type and their interactions are directly behind the complex surface F_0 trajectories in Mandarin, these communicative functions should then be the predictors. An alternative to such *functional modeling* is to simulate F_0 with predictors whose functional status is ambiguous, or whose definition includes characteristics of observed F_0 patterns, e. g., pitch accents, F_0 turning points, etc. From a theoretical perspective, functional modeling provides a powerful tool for hypothesis testing. That is, by assessing how well surface F_0 trajectories generated based on a set of hypothesized predictors, investigators can validate or falsify both general and specific theoretical assumptions about tone and intonation. Such a process is known as *analysis-by-synthesis* [8].

Parametric representation of speech often implies F_0 trajectory as a part of the model. There have been many attempts over the past decades to build a robust model capable of simulating various prosodic phenomena through F_0 modeling [9–12]. These approaches can be divided into two general categories, namely, those that model F_0 trajectories directly and those that attempt to simulate the underlying mechanisms of F_0 production. Models belonging to the first category are derived mainly based on the shape of the F_0 trajectories, with minimal consideration about the articulatory process of F_0 production.

The Fujisaki model is an effective model for approximating the trajectory of the fundamental frequency precisely for the source model of speech synthesis, representing the coarticulation of spectral frequencies making an equation for a target model of speech perception and so on [10–12].

Quantitative modeling is one of the most rigorous means of testing our understanding of a natural phenomenon. This is particularly true if the model is built directly on assumptions that closely reflect the contested view about the mechanisms underlying the phenomenon. Modeling can also help to improve our knowledge by forcing us to make our theoretical postulations as explicit as possible. Thus for improving our understanding of human speech, quantitative modeling is also indispensable. In the present paper we report the results of an attempt to simulate tone, stress, and focus in Mandarin and English with a quantitative model that generates surface F_0 trajectories through the process of target approximation TA [13]. qTA model for generating F_0 trajectories of speech. The qTA model simulates the production of tone and intonation as a process of syllable-synchronized sequential target approximation. It adopts a set of biomechanical and linguistic assumptions about the mechanisms of speech production. The communicative functions directly modeled are lexical tone in Mandarin and lexical stress in English and focus in both languages. The qTA model is evaluated by extracting function-specific model parameters from natural speech via supervised learning automatic analysis by synthesis and comparing the F_0 trajectories generated with the extracted parameters to those of natural utterances through numerical evaluation and perceptual testing. The F_0 trajectories generated by the qTA model with the learned parameters were very close to the natural trajectories in terms of root mean square error, rate of human identification of tone, and focus and judgment of naturalness by human listeners.

qTA and improving for generating F_0 trajectories of words with complex shape. In the detail, to generate F_0 trajectories of tones, we are able to use Xu model, which has been widely used for Mandarin [14; 15] to model the F_0 trajectories in the context:

$$f_0(t) \approx at + b + (ct^2 + dt + g)e^{-\lambda t}. \quad (1)$$

The linear function $t \mapsto a_{tn}t + b_{tn}$, called a «pitch target», reflects the tendency of the tone at the end of the F_0 trajectory.

The computational model used in the present study is the quantitative target approximation (qTA) model. This model simulates the production of tone and intonation as a process of syllable-synchronized sequential target approximation [15; 16]. Figure 2 illustrates the basic idea of target approximation [15]. The qTA model represents F_0 as the surface response of the target approximation process which is driven by pitch targets. A pitch target is a forcing function representing the joint force of the laryngeal muscles that control vocal fold tension. It is represented by a simple linear equation $x(t) = a^*t + b$ given by the formula (1).

Compared to Mandarin, Thai and Vietnamese tones have more complex F_0 shapes [17–20], thus the representation formula (1) should be replaced with one that can better model such complex tones. In [21], the authors present a Thai tone model based on qTA method. However, the result of the authors still has some limitations, namely:

(L1) there are no numerical computation methods for estimating automatically the coefficients of each component of the model by fitting methods.

Besides, lack of mathematical foundation to explain the use of second order polynomials in the qTA model. It is not easy to solve (L1) above because a suitable trajectory must satisfy following two conditions, given a sample of fundamental frequency trajectory of the tone:

(C1) pitch target constraint (PTC), with big enough time *t*;

(C2) fitting constraint, for any time *t*.

In this paper, we propose new computational methods to determine the pitch markers of the original speech signal based on its cumulative signal and quantitative target approximation vectors namely qTA that generate the fundamental frequency trajectories of two-syllable tones. Our methods include three numerical solutions. For the first solution, we determine the pitch markers of the original speech signal in a time domain based on its cumulative signal. The second ones is proposed to calculate the qTA parameters by fitting a given F_0 trajectory of a speech syllable. This numerical solution is a tool for determining qTA parameters by fitting a given F_0 trajectory of a speech syllable and of a multi-syllable word. The third ones calculates qTA parameters by fitting a given F_0 trajectory of a multi-syllable word with the first step is the concatenating each F_0 trajectory of each speech syllable of the given speech two-syllable word to archieve a continuous F_0 trajectory and the second step is according to each syllable, calculating qTA parameters of the its part of the F_0 trajectory by applying the second solution.

By using polynominals for the approximation component of qTA model, qTA parameters obtained by the second solution already generates a F_0 trajectory with fitting a given complex shape F_0 trajectory of a mutli syllable word is better than the results published in [16; 20]. The target and plynominal's coefficients are namely qTA vector parameters or qTA representation. By the well-known Weierstrass approximation theorem, any given F_0 trajectory of word is fitting by synthesized trajectories based on qTA parameters. In addition, it should be emphasized that qTA's parameter calculation is completely automated.

The rest of the paper is organized as follows. Section 2 presents about RAPT framework, Fujíaki model and qTA model. Section 3 presents an algorithm to determine the pitch markers of the original voice signal in a time domain based on the cumulative signal. This section also presents two algorithms to solve (L1) at one and two-tone levels respectively. Experimental results are given in section 4. Conclusions and future research direction are in section 5.

Theoretical basis

RAPT framework and instantaneous pitch estimation. This issue requests to develop algorithms in determining parameters for representing fundamental frequency trajectories of word tones of the tonal languages such as Vietnamese, Mandarin or Thai and so on.

In the tonal languages, by distinguishing the meaning of a syllable and by tone sandhi in which the tones assigned to individual syllables change based on the pronunciation of adjacent syllables, one of the basic parameters of speech is PM and the parameters generating the fundamental frequency trajectory of the word.

For determining PMs and calculating the fundamental frequencies of speech samples, there are many algorithms in the literal, such as the results published in [4; 6; 22–25].

Most of these results follow an approach with three main steps: (i) divide a voiced segment into short segments (frames), (ii) estimate the fundamental frequency value at each frame and (iii) use dynamic programming algorithms to determine the PMs taken from the peak, or valley points of the speech signal and so on.

In details, RAPT estimates overall periodicity of the analysis frame using normalized cross-correlation function (NCCF). Let $s(m)$ be a speech signal, z – step size in samples and n – window size. The NCCF $\varphi(x, k)$ of K samples length at lag k and analysis frame x is defined as [4]:

$$\varphi(x,k)=\frac{\sum_{i=m}^{m+n-1}S(i)S(i+k)}{\sqrt{e_m e_{m+k}}},$$

$$k=0,K-1;\ m=xz;\ x=0,M-1,$$

where $e_i=\sum_{l=i}^{i+n-1}s_l^2$.

The Fujisaki model is a super positional model for representing F_0 trajectory of speech. According to the model, F_0 trajectory is generated as a result of the superposition of the outputs of two second order linear filters with a base frequency value. The second order linear filters are for generating the phrase and accent components of speech. The base frequency is the minimum frequency value of the speaker. In other words, F_0 trajectory is obtained by adding base frequency, phrase components and accent components.

Fujisaki model has many parameters which described in the below formula, and currently, there is no numerical method to solve fitting problems when knowing a trajectory in advance.

$$\log F_0(t)=\log F_b+\sum_{i=1}^{I}A_{pi}G_p(t-T_{0i})+\sum_{j=1}^{J}A_{ai}\left\{G_a(t-T_{1j})-G_a(t-T_{2j})\right\},$$

$$G_p=\begin{Bmatrix}\alpha^2 t\exp(-\alpha)\text{ for }t\ge 0\\ 0\text{ for }t<0\end{Bmatrix},$$

$$G_a=\begin{Bmatrix}\min\left[1-(1+\beta t)\exp(-\beta t)\right]\text{ for }t\ge 0\\ 0\text{ for }t<0\end{Bmatrix},$$

where F_b – baseline value of fundamental frequency; I – number of phrase commands; J – number of accent commands; A_{pi} – magnitude of i phrase command; T_{0i} – timing of i phrase command; A_{aj} – amplitude of j accent command; T_{1j} – onset of j accent command; T_{2j} – offset of j accent command; α – natural angular frequency of the phrase control mechanism; β – natural angular frequency of the accent control mechanism; γ – relative ceiling level of accent components.

The NCCF is the most computationally expensive operation in RAPT and so the algorithm performs the NCCF in a two pass process. A down-sampled version of the input signal issued to estimate the first set of candidate peaks, followed by a high resolution (full sample rate) NCCF around the candidates of interest.

The algorithm is summarized below:

• periodically compute the NCCF of the down sampled signal for all lags in the range of pitch. Location so flocal maxima in this 1st pass of the NCCF are recorded;

• compute the high resolution NCCF (signal at original sampling frequency) only around the peak locations recorded in previous step;

• search for local maxima in the high resolution NCCF to obtain improved peak locations and amplitude estimates;

• dynamic programming is used to select the set of NCCF peaks or unvoiced hypothesis across all frames.

Fujisaki model. In the Fujisaki model, as illustrated in the fig. 1, the shapes of local F_0 peaks and global F_0 trends are modeled as the on- and off-ramps of step and pulse responses of a second-order linear system. These responses are assumed to be associated with accent and phrase commands that are linguistically meaningful. Thus the commands, as the hypothetical underlying components of intonation, are different from the surface F_0 trajectories. And the latter are the product the underlying commands generated by the articulatory mechanism implemented in the model. The surface F_0 trajectories are generated by a mechanism that compromises between maximum smoothness and full realization of the underlying tonal templates. Fujisaki model is also available for generating intonation trajectories of any language such as Russian, English, Vietnamese and so on. However, it is a complex model with a lot of parameters.

The qTA model is presented on the fig. 2, which will be detailed in the next section, simulates F_0 trajectories as syllable-synchronized laryngeal movements toward underlying pitch targets that are either static or dynamic.

Thus all these models assume that surface F_0 trajectories result from certain articulatory mechanisms rather than from direct acoustic manipulations.

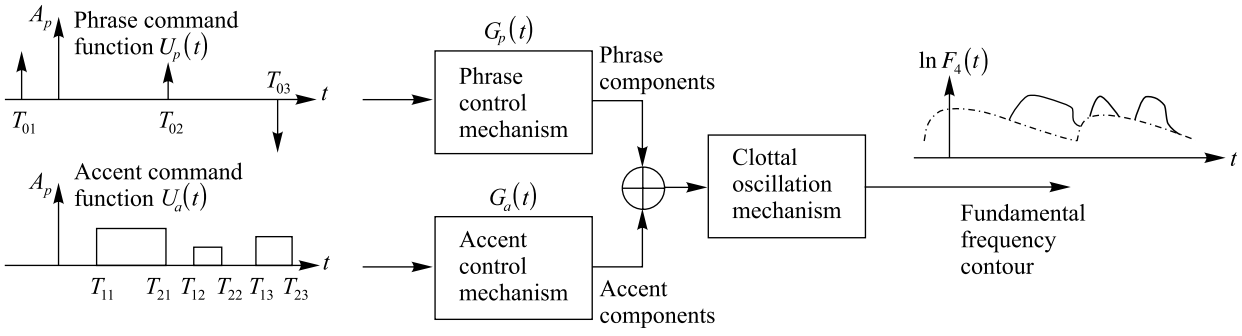


Fig. 1. Fujisaki model

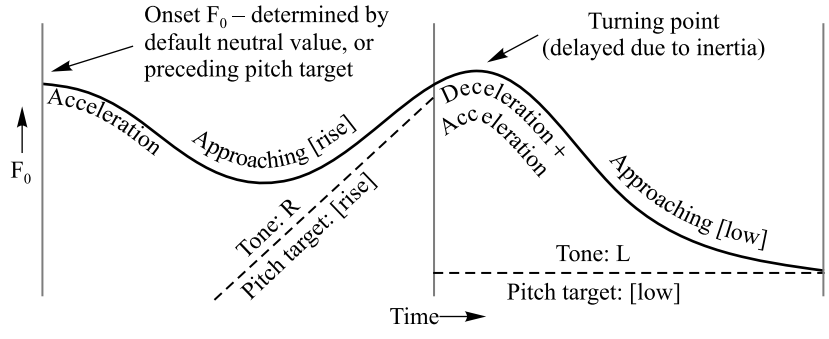


Fig. 2. The qTA model

qTA model (Xu model). In the phrase context, by the tone sandhi occuring, the number of trajectory shapes of syllables is increasing many times over the isolated syllables. Therefore, it is not easy to model these variations.

In the tonal languages, for parameterizing fundamental frequency trajectories of speech utterances, it is usual to use the Fujisaki or Xu models. For example, in [20] Hansjoerg Mixdorf and his colleagues already used the Fujisaki model to model Vietnamese fundamental frequency trajectories of syllables in the phrase context.

In the Fujisaki model, fundamental frequency trajectories are formed from the intonation trajectories and the stress trajectories. This can lead to a change in the shape of the original tone in tones, such as flat tone being converted to another tone with the fundamental frequency value falling down due to the influence of the intonation trajectories. In addition, the Fujisaki model requires a lot of parameters to represent the fundamental frequency trajectories. Therefore, it is not easy to calculate Fujisaki model parameters by fitting the given fundamental frequency trajectory and until now there are no numerical computation methods to extrace the parameters by fitting methods.

Tones can be analyzed into two components frequently combined: the pitch (the height of the base bar, referred to as the static characteristic) and the tone (direction of the high-frequency change, called dynamic features) in the process of expression. Thus, each tone can be described as a combination of the two.

The static and dynamic characteristics can be modeled using the «pitch target» concept of the Xu model [6]. This is a model that has been investigated and used by Xu and his colleagues to generate fundamental frequency trajectories for tonal languages such as Mandarin and Thai, for example Prom-on and Yi Xu [24; 26]. Advantages of the model are simple, less parameters and can be learned statistically to generate the appropriate fundamental frequency trajectories representation. About recent results using qTA representations of Xu model can be read in [21; 27; 28].

The F_0 control is implemented through a third order critically damped linear system, in which the total response is the remain component given by formula (1), where the first term $x(t)$ is the forced response of the system which is the pitch target and the second term is the natural response of the system. The transient coefficients *c*, *d* and *g* are calculated based on the initial F_0 dynamic state and the pitch target of the specified segment. The parameter λ represents the strength of the target approximation movement. In qTA, the initial F_0 dynamic state consists of initial F_0 level, $f_0(0)$, velocity $f_0'(0)$, and acceleration $f_0''(0)$. The dynamic state is transferred from one syllable to the next at the syllable boundary to ensure continuity of F_0. The three transient coefficients are computed with the formula presented on the fig. 2.

Proposed method for determining PMs and qTA representation

In this section, we propose a pitch mark detection algorithm for utterances and F_0 trajectories generation algorithms for tones in a tonal language.

PMs with cumulative signals. Let $x = \{x_j\}_{1 \le j \le N}$ be a voiced segment, without loss of generality, we assumed that the signal x is sampled from an interval $[-a, a]$ with some $a > 0$. The cumulative signal $s = \{s_j\}_{1 \le j \le N}$ of x can be defined by

$$s_1 = x_1,\ \forall j = \overline{2, N},\ s_j \overset{\text{def}}{=} s_{j-1} + x_j = \int_1^j x_t dt.$$

Example 1. Consider the following utterance extracted in the Vietnamese book «Adventures of a Cricket», where PMs (marked by small circles) of the original speech are located at the signal points changing from positive to negative, as the peaks of the cumulative signal respectively, this case is described by the fig. 3.

As we can see, there is a relationship between signal points changing from positive to negative and the peaks of the correspondent cumulative signal by the following fig. 4.

For the following utterance, it is divided automatically into 7 voiced segments by using a method for locating the silience/voiced/unvoiced part (see [21]), each segment is shown in a pair of red-blue dashed lines, this case described by the fig. 5.

The peaks of the cumulative signal are more visible than the peaks of the original voice signal as illustrated in fig. 6, *b*.

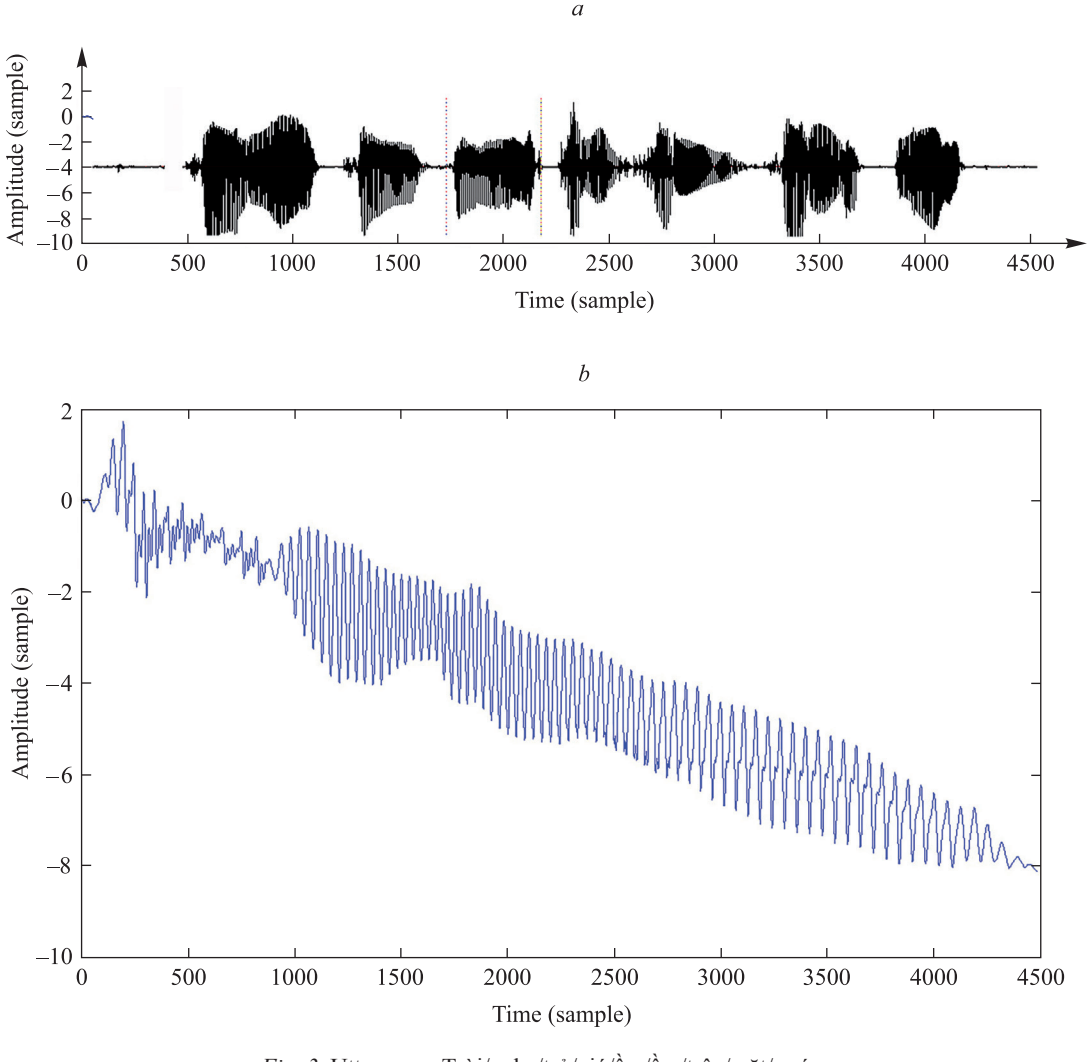


Fig. 3. Utterance «Trời/nghe/trở/gió/ầm/ầm/trên/mặt/nước»
(IPA transcription: «tɕəː˨˩ ŋɛ˧˧ tɕəː˧ zɔ˧˩ ə̰m˨˩ ə̰m˨˩ tɕen˧˧ mạʔt˨˩ niək˧˩»;
translation: «God make the rumbling wind on the water») (*a*)
and the cumulative signal of the voiced part /gió/(/zɔ˧˩/, /wind/) (*b*)

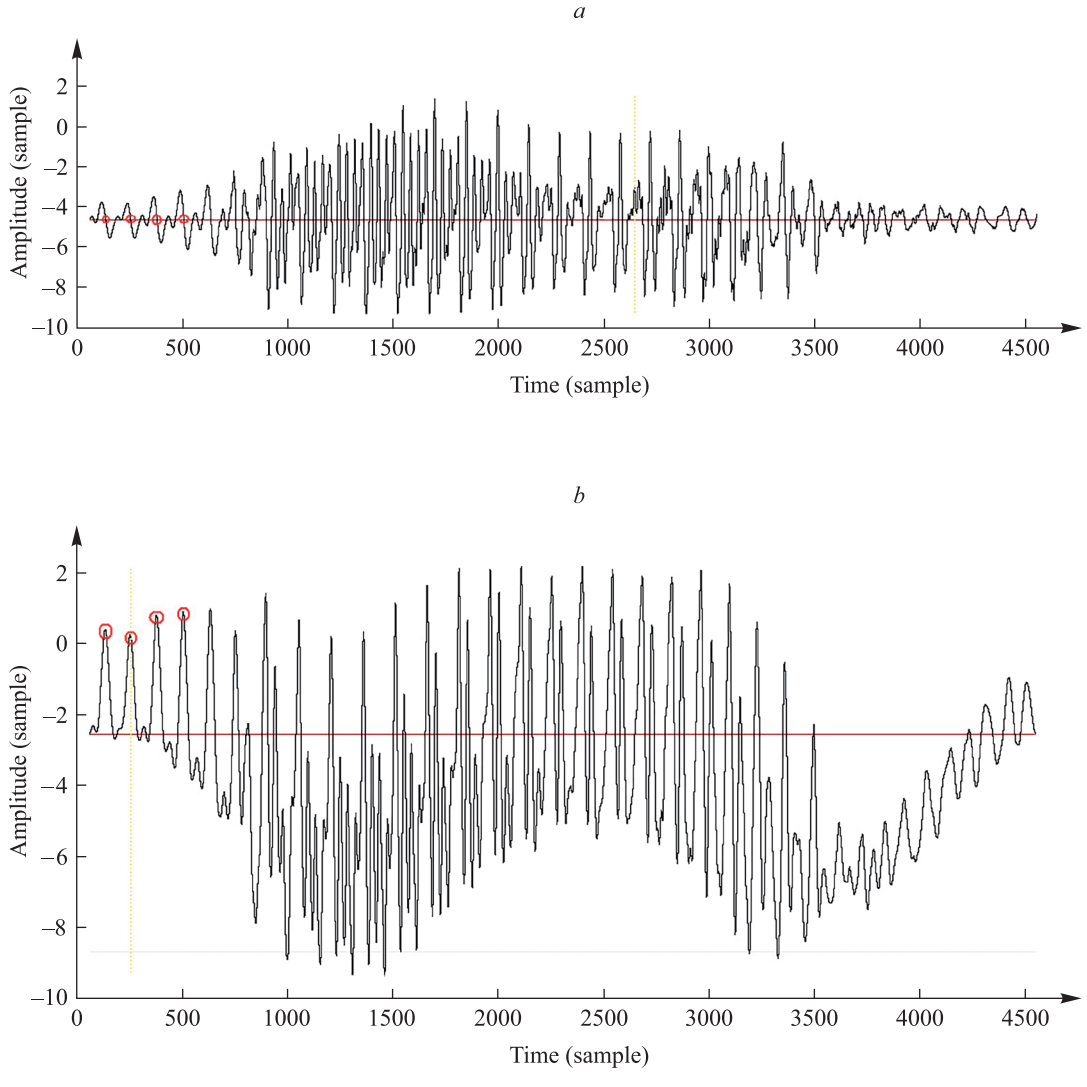


Fig. 4. PMs are located at the points
in which the voice signal changes from positive to negative (*a*)
and corresponding peaks of the cumulative signal (*b*)

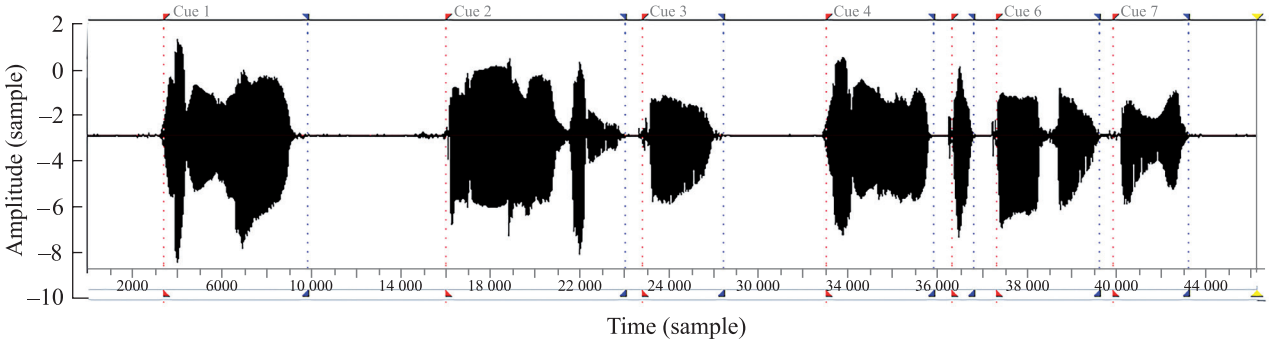


Fig. 5. «Đừng lo, xem mây vận trời đêm nay có cơ đổi gió»
(IPA transcription: «ɗi̤ŋ˨ lɔ˧˧ sɛm˧˧ məj˧˧ və̰ʔn˨ tɕə̰ːj˨ ɗem˧˧ naj˧˧ kɔ˧ kəː˧˧ ɗo̰j˨ zɔ˧»;
translation: «Do not worry, looking at the clouds, the wind may change direction tonight»)

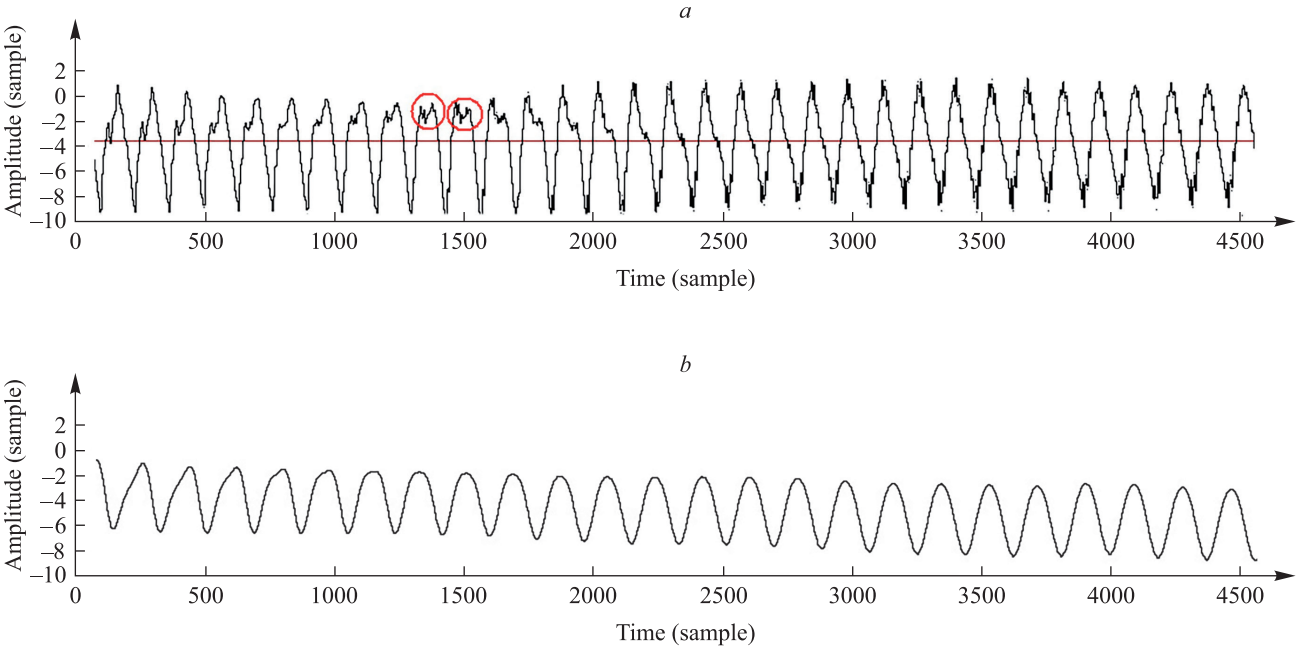


Fig. 6. A partial signal of the second segment (*a*) and corresponding cumulative signal (*b*)

The PMs located at peaks of the original voice signal are indistinguishable in amplitude from points surrounding them. In the details, the peaks having higher amplitude than the surrounding ones, usually also are PMs of the cumulative signal.

The peaks of the cumulative signal are related to the time points at which the original signal changes from positive to negative. This is the principle that if the PMs of the cumulative signal are well positioned, we will successfully locate PMs from the peak or valley points of the original voiced signal.

First of all, we will give some definitions and prove some simple properties derived from them.

Definition 1. (The sets of time points at which the original signal changes from positive to negative and vice versa.)

For $x = \{x_j\}_{1 \le j \le N}$, we let z_x^+ and z_x^- denote two sets of time points as the following:

$$z_x^+ \stackrel{\text{def}}{=} \{i \mid x_i > 0 \wedge x_{i+1} < 0\},\ z_x^- \stackrel{\text{def}}{=} \{i \mid x_i < 0 \wedge x_{i+1} > 0\}.$$

In addition, we also denote

$$x^+ \stackrel{\text{def}}{=} \{i \mid x_i > 0\}$$

and

$$x^- \stackrel{\text{def}}{=} \{i \mid x_i < 0\}$$

and $\text{peak}(\mathbf{x})$ denote the peak set of $\mathbf{x}$ (to get $\text{peak}(\mathbf{x})$, see [28]).

Proposition 1. *(i)* $\text{peak}(s) \subset z_x^+$ *and* $\text{peak}(-s) \subset z_x^-$, *where s is the cumulative signal of x.*

(ii) If $\overline{i, j} \subset x^+$ *then* $\{s_i, s_{i+1}, \ldots, s_j\}$ *is a monotonic increasing sequence, and if* $\overline{i, j} \subset x^-$ *then* $\{s_i, s_{i+1}, \ldots, s_j\}$ *is a monotonic decreasing sequence.*

Proof.

(i) $\forall i \in \text{peak}(s) \Rightarrow s_i > s_{i-1} \wedge s_i > s_{i+1} \Rightarrow (x_i = s_i - s_{i-1} > 0) \wedge (x_{i+1} = s_{i+1} - s_i < 0) \Rightarrow x_i > 0 \wedge x_{i+1} < 0 \Rightarrow$

$\Rightarrow i \in z_x^+$. So $\text{peak}(s) \subset z_x^+$. Similarly, we have $\text{peak}(-s) \subset z_x^-$.

(ii) $\overline{i, j} \subset x^+ \Rightarrow \forall k = \overline{i, j-1},\ x_k > 0 \Rightarrow s_{k+1} - s_k = x_{k+1} > 0 \Rightarrow s_{k+1} > s_k$.

Moreover, $\overline{i, j} \subset x^- \Rightarrow \forall k = \overline{i, j-1},\ x_k < 0 \Rightarrow s_{k+1} - s_k = x_{k+1} < 0 \Rightarrow s_{k+1} < s_k$.

From here, we propose a new approach, instead of locating PMs based on the original speech wave, we determine PMs in the timing of peaks of the cumulative signal of the speech. From the PMs of the cumulative signal we will locate the other PMs, such as the PMs located from the peaks or valleys of the speech signal.

Definition 2. Let $x = \{x_j\}_{1 \le j \le N}$ be a voiced segment and $s = \{s_j\}_{1 \le j \le N}$ the cumulative signal of **x**. Let denote pitch marker zeros (PMZ), $\text{PMZ}_x^+ = \{pmz_j^+\}$ as the given PMs which located from the peaks of **s**. We let PMZ_x^-, PM_x^+ and PM_x^- denote three PM sets derived from PMZ_x^+ (find in each range of two consecutive PMs of PMZ_x^+) as the following:

$$\text{PMZ}_x^- \stackrel{\text{def}}{=} \left\{k/\exists j : k = \min\left\{l/l \in \text{peak}(-s),\ pmz_{j-1}^+ \le l \le pmz_j^+\right\}\right\},$$

$$\text{PM}_x^+ \stackrel{\text{def}}{=} \left\{k/\exists j : k = \min\left\{l/l \in \text{peak}(x),\ pmz_{j-1}^+ \le l \le pmz_j^+\right\}\right\},$$

$$\text{PM}_x^- \stackrel{\text{def}}{=} \left\{k/\exists j : k = \min\left\{l/l \in \text{peak}(-x),\ pmz_{j-1}^+ \le l \le pmz_j^+\right\}\right\}.$$

Let s_k be the cumulative signal of x_k, where x_k is the k voiced segment of the utterance. To determine PMZ_k^+ of s_k, we can see that the PMs are chosen based on the following two criteria:

(i) the dependencies of the distances between consecutive PMs;

(ii) with two adjacent peaks of s_k, the peak with a greater amplitude is preferred over the other.

The process of selecting the appropriate peaks of s_k is a looping, multi-step process, consisting of appends, deletions, insertions and modifications to ensure that the criteria (i) and (ii) described above do not create redundancy and lost of PMs. With that said, we propose a simple and intuitive R1–R6 rules, to determine $\text{PMZ}_{k,x}^+ = \left\{p_{k,n} \middle| p_{k,n} \in \text{peak}(s_k)\right\}$ for s_k.

R1. (Appending the first PM.)

$\text{PMZ}_{k,x}^+ = \{p_{k,1}\}$, where $\text{mean}_k \stackrel{\text{def}}{=} \dfrac{\sum\limits_{n \in \text{peak}\{s_k\}} |s_{k,n}|}{\#\text{peak}\{s_k\}}$, $p_{k,1} = \underset{n \in \text{peak}\{s_k\}}{\arg\min}\left\{|s_{k,n}| \ge \text{mean}_k\right\}$ (the first PM $p_{k,1}$ of s_k is the first n peak whose amplitude $s_{k,n}$ is over the threshold mean_k).

R2. (Appending the next temporary PM.)

If there exist some $m \in \text{peak}(s_k)$ and $m - p_{k,j} \in \left[f_s/f_{0,\max},\ f_s/f_{0,\min}\right]$, where $p_{k,j} = \max\left\{\text{PMZ}_{k,x}^+\right\}$ then $\text{PMZ}_{k,x}^+ = \text{PMZ}_{k,x}^+ \cup \{m\}$.

R3. (Delete a temporary PM.)

If there exist some two consecutive temporary PMs, $p_{k,j-1}, p_{k,j} \in \text{PMZ}_{k,x}^+$ such that $p_{k,j} - p_{k,j-1} \notin \left[f_s/f_{0,\max},\ f_s/f_{0,\min}\right]$, then $\text{PMZ}_{k,x}^+ = \text{PMZ}_{k,x}^+ \setminus \{p_{k,j}\}$.

R4. (Delete a temporary PM.)

If there exist some three consecutive temporary PMs $p_{k,j-1}, p_{k,j}, p_{k,j+1} \in \text{PMZ}_{k,x}^+$ such that $s_{k,p_{k,j}} < \min\left\{s_{k,p_{k,j-1}}, s_{k,p_{k,j+1}}\right\} \wedge \min\left\{p_{k,j} - p_{k,j-1},\ p_{k,j+1} - p_{k,j}\right\} < 0.5^* \max\left\{p_{k,j} - p_{k,j-1},\ p_{k,j+1} - p_{k,j}\right\}$, then $\text{PMZ}_{k,x}^+ = \text{PMZ}_{k,x}^+ \setminus \{p_{k,j}\}$.

R5. (Insert a peak into the temporary PM set.)

If there exist some three consecutive temporary PMs $p_{k,j-1}, p_{k,j}, p_{k,j+1} \in \text{PMZ}_{k,x}^+$ such that $p_{k,j+1} - p_{k,j} > \alpha^*\left(p_{k,j} - p_{k,j-1}\right)$ then $\text{PMZ}_{k,x}^+ = \text{PMZ}_{k,x}^+ \cup \{m\}$, where m is $m \in \text{peak}(S_k)$: $p_{k,j} < m < p_{k,j+1}$, $\left|m - \left(p_{k,j} + p_{k,j+1}\right)/2\right| \to \min$ and α is an experimental parameter, $\alpha > 1$.

R6. (Replace value of a temporary PM.)

If there exist some three consecutive temporary PMs $p_{k,j-1}, p_{k,j}, p_{k,j+1} \in \text{PMZ}_{k,x}^+$ such that $\exists m \in \text{peak}\left(S_{\left[p_{k,j} - T/2,\ p_{k,j} + T/2\right]}\right) \wedge \left(\forall t \in \left[p_{k,j} - T/2,\ p_{k,j} + T/2\right] \Rightarrow s_{k,t} \le s_{k,m}\right)$ then reassign $p_{k,j} = m$, where $T \stackrel{\text{def}}{=} \min\left\{p_{k,j} - p_{k,j-1},\ p_{k,j+1} - p_{k,j}\right\}$.

Using R1–R6 rules, the proposed algorithm determining the PMs based on the cumulative signals includes some simple main steps, it is described by the fig. 7.

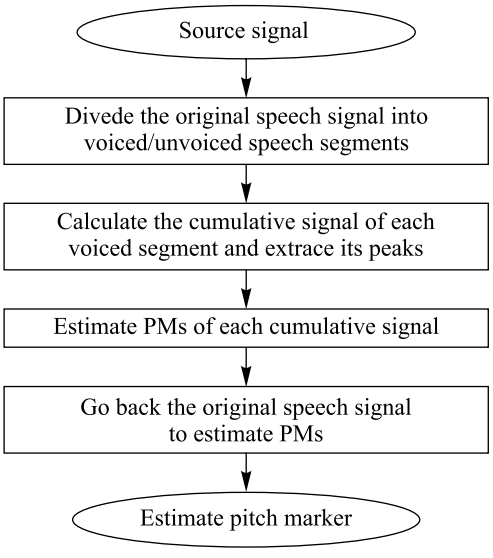


Fig. 7. Scheme of estimate PMs of a speech utterance

The algorithm of EPM is given as follows.

Algorithm 1. EPM (Estimating the PMs of speech waves.)
Input: speech signal $\{x_m\}_{1 \le m \le N}$ in time domain.

Sampling frequency value: f_s, $\left[f_{0,\min}, f_{0,\max}\right]$ is the range of F_0 values.

Output: number of voiced segments K, PMs according to four types

$$\left\{pm^{+}_{k,j}\right\}_{1 \le k \le K,\, 1 \le j \le n^{+}_{k}}, \left\{pm^{-}_{k,j}\right\}_{1 \le k \le K,\, 1 \le j \le n^{-}_{k}}, \left\{p^{-}_{k,j}\right\}_{1 \le k \le K,\, 1 \le j \le n^{-}_{k,2}}, \left\{p^{+}_{k,j}\right\}_{1 \le k \le K,\, 1 \le j \le n^{+}_{k,2}},$$

where $\left\{pm^{+}_{k,j}\right\}_{1 \le k \le K,\, 1 \le j \le n^{+}_{k}}$, $\left\{pm^{-}_{k,j}\right\}_{1 \le k \le K,\, 1 \le j \le n^{-}_{k}}$ are the two traditional PMs.

Step 1: segment the signal $\{x_m\}_{1 \le m \le N}$ into K voiced segments, $\{x_m\}_{N_{k,1} \le m \le N_{k,2}}$ and other ones (a simple method, see [14]).

Step 2: $T_{\min} = f_s / f_{0,\max}$, $T_{\max} = f_s / f_{0,\min}$.

Step 3: repeat, on each voiced segment $\{x_m\}_{N_{k,1} \le m \le N_{k,2}}$, $k = \overline{1, K}$ to determine $\mathrm{PMZ}^{+}_{x,k}$:

3.1: calculates the cumulative signal $s_k = \{s_m\}_{N_{k,1} \le m \le N_{k,2}}$, $k = \overline{1, K}$ following the formula (1).

3.2: determine the peak of s_k, compute the average amplitude at the peak of s_k:

$$\mathrm{mean}_k = \sum_{n \in \mathrm{peak}\{S_k\}} \left|s_{k,n}\right| / \#\mathrm{peak}\left\{s_{k,n}\right\}.$$

3.3: determine the first PM of $\mathrm{PMZ}^{+}_{x,k}$ by using rule R1.
3.4: repeat the substeps 3.5–3.8 when at least one of the conditions of the rules R2 – R6 is true.
3.5: using the rule R2 to extend $\mathrm{PMZ}^{+}_{x,k}$.
3.6: using the rules R3 and R4 to reduct $\mathrm{PMZ}^{+}_{x,k}$.
3.7: using the rule R5 to extend $\mathrm{PMZ}^{+}_{x,k}$.
3.8: using the rule R6 to change the element values of $\mathrm{PMZ}^{+}_{x,k}$.
3.9: stop and obtain $\mathrm{PMZ}^{+}_{x,k}$.
3.10: determine the PMs according to the local maximum point criterion of k segment $\{x_m\}_{N_{k,1} \le m \le N_{k,2}}$.

For each range of two consecutive PMs of PMZ^{+}_{x}, find $pm^{+}_{k,j} = \max\left\{\mathrm{peak}\left\{x_n\right\}_{p_{k,j} \le n \le p_{k,j+1}}\right\}$ and obtain $\mathrm{PM}^{+}_{x,k} = \left\{pm^{+}_{k,j}\right\}$.

3.11: determine PMs according to the local minimum point criterion of k segment $\{x_m\}_{N_{k,1} \le m \le N_{k,2}}$.

For each range of two consecutive PMs of PMZ_x^+, find $pm_{k,j}^- = \min\left\{\mathrm{peak}\left\{-x_n\right\}_{p_{k,j} \le n \le p_{k,j+1}}\right\}$ and obtain $\mathrm{PM}_{x,k}^- = \left\{pm_{k,j}^-\right\}$.

Step 4: determine PMs like pulse points for Praat type [24], the same as step 3 above, but taking the local minimum point of the cumulative signal, obtain $\mathrm{PMZ}_{x,k}^- = \left\{p_{k,j}^-\right\}$ on k segment $\left\{x_m\right\}_{N_{k,1} \le m \le N_{k,2}}$, $k = \overline{1, K}$.

Step 5: put $\mathrm{PMZ}_x^+ = \bigcup\limits_{1 \le k \le K} \mathrm{PMZ}_{x,k}^+$, $\mathrm{PMZ}_x^- = \bigcup\limits_{1 \le k \le K} \mathrm{PMZ}_{x,k}^-$, $\mathrm{PM}_x^+ = \bigcup\limits_{1 \le k \le K} \mathrm{PM}_{x,k}^+$ and $\mathrm{PM}_x^- = \bigcup\limits_{1 \le k \le K} \mathrm{PM}_{x,k}^-$.

Return: number of voiced segments K, PMZ_x^+, PMZ_x^-, PM_x^+ and PM_x^-.

After obtaining PMs of tonal word speech signals, the next step is to stylize F_0 trajectory of the tones and finally use an algorithm such as the PSOLA [29] to create the desired speech word from multiple input syllables.

The following proposed algorithms will focus on generating F_0 trajectories of tones by using the pitch target model.

Generating F_0 trajectories of Vietnamese isolated syllables. We will apply the method to identify PMs to synthesize tones of Vietnamese isolated syllables. To stylized tones, we use Xu model, which has been widely used for Mandarin [30] to model F_0 contours of the tones (for tonal languages). $F(t) \approx \alpha^* e^{-\lambda t} + a^* t + b$ such that a F_0 contour is created from the combination of the two components: the linear approximation $\alpha^* t + b$ and the non-linear approximation $\alpha^* e^{-\lambda t}$.

The computing of the coefficients of the model, given trend-line F_0 value also uses the least squares method, instead of finding the coefficients a, b, α, λ we determine a, b, k ($k = e^{-\lambda}$) by minimize the objective function:

$$\sum_{i=1}^{n-1}\left(F_{0,i+1} - a^*(i+1) - b - k^*\left(F_{0,i} - a^* i - b\right)\right)^2 \to \min, \tag{2}$$

where n is the number of speech frames, $\left\{F_{0,i}\right\}_{i=1}^{n}$ is a F_0 sequence of each frame corresponding. The stylized method using Xu model is built as follows.

Step 1: select syllables with level tone, drop tone with syllables ending *p-t-c*/*ch*, determine F_0 trajectory of them.

Step 2: determine the PMs of this wave of tone by algorithm 1.

Step 3: using least squares method to fit Xu model's parameters as a, b, k. Generate target F_0 trajectory by the Xu model.

Step 4: using PSOLA algorithm to synthesize a syllable with the target tone.

The algorithm of synthesis of tones is given as follows.

Algorithm 2. (Synthesis of tone for a Vietnamese syllable signal.)

Input: voice signal x_{in} in time domain of a Vietnamese syllable with any given tone {level, falling, raising, drop, curve, broken}. Sampling frequency value f_s.

Parameters [a_{tn}, b_{tn}, c_{tn}, d_{tn}, g_{tn}, k_{tn}] represent the target tone *tn* belong to {level, falling, raising, drop, curve, broken}. Need to synthesize in the form of qTA in formula (2), $0 < k_{tn} < 1$.

$\Delta > 0$ is the width parameter of the frame with measure units of milliseconds, N, M are the length of input syllable length and synthesis syllable, the calculating unit is milliseconds.

Output: x_{out}, the sound wave has the tone *tn*.

Step 1: use the value of f_s, convert N, M, Δ to the number units of sample.

Step 2: determine the set $\mathrm{PM}_{\mathrm{in}}^+$ (starting assign $\mathrm{PM}_{\mathrm{in}}^+(0) = 0$) of input sound waves using the proposed algorithm 1. Notice that on the unsound we assign:

$$\mathrm{PM}_{\mathrm{in}}^+(k) = \mathrm{PM}_{\mathrm{in}}^+(k-1) + \Delta,\ k \in (1, N_{\mathrm{PM}}).$$

Step 3: generating F_0 trajectory of target syllables using formula (2), concretely calculated as follows:

$$f_{0,\mathrm{out}}(t) = a_{tn}^* t + b_{tn} + (k_{tn})^{t^*}\left(c_{tn} t^2 + d_{tn} t + g_{tn}\right),\ t = \overline{1, T_{\mathrm{out}}},$$

where $T_{\mathrm{out}} = [M/\Delta]$.

Step 4: determine the set $\mathrm{PM}_{\mathrm{out}}$ as the following formula:

$$PM_{\mathrm{out}}(0) = 0,\ PM_{\mathrm{out}}(k) = PM_{\mathrm{out}}(k-1) + f_s / f_{0,\mathrm{out}}(k/\Delta),\ k = \overline{1, N_{\mathrm{out}}},$$

where $N_{\mathrm{out}} = \max\left\{k : PM_{\mathrm{out}}(k) \le M\right\}$.

Step 5: use the algorithm PSOLA [29] getting:

$$x_{\text{out}} = \text{PSOLA}\left(x_{\text{in}}, PM^{+}_{\text{in}}, PM_{\text{out}}\right).$$

Output: wave signal x_{out} syllable has new tone is *tn*.

Experiment

In order to experiment with proposed algorithms, we use Vietnamese voice data to illustrate. The Vietnamese language is a monosyllabic and tonal language with six tones (see table) and is the most complex lexical tone in tonal languages.

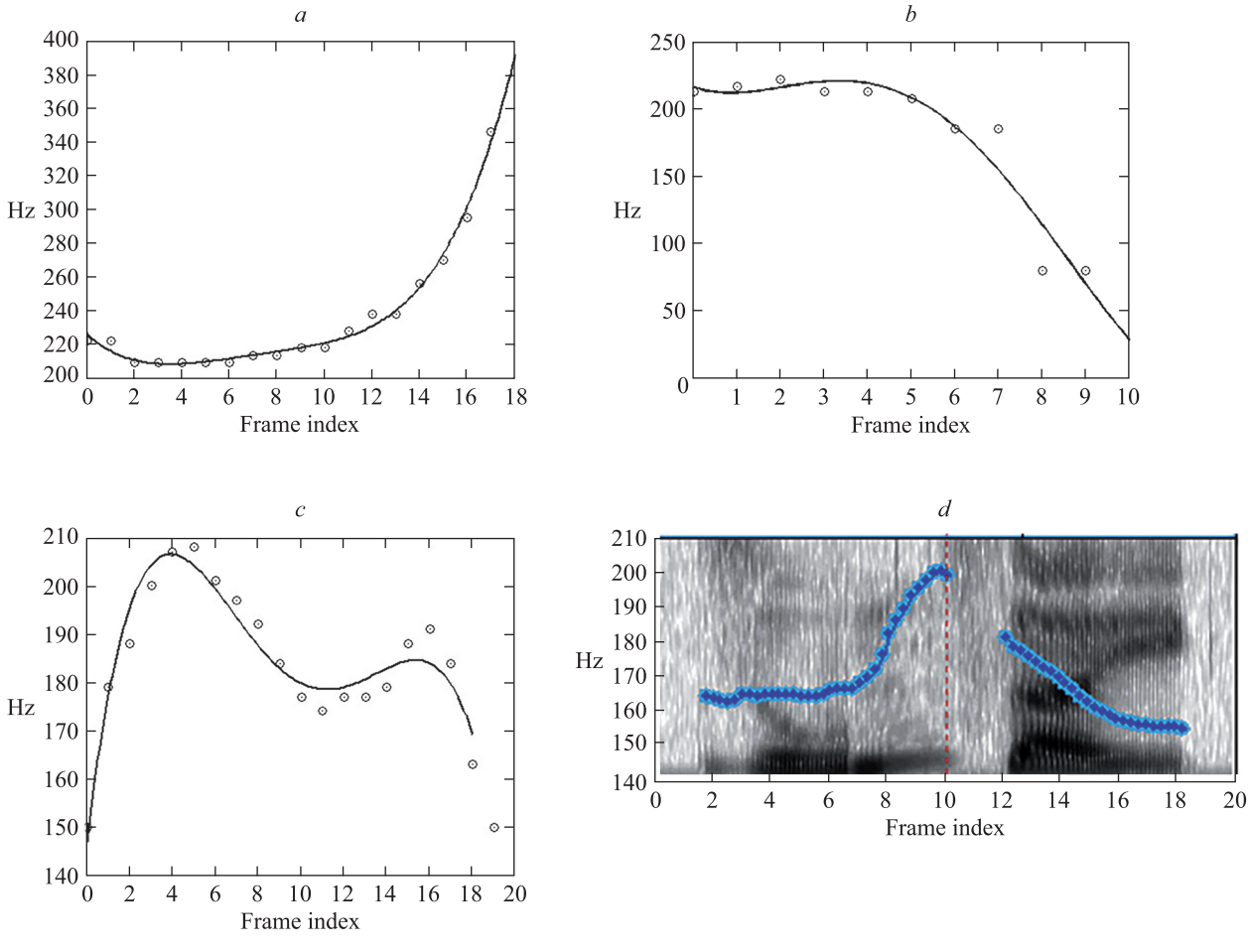


Fig. 8. The typical F_0 trajectories shape of some tones of Vietnamese isolated syllables rising tone (*a*), broken tone (*b*), drop tone (*c*) and A F_0 trajectory (*d*) of the word /*dún*/*dây*/ (z unˀ z ajˋ) with tone sandhi

Experimental data. In order to experiment the algorithms, a single speaker story reading corpus was created, uttered by a female speaker of standard Vietnamese voice. Sentences are extracted in the Vietnamese book «Adventures of a Cricket».

Experiment to extract the PM points. The formulas show that algorithm 1 has a smaller computational complexity than dynamic programming-type algorithms [4] because it does not require the steps to segment the whole speech utterance into short time frames and choose a suitable time point of each short time frame that gives high autocorrelation value. For the reliability of algorithm 1, we will compare algorithm 1 with the Talkin-type algorithm implemented in software Praat [23]. The parameter $f_{0,\,\min} = 50$ Hz, $f_{0,\,\max} = 550$ Hz and $a = 1.6$ for the R5 rule.

To compare the similarity between the two PM sequences of the same voiced segment, we give the following objective indexes that is based on the edit-distance (about a related work, see the algorithm for alignment of the reference epochs (EGG epochs) to the test epochs [18]).

Firstly, let $\text{PM}_I = \left\{pm_i\right\}_{i=1}^{m}$ and $\text{PM}_J = \left\{pm'_j\right\}_{j=1}^{n}$, then we define the measured value $D_{\text{PM}}\left(\text{PM}_I, \text{PM}_J\right)$ by:

$$D_{\mathrm{PM}}\left(\mathrm{PM}_I, \mathrm{PM}_J\right) \stackrel{\text{def}}{=} D_{m,n}\left(\left\{pm_i\right\}_{i=1}^{m}, \left\{pm'_j\right\}_{j=1}^{n}\right)/\min\{m, n\},$$

where $D_{1,1} = \left|pm_1 - pm'_1\right|$ and $D_{i,j} = \left|pm_i - pm'_j\right| + \min\left\{D_{i,j-1}, D_{i-1,j}, D_{i-1,j-1}\right\} \forall i,\ j \geq 2.$

Secondly, over the whole utterance, we get the average of the D_{PM} values calculated from the same voiced segment sequence of the utterance. The average $\overline{ED}$ is defined as follows:

$$\overline{ED} = \sum_{k=1}^{K} D_{\mathrm{PM}}\left(\mathrm{PM}_{I,k}, \mathrm{PM}_{J,k}\right)/K,$$

where $\left\{\mathrm{PM}_{I,k}\right\}_{k=1}^{K}$ and $\left\{\mathrm{PM}_{J,k}\right\}_{k=1}^{K}$ are PM sequences of k voiced segment of the utterance that have K voice segments total.

To compare with another PM estimation method such as Praat [23], we use the algorithm 1 to obtain the PMs PM_x^- with valley type for each voiced segment received by Praat, then we calculate $\overline{ED}$ values. Table below shows the similarity between the estimated PM_x^- and PMs (called pulse points, PPs) of the Praat type.

Measuring the similarity between PM_x^- and PPs of Praat

Utterance	Content	K	$\overline{ED}$, ms
#1	*«Đừng lo xem mây vận trời đêm nay có cơ đổi gió»* «ɗiŋ˧˩ lɔ˧˥ sɛm˧˥ məj˧˥ vəʔn˧˩ tɕəːj˧˩ ɗɛm˧˥ naj˧˥ kɔ˧˩ kəː˧˥ ɗoj˧˩˧ zɔ˧˩» «Do not worry, looking at the clouds, the wind may change direction tonight»	7	0.502 2
#2	*«Từ chỗ này muốn qua chỗ khác chúng tôi chỉ lách nhích từng tẹo»* «tḭ˧˩ tɕoʔo˧˩˩ʔnaj˧˩˧amuən˧˩˧ləkwaː˧˥ tɕoʔo˧˩˩ʔxaːk˧˩˧atɕuŋ˧˩˩ŋtoj˧˥ tɕḭ˧˩˧ lajk˧˩ ɲïk˧˩˧kti˧˩˧ŋ˧˩˧t̰ɕʔw˧˩˧ʔ» «To move from one place to another, we have to move little by little»	13	0.323 4
#3	*«Chũi bảo chũi không nhìn thấy»* «tɕuʔuj˧˩˩ʔɓaːʔujthtɕuʔuj˧˩˩ʔxəwŋ˧˥wɲiɲ˧˩˧t͡ɬʰəj˧˩˧ləj» «Chui claimed he could not see anything»	5	0.367 1
#4	*«Trời nghe trở gió ầm ầm trên mặt nước»* «tɕəːj˧˩ ŋɛ˧˥ tɕəː˧˩˧ zɔ˧˩ əm˧˩ əm˧˩ tɕen˧˥ maʔt˧˩ niək˧˩» «God makes the rumbling wind on the water»	4	0.275 1
#5	*«Thì ra bè chúng tôi từ lúc nào đã trôi vào gần một bờ cỏ»* «tʰḭ˧˩ zaː˧˥ ɓɛ˧˩ tɕuŋ˧˩ toj˧˥ tḭ˧˩ luk˧˩ naːw˧˩ ɗaʔa˧˩ tɕoj˧˥ vaːw˧˩ ɣən˧˩ moʔt˧˩ ɓəː˧˩ kɔ˧˩˧» «*Turns out our boat has drifted toward the grasslands*»	13	0.229 2
#6	*«Ấy vậy mà lúc đó chén ngon đáo để»* «əj˧˩ vəʔj˧˩ maː˧˩ luk˧˩ ɗɔ˧˩ tɕɛn˧˩ ŋɔn˧˥ ɗaːw˧˩ ɗẹ˧˩˧» «*The food was surprisingly yummy to me though*»	7	0.221 7

As we can see, the PMs determined by the algorithm 1 are more noticeable than the result of Praat when directly observing by eyes the speech signals as illustrated in fig. 9, *a*, and fig. 9, *b*, below.

However, Praat can ignore some pulse points, this case is described by the fig. 10, *a*, and fig. 10, *b* (whereas algorithm 1 does not).

Conclusion

In this paper, we propose two algorithms to determine the pitch markers of the original voice signal based on the cumulative signal and generate F_0 trajectories of tones.

The first algorithm is effective, with no need to divide a voiced segment into short segments (frames) as other methods, yet still achieving high accuracy. With the Vietnamese speech data of the lexical tones and phonetics tested (the full coverage of the Vietnamese phonetics was included), the results of calculating the pitch markers according to the new approach proved to be correct. The second algorithm used for generating F_0 trajectories of tones with qTA parameters of Xu model.

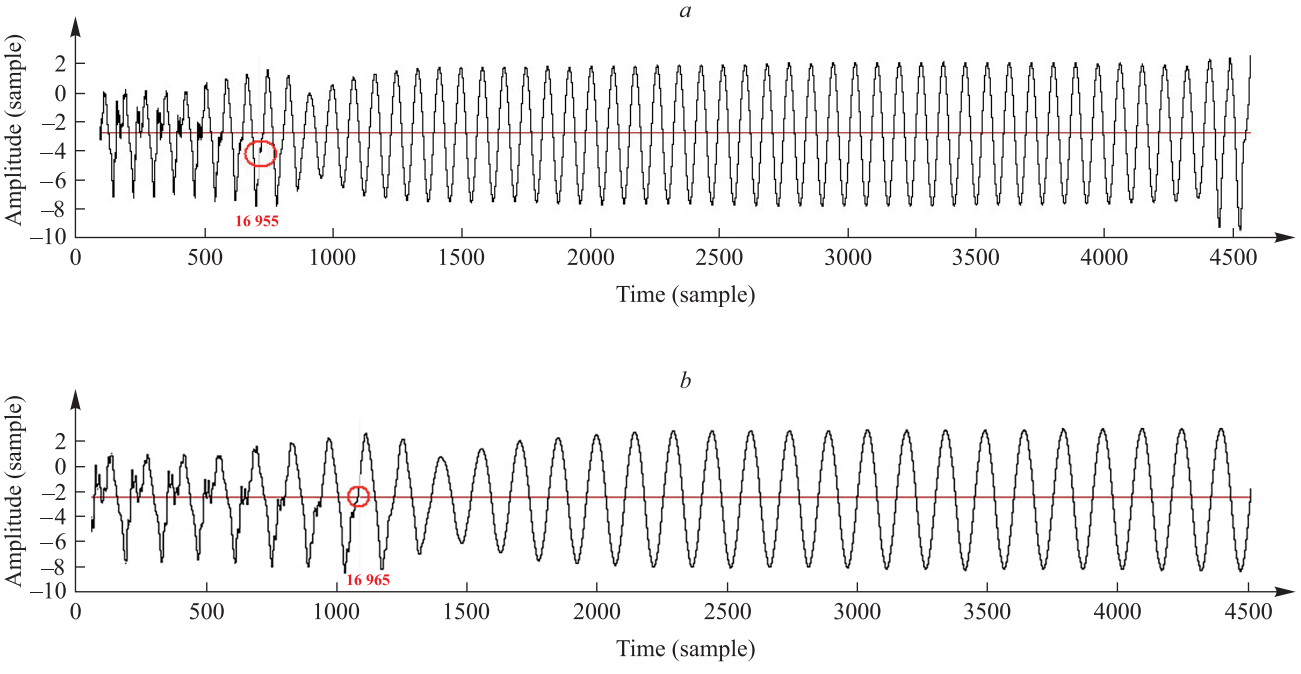


Fig. 9. One PM is determined by Praat (*a*); one PM is determined by the algorithm 1 (*b*).
Source: [8]

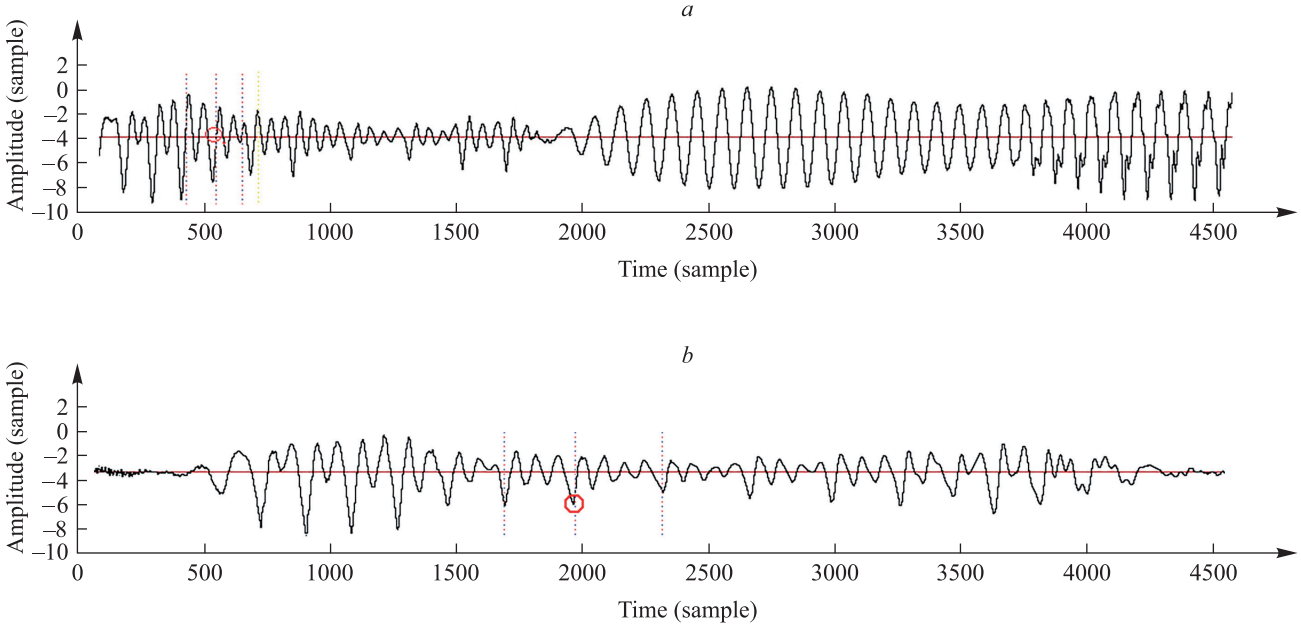


Fig. 10. With the utterance #2 (see table)
in the second voice segment, mising a PP around 355th sample (*a*),
and in the five voice segment, mising a PP around 497th sample (*b*)

Yi Xu has focused on how lexical tones of Mandarin were produced and perceived in continuous speech and has proposed the qTA model which considers the segmental phonemes, tones, and pitch accents as abstract units called pitch targets. In Mandarin, pitch targets are separated into static targets-[high] and targets-[low], and dynamic ones-[rise] and ones-[fall], which are associated with the four lexical tones respectively. This model gives a more accurate description of lexical tone variations in the syllable than the Fujisaki model. However, the qTA model needs labels on the onset and offset of the pitch target, and is difficult to implement on training speaker dependent prosodic styles. Prosody is employed to express attitude, assumptions and attention in daily speech communication and has been studied by linguists, phoneticians, speech therapists. In recent artificial intelligence developments, people seek to communicate effectively with intelligent machines

on a more personal and human level. To synthesize natural and human-sounding speech by computers, prosody plays an important role, which related to pause, pitch, speech rate and loudness. Among the factors which weave the prosody, pitch or fundamental frequency (in this paper we consider pitch and fundamental frequency (F_0) as the same) is the most characteristic.

References

1. Kovacs MD, Cho MY, Burchett PF, Trambert M. Benefits of integrated RIS/PACS/Reporting due to automatic population of templated reports. *Current Problems in Diagnostic Radiology.* 2019;48(1):37–39. DOI: 10.1067/j.cpradiol.2017.12.002.

2. Plonkowski M, Urbanovich P. The use of pitch in large-vocabulary continuous speech recognition system. *Przeglad Elektrotechniczny.* 2016;92(8):78–81.

3. Wang D, Hansen JHL. F_0 estimation for noisy speech by exploring temporal harmonic structures in local time frequency spectrum segment. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2016 March 20–25; Shanghai, China.* [S. l.]: IEEE; 2016. p. 6510–6514. DOI: 10.1109/ICASSP.2016.7472931.

4. Talkin D. A Robust Algorithm for Pitch Tracking (RAPT). In: Kleijn WB, Paliwal KK, editors. *Speech Coding & Synthesis.* [S. l.]: Elsevier Science B. V.; 1995. p. 495–518.

5. Xu Yi, Prom-on S. Articulatory-functional modeling of speech prosody: a review. In: Kobayashi T, Hirose K, Nakamura S. *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH-2010); 2010 September 26–30; Makuhari, Chiba, Japan.* [S. l.]: International Speech Communication Association; 2010. p. 46–49.

6. Kounoudes A, Naylor PA, Brookes M. The DYPSA algorithm for estimation of glottal closure instants in voiced speech. In: *Proceedings of International Conference on Acoustics, Speech and Signal Processing (CASSP'02); 2002 May 13–17; Orlando, FL, USA.* [S. l.]: IEEE; 2002. p. I349–I352. DOI: 10.1109/ICASSP.2002.5743726.

7. Smits R, Yegnanarayana B. Determination of instants of significant excitation in speech using group delay function. *IEEE Transactions on Speech and Audio Processing.* 1995; 3(5):325–333. DOI: 10.1109/89.466662.

8. Prom-on S, Liu F, Xu Y. Functional modeling of tone, focus and sentence type in mandarin Chinese. *Proceedings of the 17th International Congress of Phonetic Sciences; 2011 August 17–21; Hong Kong, China.* Hong Kong: City University of Hong Kong; 2011. p. 1638–1641.

9. Bailly G, Holm B. SFC: a trainable prosodic model. *Speech Communication.* 2005;46(3–4):348–364.

10. Fujisaki H. dynamic characteristics of voice fundamental frequency in speech and singing. In: MacNeilage PF, editor. *The Production of Speech.* New York: Springer; 1983. p. 39–55. DOI: 10.1007/978-1-4613-8202-7_3.

11. Kochanski G, Shih C. Prosody modeling with soft templates. *Speech Communication.* 2003;39(3–4):311–352. DOI: 10.1016/S0167-6393(02)00047-X.

12. Fujisaki H, Hirose K. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan.* 1984;5(4):233–242.

13. Xu Y, Wang QE. Pitch targets and their realization: evidence from Mandarin. *Speech Communication.* 2001;33(4):319–337. DOI: 10.1016/S0167-6393(00)00063-7.

14. Thai TY, Hung NV, Tuyet DV, Huy NHo, Ablameyko S. An effective algorithm for determining pitch markers of Vietnamese speech sentences. In: Huang T, Lv J, Sun C, Tuzikov A, editors. *Advances in Neural Networks – ISNN'2018. Proceedings of the 15th International Symposium on Neural Networks, ISNN'2018; 2018 June 25–28; Minsk, Belarus.* Cham: Springer; 2018. p. 628–636. (Lecture Notes in Computer Science; volume 10878).

15. Brookes M. Voicebox: speech processing toolbox for MATLAB [Internet; cited 2019 April 24]. Available from: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

16. Xu Y, Prom-on S. Toward invariant functional representations of variable surface fundamental frequency trajectories: synthesizing speech melody via model-based stochastic learning. *Speech Communication.* 2014;57:181–208. DOI: 10.1016/j.specom.2013.09.013.

17. Weierstrass K. *Über die analytische Darstellbarkeit sogenannter willkürlicher Funktionen einer reellen Veränderlichen Sitzungsberichteder.* Berlin: Königlich Preussischen Akademie der Wissenschaften zu Berlin; 1885. p. 633–639.

18. Cabral JP, Kane J, Gobl C, Carson-Berndsen J. Evaluation of glottal epoch detection algorithms on different voice types. In: *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH-2011); 2011 August 27–31; Florence, Italy.* [S. l.]: International Speech Communication Association; 2011. p. 1989–1992.

19. Optimizing Nonlinear Functions – MATLAB and Simulink [Internet; cited 2019 April 20]. Available from: https://www.mathworks.com/help/matlab/math/optimizing-nonlinear-functions.html.

20. Xu Y, Prom-on S. What is PENTAtrainer2? [Internet; cited 2019 April 20]. Available from: http://www.homepages.ucl.ac.uk/~uclyyix/PENTAtrainer2/.

21. Prom-on S, Xu Yi. The qTA toolkit for prosody: learning underlying parameters of communicative functions through modeling. In: Hasegawa-Johnson M, editor. *Proceedings of Speech Prosody 2010.* 2010;100034:1–4.

22. Chen JH, Kao YA. Pitch marking based on an adaptable filter and a peak-valley estimation method. *Computational Linguistics and Chinese Language Processing.* 2001;6(2):31–42.

23. Boersma P, Weenink D. Praat: Doing phonetics by computer [Internet; cited 2019 May 3]. Available from: http://www.fon.hum.uva.nl/praat/.

24. Babacan O, Drugman T, d'Alessandro N, Henrich N, Dutoit T. A comparative study of pitch extraction algorithms on a large variety of singing sounds. *Proceedings of International Conference on Acoustics, Speech and Signal Processing (CASSP'13); 2013 May 26–31; Vancouver, BC, Canada.* [S. l.]: IEEE; 2013. p. 7815–7819. DOI: 10.1109/ICASSP.2013.6639185.

25. Yin pitch estimator [Internet]. 2012 November 27 [cited 2019 August 28]. Available from: http://audition.ens.fr/adc/sw/yin.zip.

26. Prom-on S, Xu Yi. Discovering underlying tonal representations by computational modeling: a case study of thai. *Phonology Journal.* 2015;32(3):505–535.

27. Li Y, Tao J, Lai W, Xu X. Quantitative intonation modeling of interrogative sentences for Mandarin speech synthesis. *Speech Communication.* 2017;89:92–102. DOI: 10.1016/j.specom.2017.03.002.

28. Wang B, Xu Y, Ding Q. Interactive prosodic marking of focus, boundary and newness in Mandarin. *Phonetica.* 2018;75(1): 24–56. DOI: 10.1159/00045308.

29. Charpentier F, Stella M. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In: *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP'86); 1986 April 7–11; Tokyo, Japan.* [S. l.]: IEEE; 1986. p. 2015–2018. DOI: 10.1109/ICASSP.1986.1168657.

30. Ching XXu, Yi Xu, Li-Shi Luo. A pitch target approximation model for F_0 trajectories in Mandarin. In: Ohala JJ, editor. *Proceedings of the 14th International Congress of Phonetic Sciences (ICPHS'99).* San Francisco: University of California; 1999. p. 2359–2362.

Received by editorial board 04.09.2019.

Краткие сообщения

Short communications

УДК 513.88

ФОРМУЛЫ *t*-ЭНТРОПИИ ДЛЯ КОНКРЕТНЫХ КЛАССОВ ТРАНСФЕР-ОПЕРАТОРОВ

К. БАРДАДИН [1)]***, Б. К. КВАСЬНЕВСКИЙ*** [1)]***,***
К. С. КУРНОСЕНКО [2)]***, А. В. ЛЕБЕДЕВ*** [2)]

[1)]*Университет Белостока, ул. К. Циолковского, 1М, 15-245, г. Белосток, Польша*
[2)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*

t-Энтропия является принципиальным объектом спектральной теории операторов, порожденных динамическими системами (операторов взвешенного сдвига и трансфер-операторов). По существу она представляет собой преобразование Фенхеля – Лежандра от спектрального потенциала оператора, и получение явных формул для ее вычисления – нетривиальная задача. В работе такие формулы получены для *t*-энтропии двух наиболее часто используемых в приложениях классов трансфер-операторов: порожденных обратимыми отображениями (т. е. операторов взвешенного сдвига) и порожденных локальными гомеоморфизмами (т. е. операторов Перрона – Фробениуса). В первом случае *t*-энтропия вычисляется с помощью интегралов по инвариантным мерам, во втором – с использованием интегралов по инвариантным мерам и энтропии Колмогорова – Синая.

Ключевые слова: трансфер-оператор; спектральный потенциал; *t*-энтропия; инвариантная мера; метрическая энтропия.

Образец цитирования:
Бардадин К, Квасьневский БК, Курносенко КС, Лебедев АВ. Формулы *t*-энтропии для конкретных классов трансфер-операторов. *Журнал Белорусского государственного университета. Математика. Информатика.* 2019;3:122–128.
https://doi.org/10.33581/2520-6508-2019-3-122-128

For citation:
Bardadyn K, Kwasniewski BK, Kurnosenko KS, Lebedev AV. *t*-Entropy formulae for concrete classes of transfer operators. *Journal of the Belarusian State University. Mathematics and Informatics.* 2019;3:122–128. Russian.
https://doi.org/10.33581/2520-6508-2019-3-122-128

Авторы:
Кшыштоф Бардадин – ассистент кафедры анализа математического факультета.
Бартош Косма Квасьневский – доктор математических наук; заведующий кафедрой анализа математического факультета.
Кирилл Сергеевич Курносенко – аспирант кафедры функционального анализа и аналитической экономики механико-математического факультета. Научный руководитель – А. В. Лебедев.
Андрей Владимирович Лебедев – доктор физико-математических наук, профессор; заведующий кафедрой функционального анализа и аналитической экономики механико-математического факультета.

Authors:
Krzysztof Bardadyn, assistant at the department of analysis, faculty of mathematics.
kbardadyn@math.uwb.edu.pl
Bartosz Kosma Kwasniewski, doctor of science (mathematics); head of the department of analysis, faculty of mathematics.
bartoszk@math.uwb.edu.pl
Kirill S. Kurnosenko, postgraduate student at the department of functional analysis and analytic economy, faculty of mechanics and mathematics.
kurn.ne@gmail.com
Andrei V. Lebedev, doctor of science (physics and mathematics), full professor; head of the department of functional analysis and analytic economy, faculty of mechanics and mathematics.
lebedev@bsu.by

t-ENTROPY FORMULAE FOR CONCRETE CLASSES OF TRANSFER OPERATORS

K. BARDADYN [a], ***B. K. KWASNIEWSKI*** [a],
K. S. KURNOSENKO [b], ***A. V. LEBEDEV*** [b]

[a] *University of Bialystok, 1M K. Ciolkowskiego Street, Bialystok 15-245, Poland*
[b] *Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*
Corresponding author: A. V. Lebedev (lebedev@bsu.by)

t-Entropy is a principal object of the spectral theory of operators, generated by dynamical systems, namely, weighted shift operators and transfer operators. In essence *t*-entropy is the Fenchel – Legendre transform of the spectral potential of an operator in question and derivation of explicit formulae for its calculation is a rather nontrivial problem. In the article explicit formulae for *t*-entropy for two the most exploited in applications classes of transfer operators are obtained. Namely, we consider transfer operators generated by reversible mappings (i. e. weighted shift operators) and transfer operators generated by local homeomorphisms (i. e. Perron – Frobenius operators). In the first case *t*-entropy is computed by means of integrals with respect to invariant measures, while in the second case it is computed in terms of integrals with respect to invariant measures and Kolmogorov – Sinai entropy.

Keywords: transfer operator; spectral potential; *t*-entropy; invariant measure; metric entropy.

Введение

В данной работе получены явные формулы для основных компонент вариационных принципов, используемых при вычислении спектральных потенциалов двух конкретных типов трансфер-операторов.

Пусть (X, α) – динамическая система, где X – компактное пространство; $\alpha : X \to X$ – непрерывное отображение.

Линейный оператор $A : C(X) \to C(X)$ называется *трансфер-оператором* для (X, α), если он:

а) является положительным оператором (отображает неотрицательные функции в неотрицательные функции);

б) удовлетворяет *гомологическому тождеству*

$$A(f \circ \alpha \cdot g) = fAg, \quad f, g \in C(X).$$

По заданному трансфер-оператору A определим семейство операторов $A_\varphi : C(X) \to C(X)$, зависящих от функционального параметра $\varphi \in C(X)$, с помощью формулы

$$A_\varphi f := A(e^\varphi f).$$

Операторы такого типа являются принципиальными объектами спектральной теории динамических систем, эллиптической теории функционально-дифференциальных уравнений и т. д.

Обозначим через $\lambda(\varphi)$ логарифм спектрального радиуса оператора A_φ. Обычно $\lambda(\varphi)$ называют *спектральным потенциалом* трансфер-оператора A. В соответствии с формулой Гельфанда имеем

$$\lambda(\varphi) = \lim_{n \to \infty} \frac{1}{n} \ln \left\| A_\varphi^n \right\|. \tag{1}$$

Положительность оператора A_φ^n означает, что

$$\lambda(\varphi) = \lim_{n \to \infty} \frac{1}{n} \ln \left\| A_\varphi^n 1 \right\|, \tag{2}$$

где 1 – единичная функция на X; норма $\|f\|$ – равномерная норма функции $f \in C(X)$: $\|f\| = \max_{x \in X} |f(x)|$.

Непосредственно из гомологического тождества вытекает равенство

$$A_\varphi^n f = A^n (e^{S_n \varphi} f), \tag{3}$$

где

$$S_n \varphi := \varphi + \varphi \circ \alpha + \ldots + \varphi \circ \alpha^{n-1}. \tag{4}$$

Известен вариационный принцип для вычисления $\lambda(\varphi)$ [1]:

$$\lambda(\varphi) = \max_{\mu \in M_\alpha} \left(\mu(\varphi) + \tau(\mu)\right), \tag{5}$$

где M_α – множество α-инвариантных вероятностных мер; $\mu(\varphi) = \int_X \varphi d\mu$; $\tau(\mu)$ – t-энтропия (довольно сложно вычисляемый динамический инвариант). Конкретнее для определения $\tau(\mu)$ нужно выполнить следующие три шага [2]:

1) для каждого $n \in \mathbb{N}$ и каждого разбиения единицы G, т. е. набора $G = \{g_1, \ldots, g_k\}$ неотрицательных функций $g_i \in C(X)$, удовлетворяющих тождеству $g_1 + \ldots + g_k \equiv 1$, вводится число

$$\tau_n(\mu, G) := \sum_{g_i \in G} \mu(g_i) \ln \frac{\mu\left(A^n g_i\right)}{\mu(g_i)},$$

здесь полагается $\ln(0) := -\infty$, а если $\mu(g_i) = 0$, то соответствующее слагаемое приравнивается к 0 независимо от значения $\mu\left(A^n g_i\right)$;

2) полагается

$$\tau_n(\mu) := \inf_G \tau_n(\mu, G),$$

где инфимум берется по всем разбиениям единицы G в $C(X)$;

3) окончательно t-энтропия $\tau(\mu)$ определяется как

$$\tau(\mu) := \inf_{n \in \mathbb{N}} \frac{\tau_n(\mu)}{n}.$$

В настоящей работе получены явные формулы t-энтропии для двух конкретных классов трансфер-операторов.

Отметим также, что связь процедуры вычисления t-энтропии и дивергенции Кульбака – Лейблера обсуждается в работе [3].

Трансфер-операторы и положительные функционалы

Для начала дадим более явное описание трансфер-операторов, связав их со специальным семейством положительных функционалов.

Для каждой точки $x \in X$ определим функционал φ_x по формуле

$$\varphi_x(f) := [Af](x), \quad f \in C(X). \tag{6}$$

Очевидно, φ_x – положительный функционал.

Для точки x возможны две ситуации:

1) $[A1](x) = 0$. Это значит, что $\varphi_x(1) = 0$, следовательно, $\varphi_x = 0$, так как φ_x есть положительный функционал;

2) $[A1](x) \neq 0$. В этом случае $\varphi_x \neq 0$ и φ_x определяет некоторую меру ν_x на X.

Гомологическое тождество означает также, что для любой функции $f \in C(X)$ справедливо

$$\left[A(f \circ \alpha)\right](x) = \left[A(f \circ \alpha \cdot 1)\right](x) = f(x) \cdot A1(x).$$

Следовательно,

$$\frac{1}{A1(x)} \varphi_x(f \circ \alpha) = f(x),$$

и, значит,

$$\operatorname{supp} \nu_x \subset \alpha^{-1}(x). \tag{7}$$

Ясно, что отображение $x \to \varphi_x$ является *-слабо непрерывным.

Отметим также, что если $x \notin \alpha(X)$, то $A1(x) = 0$. В самом деле если $A1(x) \neq 0$, то можно выбрать функцию $f \in C(X)$ такую, что $f\big|_{\alpha(X)} = 0$ и $f(x) = 1$. Тогда

$$0 = \frac{1}{A1(x)}\big[A(f \circ \alpha)\big](x) = f(x) = 1,$$

что приводит к противоречию.

Рассмотренные выше объекты дают полное описание трансфер-операторов, так как, очевидно, любое *-слабо непрерывное отображение $x \to \varphi_x$, где φ_x – положительные функционалы такие, что справедливо:

• $\varphi_x = 0$, когда $x \notin \alpha(X)$;

• φ_x удовлетворяет соотношению (7), когда $x \in \alpha(X)$ (здесь может быть также $\varphi_x = 0$),

задает трансфер-оператор, действующий по формуле (6).

Замечание. Из данного описания трансфер-операторов вытекают два наблюдения:

1) если $\alpha : X \to X$ является гомеоморфизмом, то трансфер-оператор имеет вид

$$Af(x) = \rho(x) f\left(\alpha^{-1}(x)\right),$$

где $\rho \in C(X)$ – некоторая неотрицательная функция. Иными словами, трансфер-оператор есть оператор взвешенного сдвига;

2) если $\alpha : X \to X$ – локальный гомеоморфизм (в этом случае прообраз каждой точки x содержит конечное число точек), то трансфер-оператор имеет вид

$$Af(x) = \sum_{y \in \alpha^{-1}(x)} \rho(y) f(y),$$

где $\rho \in C(X)$ – некоторая неотрицательная функция. В этом случае трансфер-оператор является оператором Перрона – Фробениуса.

Именно для таких двух типов трансфер-операторов будут получены основные результаты в данной работе.

Спектральный потенциал, *t*-энтропия и преобразование Фенхеля – Лежандра

Отметим еще одно наблюдение, которое будем использовать в дальнейшем.

В [1, предложения 8.4, 8.6] доказано, что $\tau(\mu)$ является вогнутой и полунепрерывной сверху в *-слабой топологии функцией от μ. Поэтому формула (5) означает, что спектральный потенциал $\lambda(\varphi)$ (2) есть не что иное, как преобразование Фенхеля – Лежандра от $-\tau(\mu)$. Более того, в силу двойственности Фенхеля – Лежандра – Моро из полунепрерывности сверху $\tau(\mu)$ следует, что $-\tau(\mu) = \lambda^*$ (здесь через λ^* обозначен двойственный к $\lambda(\cdot)$ по Фенхелю – Лежандру функционал), т. е. $\tau(\mu)$ однозначно определяется спектральным потенциалом λ. В силу уже упомянутой двойственности мы также заключаем, что если $S(\mu)$ – некоторая вогнутая полунепрерывная сверху в *-слабой топологии функция от μ такая, что

$$\lambda(\varphi) = \sup_{\mu \in M_\alpha} \left(\mu(\varphi) + S(\mu)\right) \tag{8}$$

(т. е. $\lambda(\varphi)$ есть преобразование Фенхеля – Лежандра от $-S(\mu)$), то $-S(\mu) = \lambda^*$, и, значит,

$$S(\mu) = \tau(\mu). \tag{9}$$

Следующий результат относится к обратимым динамическим системам.

Теорема 1. *Пусть* (X, α) *– обратимая динамическая система,* $\psi \in C(X)$ *и трансфер-оператор имеет вид*

$$(Af)(x) := e^{\psi(x)} f\left(\alpha^{-1}(x)\right)$$

(любой трансфер-оператор для обратимой динамической системы имеет такой вид). Тогда $\tau(\mu) = \mu(\psi)$.

Доказательство. В рассматриваемом случае справедливо соотношение

$$A_\varphi f = A\left(e^\varphi f\right) = e^\psi e^{\varphi \circ \alpha^{-1}} f \circ \alpha^{-1} = e^{\psi + \varphi \circ \alpha^{-1}} f \circ \alpha^{-1}.$$

Из этого равенства и вариационного принципа для спектрального радиуса оператора взвешенного сдвига, порожденного гомеоморфизмом компакта [4; 5], следует, что

$$\lambda(\varphi) = \max_{\mu \in M_\alpha} \mu\left(\varphi \circ \alpha^{-1} + \psi\right) = \max_{\mu \in M_\alpha} \left[\mu\left(\varphi \circ \alpha^{-1}\right) + \mu(\psi)\right] = \max_{\mu \in M_\alpha} \left[\mu(\varphi) + \mu(\psi)\right], \tag{10}$$

где в последнем выражении использовалось равенство $\mu\left(\varphi \circ \alpha^{-1}\right) = \mu(\varphi)$, вытекающее из α-инвариантности меры μ.

Так как $\mu(\psi)$ линейно и непрерывно в *-слабой топологии зависит от параметра μ, то равенство (10) с учетом наблюдений (8) и (9) означает, что $\tau(\mu) = \mu(\psi)$. Теорема 1 доказана.

Следующий результат относится к динамическим системам, порожденным растягивающими отображениями.

Теорема 2. *Пусть X – метрический компакт, α – растягивающее непрерывное отображение, для которого $\alpha^{-1}(x) \equiv \text{const}$, $\psi \in C(X)$ и трансфер-оператор имеет вид*

$$(Af)(x) := \sum_{y \in \alpha^{-1}(x)} e^{\psi(y)} f(y).$$

Тогда

$$\tau(\mu) = \mu(\psi) + h(\mu),$$

где $h(\mu)$ – энтропия Колмогорова – Синая.

Доказательство. Из [6; 7] следует, что в рассматриваемой ситуации (т. е. для растягивающих отображений) справедливо равенство

$$\lambda(\varphi) = P\left((\varphi + \psi), \alpha\right), \tag{11}$$

где $P(c, \alpha)$ – топологическое давление, ассоциированное с динамической системой (X, α) и непрерывной функцией $c \in C(X)$ (определение топологического давления см. в [8; 9]).

Для топологического давления и любой динамической системы (X, α) известен следующий вариационный принцип [8; 9]:

$$P(c, \alpha) = \sup_{\mu \in M_\alpha} \left(\mu(c) + h(\mu)\right), \tag{12}$$

где $h(\mu)$ – энтропия Колмогорова – Синая.

Равенства (11) и (12) означают, что

$$\lambda(\varphi) = \sup_{\mu \in M_\alpha} \left(\mu(\varphi + \psi) + h(\mu)\right) = \sup_{\mu \in M_\alpha} \left(\mu(\varphi) + \left[\mu(\psi) + h(\mu)\right]\right). \tag{13}$$

Энтропия $h(\mu)$ является вогнутой функцией от μ. При этом так как α – растягивающее отображение, то согласно [10, theorem 8.2] $h(\mu)$ есть полунепрерывная сверху в *-слабой топологии функция от μ. Кроме того, $\mu(\psi)$ – линейная и непрерывная в *-слабой топологии функция от μ. Значит, $\left[\mu(\psi) + h(\mu)\right]$ является вогнутой и полунепрерывной сверху в *-слабой топологии функцией от μ.

Теперь равенство (13) с учетом наблюдений (8) и (9) означает, что $\tau(\mu) = \left[\mu(\psi) + h(\mu)\right]$. Теорема 2 доказана.

Ниже приводится еще один результат, который можно использовать при вычислении спектральных потенциалов для трансфер-операторов, ассоциированных с произведениями динамических систем.

Пусть (X, α) и (Y, β) – некоторые динамические системы и $A_X : C(X) \to C(X)$ – фиксированный трансфер-оператор для (X, α), $A_Y : C(Y) \to C(Y)$ – фиксированный трансфер-оператор для (Y, β). Как отмечено в разделе «Трансфер-операторы и положительные функционалы», операторам A_X и A_Y отвечают семейства положительных функционалов (мер) $\{\varphi_x\}_{x \in X}$ и $\{\nu_y\}_{y \in Y}$ соответственно. Эти семейства мер задают трансфер-оператор $A_{X \times Y} : C(X \times Y) \to C(X \times Y)$ для динамической системы $\left(X \times Y, (\alpha, \beta)\right)$ по формуле

$$\left(A_{X \times Y} f\right)(x, y) := \iint\limits_{\alpha^{-1}(x) \times \beta^{-1}(y)} f d\varphi_x \otimes d\nu_y. \tag{14}$$

То, что $A_{X \times Y}$ действительно трансфер-оператор, следует из рассуждений раздела «Трансфер-операторы и положительные функционалы» и того факта, что $(\alpha, \beta)^{-1}(x, y) = \alpha^{-1}(x) \times \beta^{-1}(y)$.

Теорема 3. *Пусть A_X, A_Y и $A_{X \times Y}$ – вышеупомянутые трансфер-операторы; $\varphi \in C(X)$, $\psi \in C(X)$ и $\theta(x, y) = \varphi(x) + \psi(y) \in C(X \times Y)$. Тогда*

$$\lambda(\theta) = \lambda(\varphi) + \lambda(\psi),$$

где $\lambda(\varphi)$ *– спектральный потенциал оператора* $A_{X\varphi}$*;* $\lambda(\psi)$ *– спектральный потенциал оператора* $A_{Y\psi}$*;* $\lambda(\theta)$ *– спектральный потенциал оператора* $A_{X\times Y\theta}$.

Доказательство. Для сокращения записи обозначим

$$A_\theta := A_{X\times Y\theta},\ A_\varphi := A_{X\varphi},\ A_\psi := A_{Y\psi}.$$

Проверим, что

$$\left\|A_\theta^n\right\| = \left\|A_\varphi^n\right\|\left\|A_\psi^n\right\|. \tag{15}$$

Действительно, из формулы (14) следует

$$A_\theta 1(x,y) = A\left(e^\theta 1\right)(x,y) = \iint\limits_{\alpha^{-1}(x)\times\beta^{-1}(y)} \left(e^\theta 1\right) d\varphi_x \otimes d\nu_y = \int\limits_{\alpha^{-1}(x)} \left(\int\limits_{\beta^{-1}(y)} e^\theta 1\, d\nu_y \right) d\varphi_x =$$

$$= \int\limits_{\alpha^{-1}(x)} \left(\int\limits_{\beta^{-1}(y)} e^\varphi e^\psi\, d\nu_y \right) d\varphi_x = \int\limits_{\alpha^{-1}(x)} e^\varphi \left(\int\limits_{\beta^{-1}(y)} e^\psi d\nu_y \right) d\varphi_x = \left(\int\limits_{\alpha^{-1}(x)} e^\varphi d\varphi_x \right)\left(\int\limits_{\beta^{-1}(y)} e^\psi d\nu_y \right) =$$

$$= A_\varphi 1(x)\cdot A_\psi 1(y).$$

Отсюда с учетом положительности операторов A_θ, A_φ, A_ψ имеем

$$\left\|A_\theta\right\| = \left\|A_\theta 1\right\| = \left\|A_\varphi 1\right\|\left\|A_\psi 1\right\| = \left\|A_\varphi\right\|\left\|A_\psi\right\|.$$

Рассуждая аналогично и применяя равенства (3) и (4), можно показать, что

$$A_\theta^n 1(x,y) = A_{X\times Y}^n \left(e^{S_n\theta} 1\right)(x,y) = A_X^n \left(e^{S_n\varphi} 1\right)(x)\cdot A_Y^n \left(e^{S_n\psi} 1\right)(y) = A_\varphi^n 1(x)\cdot A_\psi^n 1(y).$$

Из последних соотношений следует равенство (15).

Теперь из (1) и (15) получим

$$\lambda(\theta) = \lim_{n\to\infty}\frac{1}{n}\ln\left\|A_\theta^n\right\| = \lim_{n\to\infty}\frac{1}{n}\ln\left(\left\|A_\varphi^n\right\|\left\|A_\psi^n\right\|\right) =$$

$$= \lim_{n\to\infty}\frac{1}{n}\ln\left\|A_\varphi^n\right\| + \lim_{n\to\infty}\frac{1}{n}\ln\left\|A_\psi^n\right\| = \lambda(\varphi) + \lambda(\psi).$$

Теорема 3 доказана.

Библиографические ссылки

1. Antonevich AB, Bakhtin VI, Lebedev AV. On *t*-entropy and variational principle for the spectral radii of transfer and weighted shift operators. *Ergodic Theory and Dynamical Systems.* 2011;31(4):995–1042. DOI: 10.1017/S0143385710000210.

2. Bakhtin VI, Lebedev AV. A New Definition of *t*-Entropy for Transfer Operators. *Entropy.* 2017;19(11):573. DOI: 10.3390/e19110573.

3. Сокол ЭЭ. Введение информационной функции Кульбака – Лейблера с помощью разбиений вероятностного пространства. *Журнал БГУ. Математика. Информатика.* 2018;1:59–67.

4. Китовер АК. О спектре автоморфизмов с весом и теореме Камовица – Шайнберга. *Функциональный анализ и его приложения.* 1979;13(1):70–71.

5. Лебедев АВ. Об обратимости элементов в C^*-алгебрах, порожденных динамическими системами. *Успехи математических наук.* 1979;34(4):199–200.

6. Латушкин ЮД, Степин АМ. Оператор взвешенного сдвига на топологической марковской цепи. *Функциональный анализ и его приложения.* 1988;22(4):86–87.

7. Латушкин ЮД, Степин АМ. Операторы взвешенного сдвига и линейные расширения динамических систем. *Успехи математических наук.* 1991;46(2):85–143.

8. Ruelle D. Statistical mechanics on a compact set with Z^ν action satisfying expansiveness and specification. *Transactions of the American Mathematical Society.* 1973;185:237–251. DOI: 10.2307/1996437.

9. Walters P. A variational principle for the pressure on continuous transformations. *American Journal of Mathematics.* 1975;97(4):937–971. DOI: 10.2307/2373682.

10. Walters P. *An introduction to ergodic theory.* New York: Springer-Verlag; 1982. 250 p.

References

1. Antonevich AB, Bakhtin VI, Lebedev AV. On *t*-entropy and variational principle for the spectral radii of transfer and weighted shift operators. *Ergodic Theory and Dynamical Systems.* 2011;31(4):995–1042. DOI: 10.1017/S0143385710000210.

2. Bakhtin VI, Lebedev AV. A New Definition of *t*-Entropy for Transfer Operators. *Entropy.* 2017;19(11):573. DOI: 10.3390/e19110573.

3. Sokal EE. Introduction of the Kullback – Leibler information function by means of partitions of the probability space. *Journal of the Belarusian State University. Mathematics and Informatics.* 2018;1:59–67. Russian.

4. Kitover AK. [Spectrum of automorphisms with weight and the Kamowitz – Scheinberg theorem]. *Funktional'nyi analiz i ego prilozheniya.* 1979;13(1):70–71. Russian.

5. Lebedev AV. [On the invertibility of elements in C^*-algebras generated by dynamical systems]. *Uspekhi matematicheskikh nauk.* 1979;34(4):199–200. Russian.

6. Latushkin JuD, Stepin AM. [Weighted shift operator on a topological Markov chain]. *Funktional'nyi analiz i ego prilozheniya.* 1988;22(4):86–87. Russian.

7. Latushkin JuD, Stepin AM. [Weighted translation operators and linear extensions of dynamical systems]. *Uspekhi matematicheskikh nauk.* 1991;46(2):85–143. Russian.

8. Ruelle D. Statistical mechanics on a compact set with Z^{ν} action satisfying expansiveness and specification. *Transactions of the American Mathematical Society.* 1973;185:237–251. DOI: 10.2307/1996437.

9. Walters P. A variational principle for the pressure on continuous transformations. *American Journal of Mathematics.* 1975;97(4):937–971. DOI: 10.2307/2373682.

10. Walters P. *An introduction to ergodic theory.* New York: Springer-Verlag; 1982. 250 p.

Статья поступила в редколлегию 22.09.2019.
Received by editorial board 22.09.2019.

УДК 519.719.2

РАЗДЕЛЕНИЕ СЕКРЕТА В КОЛЬЦАХ МНОГОЧЛЕНОВ ОТ НЕСКОЛЬКИХ ПЕРЕМЕННЫХ С ИСПОЛЬЗОВАНИЕМ КИТАЙСКОЙ ТЕОРЕМЫ ОБ ОСТАТКАХ

Г. В. МАТВЕЕВ[1)]

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*

Обобщено разделение целочисленного секрета, использующего алгоритм китайской теоремы об остатках на случай кольца многочленов от нескольких переменных над конечным полем. Для генерации частичных секретов вместо целочисленных модулей применяются идеалы и их базисы Грёбнера. Этот подход предложен нами ранее. В настоящей работе показано, что любую пороговую структуру доступа можно реализовать идеально. Это является одним из преимуществ предлагаемого подхода. В кольце целых чисел никакую структуру доступа нельзя осуществить идеально, поскольку частичные секреты всех участников имеют различные размеры.

Ключевые слова: китайская теорема об остатках; разделение секрета; равноостаточные идеалы; эквипроективные множества.

Благодарность. Автор выражает благодарность Т. Галибус и Н. Шенецу за их ценные замечания, а также В. Матулису за помощь при подготовке рукописи к печати.

CHINESE REMAINDER THEOREM SECRET SHARING IN MULTIVARIATE POLYNOMIALS

G. V. MATVEEV[a]

[a]*Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*

This paper deals with a generalization of the secret sharing using Chinese remainder theorem over the integers to multivariate polynomials over a finite field. We work with the ideals and their Gröbner bases instead of integer moduli. Therefore, the proposed method is called GB secret sharing. It was initially presented in our previous paper. Now we prove that any threshold structure has ideal GB realization. In a generic threshold modular scheme in ring of integers the sizes of the share space and the secret space are not equal. So, the scheme is not ideal and our generalization of modular secret sharing to the multivariate polynomial ring is more secure.

Keywords: Chinese remainder theorem; secret sharing; equiresidual ideals; equiprojectable sets.

Acknowledgements. I thank T. Galibus and N. Shenets for their valuable comments. I also want to thank to V. Matulis for his help in preparation the paper.

Образец цитирования:
Матвеев ГВ. Разделение секрета в кольцах многочленов от нескольких переменных с использованием китайской теоремы об остатках. *Журнал Белорусского государственного университета. Математика. Информатика.* 2019;3:129–133 (на англ.).
https://doi.org/10.33581/2520-6508-2019-3-129-133

For citation:
Matveev GV. Chinese remainder theorem secret sharing in multivariate polynomials. *Journal of the Belarusian State University. Mathematics and Informatics.* 2019;3:129–133.
https://doi.org/10.33581/2520-6508-2019-3-129-133

Автор:
Геннадий Васильевич Матвеев – кандидат физико-математических наук; доцент кафедры высшей математики факультета прикладной математики и информатики.

Author:
Gennadii V. Matveev, PhD (physics and mathematics); associate professor at the department of higher mathematics, faculty of applied mathematics and informatics.
matveev@bsu.by
https://orcid.org/0000-0002-1372-0117

Introduction

Secret sharing enables a group of l participants to share a secret. Each of them is provided a share. The sharing scheme has a threshold t if any t-subset of participants with t out of l shares enables the secret to be recovered.

The basic idea of the modular secret sharing is as follows. Let $s \in \mathbb{Z}$ be the secret value, and let the residue $s_i = s \bmod m_i$, where m_i is the public key, be the share of the i participant. It is necessary to choose the secret s and moduli m_i so that only the authorized groups of participants can compute the secret. For more details, see [1]. However, in a generic (t, l)-threshold modular scheme in $\mathbb{Z}$, the sizes of the share space and the secret space are not equal. So, the scheme is not ideal.

In this paper, the modular constructions in the ring of integers are transformed into the modular constructions in the multivariate polynomial ring $F_q[x]$, where $x = (x_1, x_2, \ldots, x_n)$. We prove that any threshold structure has the ideal GB realization. So, our generalization of modular secret sharing to the multivariate polynomial ring is more secure.

The modular secret sharing in the ring $F_q[x]$ is based on the following facts:

• first, given a monomial ordering, we can compute the residue of a secret polynomial $s(x) \in F_q[x]$ modulo any zero-dimensional ideal;

• second, there is the CRT-algorithm for computing the secret [2].

Our approach can be generalized to other commutative rings with the effective Gröbner basis theory. We studied the univariate case and its verification protocols in our previous papers [3–6]. GB secret sharing was presented in [7].

The paper is organized as follows. In the second section we construct the special zero-dimensional ideals of $F_q[x]$. They provide the security of the proposed scheme. Our construction is based on the triangular ideals' characterization (see [8]). In the third section, we present ideal threshold schemes in the ring $F_q[x]$.

Equiresidual ideals

The results of this section are essentially inspired by the concept of equiprojectivity (see [8]). Following their notation, we say that an ideal of $F_q[x]$ is a triangular ideal if it admits a separable triangular set of generators. Throughout the paper, we consider the Gröbner bases in the ring $F_q[x]$, where $x = (x_1, x_2, \ldots, x_n)$, $x_1 < x_2 < \ldots < x_n$.

Let I be a triangular zero-dimensional ideal of $F_q[x]$. It has the reduced Gröbner basis $\{f_1, f_2, \ldots, f_n\}$:

$$f_i = x_i^{d_i} + a_{i,\,d_i-1}x_i^{d_i-1} + \ldots + a_{i,1}x_i + a_{i,0},\ a_{i,\,d_i-1}, \ldots, a_{i,1}, a_{i,0} \in F_q[x_1, x_2, \ldots, x_{i-1}],$$

and its zero-set $V(I)$ in the algebraic closure of F_q is equiprojectable (see theorem 4.5 in [8]). In this case, the vector of fiber cardinalities is defined as

$$FC(I) = \left(\operatorname{card}\pi_1^{-1}(M), \operatorname{card}\pi_2^{-1}(M), \ldots, \operatorname{card}\pi_{n-1}^{-1}(M)\right) = (d_2 \cdots d_n, d_3 \cdots d_n, \ldots, d_n),$$

where $\pi_i(\alpha_1, \alpha_2, \ldots, \alpha_n) = (\alpha_1, \alpha_2, \ldots, \alpha_i)$ (see [8, p. 640]). $FC(I)$ does not depend on the choice of the point $M \in V(I)$.

The set of all reduced terms modulo I is denoted by $RT(I)$. The set of all reduced polynomials is denoted by $RP(I)$. Let

$$D(I) = (d_1, d_2, \ldots, d_n),\ d = d_1 d_2 \cdots d_n.$$

Definition 1. We say that zero-dimensional ideals $I_1, I_2, \ldots, I_l$ are equiresidual if

$$RT(I_1) = RT(I_2) = \ldots = RT(I_l).$$

In this case, it is convenient to use the notation:

$$RT(I_1) = RT(I_2) = \ldots = RT(I_l) = RT_1.$$

Obviously, zero-dimensional triangular ideals $I_1, I_2, \ldots, I_l$ are equiresidual if and only if (*ER condition*)

$$D(I_1) = D(I_2) = \ldots = D(I_l).$$

Remark 1. Let I be a zero-dimensional triangular ideal I. According to theorem 4.5 in [8], $d_2, \ldots, d_n$ (not d_1) are uniquely determined by $FC(I)$. It will be used in the proof of theorem 2.

Definition 2. We say that zero-dimensional ideals are strongly equiresidual if

$$RT\left(I_{i_1} I_{i_2} \cdots I_{i_k}\right) = RT\left(I_{j_1} I_{j_2} \cdots I_{j_k}\right),$$

where $1 \le i_1 < i_2 < \ldots < i_k \le l$, $1 \le j_1 < j_2 < \ldots < j_k \le l$, for each $k \in [1, l]$.

Obviously, we have

$$RT\left(I_{i_1} I_{i_2} \cdots I_{i_k}\right) = RT\left(I_{j_1} I_{j_2} \cdots I_{j_k}\right) \Leftrightarrow RP\left(I_{i_1} I_{i_2} \cdots I_{i_k}\right) = RP\left(I_{j_1} I_{j_2} \cdots I_{j_k}\right).$$

In this case, it is convenient to introduce a simpler notation:

$$RT_k = RT\left(I_{i_1} I_{i_2} \cdots I_{i_k}\right),\ RP_k = RP\left(I_{i_1} I_{i_2} \cdots I_{i_k}\right),\ 1 \le i_1 < i_2 < \ldots < i_k \le l.$$

Definition 3. (SDNI condition.) We say that zero-sets $V(I_1)$ and $V(I_2)$ strongly don't intersect if

$$(\alpha_1, \alpha_2, \ldots, \alpha_n) \in V(I_1),\ (\beta_1, \beta_2, \ldots, \beta_n) \in V(I_2) \Rightarrow \alpha_i \ne \beta_j,\ 1 \le i,\ j \le n.$$

Remark 2. The motivation of SDNI is to provide the following property of ER zero-dimensional triangular ideals I_1, I_2:

$$FC(I_1) = FC(I_2) = FC(I_1 I_2).$$

Theorem 1. *Let zero-dimensional triangular ideals $I_1, I_2, \ldots, I_k$ satisfy ER and SDNI conditions. Then their product $I = I_1 I_2 \cdots I_k$ is a triangular ideal.*

P r o o f. ER implies:

$$FC(I_1) = FC(I_2) = \ldots = FC(I_k).$$

SDNI implies that $V(I)$ is equiprojectable with

$$FC(I) = FC(I_j),\ \text{for each}\ j \in [1, k].$$

It follows from theorem 4.5 in [8] that I is a triangular ideal. The theorem 1 is proved.

Theorem 2. *For any integer $l > 0$ there exist strongly equiresidual ideals $I_1, I_2, \ldots, I_l$ of $F_q[x]$.*

P r o o f. If $n = 1$ and $f_1(x),\ f_2(x), \ldots, f_l(x)$ are pairwise different of given degree m then the ideals $\langle f_1(x)\rangle$, $\langle f_2(x)\rangle, \ldots, \langle f_l(x)\rangle$ are strongly equiresidual and $RT_k = \left\{1, x, \ldots, x^{km-1}\right\}$.

In general case pick triangular $I_1, I_2, \ldots, I_l$ under ER and SDNI conditions. According to theorem 1 the product $I = I_1 I_2 \cdots I_k$, $k \le l$, is triangular. Let us calculate $D(I)$. According to CRT, there is a ring isomorphism:

$$F_q[x]/I \cong F_q[x]/I_1 \times F_q[x]/I_2 \times \ldots \times F_q[x]/I_k.$$

Hence,

$$\left|F_q[x]/I\right| = k\left|RP_1\right|.$$

It is the first observation. Secondly,

$$D(I_1) = \ldots = D(I_k) = (d_1, d_2, \ldots, d_n),\ FC(I_1) = \ldots = FC(I_k) = FC(I)$$

implies

$$D(I) = (d_1', d_2, \ldots, d_n).$$

In summary, $d_1' = kd_1$, and

$$D(I) = (kd_1, d_2, \ldots, d_n).$$

The same holds for each product $I_{j_1} I_{j_2} \cdots I_{j_k}$, $1 \le j_1 < j_2 < \ldots < j_k \le l$. The theorem 2 is proved.

Remark 3. Ideals of symmetric relations are strongly equiresidual ideals if their separable polynomials are pairwise coprime and of the same degree (see [8]).

Ideal threshold schemes

We propose the following generalization of Asmuth – Bloom (t, l)-threshold scheme [1]. Pick strongly equiresidual ideals $I_0, I_1, \ldots, I_l$. Let $S(x)$ be a uniformly distributed intermediate secret value, $S(x) \in RP_t$. We identify $RP\left(I_{i_1} I_{i_2} \cdots I_{i_k}\right) = RP_k$ with $F_q[x]/I_{i_1} I_{i_2} \cdots I_{i_k}$. Then we define the secret $s(x)$ and the shares $s_i(x)$, $i = 1, 2, \ldots, l$, as follows:

$$s(x) = S(x) \bmod I_0,\ s_i(x) = S(x) \bmod I_i.$$

Hence, the common space of the secret and secret shares is

$$RP(I_0) = \ldots = RP(I_l) = F_q[x]/I_0 = RP_1.$$

That's why the proposed modular scheme is potentially ideal. The space of $S(x)$ is RP_t.

If k shares $k \geq t$ are known, we uniquely determine $S(x)$ using the CRT-algorithm [2], as $S(x) \in RP_t$. After that we evaluate $s(x)$.

We will use below the following simple fact.

It is well-known that the image of a function $s = f(S)$ has the uniform distribution if the cardinalities of all fibres $f^{-1}(s)$ are the same (*EP condition*).

Theorem 3. *The generalized (t, l)-threshold Asmuth – Bloom scheme with strongly equiresidual ideals is ideal.*

P r o o f. We only need to prove the perfectness. The proof is based on the following ring isomorphism:

$$S(x) \in RP_t = F_q[x]/I_0 I_1 \cdots I_{t-1} \cong F_q[x]/I_0 \times F_q[x]/I_1 \times \ldots \times F_q[x]/I_{t-1}.$$

Therefore, we may put

$$S(x) = \left(s(x), s_1(x), \ldots, s_{t-1}(x)\right).$$

The secret $s(x)$ is the projection of $S(x)$ onto the first component, and the cardinality of every fibre is equal to

$$\left|F_q[x]/I_1\right|\left|F_q[x]/I_2\right| \cdots \left|F_q[x]/I_{t-1}\right|.$$

As $\dim_{F_q} F_q[x]/I_i = d = d_1 d_2 \cdots d_n$, then all cardinalities of the fibers are equal to $q^{d(t-1)}$. Hence, $s(x)$ is uniformly distributed on RP_1.

What happens if a group of $k < t$ participants attempt to compute $s(x)$? Let $I_1, I_2, \ldots, I_k$ be their moduli and $s_1(x), \ldots, s_k(x)$ be their shares. In this case, $S(x)$ is uniformly distributed on the direct product

$$RP_1 \times s_1(x) \times \ldots \times s_k(x) \times \ldots \times (RP_1) \subset RP_t.$$

The map $S(x) \to s(x)$ is EP with $q^{d(t-k-1)}$ being the cardinality of the fibres. Hence, our scheme is perfect.

Example. Shamir's scheme [9] is a particular case of the proposed scheme, which we can see as follows. Take the univariate case and consider different polynomials of degree 1: $x - x_0, x - x_1, x - x_2, \ldots, x - x_l$. The ideals generated by these polynomials are strongly equiresidual. Now if one goes over the construction in theorem 3, one would first construct polynomial of degree at most t. Now taking this polynomial modulo $x - x_i$ is exactly evaluating it in x_i.

Remark 4. The ideals of symmetric relations are suitable for the construction of the ideal secret sharing in the general case $n \geq 1$.

Conclusion

Ideal threshold modular secret sharing schemes in the multivariate polynomial ring over a finite field are presented. The existence of the strongly residual ideals is proved.

Библиографические ссылки

1. Asmuth C, Bloom J. A modular approach to key safeguarding. *IEEE Transactions on Information Theory.* 1983;29(2):208–210. DOI: 10.1109/TIT.1983.1056651.

2. Becker T, Weispfenning V. *Gröbner Bases. A Computational Approach to Commutative Algebra.* New York: Springer-Verlag; 1993. 576 p. (Graduate Texts in Mathematics; volume 141). DOI: 10.1007/978-1-4612-0913-3.

3. Galibus T, Matveev G, Shenets N. Some structural and security properties of the modular secret sharing. In: *Symbolic and Numeric Algorithms for Scientific Computing, 2008. SYNASC 2008. 10th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing; 2008 September 26–29; Timisoara, Romania.* Los Alamitos: IEEE Computer Society Press; 2009. p. 197–200. DOI: 10.1109/SYNASC.2008.14.

4. Galibus T, Matveev G. Generalized mignotte's sequences over polynomial rings. *Electronic Notes in Theoretical Computer Science.* 2007;186(14):43–48. DOI: 10.1016/j.entcs.2006.12.044.

5. Васьковский ММ, Матвеев ГВ. Верификация модулярного разделения секрета. *Журнал Белорусского государственного университета. Математика. Информатика.* 2017;2:17–22.

6. Матвеев ГВ, Матулис ВВ. Совершенная верификация модулярной схемы. *Журнал Белорусского государственного университета. Математика. Информатика.* 2018;2:4–9.

7. Galibus T, Matveev G. Finite Fields. Gröbner Bases and Modular Secret Sharing. *Journal of Discrete Mathematical Sciences and Cryptography.* 2012;15(6):339–348. DOI: 10.1080/09720529.2012.10698386.

8. Aubry P, Valibouze A. Using galois ideals for computing relative resolvents. *Journal of Symbolic Computations.* 2000;30(6): 635–651. DOI: 10.1006/jsco.2000.0376.

9. Shamir A. How to share a secret. *Communications of the ACM.* 1979;22(11):612–613. DOI: 10.1145/359168.359176.

References

1. Asmuth C, Bloom J. A modular approach to key safeguarding. *IEEE Transactions on Information Theory.* 1983;29(2):208–210. DOI: 10.1109/TIT.1983.1056651.

2. Becker T, Weispfenning V. *Gröbner Bases. A Computational Approach to Commutative Algebra.* New York: Springer-Verlag; 1993. 576 p. (Graduate Texts in Mathematics; volume 141). DOI: 10.1007/978-1-4612-0913-3.

3. Galibus T, Matveev G, Shenets N. Some structural and security properties of the modular secret sharing. In: *Symbolic and Numeric Algorithms for Scientific Computing, 2008. SYNASC 2008. 10th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing; 2008 September 26–29; Timisoara, Romania.* Los Alamitos: IEEE Computer Society Press; 2009. p. 197–200. DOI: 10.1109/SYNASC.2008.14.

4. Galibus T, Matveev G. Generalized mignotte's sequences over polynomial rings. *Electronic Notes in Theoretical Computer Science.* 2007;186(14):43–48. DOI: 10.1016/j.entcs.2006.12.044.

5. Vaskouski MM, Matveev GV. Verification of modular secret sharing. *Journal of the Belarusian State University. Mathematics and Informatics.* 2017;2:17–22. Russian.

6. Matveev GV, Matulis VV. Perfect verification of modular scheme. *Journal of the Belarusian State University. Mathematics and Informatics.* 2018;2:4–9. Russian.

7. Galibus T, Matveev G. Finite Fields. Gröbner Bases and Modular Secret Sharing. *Journal of Discrete Mathematical Sciences and Cryptography.* 2012;15(6):339–348. DOI: 10.1080/09720529.2012.10698386.

8. Aubry P, Valibouze A. Using galois ideals for computing relative resolvents. *Journal of Symbolic Computations.* 2000;30(6): 635–651. DOI: 10.1006/jsco.2000.0376.

9. Shamir A. How to share a secret. *Communications of the ACM.* 1979;22(11):612–613. DOI: 10.1145/359168.359176.

Received by editorial board 23.08.2019.

Журнал Белорусского государственного университета. Математика. Информатика. 2019;3:134–138
Journal of the Belarusian State University. Mathematics and Informatics. 2019;3:134–138

УДК 519.854

SEMIONLINE-ВЕРСИЯ ЗАДАЧИ ТЕОРИИ РАСПИСАНИЙ С ДВУМЯ ГРУППАМИ ПРЕДМЕТОВ

В. М. КОТОВ[1)]***, Н. С. БОГДАНОВА***[2)]

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*
[2)]*Гомельский государственный технический университет им. П. О. Сухого, пр. Октября, 48, 246746, г. Гомель, Беларусь*

Предложен метод упаковки для задачи *semionline* с двумя группами предметов. Алгоритмом решения этой задачи является распределение предметов из первой группы с использованием групповой технологии, после чего применяется LS-алгоритм для назначения предметов из второй группы. Чтобы доказать оценку алгоритма, введены разные типы упаковок. В соответствии с весами предметов определены классы предметов. Предложен алгоритм распределения предметов из первой группы для получения необходимых упаковок. На втором этапе применяется алгоритм «в минимально загруженный» с наихудшей оценкой $\frac{17}{9}$.

Ключевые слова: метод упаковки; *semionline*; разбиение; планирование; наихудшая оценка.

BUNCH TECHNIQUE FOR *SEMIONLINE* WITH TWO GROUPS OF ITEMS

V. M. KOTOV[a]***, N. S. BOGDANOVA***[b]

[a]*Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*
[b]*Sukhoi State Technical University of Gomel, 48 Kastryčnika Avenue, Gomel 246746, Belarus*
Corresponding author: V. M. Kotov (kotovvm@bsu.by)

Bunch technique for *semionline* with two groups of items is proposed in this paper. Algorithm to solve this problem is to distribute items from the first group bunch approach and after that apply LS-algorithm to assign items from the second group. In order to prove the estimation of our algorithm is introduced different types of bunches to distribute all items from the first group such a way that only one of the entered types of bunches are obtained. During the second stage we use LS with worst case performance is at most $\frac{17}{9}$.

Keywords: bunch technique; *semionline*; partition; scheduling; worst case performance.

Образец цитирования:
Котов ВМ, Богданова НС. *Semionline*-версия задачи теории расписаний с двумя группами предметов. *Журнал Белорусского государственного университета. Математика. Информатика.* 2019;3:134–138.
https://doi.org/10.33581/2520-6508-2019-3-134-138

For citation:
Kotov VM, Bogdanova NS. Bunch technique for *semionline* with two groups of items. *Journal of the Belarusian State University. Mathematics and Informatics.* 2019;3:134–138. Russian.
https://doi.org/10.33581/2520-6508-2019-3-134-138

Авторы:
Владимир Михайлович Котов – доктор физико-математических наук, профессор; заведующий кафедрой дискретной математики и алгоритмики факультета прикладной математики и информатики.
Наталья Сергеевна Богданова – старший преподаватель кафедры информатики факультета автоматизированных и информационных систем.

Authors:
Vladimir M. Kotov, doctor of science (physics and mathematics), full professor; head of the department of discrete mathematics and algorithms, faculty of applied mathematics and computer science.
kotovvm@bsu.by
Natallia S. Bogdanova, senior lecturer at the department of computer science, faculty of automation and information systems.
bogdanova@gstu.by

Задачи разбиения *online* часто находят реальное применение и тем самым имеют огромное практическое значение. В основном возникают не чисто *online*-задачи, а с доступной некоторой дополнительной информацией о последующих значениях или с дополнительными алгоритмическими возможностями, что позволяет улучшить эффективность алгоритма их решения. Такие задачи являются задачами *semionline* [1–5]. Таким образом, весьма актуально исследование указанных задач и их решений.

Рассмотрим задачу *semionline* с двумя группами предметов. Веса предметов первой группы известны заранее, а предметы второй группы недоступны до тех пор, пока не будут назначены предметы в первой группе. Как только предмет назначается на один из m компьютеров, он больше не может перемещаться. Предметы из второй группы появляются в порядке невозрастания их веса. Цель задачи – минимизировать суммарную загрузку самого загруженного компьютера.

Наиболее естественным алгоритмом для решения рассматриваемой задачи является распределение предметов из первой группы с использованием LPT-алгоритма [1] и с последующим применением LS-алгоритма для назначения предметов из второй группы [3; 4]. Нетрудно понять, что наихудшей оценкой этого алгоритма будет величина $2 - \frac{1}{m}$, т. е. оценка для версии *online* [4; 5]. В данной работе предлагается алгоритм с наихудшей оценкой $\frac{17}{9}$.

Пусть $L = \frac{w_1 + w_2 + \ldots + w_n}{m}$, где w_i – веса предметов первой группы, $i = \overline{1, n}$. Легко видеть, что L – нижняя граница оптимального решения.

Для требуемого распределения по компьютерам введем различные типы упаковок.

*Первый тип упаковок BU*2 состоит из двух компьютеров ($A1$, $A2$), причем $A1$ имеет общий вес не более $\frac{16L}{9}$, а $A2$ – не более $\frac{8L}{9}$. Кроме того, общий вес упаковки составляет не менее $2L$.

Далее через $W(A)$ будем обозначать общий вес предметов, назначенных на компьютер A.

Лемма 1. *Если есть распределение предметов из первой группы таким образом, что получаются только упаковки типа BU*2*, то наихудшая оценка LS составляет не более* $\frac{17}{9}$.

Доказательство. Пусть X – вес текущего предмета и $W(A2) \le \frac{8L}{9}$. Легко видеть, что $\frac{W(A2) + X}{\max\{X, L\}} \le \frac{17}{9}$. Поэтому можно добавить текущий предмет на компьютер $A2$, обеспечив нужную точность.

Добавим текущие предметы на $A2$ в каждой из $BU2$. Это означает, что общий вес добавленных предметов составит не менее kX, где k – количество упаковок данного типа. Поэтому общий вес имеющихся предметов не меньше $kX + mL$. Следовательно, если $k \ge \frac{m}{2}$, то новая нижняя оценка будет как минимум $L + \frac{X}{2}$.

После этого выполняется неравенство $\frac{W(A1) + X}{\max\left\{X, L + \frac{X}{2}\right\}} \le \frac{17}{9}$. Действительно,

$$\frac{W(A1) + X}{\max\left\{X, L + \frac{X}{2}\right\}} \le \frac{\frac{16L}{9} + \frac{8X}{9} + \frac{X}{9}}{\max\left\{X, L + \frac{X}{2}\right\}} \le \frac{16}{9} + \frac{1}{9} = \frac{17}{9}.$$

Лемма 1 доказана.

*Второй тип упаковок BU*3 состоит из трех компьютеров ($B1$, $B2$, $B3$), причем $W(B1) \le \frac{17L}{9}$, $W(B2) \le \frac{4L}{3}$ и $W(B3) \le \frac{8L}{9}$. Кроме того, общий вес упаковки не менее $3L$.

Лемма 2. *Если есть распределение предметов из первой группы таким образом, что получаются только упаковки типа BU*3*, то наихудшая оценка LS составляет не более* $\frac{17}{9}$.

Доказательство. Пусть X – вес текущего предмета. Очевидно, что этот предмет можно добавить на $B3$ в силу $\frac{W(B3)+X}{\max\{X, L\}} \le \frac{17}{9}$. Добавим текущие предметы на $B3$ в каждой из $BU3$. Это означает, что общий вес добавленных предметов составит не менее kX, где k – количество упаковок данного типа. Поэтому общий вес всех имеющихся предметов не меньше $kX + mL$. Следовательно, если $k \ge \frac{m}{3}$, то новая нижняя оценка будет как минимум $L + \frac{X}{3}$.

Докажем, что после этого выполняется соотношение $\frac{W(B2)+X}{\max\left\{X, L+\frac{X}{3}\right\}} \le \frac{17}{9}$. Действительно,

$$\frac{W(B2)+X}{\max\left\{X, L+\frac{X}{3}\right\}} \le \frac{\frac{4L}{3}+\frac{4X}{9}+\frac{5X}{9}}{\max\left\{X, L+\frac{X}{3}\right\}} \le \frac{4}{3}+\frac{5}{9}=\frac{17}{9}.$$

Добавим текущие предметы на $B2$ в каждой из $BU3$. Это означает, что общий вес имеющихся предметов не меньше $2kX + mL$, где k – количество упаковок. Следовательно, если $k \ge \frac{m}{3}$, то новая нижняя оценка будет как минимум $L + \frac{2X}{3}$.

Докажем, что после этого выполняется неравенство $\frac{W(B1)+X}{\max\left\{X, L+\frac{2X}{3}\right\}} \le \frac{17}{9}$. Действительно,

$$\frac{W(B1)+X}{\max\left\{X, L+\frac{2X}{3}\right\}} = \frac{\frac{17L}{9}+\frac{34X}{27}-\frac{7X}{27}}{\max\left\{X, L+\frac{2X}{3}\right\}} \le \frac{17}{9}.$$

Лемма 2 доказана.

*Третий тип упаковок BU*4 состоит из четырех компьютеров ($C1$, $C2$, $C3$, $C4$), причем $W(C1) \le \frac{17L}{9}$, $W(C2) \le \frac{16L}{9}$, $W(C3) \le \frac{10L}{9}$ и $W(C4) \le \frac{8L}{9}$. Кроме того, общий вес упаковки составляет не менее $4L$.

Лемма 3. *Если есть распределение предметов из первой группы таким образом, что получаются только упаковки BU*4*, то наихудшая оценка LS составляет не более* $\frac{17}{9}$.

Доказательство аналогично доказательству лемм 1 и 2.

Теорема 1. *Если все предметы первой группы распределены таким образом, что образованы* k_1 *упаковок первого типа,* k_2 *упаковок второго типа и* k_3 *упаковок третьего типа, причем* $2k_1 + 3k_2 + 4k_3 = m$*, то наихудшая оценка LS составляет не более* $\frac{17}{9}$.

Доказательство. Сначала текущие предметы могут добавляться на компьютеры вида $A2$, $B3$ и $C4$, количество которых не меньше чем $\frac{m}{4}$. Затем текущие предметы могут добавляться на компьютеры вида $C3$. Общее количество $A2$, $B3$, $C3$ и $C4$ не меньше чем $\frac{m}{3}$. Поэтому новые текущие предметы можно добавлять на $B2$. Общее количество компьютеров вида $A2$, $B2$, $B3$, $C3$ и $C4$ не меньше чем $\frac{m}{2}$, поэтому следующие текущие предметы можно добавлять на $A1$ и $C2$, затем на $B1$ и $C1$.

Теорема 1 доказана.

Теперь покажем, как можно распределить предметы из первой группы, чтобы получить необходимые упаковки. В соответствии с весами предметов введем следующие *классы предметов*:

• $I1$ – предметы с $\frac{2L}{3} < X \le \frac{8L}{9}$;

• $I2$ – предметы с $0,6L < X \le \frac{4L}{3}$;

• $I3$ – предметы с $0,5L < X \le 0,6L$;

• $I4$ – предметы с $\frac{4L}{9} < X \le 0,5L$;

• $I5$ – предметы с $X \le \frac{4L}{9}$.

Пусть H является классом предметов с весом больше $\frac{8L}{9}$.

Следствие. *Для достаточного количества элементов из каждого класса $I1$, $I2$, $I3$, $I4$, $I5$ можно составить базовую упаковку одного из типов $BU2$ – $BU4$:*

• *для 3 предметов из $I1$ упаковка $BU2$ имеет вид* $A1 = \{a1, a2\}$; $A2 = \{a3\}$;

• *для 5 предметов из $I2$ упаковка $BU3$ имеет вид* $B1 = \{a1, a2\}$; $B2 = \{a3, a4\}$; $B3 = \{a5\}$;

• *для 6 предметов из $I3$ упаковка $BU3$ имеет вид* $B1 = \{a1, a2, a3\}$; $B2 = \{a4, a5\}$; $B3 = \{a6\}$;

• *для 9 предметов из $I4$ упаковка $BU4$ имеет вид* $C1 = \{a1, a2, a3\}$; $C2 = \{a4, a5, a6\}$; $C3 = \{a7, a8\}$; $C4 = \{a9\}$.

Легко заметить, что из имеющихся предметов класса $I5$ можно составить упаковку $BU3$ следующим образом. Будем помещать предметы в $B1$ и $B2$ пока позволяют ограничения, а 2 предмета, которые не поместились ни в $B1$ и ни в $B2$, поместим в $B3$.

Пусть теперь $|H| > 0$. Следует отметить, что в H могут быть предметы, вес которых превосходит L. В этом случае каждый из них назначается на отдельный компьютер. Для предметов из H, вес которых не превосходит L, упаковки требуемого типа следующие:

• упаковка $BU2$ имеет вид $A1 = \{h, i\_1$ или i_2 или $i\_3\}$, $A2 = \{i\_1$ или i_2 или $i\_3\}$ при $W(i\_3) \ge \frac{5}{9}$;

• упаковка $BU4$ имеет вид $C1 = \{h, i\_3$ или $i\_4\}$, $C2 = \{h, i\_3$ или $i\_4\}$, $C3 = \{i\_3$ или i_4, i_3 или $i\_4\}$, $C4 = \{i\_3$ или $i\_4\}$ при $W(i\_3) < \frac{5}{9}$, $W(h) > \frac{8}{9}$;

• упаковка $BU3$ строится для предметов из $I4$ и имеет вид $B1 = \{h,$ предметы $i\_5\}$, $B2 = \{h,$ предметы $i\_5\}$, $B3 = \{2$ предмета $i\_5\}$.

(Здесь i_j соответствует предмету из класса Ij, $j = 1, 2, \ldots, 5$.) При этом на пустые компьютеры сначала назначаются все предметы из H.

В случае $|H| > m$ нижняя граница оптимального решения может быть уточнена, так как на один компьютер будет назначено минимум 2 предмета из класса H.

Если оказалось, что для предметов из классов $I1$, $I2$, $I3$, $I4$, $I5$ нет пустых компьютеров для формирования компьютера вида $A2$ в упаковке $BU2$, компьютера вида $B3$ в упаковке $BU3$ или компьютеров вида $C3$, $C4$ в упаковке $BU4$, то это означает, что на оставшиеся компьютеры (не входящие в требуемые упаковки) уже назначены предметы из H.

Если еще остались не назначенные предметы из классов $I1$, $I2$, $I3$, $I4$, то нижняя граница оптимального решения может быть уточнена, а все оставшиеся предметы из $I1$, $I2$, $I3$, $I4$ назначены по одному к предмету из класса H. Если остались предметы из класса $I5$, то они будут назначаться на компьютер с предметом из H до тех пор, пока суммарная загрузка компьютера не превосходит $\frac{4L}{3}$. В силу определения величины L все предметы первой группы будут распределены.

После распределения всех предметов из первой группы на втором этапе будет выполняться алгоритм «в минимально загруженный». При этом новая нижняя граница оптимального решения $L1$ пересчитается по следующему правилу.

Пусть $k_1 = |H|$, k_2 соответствует количеству предметов, поступивших из второй группы, $k_3 = m - k_1 - k_2$. Кроме того, пусть $Z1$ соответствует весу максимального предмета из первой группы, $X1$ – весу первого

поступившего предмета из второй группы. Отсортируем предметы из первой группы в порядке невозрастания весов, и пусть $Z2$ соответствует весу предмета на позиции $2k_3 + 1$ (если предметов меньше, то полагаем $Z2 = 0$).

Утверждение. Если поступил предмет из второй группы, вес которого не больше $\frac{8L}{9}$, то алгоритм «в минимально загруженный» обеспечивает оценку $\frac{17}{9}$.

Доказательство очевидно, так как всегда существует компьютер, суммарная загрузка которого не превосходит величины нижней границы оптимального решения.

Поэтому в дальнейшем будем полагать, что для любого предмета из второй группы его вес не меньше $\frac{8L}{9}$.

Лемма 4. *В качестве новой границы оптимального решения* $L1$ *будем выбирать максимальное значение из трех величин: S,* $Z1$ *или* $\min\left\{\frac{8L}{9} + Z2, 3 \cdot Z2\right\}$ *в случае* $k_3 > 0$*; S,* $Z1$ *или* $\frac{8L}{9} + Z2$ *в случае* $k_3 = 0$*. Здесь mS соответствует сумме весов всех имеющихся предметов из первой и второй групп. Кроме того, в качестве* $L1$ *можно полагать* $Z3 + Z4$*, где* $Z3$ *и* $Z4$ *соответствуют значениям весов, стоящих на позициях m и* $m + 1$ *в отсортированном по невозрастанию массиве весов всех имеющихся предметов (первой группы и поступивших из второй группы).*

Доказательство следует из того факта, что при $k_3 > 0$ либо 3 предмета с весом как минимум $Z2$ должны назначаться на один компьютер в оптимальном решении, либо один из них должен назначаться с предметом из H или с поступившим предметом второй группы. Если $k_3 = 0$, то предмет с весом $Z2$ должен назначаться с предметом из H или с поступившим предметом второй группы. Кроме того, два предмета из $m + 1$ предметов с наибольшим весом должны назначаться на один компьютер в оптимальном решении.

Теорема 2. *Наихудшая оценка алгоритма* $\frac{17}{9}$.

Доказательство следует из лемм 1–4, следствия и утверждения.

Библиографические ссылки / References

1. Kellerer H, Kotov V, Speranza MG, Tuza Z. Semi on-line algorithms for the partition problem. *Operations Research Letters.* 1997;21(5):235–242. DOI: 10.1016/ S0167-6377(98)00005-4.

2. Albers S, Hellwig M. Semi-online scheduling revisited. *Theoretical Computer Science.* 2012;443:1–9. DOI: 10.1016/jagm.2012.03.031.

3. He Y, Zhang G. Semi on-line scheduling on two identical machines. Computing. 1999;62(3):179–187. DOI: 10.1007/s006070050020.

4. Gabay M, Kotov V, Brauner N. Semi-online bin stretching with bunch techniques. *Theoretical Computer Science.* 2015;602:103–113. DOI: 10.1016/j.tcs.2015.07.065.

5. Kellerer H, Kotov V, Gabay M. An efficient algorithm for semi-online multiprocessor scheduling with given total processing time. *Journal of Scheduling.* 2015;18(6):623–630. DOI: 10.1007/s10951-015-0430-4.

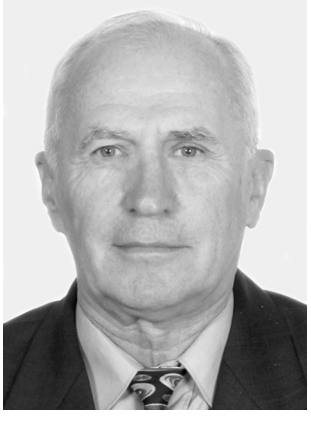
Статья поступила в редколлегию 24.10.2019.
Received by editorial board 24.10.2019.

Юбилеи

Jubilees

Анатолий Афанасьевич ЛЕВАКОВ

Anatoly Afanasyevich LEVAKOV



Исполнилось 70 лет известному белорусскому математику – профессору кафедры высшей математики факультета прикладной математики и информатики доктору физико-математических наук, профессору Анатолию Афанасьевичу Левакову.

А. А. Леваков родился 14 сентября 1949 г. в г. Минске в семье военнослужащего. В 1966 г. окончил с серебряной медалью среднюю школу № 5 в г. Борисове Минской области и в том же году поступил на математический факультет Белорусского государственного университета. В 1970 г. после открытия факультета прикладной математики был переведен на этот факультет, который успешно окончил в 1971 г. В том же году приглашен на кафедру высшей математики в качестве ассистента, где продолжает работать и в настоящее время, пройдя путь от ассистента до профессора. С 1990 по 1995 г. был заведующим указанной кафедрой.

Научные исследования Анатолий Афанасьевич начал еще в студенческие годы. Его первая работа, написанная совместно с Э. С. Шпигельманом под руководством тогда аспиранта, а ныне академика НАН Беларуси Н. А. Изобова, посвящена исследованию двумерных дифференциальных систем с квадратичными правыми частями. Становление А. А. Левакова как математика проходило в рамках минского городского семинара по дифференциальным уравнениям под руководством известного советского ученого профессора Ю. С. Богданова. После учебы в аспирантуре в 1978 г. Анатолий Афанасьевич защитил кандидатскую диссертацию (научный руководитель – член-корреспондент Российской академии образования Н. Х. Розов).

В 1976 г. А. А. Левакову поручено ведение курса «Математический анализ» для студентов факультета прикладной математики, который он читает постоянно до настоящего времени. За прошедшие годы около 2 тыс. студентов прослушали этот курс у Анатолия Афанасьевича, и каждому из них отдана частица его душевной теплоты. В настоящее время

обучение студентов на факультете прикладной математики и информатики осуществляется на основе учебного пособия А. А. Левакова «Математический анализ», изданного в 2014 г. В начале 1990-х гг. в качестве эксперимента он реализовал предложение Ю. С. Богданова о прочтении всех курсов по фундаментальной математике одним лектором. Анатолий Афанасьевич являлся председателем оргкомитета шести Богдановских чтений по дифференциальным уравнениям – международных математических конференций, первая из которых состоялась в 1990 г.

В начале 1990-х гг. А. А. Леваков начал исследование стохастических дифференциальных уравнений – нового и перспективного в теоретическом и прикладном аспектах направления в теории дифференциальных уравнений. За прошедшие годы им были построены общая и асимптотическая теории стохастических дифференциальных уравнений с разрывными правыми частями. Они составили основу его докторской диссертации «Существование и устойчивость дифференциальных и стохастических дифференциальных включений» (научный консультант – академик НАН Беларуси Н. А. Изобов), защищенной в 2004 г. Результаты, полученные Анатолием Афанасьевичем до 2008 г., вошли в его монографию «Стохастические дифференциальные уравнения» (2009), а исследования последнего десятилетия – в книгу «Стохастические дифференциальные уравнения и включения», написанную совместно с его учеником М. М. Васьковским (издана в 2019 г.).

В настоящее время профессор А. А. Леваков продолжает активную научную и педагогическую деятельность: читает лекции по математическому анализу для студентов, работает заместителем заведующего кафедрой высшей математики по научной работе, является членом совета по защите докторских диссертаций БГУ.

Коллектив кафедры высшей математики и студенты факультета прикладной математики и информатики сердечно поздравляют Анатолия Афанасьевича с юбилеем и желают ему крепкого здоровья и новых творческих успехов.

Хроника, информация

Chronicle, information

XII МЕЖДУНАРОДНАЯ НАУЧНАЯ КОНФЕРЕНЦИЯ «КОМПЬЮТЕРНЫЙ АНАЛИЗ ДАННЫХ И МОДЕЛИРОВАНИЕ: СТОХАСТИКА И НАУКА О ДАННЫХ»

XII INTERNATIONAL SCIENTIFIC CONFERENCE «COMPUTER DATA ANALYSIS AND MODELING: STOCHASTICS AND DATA SCIENCE»

В соответствии со сводным тематическим планом проведения научных и научно-технических мероприятий Республики Беларусь на 2019 г. и перечнем конференций Международного института математической статистики (International Statistical Institute) на базе Белорусского государственного университета с 18 по 22 сентября 2019 г. прошла XII Международная научная конференция «Компьютерный анализ данных и моделирование: стохастика и наука о данных», приуроченная к 100-летию БГУ.

Организаторами конференции выступили НИИ прикладных проблем математики и информатики, факультет прикладной математики и информатики БГУ и Венский университет технологий (Австрия). Конференция проходила при финансовой поддержке Министерства образования Республики Беларусь, Белорусского республиканского фонда фундаментальных исследований, Венского университета технологий, научно-технологической ассоциации «Инфопарк», ЗАО «Итранзишэн» и ЗАО «БСБ Банк».

Тематика конференции сочетает такую фундаментальную науку, как математическая статистика (в качестве теоретической основы компьютерного анализа данных), с актуальными прикладными исследованиями. Она охватывает математические модели, методы, алгоритмы и программные средства анализа стохастических данных и моделирования в целях получения оптимальных статистических выводов (решений, оценок, прогнозов), а также разработки компьютерных моделей сложных стохастических систем и процессов.

Прикладную направленность тематики отражает новая наука на стыке статистики и информатики – data science (наука о данных), имеющая целью создание компьютерных методов и информационных технологий для извлечения значимой информации из массивов данных произвольной природы. Появление этой науки обусловлено интенсивной цифровизацией человеческого общества и экспоненциальным ростом количества данных.

Форум является крупнейшим на постсоветском пространстве и организуется раз в три года. Среди более чем 120 участников нынешней конференции – известные в мире ученые и молодые исследователи из 20 стран: Австралии, Австрии, Беларуси, Болгарии, Великобритании, Вьетнама, Германии, Индии, Италии, Литвы, Нидерландов, Норвегии, Польши, России, США, Узбекистана, Украины, Финляндии, Чехии и Эстонии. В конференции приняли участие ученые Института математики и Объединенного института проблем информатики НАН Беларуси. Издан сборник материалов конференции на английском языке. Лучшие доклады будут опубликованы в специальных выпусках зарубежных журналов «Austrian Journal of Statistics», «Applied Econometrics», «Informatica».

Проведение подобного форума в БГУ закономерно, поскольку университет является ведущим учреждением высшего образования страны по ряду приоритетных научных направлений, связанных с разработкой и использованием новых информационных технологий, включая компьютерный анализ данных и моделирование сложных

Участники заседания XII Международной научной конференции
«Компьютерный анализ данных и моделирование: стохастика и наука о данных»

Participants in the meeting of the XII International Scientific Conference
«Computer data analysis and modeling: stochastics and data science»

систем в различных областях: научных исследованиях, экономике, социологии, защите информации, медицине, проектировании транспортных и производственных систем, банковской деятельности, бизнесе и др. Среди актуальных для страны практических достижений НИИ прикладных проблем математики и информатики БГУ наиболее востребованы программные комплексы моделирования и прогнозирования микро- и макроэкономических показателей по заказу Национального банка Республики Беларусь, научно обоснованные методики и эталонные программные реализации алгоритмов криптографической защиты электронного документооборота страны, формирование учебно-научно-производственного кластера в области прикладной математики и информатики.

Важность конференции определяется также и участием в ней молодых ученых, аспирантов, магистрантов, студентов. На факультете прикладной математики и информатики БГУ в 2016–2018 гг. в рамках программы Евросоюза *TEMPUS* совместно с пятью европейскими университетами открыта пользующаяся большим спросом магистратура «Прикладной компьютерный анализ данных». Молодежь привлекается для исследований в студенческих научных кружках, четырех студенческих научно-исследовательских лабораториях, а также занимается анализом данных и моделированием в рамках реальных научных проектов в НИИ прикладных проблем математики и информатики. Студенческая научно-исследовательская лаборатория кафедры математического моделирования и анализа данных за последние 10 лет трижды удостаивалась финансовой поддержки специального фонда Президента Республики Беларусь за высокие показатели в научной деятельности.

Организация и проведение конференции получили высокую оценку со стороны ее участников. На заключительном заседании поступило предложение о проведении очередного форума «Компьютерный анализ данных и моделирование».

Ю. С. Харин[1]

[1]*Юрий Семенович Харин* – доктор физико-математических наук, профессор, член-корреспондент НАН Беларуси; директор Научно-исследовательского института прикладных проблем математики и информатики БГУ, научный руководитель кафедры математического моделирования и анализа данных факультета прикладной математики и информатики БГУ.

Yurii S. Kharin, doctor of science (physics and mathematics), full professor, corresponding member of National Academy of Sciences of Belarus; director of the Research Institute for Applied Problems of Mathematics and Informatics, Belarusian State University, supervisor of the department mathematical modeling and data analysis, faculty of applied mathematics and computer science, Belarusian State University.

E-mail: *kharin@bsu.by*

АННОТАЦИИ ДЕПОНИРОВАННЫХ В БГУ РАБОТ
INDICATIVE ABSTRACTS OF THE PAPERS DEPOSITED IN BSU

УДК 519.6(075.8)

Методы вычислений [Электронный ресурс] : электрон. учеб.-метод. комплекс для спец. 1-31 03 07 «Прикладная информатика (по направлениям)» / сост. Б. В. Фалейчик ; БГУ. Электрон. текстовые дан. Минск, 2019. 293 с. : ил. Библиогр.: с. 291–292. Режим доступа: http://elib.bsu.by/handle/123456789/226305. Загл. с экрана. Деп. 06.08.2019, № 008206082019.

Настоящий электронный учебно-методический комплекс разработан в соответствии с образовательным стандартом первой ступени высшего образования для специальности 1-31 03 07 «Прикладная информатика (по направлениям)» и предназначен для информационно-методического обеспечения преподавания дисциплины «Методы вычислений» для студентов данной специальности.

СОДЕРЖАНИЕ

CONTENTS

Журнал включен Высшей аттестационной комиссией Республики Беларусь в Перечень научных изданий для опубликования результатов диссертационных исследований по физико-математическим наукам (в области математики и информатики).

Журнал включен в библиографические базы данных научных публикаций «Российский индекс научного цитирования» (РИНЦ), Mathematical Reviews, Ulrichsweb, Google Scholar, zbMath.

Журнал Белорусского государственного университета. Математика. Информатика. № 3. 2019

Учредитель:
Белорусский государственный университет

Юридический адрес: пр. Независимости, 4, 220030, г. Минск.
Почтовый адрес: пр. Независимости, 4, 220030, г. Минск.
Тел. (017) 259-70-74, (017) 259-70-75.
E-mail: jmathinf@bsu.by
URL: https://journals.bsu.by/index.php/mathematics

«Журнал Белорусского государственного университета. Математика. Информатика» издается с января 1969 г.
До 2017 г. выходил под названием «Вестник БГУ. Серия 1, Физика. Математика. Информатика» (ISSN 1561-834X).

Редактор *Т. Р. Джум*
Технический редактор *В. В. Пишкова*
Корректор *Л. А. Меркуль*

Подписано в печать 29.11.2019.
Тираж 100 экз. Заказ 469.

Республиканское унитарное предприятие «Информационно-вычислительный центр Министерства финансов Республики Беларусь».
ЛП № 02330/89 от 03.03.2014.
Ул. Кальварийская, 17, 220004, г. Минск.

Journal of the Belarusian State University. Mathematics and Informatics. No. 3. 2019

Founder:
Belarusian State University

Registered address: 4 Niezaliežnasci Ave., Minsk 220030.
Correspondence address: 4 Niezaliežnasci Ave., Minsk 220030.
Tel. (017) 259-70-74, (017) 259-70-75.
E-mail: jmathinf@bsu.by
URL: https://journals.bsu.by/index.php/mathematics

«Journal of the Belarusian State University. Mathematics and Informatics» published since January, 1969.
Until 2017 named «Vestnik BGU. Seriya 1, Fizika. Matematika. Informatika» (ISSN 1561-834X).

Editor *T. R. Dzhum*
Technical editor *V. V. Pishkova*
Proofreader *L. A. Merkul'*

Signed print 29.11.2019.
Edition 100 copies. Order number 469.

Republican Unitary Enterprise «Informatsionno-vychislitel'nyi tsentr Ministerstva finansov Respubliki Belarus'».
License for publishing No. 02330/89, 3 March, 2014.
17 Kal'varyjskaja Str., Minsk 220004.