ISSN 1812-5093

ТРУДЫ ИНСТИТУТА МАТЕМАТИКИ НАН БЕЛАРУСИ

Государственное научное учреждение
«Институт математики Национальной академии наук Беларуси»

Журнал основан в 1998 г.
До 2024 г. выходил под названием
«Труды Института математики».

PROCEEDINGS OF THE INSTITUTE OF MATHEMATICS OF THE NAS OF BELARUS

State Scientific Institute
«The Institute of Mathematics
of the National Academy of Sciences
of Belarus»

The journal was founded in 1998.
Before 2024, the name of the journal was
«Proceedings of the Institute of Mathematics».

ТОМ 32, № 2
Минск, 2024

ОТ РЕДАКЦИИ

В журнале «Труды Института математики НАН Беларуси» публикуются оригинальные статьи фундаментальной и прикладной математики. Обзорные статьи публикуются по решению редакционной коллегии. Журнал входит в «Перечень научных изданий Республики Беларусь для опубликования результатов диссертационных исследований», утвержденный Высшей аттестационной комиссией Республики Беларусь.

До 2024 г. журнал издавался под названием «Труды Института математики».

Основные разделы журнала:
Алгебра и теория чисел
Вещественный, комплексный и функциональный анализ
Вычислительная математика
Дискретная математика и математическая кибернетика
Дифференциальные уравнения, динамические системы и оптимальное управление
Математическое моделирование и численные методы
Математические методы защиты информации
Краткие сообщения

Адрес редакции:	**Address:**
ул. Сурганова, 11, к. 45	11, Surganov Str., room 45,
220072, г. Минск, Беларусь	Minsk, Belarus, 220072
Тел. +375-17-379-17-84	Phone: +375-17-379-17-84

e-mail: tbusel@im.bas-net.by

EDITOR'S NOTE

The journal "Proceedings of the Institute of Mathematics of the NAS of Belarus" publishes original articles in fundamental and applied mathematics. Publication of review articles requires approval of the editorial board. The journal is approved by the Higher Attestation Commission of the Republic of Belarus for publishing results of dissertation research, and is included in the official list of such publications.

Before 2024, the name of the journal was "Proceedings of the Institute of Mathematics" (Trudy Instituta Matematiki).

Main sections of the journal:
Algebra and number theory
Real, complex and functional analysis
Computational mathematics
Discrete mathematics and mathematical cybernetics
Differential equations, dynamic systems and optimal control
Mathematical modeling and numerical methods
Mathematical methods for information security
Brief communications

ТРУДЫ
ИНСТИТУТА МАТЕМАТИКИ НАН БЕЛАРУСИ
Минск. 2024. Т. 32. № 2

СОДЕРЖАНИЕ

P R O C E E D I N G S

OF THE INSTITUTE OF MATHEMATICS OF THE NAS OF BELARUS

Minsk. 2024. Vol. 32. N 2

CONTENTS

Национальная академия наук Беларуси
Труды Института математики НАН Беларуси. 2024. Том 32. № 2. С. 7–16

АЛГЕБРА И ТЕОРИЯ ЧИСЕЛ
ALGEBRA AND NUMBER THEORY

УДК 512.71

ФУНДАМЕНТАЛЬНАЯ ОБЛАСТЬ В СПЕЦИАЛЬНОЙ ЛИНЕЙНОЙ ГРУППЕ $SL_2(\mathbb{F}_p[x])$ И СХЕМА РАЗДЕЛЕНИЯ СЕКРЕТА НА ЕЕ ОСНОВЕ

Г. В. Матвеев[1]**, А. А. Осиновская**[2]**, В. И. Янчевский**[2]

[1]*Белорусский государственный университет, Минск, Беларусь*
[2]*Институт математики НАН Беларуси, Минск, Беларусь*
e-mail: matveev47@bsu.by, anna@im.bas-net.by, yanch@im.bas-net.by

Поступила: 16.10.2024 **Исправлена:** 23.11.2024 **Принята:** 12.12.2024

Ключевые слова: специальная линейная группа, конгруэнц-подгруппа, фундаментальная область, модулярное разделение секрета, пороговая структура доступа.

Аннотация. Решается задача по разработке математических основ модулярного разделения секрета в специальной линейной группе над кольцом многочленов от одной переменной над конечным полем Галуа из p элементов. К схемам разделения секрета предъявляется большое число требований: совершенность и идеальность схемы, возможность проведения верификации, изменение порога без участия дилера, реализация непороговой структуры доступа и некоторые другие. Каждая разработанная к настоящему времени схема разделения секрета не в полной мере удовлетворяет всем этим требованиям. Разработка схемы на новой математической основе призвана расширить список этих конфигураций, что создает для пользователя больше возможностей в выборе оптимального варианта. В специальной линейной группе размерности 2 над кольцом многочленов строится фундаментальная область относительно действия главной конгруэнц-подгруппы правыми сдвигами. На этой основе предложены способы модулярного порогового разделения секрета и его восстановления.

A FUNDAMENTAL DOMAIN IN THE SPECIAL LINEAR GROUP $SL_2(\mathbb{F}_p[x])$ AND SECRET SHARING ON ITS BASIS

G. V. Matveev[1]**, A. A. Osinovskaya**[2]**, V. I. Yanchevskiĭ**[2]

[1]*Belarusian State University, Minsk, Belarus*
[2]*Institute of Mathematics of the National Academy of Sciences of Belarus, Minsk, Belarus*
e-mail: matveev47@bsu.by, anna@im.bas-net.by, yanch@im.bas-net.by

Received: 16.10.2024 **Revised:** 23.11.2024 **Accepted:** 12.12.2024

Keywords: a special linear group, a congruence subgroup, a fundamental domain, modular secret sharing, a threshold access structure.

Abstract. The problem of developing the mathematical foundations of modular secret sharing in the special linear group over the ring of polynomials in one variable over the finite Galois field with p elements is being solved. Secret sharing schemes should meet a large number of requirements: perfectness and ideality of a scheme, possibility of verification, changing a threshold without participation of a dealer, implementation of a non-threshold access structure and some others. Every secret sharing scheme developed to date does not fully satisfy all these requirements. The development of a scheme on a new mathematical basis is intended to expand the list of these configurations, thereby creating more possibilities for a user to choose the optimal option. A fundamental domain with respect to the action of the main congruence subgroup by right shifts in the special linear group of dimension 2 over the ring of polynomials is constructed. On this basis, methods for modular threshold secret sharing and its reconstruction are proposed.

1. Введение

В последнее время все большее значение приобретает организация схем доступа к тем или иным информационным ресурсам. Подобного рода задачи призваны решать схемы разделения секрета, относящиеся к числу важных криптографических протоколов. Они используются в системах электронного голосования [1], шифрования на основе атрибутов [2] и в распределенных конфиденциальных вычислениях [3].

Схема разделения секрета решает следующую задачу. Пусть имеется некоторая важная информация (секрет) s и множество $P = \{1, 2, \ldots, k\}$ пользователей. Требуется сообщить каждому пользователю i некоторую информацию s_i (частичный секрет) таким образом, чтобы только заранее определенные группы участников могли, объединяя свои частичные секреты, восстановить секрет s, а для остальных групп эта задача являлась бы трудноразрешимой. Как правило, под этим понимается, что задача восстановления секрета неразрешенной группой участников должна быть эквивалентна полному перебору.

Настоящая работа посвящена модулярному подходу в теории разделения секрета. Основы этого подхода и теории в целом были заложены в работе А. Шамира [4], а собственно модулярный подход получил развитие в работах К. Асмута и Дж. Блума [5] и М. Миньотта [6].

В дальнейшем модулярный подход был развит в работах [7–9]. В частности, он был обобщен на кольца многочленов от одной и нескольких переменных над полем Галуа. Было показано, что любая структура доступа допускает модулярную реализацию в кольцах целых чисел и многочленов над полями Галуа. В работе [7] было доказано, что модулярный подход в кольце многочленов от одной переменной над полем Галуа позволяет реализовать пороговую структуру доступа совершенно и идеально. Более того, модулярная пороговая схема в кольце многочленов от одной переменной над полем Галуа легла в основу стандарта Республики Беларусь 12.34.101.60–2014 «Информационные технологии и безопасность. Алгоритмы разделения секрета». В работах [10; 11] были предложены методы верификации модулярных схем.

В настоящее время для схем разделения секрета разработано много критериев качества, таких как совершенность, идеальность, верифицируемость, пригодность для реализации предпороговых структур доступа и ряд других. Схем разделения секрета, удовлетворяющих всем известным критериям качества, еще нет. Вот почему представляет интерес построение новых схем, основанных на принципиально иной алгебраической базе. В настоящей работе в качестве такой базы предлагается специальная линейная группа над кольцом многочленов от одной переменной над полем Галуа и модулярное разделение секрета в ней.

2. Модулярное разделение секрета

Определение 2.1. Под структурой доступа Γ множества $P = \{1, 2, \ldots, k\}$ пользователей понимают монотонное семейство подмножеств, т. е. семейство, для элементов которого выполняется условие

$$A \in \Gamma,\ A \subset B \subset P \Rightarrow B \in \Gamma.$$

Эти подмножества называют разрешенными, а остальные – запрещенными. Под реализацией структуры доступа будем понимать построение соответствующей схемы разделения секрета.

Структура доступа, когда разрешенными считаются подмножества A с условием $|A| \geqslant t$, называется (t, k)-*пороговой*, а число t, $1 \leqslant t \leqslant k$, называется ее *порогом*.

Схемой разделения секрета (СРС) называют алгоритмы распределения частичных секретов и восстановления исходного секрета. Они, в частности, должны обеспечивать правильное восстановление секрета разрешенными группами участников. Схему разделения секрета называют *совершенной*, если запрещенное множество участников не получает никакой информации о секрете, кроме априорной.

Схему разделения секрета называют *идеальной*, если ключи всех участников и ключ s имеют один и тот же размер. Иногда в условие идеальности включают и совершенство схемы.

В самых общих чертах СРС позволяет распределить секрет между t участниками таким образом, чтобы заданные разрешенные множества участников могли однозначно восстановить секрет, а неразрешенные – не получили бы никакой дополнительной к имеющейся априорной информации о возможном значении секрета.

Модулярное разделение секрета основано на следующем простом наблюдении (схема Миньотта [6]). Пусть $m_1 < m_2 < \ldots < m_k$ – система попарно взаимно простых натуральных модулей. Если секретом является некоторое натуральное число s, а секретом i-го участника, $i \in P = \{1, 2, .., k\}$, является наименьший неотрицательный вычет s по модулю m_i, т. е. $s_i \equiv s \pmod{m_i}$, то группа участников $A \subset P$ восстанавливает исходный секрет s путем решения системы сравнений $x \equiv \equiv s_i (\bmod\ m_i)$, $i \in A$. Это можно сделать, например, с помощью китайской теоремы об остатках. При этом правильно найдет секрет s лишь та группа участников A, для которой выполнено условие $s < \prod\limits_{i \in A} m_i$. Тот же принцип используется при построении схемы разделения секрета над кольцом многочленов от одной и нескольких переменных [7–9].

Замечание. В схеме Асмута–Блума [5] пользователи находят вспомогательный секрет как указано выше. Хранимым секретом является вычет вспомогательного по некоторому несекретному модулю m_0.

3. Фундаментальная область в специальной линейной группе

Целью статьи является построение модулярной схемы разделения секрета в специальной линейной группе $SL_2(\mathbb{F}_p[x])$, где SL_2 – множество квадратных матриц размерности 2 с определителем 1, $\mathbb{F}_p[x]$ – кольцо многочленов от одной переменной над конечным полем $\mathbb{F}_p$. Мы хотим найти в этой группе все необходимое для построения схем подобно тому, как это происходит в кольце целых чисел $\mathbb{Z}$ [5; 6], в кольце многочленов $\mathbb{F}_p[x]$ [7–9] и в группе $SL_2(\mathbb{Z})$ [12].

Кольцо $\mathbb{F}_p[x]$ имеет много общих свойств с кольцом целых чисел $\mathbb{Z}$ (см. [13, глава 1]). Оба кольца являются областями главных идеалов, оба имеют конечную группу единиц, и оба обладают тем свойством, что у каждого кольца класс вычетов по модулю ненулевого идеала имеет конечное число элементов. Обратимые элементы $\mathbb{F}_p[x]$ – это в точности ненулевые константы (элементы $\mathbb{F}_p^*$). Кольцо $\mathbb{F}_p[x]$, как и $\mathbb{Z}$, является евклидовым, при этом евклидова норма – это степень многочлена deg, т. е. для любых двух многочленов $f(x), g(x) \in \mathbb{F}_p[x]$, $g(x) \neq 0$, имеется однозначное представление в виде $f(x) = g(x)q(x) + r(x)$, для которого $\deg r(x) < \deg g(x)$ или $r(x) = 0$.

Напомним, что в евклидовом кольце каждый необратимый элемент представим в виде конечного произведения простых элементов, и притом однозначно (с точностью до их перестановки и умножения на обратимые элементы). Многочлен $f(x)$ называется нормированным, если его старший коэффициент равен 1. Как и в кольце целых чисел, в кольце многочленов справедлива следующая лемма [13, предложение 1.4], называемая иначе китайской теоремой об остатках.

Лемма 3.1. *Пусть многочлены* $f_1(x)$, $f_2(x)$, …, $f_r(x) \in \mathbb{F}_p[x]$ *– попарно взаимно просты,* $f(x) = f_1(x)f_2(x)\ldots f_r(x)$ *и* $\varphi_i : \mathbb{F}_p[x]/(f(x)) \to \mathbb{F}_p[x]/(f_i(x))$ *– естественные гомоморфизмы. Тогда отображение*

$$\varphi : \mathbb{F}_p[x]/(f(x)) \to \mathbb{F}_p[x]/(f_1(x)) \times \ldots \times \mathbb{F}_p[x]/(f_r(x)), \text{ где для } a \in \mathbb{F}_p[x]/(f(x))$$

$$a \mapsto (\varphi_1(a), \varphi_2(a), \ldots, \varphi_r(a)),$$

является изоморфизмом колец.

Возьмем некоторый многочлен $m(x) \in \mathbb{F}_p[x]$ ненулевой степени. Многочлены $f(x)$ и $g(x)$ сравнимы по модулю $m(x)$ (что записывается $f(x) \equiv \underline{m(x)(\bmod\ m(x)))}$, если их остатки при делении на $m(x)$ совпадают. Это эквивалентно тому, что $\overline{f(x)} = \overline{g(x)}$, где $\overline{f(x)}$ – образ многочлена $f(x)$ относительно канонического гомоморфизма $\mathbb{F}_p[x] \to \mathbb{F}_p[x]/(m(x))$. Заметим, что $|\mathbb{F}_p[x]/(m(x))| = = p^{\deg m(x)}$.

Определение 3.2. Для ненулевого многочлена $g(x)$ положим $|g(x)| = p^{\deg g(x)}$.

Очевидно, для любого натурального числа e и любого многочлена $g(x) \in \mathbb{F}_p[x]$ справедливо свойство

$$|g^e(x)| = p^{\deg(g^e(x))} = p^{e \deg g(x)} = |g(x)|^e.$$

Введем следующие подгруппы группы $SL_2(\mathbb{F}_p[x])$:

$$\Gamma_0(m(x)) = \left\{ \begin{pmatrix} a(x) & b(x) \\ c(x) & d(x) \end{pmatrix} \in SL_2(\mathbb{F}_p[x]) \ : \ \begin{pmatrix} a(x) & b(x) \\ c(x) & d(x) \end{pmatrix} \equiv \begin{pmatrix} a(x) & b(x) \\ 0 & d(x) \end{pmatrix} \ (\bmod\ m(x)) \right\},$$

$$\Gamma_1(m(x)) = \left\{ \begin{pmatrix} a(x) & b(x) \\ c(x) & d(x) \end{pmatrix} \in SL_2(\mathbb{F}_p[x]) \ : \ \begin{pmatrix} a(x) & b(x) \\ c(x) & d(x) \end{pmatrix} \equiv \begin{pmatrix} 1 & b(x) \\ 0 & 1 \end{pmatrix} \ (\bmod\ m(x)) \right\},$$

$$\Gamma(m(x)) = \left\{ \begin{pmatrix} a(x) & b(x) \\ c(x) & d(x) \end{pmatrix} \in SL_2(\mathbb{F}_p[x]) \ : \ \begin{pmatrix} a(x) & b(x) \\ c(x) & d(x) \end{pmatrix} \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \ (\bmod\ m(x)) \right\}.$$

Здесь и далее сравнимость матриц по модулю $m(x)$ понимают как их поэлементную сравнимость. Очевидно, что справедливы включения

$$\Gamma(m(x)) \subset \Gamma_1(m(x)) \subset \Gamma_0(m(x)) \subset SL_2(\mathbb{F}_p[x]).$$

Подгруппа $\Gamma(m(x))$ называется главной конгруэнц-подгруппой по модулю $m(x)$, остальные подгруппы – просто конгруэнц-подгруппами.

Построим сначала несколько важных гомоморфизмов.

Лемма 3.3. *Отображение* $\varphi : SL_2(\mathbb{F}_p[x]) \to SL_2(\mathbb{F}_p[x]/(m(x)))$, *где*

$$\begin{pmatrix} a(x) & b(x) \\ c(x) & d(x) \end{pmatrix} \mapsto \begin{pmatrix} \overline{a(x)} & \overline{b(x)} \\ \overline{c(x)} & \overline{d(x)} \end{pmatrix},$$

является сюрьективным гомоморфизмом с ядром $\ker\varphi = \Gamma(m(x))$.

Доказательство. Очевидно, что φ – гомоморфизм групп. Возьмем произвольный элемент $y \in SL_2(\mathbb{F}_p[x]/(m(x)))$. Пусть $M_{2\times 2}$ – множество квадратных матриц размерности 2. Тогда существует матрица

$$\gamma(x) = \begin{pmatrix} a(x) & b(x) \\ c(x) & d(x) \end{pmatrix} \in M_{2\times 2}(\mathbb{F}_p[x]),$$

такая, что

$$y = \overline{\gamma(x)} = \begin{pmatrix} \overline{a(x)} & \overline{b(x)} \\ \overline{c(x)} & \overline{d(x)} \end{pmatrix}.$$

Заметим, что поскольку $a(x)d(x) - b(x)c(x) = 1 + m(x)k(x)$ для некоторого $k(x) \in \mathbb{F}_p[x]$, мы получаем $(c(x), d(x), m(x)) = 1$. Построим матрицу

$$\gamma'(x) = \begin{pmatrix} a'(x) & b'(x) \\ c'(x) & d'(x) \end{pmatrix} \in SL_2(\mathbb{F}_p[x]),$$

для которой $\overline{\gamma'(x)} = \overline{\gamma(x)}$, она и будет искомым прообразом.

Мы утверждаем, что существуют многочлены $c'(x), d'(x)$, для которых $c'(x) \equiv c(x)(\bmod\ m(x))$, $d'(x) \equiv d(x)(\bmod\ m(x))$ и $(c'(x), d'(x)) = 1$ (они дадут нам вторую строку нашей матрицы $\gamma'(x)$). Положим

$$d'(x) = \begin{cases} d(x), & \text{если } d(x) \neq 0, \\ m(x) & \text{в противном случае.} \end{cases}$$

Если $d'(x) = d' \in \mathbb{F}_p$, можно взять $c'(x) = c(x)$. В противном случае разложим $d'(x)$ на простые нормированные множители $d'(x) = \alpha P_1^{e_1}(x) \cdot \dots \cdot \dots \cdot P_r^{e_r}(x)$, $\alpha \in \mathbb{F}_p$. Для каждого $i \in \{1, \dots, r\}$ определим $t_i(x) = 1$, если $P_i(x) | c(x)$, и $t_i(x) = 0$, если $P_i(x) \nmid c(x)$. Согласно лемме 3.1, существует многочлен $t(x) \in \mathbb{F}_p[x]$ такой, что $t(x) \equiv t_i(x)(\bmod\ P_i(x))$ для всех $i \in \{1, \dots, r\}$. Положим

$$c'(x) = c(x) + t(x)m(x)$$

и докажем от противного, что $(c'(x), d'(x)) = 1$.

Пусть это не так, тогда существует неприводимый многочлен $P(x) = P_i(x)$ такой, что $P(x) \mid d'(x)$ и $P(x) \mid c'(x)$. Если мы предположим, что $P(x) \nmid c(x)$, то по построению $t(x) \equiv 0 \ (\bmod\ P(x))$ и поэтому $P(x) \mid (c'(x) - t(x)m(x)) = c(x)$?! Если же мы предположим, что $P(x) \mid c(x)$, то $t(x) \equiv$ $\equiv 1(\bmod\ P(x))$ и поэтому равенство $c'(x) = c(x) + t(x)m(x)$ приводит к $0 \equiv m(x)(\bmod\ P(x))$, откуда следует $P(x) \mid m(x)$, $P(x) \mid c(x)$ и $P(x) \mid d(x)$, что противоречит равенству $(c(x),d(x),m(x)) = 1$. Значит, $c'(x)$ и $d'(x)$ взаимно просты.

Теперь заметим, что по построению $a(x)d'(x) - b(x)c'(x) = 1 + m(x)k'(x)$ для некоторого $k'(x) \in \mathbb{F}_p[x]$, а также существуют $f(x), g(x) \in \mathbb{F}_p[x]$ такие, что $f(x)c'(x) + g(x)d'(x) = 1$. Положив

$$a'(x) = a(x) - k'(x)g(x)m(x),\; b'(x) = b(x) + k'(x)f(x)m(x),$$

получаем $\det\gamma'(x) = 1$, а потому $\gamma'(x)$ – искомая матрица. Сюръективность доказана.

Утверждение $\ker\varphi = \Gamma(m(x))$ очевидно. $\square$

Лемма 3.4. *Отображение* $\psi : \Gamma_0(m(x)) \to (\mathbb{F}_p[x]/(m(x)))^*$, *задаваемое формулой*

$$\begin{pmatrix} a(x) & b(x) \\ m(x)c'(x) & d(x) \end{pmatrix} \mapsto \overline{d(x)},$$

является сюръективным гомоморфизмом групп и $\ker\psi = \Gamma_1(m(x))$.

Доказательство. Это гомоморфизм групп, поскольку

$$\begin{pmatrix} a(x) & b(x) \\ m(x)c'(x) & d(x) \end{pmatrix} \begin{pmatrix} a_1(x) & b_1(x) \\ m(x)c_1'(x) & d_1(x) \end{pmatrix} \mapsto \overline{m(x)c'(x)b_1(x) + d(x)d_1(x)} = \overline{d(x)}\,\overline{d_1(x)}.$$

Докажем, что он сюръективен. Возьмем произвольный элемент $y \in (\mathbb{F}_p[x]/(m(x)))^*$. Тогда существует $d(x) \in \mathbb{F}_p[x]$: $y = \overline{d(x)}$. Так как $\overline{d(x)}$ обратим, то существует $a(x) \in \mathbb{F}_p[x]$ такой, что $\overline{a(x)}\,\overline{d(x)} = \overline{1}$. Это равносильно равенству $a(x)d(x) = 1 + m(x)k(x)$ для некоторого $k(x) \in \mathbb{F}_p[x]$. Матрица

$$\begin{pmatrix} a(x) & k(x) \\ m(x) & d(x) \end{pmatrix} \in \Gamma_0(m(x))$$

является искомым прообразом. Равенство $\ker\psi = \Gamma_1(m(x))$ очевидно. $\square$

Лемма 3.5. *Отображение* $\xi : \Gamma_1(m(x)) \to \mathbb{F}_p[x]/(m(x))$, *из мультипликативной группы* $\Gamma_1(m(x))$ *в аддитивную группу кольца* $\mathbb{F}_p[x]/(m(x))$, *которое задается формулой*

$$\begin{pmatrix} 1 + m(x)a'(x) & b(x) \\ m(x)c'(x) & 1 + m(x)d'(x) \end{pmatrix} \mapsto \overline{b(x)},$$

является сюръективным гомоморфизмом групп с ядром $\ker\xi = \Gamma(m(x))$.

Доказательство. Очевидно, что отображение корректно определено и является гомоморфизмом. Докажем его сюръективность. Выберем произвольный элемент $y \in \mathbb{F}_p[x]/(m(x))$, для него существует $b(x) \in \mathbb{F}_p[x]$: $y = \overline{b(x)}$. Тогда матрица

$$\begin{pmatrix} 1 + m(x)b(x) & b(x) \\ m(x) & 1 \end{pmatrix} \in \Gamma_1(m(x))$$

является прообразом y. Теперь легко видеть, что $\ker\xi = \Gamma(m(x))$. $\square$

Значок $\lhd$ как обычно обозначает нормальность подгруппы в группе.

Следствие 3.6. $\Gamma(m(x)) \lhd SL_2(\mathbb{F}_p[x])$ *и* $\Gamma_1(m(x)) \lhd \Gamma_0(m(x))$.

Доказательство. Утверждение следует из лемм 3.3 и 3.4. $\square$

Замечание 3.7. *Однако* $\Gamma_1(m(x)) \ntriangleleft SL_2(\mathbb{F}_p[x])$ *и* $\Gamma_0(m(x)) \ntriangleleft SL_2(\mathbb{F}_p[x])$.

Действительно,

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \in \Gamma_1(m(x)) \subset \Gamma_0(m(x)),$$

но

$$\begin{pmatrix} 1 & 0 \\ p-1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ p-1 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 2 & 1 \\ p-1 & 0 \end{pmatrix} \notin \Gamma_0(m(x)).$$

Поскольку $SL_2(\mathbb{F}_p[x]/(m(x)))$ – конечная группа, то по лемме 3.3 индекс $[SL_2(\mathbb{F}_p[x]) : \Gamma(m(x))]$ конечен. Поэтому подгруппа $\Gamma_0(m(x))$ также имеет конечный индекс в $SL_2(\mathbb{F}_p[x])$. Найдем эти индексы. Для этого введем функцию $\Phi(f(x))$ – число ненулевых многочленов степени, меньшей $\deg f(x)$, взаимно простых с $f(x)$. Справедливо

Предложение 3.8 [13, предложение 1.7]. *Пусть $P(x)$ – неприводимый многочлен, тогда*

$$\Phi(m(x)) = |m(x)| \prod_{P(x)|m(x)} \left(1 - \frac{1}{|P(x)|}\right).$$

Заметим, что $\Phi(P(x)) = |P(x)| - 1$, а $\Phi(P^e(x)) = |P^e(x)| - |P^{e-1}(x)|$.

Лемма 3.9. *Каноническое разложение многочлена $m(x)$ на простые множители:*

$$m(x) = \alpha P_1^{e_1}(x) \times \ldots \times P_r^{e_r}(x), \ \alpha \in \mathbb{F}_p,$$

индуцирует естественный изоморфизм

$$\psi : SL_2(\mathbb{F}_p[x]/(m(x))) \to SL_2(\mathbb{F}_p[x]/(P_1^{e_1}(x))) \times \ldots \times SL_2(\mathbb{F}_p[x]/(P_r^{e_r}(x))).$$

Доказательство. Инъективность отображения ψ следует из леммы 3.1. В нашей ситуации, если для матрицы $\gamma(x) \in SL_2(\mathbb{F}_p[x])$ справедливо $\gamma(x) \equiv I_2 \pmod{P_i^{e_i}(x)}$ для всех $i \in \{1, \ldots, r\}$, то $\gamma(x) \equiv I_2 \pmod{m(x)}$, где I_2 – единичная матрица размерности 2.

Теперь докажем сюръективность ψ. Пусть

$$(\overline{\gamma^{(1)}(x)}, \ldots, \overline{\gamma^{(r)}(x)}) \in SL_2(\mathbb{F}_p[x]/(P_1^{e_1}(x))) \times \ldots \times SL_2(\mathbb{F}_p[x]/(P_r^{e_r}(x))).$$

Снова применяя лемму 3.1, мы можем найти матрицу $\gamma(x) \in M_{2\times 2}(\mathbb{F}_p[x])$ такую, что $\gamma(x) \equiv \gamma^{(i)}(x) \pmod{P_k^{e_i}(x)}$ для всех $i \in \{1, \ldots, r\}$. Поскольку $\det\gamma(x) \equiv \det\gamma^{(k)}(x) \equiv 1 \pmod{P_k^{e_i}(x)}$ для всех k, то $\det\gamma(x) \equiv 1 \pmod{m(x)}$. Значит, искомый прообраз – это $\overline{\gamma(x)} \in SL_2(\mathbb{F}_p[x]/(m(x)))$. $\square$

Предложение 3.10. *(i)* $[\Gamma_1(m(x)) : \Gamma(m(x))] = |m(x)|$;

(ii) $[\Gamma_0(m(x)) : \Gamma_1(m(x))] = \Phi(m(x))$;

(iii) $[SL_2(\mathbb{F}_p[x]) : \Gamma(m(x))] = |m(x)|^3 \prod_{P(x)|m(x)} \left(1 - \frac{1}{|P(x)|^2}\right)$;

(iv) $[SL_2(\mathbb{F}_p[x]) : \Gamma_0(m(x))] = |m(x)| \prod_{P(x)|m(x)} \left(1 + \frac{1}{|P(x)|}\right)$.

Доказательство. *(i)* Из леммы 3.5 вытекает, что

$$[\Gamma_1(m(x)) : \Gamma(m(x))] = |\mathbb{F}_p[x]/(m(x))| = |m(x)|.$$

(ii) По лемме 3.4,

$$[\Gamma_0(m(x)) : \Gamma_1(m(x))] = |(\mathbb{F}_p[x]/(m(x)))^*| = \Phi(m(x)).$$

(iii) Применяя лемму 3.9, получаем, что нам достаточно доказать утверждение при $m(x) = P^e(x)$. Из леммы 3.3 следует, что

$$[SL_2(\mathbb{F}_p[x]) : \Gamma(P^e(x))] = |SL_2(\mathbb{F}_p[x]/(P^e(x)))|.$$

Символом GL_2 обозначим множество обратимых квадратных матриц размерности 2. Поскольку гомоморфизм групп $\det : GL_2(\mathbb{F}_p[x]/(P^e(x))) \to (\mathbb{F}_p[x]/(P^e(x)))^*$ сюръективен и его ядро равно $SL_2(\mathbb{F}_p[x]/(P^e(x)))$, имеем

$$|SL_2(\mathbb{F}_p[x]/(P^e(x)))| = \frac{|GL_2(\mathbb{F}_p[x]/(P^e(x)))|}{\Phi(P^e(x))} = \frac{|GL_2(\mathbb{F}_p[x]/(P^e(x)))|}{|P^e(x)| - |P^{e-1}(x)|},$$

и нам осталось определить порядок группы $GL_2(\mathbb{F}_p[x]/(P^e(x)))$.

Естественный гомоморфизм групп

$$\zeta : GL_2(\mathbb{F}_p[x]/(P^e(x))) \to GL_2(\mathbb{F}_p[x]/(P(x)))$$

сюръективен. Действительно, если матрица $\gamma(x) \in M_{2\times 2}(\mathbb{F}_p[x])$ переходит в матрицу $\overline{\gamma(x)} \in GL_2(\mathbb{F}_p[x]/(P(x)))$, то из $\det\gamma(x) \not\equiv 0 \pmod{P(x)}$ следует, что $\det\gamma(x) \not\equiv 0 \pmod{P^e(x)}$, и поэтому матрица $\gamma(x) \pmod{P^e(x)} \in GL_2(\mathbb{F}_p[x]/(P^e(x)))$ и переходит при отображении ζ в $\overline{\gamma(x)}$. Получаем точную последовательность:

$$1 \to K \to GL_2(\mathbb{F}_p[x]/(P^e(x))) \to GL_2(\mathbb{F}_p[x]/(P(x))) \to 1,$$

где $K = \{I_2 + A(x) \mid A(x) \in P(x)M_{2\times 2}(\mathbb{F}_p[x]/(P^e(x)))\}$. Поэтому

$$|GL_2(\mathbb{F}_p[x]/(P^e(x)))| = |K| \cdot |GL_2(\mathbb{F}_p[x]/(P(x)))|.$$

Заметим, что $|\mathbb{F}_p[x]/(P(x))| = |P(x)|$. Кроме того, поскольку $P(x)$ неприводим, любой ненулевой многочлен $f(x) \in \mathbb{F}_p[x]$ либо делится на $P(x)$, либо взаимно прост с ним, т. е. существуют многочлены $q(x)$, $r(x) \in \mathbb{F}_p[x]$ такие, что $f(x)q(x) + P(x)r(x) = 1$, что равносильно сравнению $f(x)q(x) \equiv 1(\bmod\ P(x))$. Отсюда следует, что вычет $\overline{f(x)}$ по модулю $P(x)$ обратим и $\mathbb{F}_p[x]/(P(x))$ является полем.

Заметим, что $GL_2(\mathbb{F}_p[x]/(P(x))) \cong GL_2(\mathbb{F}_q)$, где $q = |P(x)|$. Поэтому [14, §4, с. 19]

$$|GL_2(\mathbb{F}_p[x]/(P(x)))| = (|P(x)|^2 - 1)(|P(x)|^2 - |P(x)|) = (|P(x)| - 1)^2(|P(x)| + 1)|P(x)|.$$

С другой стороны, число многочленов, степень которых меньше $\deg P^e(x)$ и которые делятся на $P(x)$, равно $|P^e(x)| - \Phi(P^e(x)) = |P^e(x)| - (|P^e(x)| - |P^{e-1}(x)|) = |P^{e-1}(x)|$. Поэтому

$$|K| = |M_{2\times 2}(\mathbb{F}_p[x]/(P^e(x)))| = |P^{e-1}(x)|^4 = |P(x)|^{4e-4}.$$

Отсюда получаем, что

$$|GL_2(\mathbb{F}_p[x]/(P^e(x)))| = (|P(x)| - 1)^2(|P(x)| + 1)|P(x)|^{4e-3},$$

$$|SL_2(\mathbb{F}_p[x]/(P^e(x)))| = \frac{(|P(x)| - 1)^2(|P(x)| + 1)||P(x)|^{4e-3}}{(|P(x)| - 1)|P(x)|^{e-1}} = |P^e(x)|^3\left(1 - \frac{1}{|P(x)|^2}\right).$$

(iv) Следует из пунктов (i)–(iii), поскольку

$$[SL_2(\mathbb{F}_p[x]) : \Gamma(m(x))] = |m(x)|^2\Phi(m(x)) \prod_{P(x)|m(x)} \left(1 + \frac{1}{|P(x)|}\right).$$

□

Заметим, что несмотря на глубокое сходство между $\mathbb{Z}$ и $\mathbb{F}_p[x]$, а также между $SL_2(\mathbb{Z})$ и $SL_2(\mathbb{F}_p[x])$, между ними есть и существенное различие. Так, группа $SL_2(\mathbb{Z})$ порождается двумя элементами, например,

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \text{ и } \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

а для $SL_2(\mathbb{F}_p[x])$ это неверно. Х. Нагао показал в [15], что группа $SL_2(\mathbb{F}_p[x])$ не является конечнопорожденной.

Также при $e > 1$ существенно отличается строение групп $(\mathbb{Z}/p^e\mathbb{Z})^*$ и $(\mathbb{F}_p[x]/(P^e(x)))^*$ (см. [13, предложение 1.6]). Если p нечетно, что группа $(\mathbb{Z}/p^e\mathbb{Z})^*$ циклическая. Если $p = 2$ и $e \geqslant 3$, то $(\mathbb{Z}/2^e\mathbb{Z})^*$ – это прямое произведение циклической группы порядка 2 и циклической группы порядка 2^{e-2}. В то же время справедливо

Предложение 3.11. *Пусть $P(x)$ – неприводимый многочлен, а e – целое число. Пусть $(\mathbb{F}_p[x]/(P^e(x)))^{(1)}$ – ядро естественного гомоморфизма из $(\mathbb{F}_p[x]/(P^e(x)))^*$ в $(\mathbb{F}_p[x]/(P(x)))^*$. Тогда это p-группа порядка $|P(x)|^{e-1}$. Если e стремится к бесконечности, то минимальное число образующих $(\mathbb{F}_p[x]/(P^e(x)))^{(1)}$ тоже стремится к бесконечности.*

Для реализации пороговой модулярной схемы в группе $SL_2(\mathbb{F}_p[x])$ нужно получить явное описание фундаментальной области относительно подгруппы $\Gamma(m(x))$ при ее действии правыми сдвигами, что является аналогом полной системы вычетов $\{0, 1, \ldots, n-1\}$ по некоторому модулю n в кольце $\mathbb{Z}$. Ввиду леммы 3.3 это можно было бы сделать с помощью техники подъема (лифтинга), однако эта задача пока решена не в полной мере. В [16, с. 438–439] указывается, что установить возможность подъема решений целочисленного уравнения $xy - zt = 1$ по некоторой системе модулей до целочисленного решения не так-то легко. Еще больше сложностей возникает при решении такого уравнения в кольце $\mathbb{F}_p[x]$.

Нам удалось построить значительную часть этой области, что достаточно для наших целей. Дадим явное описание части фундаментальной области группы $SL_2(\mathbb{F}_p[x])$ при действии на ней правыми сдвигами группой $\Gamma(m(x))$. Укажем семейство попарно несравнимых элементов группы $\Gamma_0(m(x))$ по модулю $m(x)$ в количестве, равном $[\Gamma_0(m(x)) : \Gamma_1(m(x))] = \Phi(m(x))$. Возьмем произвольный элемент

$\varepsilon_i(x) \in \mathbb{F}_p[x]$ такой, что $\overline{\varepsilon_i(x)} \in (\mathbb{F}_p[x]/(m(x)))^*$. Тогда существует $\varepsilon_i'(x) \in \mathbb{F}_p[x]$:

$$\varepsilon_i(x)\varepsilon_i'(x) = 1 + m(x)k_i(x)$$

и $\varepsilon_i'(x)$, $k_i(x)$ однозначно определяются выбором $\varepsilon_i(x)$. Имеем

$$A_i(x) = \begin{pmatrix} \varepsilon_i(x) & k_i(x) \\ m(x) & \varepsilon_i'(x) \end{pmatrix} \in \Gamma_0(m(x)).$$

Таких матриц имеется ровно $\Phi(m(x))$, а значит, можно считать, что $i = 1, \ldots, \Phi(m(x))$. Образуем из матриц $A_i(x)$ новые матрицы:

$$B_{i,j(x)}(x) = \begin{pmatrix} \varepsilon_i(x) + j(x)m(x) & k_i(x) + j(x)\varepsilon_i'(x) \\ m(x) & \varepsilon_i'(x) \end{pmatrix} \in \Gamma_0(m(x)),$$

где $i = 1, \ldots, \Phi(m(x))$, $\deg j(x) < \deg m(x)$. Матрицы $B_{i,j(x)}(x)$ попарно различны при различных i, $j(x)$. Действительно, если $i_1 \neq i_2$, то

$$\varepsilon_{i_1}(x) + j_1(x)m(x) \equiv \varepsilon_{i_2}(x) + j_2(x)m(x) \ (\bmod\ m(x)) \Leftrightarrow \varepsilon_{i_1}(x) \equiv \varepsilon_{i_2}(x) \ (\bmod\ m(x))?!$$

Если $i_1 = i_2 = i$, но $j_1(x) \neq j_2(x)$, то

$$k_i(x) + j_1(x)\varepsilon_i'(x) \equiv k_i(x) + j_2(x)\varepsilon_i'(x) \ (\bmod\ m(x)) \Leftrightarrow j_1(x)\varepsilon_i'(x) \equiv j_2(x)\varepsilon_i'(x) \ (\bmod\ m(x)) \Leftrightarrow$$

$$\Leftrightarrow j_1(x) \equiv j_2(x) \ (\bmod\ m(x))?!$$

Тем самым доказана

Теорема 3.12. *Матрицы $A_i(x)$ при $i = 1, \ldots, \Phi(m(x))$, составляют фундаментальную область группы $\Gamma_0(m(x))$ относительно подгруппы $\Gamma_1(m(x))$, а матрицы $B_{i,j(x)}(x)$ при $i = 1, \ldots$ $\ldots, \Phi(m(x))$, $\deg j(x) < \deg m(x)$, – фундаментальную область группы $\Gamma_0(m(x))$ относительно подгруппы $\Gamma(m(x))$.*

Таким образом мы построили часть фундаментальной области, необходимую для реализации схемы разделения секрета.

4. Пороговое модулярное разделение секрета в группе $SL_2(\mathbb{F}_p[x])$

Построим в группе $SL_2(\mathbb{F}_p[x])$ модулярное пороговое разделение секрета. Напомним, что необходимо подобрать семейство $m_1(x), m_2(x), \ldots, m_k(x) \in \mathbb{F}_p[x]$ попарно взаимно простых модулей участников. Условия $m_1 < m_2 < \ldots < m_k$ и

$$M_1 = m_{k-t+2}m_{k-t+3}\ldots m_k < m_1m_2\ldots m_t = M_2$$

из схемы Миньотта трансформируются в неравенство для сумм степеней:

$$M_1 = \max_{A\in\overline{\Gamma}} \sum_{j\in A} \deg m_j(x) < \min_{A\in\Gamma} \sum_{i\in A} \deg m_i(x) = M_2. \tag{4.1}$$

Очевидно, что если модули участников имеют одинаковую степень n, то неравенство (4.1) для (t,k)-пороговой схемы выполняется автоматически, поскольку оно равносильно неравенству $(t-1)n < tn$.

Тогда если многочлен $s(x)$ выбран так, что $M_1 < \deg s(x) < M_2$, то он однозначно определяется своими вычетами по любым t или более модулям $m_i(x)$ с помощью китайской теоремы об остатках (лемма 3.1). Если же вычетов меньше, чем t, то решение соответствующей системы сравнений будет отличаться от искомого $s(x)$.

Предложим следующую схему разделения секрета в группе $SL_2(\mathbb{F}_p[x])$.

1) Выбор открытых ключей (модулей) участников схемы.

В качестве открытых ключей берутся главные конгруэнц-подгруппы $\Gamma(m_1(x)), \ldots, \Gamma(m_k(x))$, где модули $m_1(x), \ldots, m_k(x)$ попарно взаимно просты и имеют степень n.

2) Выбор секрета S и частичных секретов участников.

Секретом S является матрица из фундаментальной области

$$S = \begin{pmatrix} s_i(x) + j(x)m(x) & k_i(x) + j(x)s_i'(x) \\ m(x) & s_i'(x) \end{pmatrix},$$

где $m(x) = m_1(x)\dots m_k(x)$, $(s_i(x), m(x)) = 1$, $i = 1, \dots, \Phi(m(x))$, причем $M_1 < \deg s_i(x) < M_2$,

$$s_i(x)s_i'(x) = 1 + m(x)k(x), \tag{4.2}$$

$\deg j(x) < M_2$.

Частичными секретами участников являются поэлементные вычеты этой матрицы по модулям $m_1(x), \dots, m_k(x)$. Например, частичным секретом первого участника будет образ матрицы S при каноническом эпиморфизме

$$SL_2(\mathbb{F}_p[x]) \to SL_2(\mathbb{F}_p[x])/\Gamma(m_1(x)) \cong SL_2(\mathbb{F}_p[x]/(m_1(x))),$$

что является аналогом обычного частичного секрета в схеме Миньотта.

3) Восстановление секрета S по частичным секретам подмножества участников A, где $|A| \geqslant t$.

- $m(x)$ находится автоматически.
- Нам известны $s_i(x) + j(x)m(x) \equiv s_i(x) \pmod{m_r(x)}$, $r \in A$, по китайской теореме об остатках (лемма 3.1) находим $s_i(x) \pmod{\prod\limits_{r\in A} m_r(x)}$. Найденное решение в силу выбора $s_i(x)$ будет одним и тем же по модулям $\prod\limits_{l\in A} m_l(x)$ и $m(x)$, так как $\deg s_i(x) < \sum\limits_{l\in A} \deg m_l(x)$.
- Решив сравнение $s_i(x)s_i'(x) \equiv 1 \pmod{m(x)}$, находим $s_i'(x)$. Напомним, что все модули $m_1(x), \dots$ $\dots, m_k(x)$ известны участникам.
- Учитывая равенство (4.2), многочлен $k_i(x)$ однозначно восстанавливается по формуле $k_i(x) =$ $= \frac{s_i(x)s_i'(x)-1}{m(x)}$.
- Нам известны $k_i(x) + j(x)s_i'(x) \pmod{m_r(x)}$, $r \in A$, Используя лемму 3.1, находим $k_i(x) +$ $+ j(x)s_i'(x) \pmod{\prod\limits_{r\in A} m_r(x)}$. Поскольку $(s_i'(x), m(x)) = 1$, то значение $j(x) \pmod{\prod\limits_{l\in A} m_l(x)}$ восстанавливается однозначно. Так как $\deg j(x) < M_2$, отсюда получаем $j(x)$.

Таким образом, матрица S корректно восстановлена.

Работа выполнена при поддержке Государственной программы научных исследований «Конвергенция-2025», задание 1.1.01.

Литература

1. *Cramer R., Damgard I., Nielsen J.* Multiparty computation from threshold homomorphic encryption // LNCS. 2001. Vol. 2045. P. 280–300. https://doi.org/10.1007/3-540-44987-6_18

2. *Bethencourt J., Sahai A., Waters B.* Ciphertext-policy attribute-based encryption // 2007 IEEE Symposium on Security and Privacy (SP'07), IEEE, 2007. P. 321–334. https://doi.org/10.1109/SP.2007.11

3. *Benaloh J.* Secret sharing homomorphisms: keeping shares of a secret sharing // LNCS. 1987. Vol. 263. P. 251–260. https://doi.org/10.1007/3-540-47721-7_19

4. *Shamir A.* How to share a secret // Communications of the ACM. 1979. Vol. 22. P. 612–613. https://doi.org/10.1145/359168.359176

5. *Asmuth C., Bloom J.* A modular approach to key safeguarding // IEEE Transactions on Information Theory. 1983. Vol. 29. P. 156–169. https://doi.org/10.1109/TIT.1983.1056651

6. *Mignotte M.* How to share a secret // LNCS. 1983. Vol. 149. P. 371–375. https://doi.org/10.1007/3-540-39466-4_27

7. *Galibus T., Matveev G., Shenets N.* Some structural and security properties of the modular secret sharing // Proceedings of SYNASC'08, IEEE, Los Alamitos, 2009. P. 197–200. https://doi.org/10.1109/SYNASC.2008.14

8. *Galibus T., Matveev G.* Generalized Mignotte's sequences over polynomial rings // Electronic Notes in Theoretical Computer Science. 2007. Vol. 186. P. 43–48. https://doi.org/10.1016/j.entcs.2006.12.044

9. *Galibus T., Matveev G.* Finite fields, Gröbner bases and modular secret sharing // Journal of Discrete Mathematical Sciences and Cryptography. 2012. Vol. 15. P. 339–348. https://doi.org/10.1080/09720529.2012.10698386

10. *Васьковский М. М., Матвеев Г. В.* Верификация модулярного разделения секрета // Журн. Белорус. гос. ун-та. Математика. Информатика. 2017. № 2. С. 17–22.

11. *Матвеев Г. В., Матулис В. В.* Совершенная верификация модулярной схемы // Журн. Белорус. гос. ун-та. Математика. Информатика. 2018. № 2. С. 4–9.

12. *Янчевский В. И., Говорушко И. О., Матвеев Г. В.* Разделение секрета в специальной линейной группе // Информатика. 2024. Т. 21, № 3. С. 39–47. https://doi.org/10.37661/1816-0301-2024-21-3-39-47

13. *Rosen M.* Number theory in function fields. New York: Springer-Verlag, 2002. 358 p.

14. *Taylor D. E.* The geometry of the classical groups. Berlin: Herdelmann Verlag, 1992. 229 p.

15. *Nagao H.* On $GL(2;K[X])$ // Journal of the Institute of Polytechnics, Osaka City University. Series A: Mathematics. 1959. Vol. 10. P. 117–121.

16. *Платонов В. П., Рапинчук А. С.* Алгебраические группы и теория чисел. М.: Наука, 1991. 656 с.

References

1. Cramer R., Damgard I., Nielsen J. Multiparty computation from threshold homomorphic encryption. *LNCS*, 2001, vol. 2045, pp. 280–300. https://doi.org/10.1007/3-540-44987-6_18

2. Bethencourt J., Sahai A., Waters B. Ciphertext-policy attribute-based encryption. *2007 IEEE Symposium on Security and Privacy (SP'07)*, IEEE, 2007, pp. 321–334. https://doi.org/10.1109/SP.2007.11

3. Benaloh J. Secret sharing homomorphisms: keeping shares of a secret sharing. *LNCS*, 1987, vol. 263, pp. 251–260. https://doi.org/10.1007/3-540-47721-7_19

4. Shamir A. How to share a secret. *Communications of the ACM*, 1979, vol. 22, pp. 612–613. https://doi.org/10.1145/359168.359176

5. Asmuth C., Bloom J. A modular approach to key safeguarding. *IEEE Transactions on Information Theory*, 1983, vol. 29, pp. 156–169. https://doi.org/10.1109/TIT.1983.1056651

6. Mignotte M. How to share a secret. *LNCS*, 1983, vol. 149, pp. 371–375. https://doi.org/10.1007/3-540-39466-4_27

7. Galibus T., Matveev G., Shenets N. Some structural and security properties of the modular secret sharing. *Proceedings of SYNASC'08*, IEEE, Los Alamitos, 2009, pp. 197–200. https://doi.org/10.1109/SYNASC.2008.14

8. Galibus T., Matveev G. Generalized Mignotte's sequences over polynomial rings. *Electronic Notes in Theoretical Computer Science*, 2007, vol. 186, pp. 43–48. https://doi.org/10.1016/j.entcs.2006.12.044

9. Galibus T., Matveev G. Finite fields, Gröbner bases and modular secret sharing. *Journal of Discrete Mathematical Sciences and Cryptography*, 2012, vol. 15, pp. 339–348. https://doi.org/10.1080/09720529.2012.10698386

10. Vaskouski M. M., Matveev G. V. Verification of modular secret sharing. *Journal of the Belarusian State University. Mathematics and Informatics*, 2017, no. 2, pp. 17–22 (in Russian).

11. Matveev G. V., Matulis V. V. Perfect verification of modular scheme. *Journal of the Belarusian State University. Mathematics and Informatics*, 2018, no. 2, pp. 4–9 (in Russian).

12. Yanchevskiĭ V. I., Havarushka I. A., Matveev G. V. Secret sharing in a special linear group. *Informatics*, 2024, vol. 21, no. 3, pp. 23–31 (in Russian). https://doi.org/10.37661/1816-0301-2024-21-3-23-31

13. Rosen M. *Number theory in function fields*. New York, Springer-Verlag, 2002, 358 p.

14. Taylor D. E. *The geometry of the classical groups*. Berlin, Herdelmann Verlag, 1992, 229 p.

15. Nagao H. On $GL(2;K[X])$. *Journal of the Institute of Polytechnics, Osaka City University. Series A: Mathematics*, 1959, vol. 10, pp. 117–121.

16. Platonov V. P., Rapinchuk A. S. *Algebraic groups and number theory*. Moscow, Nauka, 1991, 656 p. (in Russian).

Национальная академия наук Беларуси
Труды Института математики НАН Беларуси. 2024. Том 32. № 2. С. 17–30

УДК 512.542

О РАЗРЕШИМОСТИ И ФАКТОРИЗАЦИИ НЕКОТОРЫХ π-РАЗРЕШИМЫХ НЕПРИВОДИМЫХ ЛИНЕЙНЫХ ГРУПП ПРИМАРНОЙ СТЕПЕНИ. ЧАСТЬ IV

А. А. Ядченко

Институт математики НАН Беларуси, Минск, Беларусь
e-mail: yadchenko_56@mail.ru

Поступила: 19.11.2024 **Исправлена:** 25.11.2024 **Принята:** 12.12.2024

Ключевые слова: конечные группы, характеры, факторизация групп.

Аннотация. Работа является четвертой из серии статей, где для множества π, состоящего из нечетных простых чисел, исследуются конечные π-разрешимые неприводимые комплексные линейные группы степени $2|H|+1$, у которых холловы π-подгруппы H являются TI-подгруппами и не являются нормальными в группах. Цель серии – доказать разрешимость и определить условия факторизации таких групп.

ON SOLVABILITY AND FACTORIZATION OF SOME π-SOLVABLE IRREDUCIBLE LINEAR GROUPS OF PRIMARY DEGREE. PART IV

A. A. Yadchenko

Institute of Mathematics of the National Academy of Sciences of Belarus, Minsk, Belarus
e-mail: yadchenko_56@mail.ru

Received: 19.11.2024 **Revised:** 25.11.2024 **Accepted:** 12.12.2024

Keywords: finite groups, characters, factorizations of groups.

Abstract. The article is the fourth in a series of papers, where for a set π consisting of odd primes, finite π-solvable irreducible complex linear groups of degree $2|H|+1$ are investigated, for which Hall π-subgroups are TI-subgroups and are not normal in groups. The purpose of the series is to prove solvability and to determine the conditions for factorization of such groups.

1. Введение

Исследуются конечные π-разрешимые неприводимые комплексные линейные группы степени $n = 2|H|+1$, у которых холлова π-подгруппа H является TI-подгруппой.

В первой части работы [1] были доказаны некоторые предварительные результаты, а также получены некоторые свойства минимального контрпримера Γ к теореме $(*)$, которая является основой доказательства главной теоремы. Во второй [2] и в третьей [3] ее частях продолжено изучение свойств минимального контрпримера к теореме $(*)$.

Условие (*). Скажем, что для Γ, A, B, C, χ и n выполнено условие (*), если $\Gamma = BA$, где B – нормальная в Γ подгруппа, $(|B|,|A|) = 1$, A – группа нечетного порядка, большего 3, которая не является нормальной в группе Γ, $C_B(a) = C_B(A) = C$ для каждого элемента $a \in A^{\#}$, и B имеет точный неприводимый характер χ степени n, который является a-инвариантным хотя бы для одного элемента $a \in A^{\#}$.

Теорема (*). *Пусть для* Γ, A, B, C, χ *и* $n = 2|A|+1$ *выполнено условие (*). Тогда группа* Γ *разрешима,* n *является степенью простого числа* q, *подгруппа* $C_{2'}$ *абелева и, если подгруппа* C *не абелева, то в обозначениях леммы* 2.7 *[1] характер* $\chi_C = \beta + |A|\beta_1$, *где* $\beta(1) = |A|+1$ *и* $\beta_1(1) = 1$ *либо* $\beta(1) = 1$ *и* $\beta_1(1) = 2$. *Также, если* $|\pi| > 1$ *и при* $|\pi| = 1$ *характер* $\widehat{\chi}$ *примитивный, то* $G = O_q(B)C$.

Условие 1. Пусть π – множество нечетных простых чисел и G – конечная не π-замкнутая π-разрешимая группа с π-холловой TI-подгруппой H, имеющая точный неприводимый характер степени n.

Теорема. *Пусть группа G удовлетворяет условию* 1 *и* $n = 2|H|+1 > 7$. *Тогда n – степень простого числа q, группа G разрешима и, если при* $|\pi| = 1$ *характер* χ *примитивный, то* $G = N_G(H)O_q(G)$.

2. Некоторые определения, обозначения и предварительные результаты

$\mathbf{Z}_+$ – множество целых неотрицательных чисел; $i = \overline{1,n} = 1,2,...n$; если ψ – характер некоторой подгруппы $X \subseteq G$, то $\mathrm{Irr}(\psi)$ обозначает множество всех неприводимых компонент характера ψ; $\pi = \pi(H)$; $\pi' = \pi(X)\setminus\pi$; $X_{\pi'}$ – холлова π'-подгруппа группы X. Если $X \triangleleft G$ и φ – неприводимый характер подгруппы X, то условие, что φ – g-инвариантен для некоторого элемента $g \in G\setminus X$, запишем для краткости в виде $I_G(\varphi) \neq X$. Все остальные обозначения и определения обычны и их можно найти, например, в [4] или [5]. Всюду под характером группы будем понимать комплексный характер, а под группой – конечную группу.

Пусть $\Gamma = AB$ – группа с подгруппами A и B, где $B \triangleleft \Gamma$, $(|A|,|B|) = 1$ и $|A|$ нечетен (A – *группа копростых автоморфизмов группы* B). Тогда она удовлетворяет условию теоремы 13.1 [5]. Согласно этой теореме существует взаимно-однозначное соответствие $\pi(B,A) : \mathrm{Irr}_A(B) \longrightarrow \mathrm{Irr}(C_B(A))$ между множеством всех A-инвариантных неприводимых характеров группы B и множеством всех неприводимых характеров подгруппы $C_B(A)$, которое обладает рядом свойств, зависящих, в частности, от свойств подгруппы A. Пусть $\chi \in \mathrm{Irr}_A(B)$. Тогда, по лемме 13.3 [5], существует такой единственный неприводимый характер $\widehat{\chi}$ группы Γ, что $\widehat{\chi}_B = \chi$ и $A \subseteq \ker(\det\widehat{\chi})$. Он называется *каноническим продолжением характера* χ *на группу* Γ. В дальнейшем под $\widehat{\chi}$ будем понимать именно такой характер.

3. Основная часть

Продолжим нумерацию формулировок лемм, начатую в первой части работы [1]. В ней, напомним, мы получили некоторые предварительные результаты, показали, что неприводимый характер $\widehat{\chi}$ точный и что $\widehat{\chi}(1) = q^\alpha$ для некоторого нечетного простого числа q и $\alpha \in \mathbf{N}$, а также начали изучать свойства минимального контрпримера к теореме $(^*)$. Во второй части [2] мы продолжили изучение минимального контрпримера к теореме $(^*)$ и доказали свойства подгруппы C и ее характера χ_C.

Напомним, что $N = N_\Gamma(B_q)$, где $\widehat{\chi}(1) = q^\alpha$ и B_q – A-инвариантная подгруппа, т. е. $A \subseteq N$. В третьей части [3] работы мы установили, что $N \neq \Gamma$ (лемма 3.13) и что $N^{(q_1)} = N_\Gamma(B_{q_1}) \neq \Gamma$ для $q \neq q_1 \in \pi'$ (лемма 3.14).

Лемма 3.15. *Характер* $\widehat{\chi}_N$ *неприводим.*

Доказательство.

Предположим, что характер $\widehat{\chi}_N$ не является неприводимым. Так как $[B_q, A] \not\subseteq C$ по лемме 3.3 [2], то $A \not\triangleleft N$. Следовательно, мы можем применить лемму 3.2 [1]. Положим в ней, что $\Gamma_1 = N$. Тогда в обозначениях указанной леммы $\widehat{\chi}_N = \eta + \theta$.

Допустим, что $\eta = 0$. Тогда $\theta = \theta_1 + \theta_2$, где θ_1 и θ_2 – неприводимые характеры подгруппы N степеней $|A|$ и $|A|+1$ соответственно. Мы можем применить утверждения 2 леммы 3.2 [1]. По ее утверждению (2_i) $|A|+1$ – степень 2 и $|\pi| > 1$. Поскольку $(|A|,q) = 1$ и $(|A|+1,q) = 1$, то применяя теорему Клиффорда к характерам $(\theta_1)_{B_q}$ и $(\theta_2)_{B_q}$, мы убеждаемся в том, что все неприводимые компоненты характера $\widehat{\chi}_{B_q}$ являются линейными. Стало быть, подгруппа B_q абелева.

Предположим, что $B_q \not\subseteq (N_{\pi'})'$. Тогда из теоремы 7.4.4 [4] вытекает, что $|B : B'|$ делится на q. Следовательно, $O^q(B) \neq B$. Заметим, что подгруппа $O^q(B)$ не является абелевой, ибо в противном случае B содержит нормальное абелево q-дополнение. Это противоречит лемме 3.14 [3]. Поэтому характер $\widehat{\chi}_{O^q(B)}$ не содержит линейных неприводимых компонент. Поскольку все они по теореме Клиффорда имеют одинаковую степень, делящую $\widehat{\chi}(1)$, то либо характер $\widehat{\chi}_{O^q(B)}$ неприводим, либо $\psi(1) = q^{\alpha_1}$, $\alpha_1 < \alpha$, для некоторой его нелинейной неприводимой компоненты ψ. Тогда нетрудно видеть, что все неприводимые компоненты характера $\widehat{\chi}_{AO^q(B)}$ имеют степень, делящуюся на $\psi(1)$. Поэтому ни одна из этих степеней не может быть равна ни одному из чисел из леммы 2.10 [1].

Следовательно, факторгруппы группы $AO^q(B)$ по их ядрам имеют нормальные холловы π-подгруппы. И так как все эти ядра имеют тривиальное пересечение, то мы замечаем, что $A \lhd AO^q(B)$. Это означает, что $O^q(B) \subseteq C$. Отсюда вытекает, что $|B : C|$ – степень простого числа q. По лемме 2.9 [1] $B = O_q(B)C$. В этом случае теорема $(*)$ верна, что противоречит выбору группы Γ.

Поэтому мы заключаем, что $B_q \subseteq (N_{\pi'})'$. Но $(N_{\pi'})' \subseteq \ker\theta_1$ по утверждению (2_{ii}) леммы 3.2 [1]. Стало быть, $B_q \subseteq \ker\theta_1$. Отсюда вытекает, что $(\theta_1)_{B_q} = \theta_1(1)1_{B_q}$ и, значит,

$$(\widehat{\chi}_{B_q}, 1_{B_q})_{B_q} = ((\theta_1+\theta_2)_{B_q}, 1_{B_q})_{B_q} \geq \theta_1(1) = |A|.$$

С другой стороны, по теореме 3.13 [5]

$$\widehat{\chi}(1) = 2|A|+1 = q^\alpha = |\Gamma : Z(\Gamma)|_q.$$

Так как $B_q \cap Z(\Gamma) \subseteq \ker\theta_1$ и характер $\widehat{\chi}$ точный, то мы заключаем, что $B_q \cap Z(\Gamma) = 1$. Значит, $\widehat{\chi}(1) = |B_q|$ и из теорем 4.3.1 и 4.2.7 [4] вытекает, что $\widehat{\chi}_{B_q} = \rho_{B_q}$ – регулярный характер подгруппы B_q. Поэтому из леммы 2.11 [5] вытекает, что

$$(\widehat{\chi}_{B_q}, 1_{B_q})_{B_q} = 1_{B_q}(1) = 1,$$

что противоречит выше выделенному равенству.

Мы показали, что $\eta \neq 0$.

3.15.1. *Выполняется:* 1) $|\pi| = 1$, *характер* $\widehat{\chi}$ *примитивный и* $L \subseteq Z(\Gamma)$ *для абелевой и нормальной в* Γ *подгруппы* L*;*

2) $|N| = 2|A||B_q|; |B_q| = q$ *или* $|A| = 11^2 = 121$ *и* $|B_q| = 3^5$*;*

3) $N = (A \times N_2)B_q$ *и группа* B *простая.*

Убедимся вначале в том, что $|N| = 2|AB_q| = 2|A||B_q|$, $|B_q| = q$ или $|A| = 11^2 = 121$ и $|B_q| = 3^5$, $N = (A \times N_2)B_q$, а затем в том, что группа B простая.

Пусть $\theta(1) \neq |A|$, $\theta(1) \neq 2|A|$ и $\theta(1) \neq 2|A|-1$. По лемме 3.2 [1]

$$N_{\pi'} = O_2(N_{\pi'})C_{N_{\pi'}}(A),$$

т. е. $B_q \subseteq C$, что не так, ибо $[B_q, A] \nsubseteq C$.

Пусть теперь $\theta(1) = 2|A|-1 \neq 17$. Тогда по этой же лемме мы получаем, что $\theta(1) = 2|A|-1$ – степень простого числа $q_1 \in \pi'$, $q_1 \neq q$, и

$$N_{\pi'} = O_{q_1}(N_{\pi'})C_{N_{\pi'}}(A),$$

т. е. тоже $B_q \subseteq C$. Противоречие.

Рассмотрим случай, когда $\theta(1) = 2|A|-1 = 17$. Значит, $|A| = 3^2 = 9$,

$$\widehat{\chi}(1) = q^\alpha = 2|A|+1 = (2|A|-1)+2 = 19.$$

Тогда $\alpha = 1$, $q = 19$ и $\eta(1) = \widehat{\chi}(1) - \theta(1) = 2$. Поскольку $B_{19} \lhd N$ и $A\ker\eta \lhd N$, то

$$(A\ker\eta)_{19} = B_{19} \cap A\ker\eta \lhd B_{19}.$$

Поэтому

$$(A\ker\eta)_{19} \cap Z(B_{19}) \neq 1.$$

Следовательно, существует элемент

$$1 \neq z \in (A\ker\eta)_{19} \cap Z(B_{19}), \quad o(z) = 19,$$

и поэтому $(\widehat{\chi}(1), |Cl(z)|) = 1$. Отсюда и из теоремы 3.8 [5] вытекает, что $z \in Z(\widehat{\chi})$ либо $\widehat{\chi}(z) = 0$. Поскольку характер $\widehat{\chi}$ точный, то $z \notin \ker\widehat{\chi}$. Поэтому мы получаем, что $\widehat{\chi}(z) = -19$ либо $\widehat{\chi}(z) = 0$.

Если $\widehat{\chi}(z) = 0$, то $\widehat{\chi}_{<z>} = \rho_{<z>}$ – регулярный характер подгруппы $<z>$, ибо $\widehat{\chi}(1) = o(z) = 19$. Следовательно,

$$(\widehat{\chi}_{<z>}, 1_{<z>})_{<z>} = (\rho_{<z>}, 1_{<z>})_{<z>} = 1.$$

Однако мы видим, что $z \in \ker\eta$, т. е. $\eta_{<z>} = \eta(1)1_{<z>} = 2 \cdot 1_{<z>}$. Поэтому

$$(\widehat{\chi}_{<z>}, 1_{<z>})_{<z>} = ((\eta+\theta)_{<z>}, 1_{<z>})_{<z>} = 2 + (\theta_{<z>}, 1_{<z>})_{<z>} \geqslant 2.$$

Противоречие.

Пусть $\widehat{\chi}(z) = -19$. Тогда

$$(\widehat{\chi}_{<z>}, 1_{<z>})_{<z>} = 0,$$

и мы вновь получили противоречие с выражением выше.

Случай, когда $\theta(1) = 2|A| - 1 = 17$ невозможен.

Допустим теперь, что $\theta(1) = |A|$. В рассматриваемом случае $\eta(1) = |A| + 1$. Как и в случае $\theta(1) = 2|A| - 1 = 17$ убеждаемся в том, что существует такой элемент $1 \neq z \in Z(B_q) \cap \ker\eta$, что $\widehat{\chi}(z) = 0$. Применим лемму 3 [6]. По ней

$$\widehat{\chi}(z) = 0 = \alpha_{\widehat{\chi}} - \beta_{\widehat{\chi}},$$

где

$$\alpha_{\widehat{\chi}} = (\widehat{\chi}_{<z>}, 1_{<z>})_{<z>}$$

и

$$\beta_{\widehat{\chi}} = (\widehat{\chi}_{<z>}, \lambda_{<z>})_{<z>},$$

$1_{<z>} \neq \lambda \in \mathrm{Irr}(<z>)$. При этом

$$\widehat{\chi}(1) = \alpha_{\widehat{\chi}} + \beta_{\widehat{\chi}}(q-1) = \alpha_{\widehat{\chi}} q.$$

Здесь мы учли, что $\alpha_{\widehat{\chi}} = \beta_{\widehat{\chi}}$. Из того, что $z \in \ker\eta$ вытекает, что $\alpha_{\widehat{\chi}} \geqslant |A| + 1$. Значит,

$$2|A| + 1 = \widehat{\chi}(1) \geqslant (|A| + 1)q.$$

Поскольку $q > 2$, то мы получили неверное утверждение. Поэтому $\theta(1) \neq |A|$.

Предположим теперь, что $\theta(1) = 2|A|$. Тогда $\eta(1) = 1$ и $N' \subseteq \ker\eta$.

Допустим, что характер θ не является неприводимым.

По утверждению (1_{vi}) леммы 3.2 [1] $\theta_1 + \theta_2$ с $\theta_1(1) = \theta_2(1) = |A|$ и все неприводимые компоненты характера $\widehat{\chi}_{N_{\pi'}}$ линейные, т. е. подгруппа $N_{\pi'}$ абелева. По теореме Бернсайда подгруппа B содержит нормальное q-дополнение R. Так как $R \triangleleft \Gamma$ и $(\widehat{\chi}(1) = q^{\alpha}, |R|) = 1$, то с применением теоремы Клиффорда мы убеждаемся в том, что все неприводимые компоненты точного характера $\widehat{\chi}_R$ являются линейными. Тогда $R' \subseteq \ker\widehat{\chi}_R = 1$, т. е. подгруппа R абелева.

Если $R \neq 1$, то это противоречит лемме 3.14 [2].

Если же $R = 1$, то подгруппа $B = B_q$ абелева, что противоречит тому, что она имеет нелинейный неприводимый характер χ.

Поэтому мы заключаем, что характер θ неприводим. Поскольку $\eta(1) = 1$ и $\theta(1) = 2|A|$ не делятся на простое число q, то из теоремы Клиффорда вытекает, что все неприводимые компоненты характера θ_{B_q} линейные. Следовательно, все неприводимые компоненты характера $\widehat{\chi}_{B_q}$ линейные. Поскольку характер $\widehat{\chi}_{B_q}$ точный, то подгруппа B_q абелева. Тогда из теоремы 3.13 [5] следует, что $\widehat{\chi}(1) = |B_q/(Z(\Gamma))_q|$.

Допустим, что $(Z(\Gamma))_q \neq 1$. Из теоремы 7.4.4 [4] следует, что $|\Gamma : \Gamma'|$ делится на степень простого числа q. Тогда $A \subseteq O^q(\Gamma)$ и $\Gamma = O^q(\Gamma)B_q$.

Если $A \triangleleft O^q(\Gamma)$, то $(O^q(\Gamma))_{\pi'} \subseteq C$. Стало быть, $|B : C|$ – степень простого числа q и $B = O_q(B)C$ по лемме 2.9 [1]. По лемме 3.11 [2] подгруппа C разрешима и содержит абелеву холлову подгруппу $C_{2'}$. Отсюда и из леммы 3.5 [2] вытекает, что теорема (*) верна. Но это противоречит выбору группы Γ.

Пусть теперь $A \ntriangleleft O^q(\Gamma)$. Рассмотрев характер $\widehat{\chi}_{O^q(\Gamma)}$, мы с помощью теоремы Клиффорда убеждаемся в том, что он неприводим. Поскольку $|O^q(\Gamma)| < |\Gamma|$, то мы можем применить индукцию. По ней

$$(O^q(\Gamma))_{\pi'} = O_q((O^q(\Gamma))_{\pi'})C_{(O^q(\Gamma))_{\pi'}}(A).$$

Тогда нетрудно видеть, что $|B : C|$ – степень простого числа q. Мы вновь придем к противоречию с выбором группы Γ.

Поэтому $(Z(\Gamma))_q = 1$ и, значит, $\widehat{\chi}(1) = |B_q|$ и $B_q \subseteq N'$, что вытекает из теоремы 7.4.4 [4]. Также из теорем 8.17 [5] и 4.2.7 [4] вытекает, что $\widehat{\chi}_{B_q} = \rho_{B_q}$ – регулярный характер подгруппы B_q.

Пусть B_0 – такая минимальная A-инвариантная нормальная подгруппа из B_q, что $B_0 \nsubseteq C$. Так как A – TI-подгруппа, то $|B_0/C_{B_0}(A)| = k|A| + 1$ для некоторого $k \in \mathbf{Z}_+$. Но $|B_q| = \widehat{\chi}(1) = 2|A| + 1$.

Поскольку $B_0 \subseteq B_q$, то $k \leqslant 2$. Поскольку $|A|$ и B_0 являются нечетными числами, то $k \neq 1$. Так как $k \neq 0$, то $k = 2$. Поэтому $|B_0/C_{B_0}(A)| = |B_q|$, $C_{B_0}(A) = 1$ и $B_0 = B_q$. Следовательно, подгруппа B_q минимальная нормальная в AB_q и элементарная абелева.

По теореме Клиффорда

$$\theta_{N_{\pi'}} = \sum_{a \in A} \beta^a,$$

где β – неприводимый характер степени 2 подгруппы $N_{\pi'}$.

Предположим, что неприводимая компонента β характера $\theta_{N_{\pi'}}$ подгруппы $N_{\pi'}$ примитивна. Так как подгруппа $B_q \triangleleft N_{\pi'}$ нормальна в N и $(2,q) = 1$, то $B_q \subseteq Z(\beta^a)$ для каждого элемента $a \in A$. Поскольку $B_q \subseteq Z(\eta_{N_{\pi'}})$, то по лемме 5 [7] $B_q \subseteq Z(N_{\pi'})$. По теореме Бернсайда группа B содержит нормальное q-дополнение R. Рассматривая характер χ_R с помощью теоремы Клиффорда, мы убеждаемся в том, что $R' \subseteq \ker\chi_R = 1$. Это означает, что подгруппа R абелева. Если $R \neq 1$, то мы получим противоречие с леммой 3.14 [3]. Если $R = 1$, то $B = B_q$. Но в нашем случае подгруппа B_q абелева и, значит, она не может иметь нелинейный неприводимый характер χ. Стало быть, характер β не может быть примитивным.

Предположим теперь, что характер β не является примитивным, т. е. $\beta = \lambda^{N_{\pi'}}$ для линейного характера λ подгруппы $X \subseteq N_{\pi'}$, $|N_{(\pi')} : X| = 2$. Тогда $\theta = \lambda^N$.

Мы замечаем, что $(N_{\pi'})' \subseteq X$ и поэтому $(N_{\pi'})'' \subseteq X'$. Но $(N_{\pi'})'' \triangleleft N$, ибо

$$(N_{\pi'})''\mathrm{char}(N_{\pi'})'\mathrm{char}N_{\pi'}\mathrm{char}N.$$

Далее, $X' \subseteq \ker\lambda$ и

$$\ker\theta = \cap_{g \in N \setminus X}(\ker\lambda^g).$$

Поэтому $(N_{\pi'})'' \subseteq \ker\theta$. Поскольку $(N_{\pi'})'' \subseteq (N_{\pi'})' \subseteq \ker\eta$ и

$$\ker\eta \cap \ker\theta = \ker\widehat{\chi} = 1,$$

то $(N_{\pi'})'' = 1$. Поэтому подгруппа $(N_{\pi'})'$ абелева и подгруппа $N_{\pi'}$ разрешима.

Далее, по лемме 2.2 [1] $N_{\pi'} = [N_{\pi'}, A]C_{N_{\pi'}}(A)$. Поскольку факторгруппа $\Gamma/\ker\eta$ циклическая, то

$$A\ker\eta/\ker\eta \triangleleft \Gamma/\ker\eta$$

и по лемме 3 [7] $[N_{\pi'}, A] \subseteq \ker\eta$.

Предположим вначале, что $[N_{\pi'}, A] \subseteq X$. Тогда $\lambda_{[N_{\pi'},A]} \in \mathrm{Irr}(\theta_{[N_{\pi'},A]})$ и так как $[N_{\pi'}, A] \triangleleft N$, то с помощью теоремы Клиффорда мы убеждаемся в том, что все неприводимые компоненты характера $\theta_{[N_{\pi'},A]}$ и, следовательно, все неприводимые компоненты характера $\widehat{\chi}_{[N_{\pi'},A]}$ линейные. Стало быть, подгруппа $[N_{\pi'}, A]$ абелева. Поэтому

$$[N_{\pi'}, A] = [[N_{\pi'}, A], A] \times C_{[N_{\pi'},A]}(A).$$

Отсюда, из того, что $[[N_{\pi'}, A], A] = [N_{\pi'}, A]$ и из ранее установленного вытекает, что

$$N_{\pi'} = [N_{\pi'}, A] \times C_{N_{\pi'}}(A); \quad [N_{\pi'}, A] \subseteq Z(N_{\pi'}).$$

Поскольку $B_q \not\subseteq C$, то $B_q \cap [N_{\pi'}, A] \neq 1$. Следовательно, подгруппа $[N_{\pi'}, A]_q \neq 1$ и A-инвариантна. Так как B_q – минимальная нормальная в AB_q, то $[N_{\pi'}, A]_q = B_q$. Значит, $B_q \subseteq Z(N_{\pi'})$. Отсюда и из теоремы Бернсайда вытекает, что группа B содержит нормальное q-дополнение R. Чуть ранее мы показали, что это невозможно.

Пусть теперь $[N_{\pi'}, A] \not\subseteq X$. Поскольку $|N_{\pi'} : X| = 2$, то $N_{\pi'} = [N_{\pi'}, A]$.

Допустим, что $|\pi| \neq 1$. Поскольку подгруппа $N_{\pi'}$ разрешима, то $[N_{\pi'}, A] \subseteq F(N_{\pi'})$ по теореме 6 [8], т. е. подгруппа $N_{\pi'}$ нильпотентна. Это означает, что $B_q \subseteq Z(N_{\pi'})$. По теореме Бернсайда группа B содержит инвариантное q-дополнение R, которое будет абелевым, что легко следует из того, что все неприводимые компоненты характера χ_R линейные. Мы получили противоречие с леммой 3.14 [3].

Здесь мы замечаем, что в случае, когда $|\pi| > 1$, лемма 3.15 доказана.

Поэтому в дальнейшем в данной работе мы считаем, что $|\pi| = 1$, т. е. $2p^m + 1 = q^\alpha$. Применим лемму 2.8 [1]. По ней могут быть три возможности:

1) $m = 1$ и, значит, $|A| = p$ – простое число;

2) $\alpha = 1$;

3) $p = 11$, $m = 2$, $q = 3$ и $\alpha = 5$.

По условию теоремы в данном случае характер $\widehat{\chi}$ примитивный. Пусть подгруппа $L \triangleleft \Gamma$ и абелева. Тогда по следствию 6.13 [5] $L \subseteq Z(\Gamma)$.

Пусть K – такая A-инвариантная максимальная нормальная подгруппа в $N_{\pi'}$, что $(N_{\pi'})' \subseteq \subseteq K \subseteq N_{\pi'}$ и $N_{\pi'}/K$ – главный фактор группы N. Поскольку $(N_{\pi'})' \subseteq X$ и $|N_{\pi'} : X| = 2$, то мы можем положить, что $K \subseteq X$ и $K = O^2(N)$.

Тогда $\lambda_K \in \mathrm{Irr}(\theta_K)$. Поэтому из теоремы Клиффорда вытекает, что все неприводимые компоненты характера θ_K линейные. Поскольку $\eta_K(1) = 1$, то все неприводимые компоненты характера $\widehat{\chi}_K$ линейные. Следовательно, $K' \subseteq \ker\widehat{\chi}_K = 1$, т. е. подгруппа K абелева. Поскольку $K \triangleleft N$, то любая холлова подгруппа из K нормальна в N. Очевидно также, что $B_q \subseteq K$ и что $K = F(N)$.

Предположим, что $K_{q'} \neq 1$. Применим теорему D [9], по которой рациональное число

$$m = \frac{|K_{q'}||\Gamma : N|\theta(1)}{\widehat{\chi}(1)\lambda_K(1)} = \frac{|K_{q'}||\Gamma : N|2|A|}{q}$$

является целым числом. Однако простое число q не делит числитель, ибо $B_q \subseteq N$. Противоречие.

Поэтому $K_{q'} = 1$ и, значит, $K = B_q = F(N)$ и, так как $N_{\pi'}/K \cong N_2$ является главным фактором разрешимой группы N, то подгруппа N_2 является элементарной абелевой 2-группой и минимальной нормальной в AN_2. Мы видим также, что $(N_{\pi'})' \subseteq B_q$. Учитывая, что подгруппа B_q минимальная нормальная в AB_q, мы заключаем, что $(N_{\pi'})' = B_q$ порядка, напомним, $\widehat{\chi}(1) = q^\alpha$. Поскольку $C_N(B_q) \triangleleft N$, то из рассуждений выше вытекает, что $C_N(B_q) = B_q$. Следовательно, $N/B_q \cong AN_2$ и изоморфна подгруппе из $Aut(B_q)$.

1. Предположим вначале, что $m = 1$. Значит, $\widetilde{\chi}(1) = 2p + 1$. По теореме [10] группа Γ разрешима.

Заметим, что мы рассматриваем случай, когда $\eta(1) = 1$ и $\theta(1) = 2|A|$ и, значит, можем применять сведения, полученные в нем.

Пусть $M \triangleleft \Gamma$ – такая подгруппа, что B/M является главным фактором группы Γ. Так как B – разрешима, то $|B/M|$ – степень некоторого простого числа q_1.

Поскольку $M \triangleleft \Gamma$ и характер $\widehat{\chi}$ примитивен, то $\widehat{\chi}_M = e\chi_1$ для натурального числа e, делящего $|\Gamma : M|$ и неприводимого характера χ_1 степени $q_1^{\alpha_1}$, $\alpha_1 \in \mathbf{N}$, подгруппы M.

Допустим вначале, что $q_1 \neq q$. Тогда $B_q \subseteq M$, $e = 1$ и характер $\widehat{\chi}_M$ неприводим. Поэтому характер $\widehat{\chi}_{AM}$ также неприводим. Нетрудно убедиться в том, что к нему и подгруппе AM мы можем применить индукцию. По ней $M = O_q(M)C_M(A)$. Следовательно, $O_q(M) \triangleleft \Gamma$. При этом подгруппа $O_q(M)$ абелева, ибо B_q абелева. Так как характер $\widehat{\chi}$ примитивен, то $O_q(M) \subseteq Z(\Gamma)$. Следовательно, $M \subseteq C$, т. е. $B_q \subseteq C$, что противоречит лемме 3.5 [2].

Пусть теперь $q_1 = q$. Поскольку $B_q \cap M$ – q-силовская A-инвариантная подгруппа и B_q – минимальная нормальная подгруппа в AB_q, то $B_q \cap M = 1$. Следовательно, $B = MB_q$ и M – q'-группа. Следовательно, подгруппа M абелева и $M \subseteq Z(\Gamma)$. Мы получаем, что подгруппа B является абелевой, что не так.

2. Предположим теперь, что в лемме 2.8 [1] $\alpha = 1$. Тогда $|B_q| = q$ и $Aut(B_q)$ является циклической порядка $q - 1 = 2|A|$. Поэтому $AN_2 = A \times N_2$, $|N_2| = 2$ и $N = (A \times N_2)B_q$. Также мы замечаем, что N и $N_{\pi'}$ являются группами Фробениуса с ядром B_q.

3. Пусть теперь в лемме 2.8 [1] $\alpha \neq 1$. Тогда $m = 1$ или $q = 3$, $\alpha = 5$, $p = 11$ и $m = 2$.

Предположим, что $q = 3$, $\alpha = 5$, $p = 11$ и $m = 2$. Значит, B_3 – элементарная абелева 3-группа порядка 3^5 и AN_2 изоморфна некоторой подгруппе из $GL(5;3)$. Так как

$$|GL(5;3)| = \prod_{i=0}^{i=4}(3^5 - 3^i) = (3^5 - 3^1)(3^5 - 3^2)(3^5 - 3^3)(3^5 - 3^4) =$$

$$= 3^{1+2+3+4}(3^5-1)(3^4-1)(3^3-1)(3^2-1)(3-1) =$$

$$= 2^{1+4+1+3+1}\cdot 3^{10}\cdot 5\cdot 13\cdot 11^2 = 2^{10}\cdot 3^{10}\cdot 5\cdot 13\cdot 11^2.$$

Мы видим, что $|N_2| \leqslant 2^{10}$. Так как $N_2 \lhd AN_2$ и A – TI-подгруппа, то из леммы 1 [11] следует, что $|N_2/C_{N_2}(A)| \equiv 1(\mathrm{mod}|A|)$, т. е. $2^k \equiv 1(\mathrm{mod}|121|)$ для некоторого $0 \leqslant k \leqslant 10$. Непосредственной проверкой убеждаемся в том, что только $2^0 \equiv 1(\mathrm{mod}\ 121)$. Это означает, что $AN_2 = A\times N_2$. Поскольку подгруппа N_2 – минимальная нормальная в AN_2, то $|N_2| = 2$. Мы вновь замечаем, что и в этом случае N и $N_{\pi'}$ – группы Фробениуса с ядром B_3 порядка 3^5.

Итак, мы видим, что

$$|N_2| = 2, |N| = 2|AB_q| = 2|A||B_q|, \quad N = (A\times N_2)B_q,$$

где $|B_q| = q \neq 3$ или $|B_q| = 3^5$. Убедимся в том, что группа B простая.

Мы отмечали, что N является группой Фробениуса с ядром B_q, т. е. B_q – F-подгруппа в группе Γ [12, с. 8]. Также подгруппа $N_{\pi'} = N_B(B_q)$ является группой Фробениуса с ядром B_q и $|N_{\pi'} : B_q| = 2$. Поскольку B_q – F-подгруппа в группе B и $q \neq 2$, то множество S всех неприводимых характеров подгруппы $N_{\pi'}$, индуцируемых неглавными неприводимыми характерами подгруппы B_q, по теореме 1.5 [12] является τ-когерентным.

Допустим, что группа B не является простой. Тогда по лемме 4.2 [12] существует собственная подгруппа $1 \neq R \lhd B$, и для нее выполняется одно из двух утверждений: $B_q \subseteq R$ или $R\cap N_{\pi'} = 1$.

Пусть выполняется первое утверждение. Тогда $B = N_B(B_q)R = N_{\pi'}R$.

Поэтому фактор группа

$$B/R = N_{\pi'}R/R \cong N_{\pi'}/(N_{\pi'}\cap R = B_q)$$

циклическая индекса 2. Следовательно, характер χ_R неприводим. Легко видеть, что характер $\widehat{\chi}_{AR}$ неприводим. Несложно убедиться в том, что для AR, A, R, $C_R(A)$, χ_R и n выполняются условия $(^*)$. Так как $|AR| < |\Gamma|$, то по индукции $R = O_q(R)C_R(A)$ и для подгруппы $C_R(A)$ и ее характера χ_R выполняются утверждения теоремы $(^*)$. Поскольку $N_2 \subseteq C$, что отмечалось ранее, то $B = O_q(B)C$ и теорема верна. Мы получили противоречие с выбором группы Γ.

Пусть теперь $R\cap N_{\pi'} = 1$. Поскольку $(\chi(1), |R|) = 1$, то с помощью теоремы Клиффорда мы устанавливаем, что все неприводимые компоненты характера χ_R линейные. Поэтому подгруппа R абелева.

Пусть $\lambda_1 \in \mathrm{Irr}(\chi_R)$ и пусть $I_{\pi'} = I_B(\lambda_1)$. По теореме 6.11 [5] $\chi = ((\lambda_1)')^B$ для такого неприводимого характера $(\lambda_1)'$ подгруппы $I_{\pi'}$, что

$$e = (((\lambda_1)')_R, \lambda_1)_R = (\chi_R, \lambda_1)_R.$$

Отсюда вытекает, что

$$q^\alpha = \chi(1) = |B : I_{\pi'}|(\lambda_1)'(1).$$

Мы видим, что $|B : I_{\pi'}|$ – степень q и, значит, $B = B_qI_{\pi'}$.

Предположим вначале, что $|B : I_{\pi'}| = q^\alpha \neq 1$, т. е. $B \neq I_{\pi'}$. Поскольку $2|A| = |B_q| - 1$ не делит $|I_{\pi'}|$, то мы можем применить теорему 9.6 [12] по которой группа B имеет нормальное нильпотентное дополнение R_0 к $N_{\pi'}$. Это означает, что $B = N_{\pi'}R_0$, $R_0 \subseteq F(B)$ и $N_{\pi'}\cap R_0 = 1$. Так как $F(B) \lhd \Gamma$ и $B_q\cap F(B) = 1$, то подгруппа $F(B)$ абелева. Следовательно, $F(B) \subseteq Z(\Gamma)$. Поскольку $N_2 \subseteq C$, то мы видим, что $B = B_qC$, что противоречит выбору группы Γ, ибо по лемме 2.9 [1] $B = O_q(B)C$.

Пусть теперь $B = I_{\pi'}$. Тогда $\chi_R = \chi(1)\lambda_1$, т. е. $R \subseteq Z(B) = Z(\Gamma) \subseteq N_{\pi'}$. Противоречие.

Итак, группа B простая.

Утверждение 3.15.1 доказано.

Рассмотрим теперь случай, когда группа B простая. Из ранее приведенных рассуждений вытекает, что $\pi = \{p\}$ – простое число, $|A| \neq p$, $|B_q| = q$ – нечетное простое число или $|A| = 11^2$, $|B_q| = 3^5$ и подгруппа B_3 элементарна абелева. В обоих случаях $|N_{\pi'}| = 2|B_q|$, $N_2 \subseteq C$ и подгруппа B_q – минимальная нормальная в AB_q.

Пусть $B \supset Y$ – произвольная A-инвариантная подгруппа такая, что $Y \not\subseteq C$ и $\Gamma_1 = AY$. Тогда $A \subseteq \Gamma_1 \subset \Gamma$ и $A \not\triangleleft \Gamma_1$. К группе Γ_1 и ее характеру $\widehat{\chi}_{\Gamma_1}$ мы можем применить лемму 3.2 [1]. Пусть η' и θ' – компоненты характера $\widehat{\chi}_{\Gamma_1}$, которые соответствуют характерам η и θ из этой леммы.

3.15.2. *Выполняется: для любой A-инвариантной подгруппы $Y \subset B$ компонента θ' характера $(\widehat{\chi})_{\Gamma_1}$ такая, что $A\ker\theta'/\ker\theta' \not\triangleleft \Gamma_1/\ker\theta'$ может иметь только степени $\theta'(1) = |A|$ или $\theta'(1) = 2|A|$.*

Рассмотрим все различные значения характеров η' и θ' из указанной леммы.

Заметим, что при $\Gamma_1 \neq \Gamma$ характер $\widehat{\chi}_{\Gamma_1}$ не может быть неприводимым. В самом деле. Если это не так и $\widehat{\chi}_{\Gamma_1} = \theta'$ – неприводимый характер, то нам нетрудно убедиться в том, что для Γ_1, A, Y, $C_Y(A)$, $(\theta')_Y = \chi_Y$ и n выполнено условие $(^*)$. Поскольку $|\Gamma_1| < |\Gamma|$, то по индукции

$$Y = O_q(Y)C_Y(A)$$

и для подгруппы $C_Y(A)$ справедливы все утверждения доказываемой теоремы. Однако мы замечаем, что $O_q(Y) = B_q$ и, так как эта подгруппа абелева, то по теореме 6.15 [5] $\theta'(1)$ делит $|Y : B_q| = |Y|_{q'}$, что не так.

Рассмотрим вначале случай, когда $\eta' \neq 0$ в лемме 3.2 [1]. По этой лемме $A\ker\eta' \triangleleft \Gamma_1$, а по ее утверждению (1_i) характер $\theta'_{A\ker\eta'}$ точен степени $\theta'(1) = |A|-1$, $|A|$, $|A|+1$, $2(|A|-1)$, $2|A|-1$ или $2|A|$ и $\theta'(1)$ – степень некоторого простого числа r за исключением случая, когда $\theta'(1) = 2|A|$. При этом с учетом леммы 3 [7]

$$Y = (\ker\eta')_{\pi'}C_Y(A), \quad [Y,A] \subseteq \ker(\ker\eta')_{\pi'}.$$

Также по утверждению (1_{ii}) леммы 3.2 [1] при $\theta'(1) \neq 2(|A|-1)$ и при $\theta'(1) \neq 2|A|$ характер $\theta'_{A\ker\eta'}$ неприводим.

Предположим, что $\theta'(1) = |A|-1$ или $\theta'(1) = 2(|A|-1)$. Тогда $2^d = p^m - 1$ для $d,m \in \mathbf{N}$, т. е. $p^m = 2^d + 1$. Следовательно, $|A| = p$ или $|A| = 9$. Случай $|A| = p$ нами рассмотрен ранее. Поэтому $|A| = 9$, т. е. $\pi = \{3\}$. Этот случай мы рассмотрим чуть ниже.

Предположим теперь, что $\theta'(1) = |A|+1 = 2^d$. Тогда

$$q^\alpha = \widehat{\chi}(1) = 2|A|+1 = 2(|A|+1)-1 = 2^{d+1}-1 = q$$

– простое число Мерсенна. Понятно, что в этом случае $d = 2d_1$ – четное число, $d_1 \in \mathbf{N}$, и $q \neq 3$. Убедимся в том, что на 3 делится

$$|A| = (q-1)/2 = (2^{2d_1+1}-2)/2 = 2^{2d_1}-1.$$

В самом деле. Пусть $d_1 = 1$. Тогда $|A| = 2^2 - 1 = 3$. Предположим, что на 3 делится натуральное число $2^{2k}-1$ для натурального числа k. Так как $2^{2(k+1)} - 1 = (2^{k+1}-1)(2^{k+1}+1)$ и 2^{k+1} не делится на 3, то на 3 делится один из сомножителей в скобках. Поэтому $|A| = 2^{2(k+1)}-1$ делится на 3, т. е. $\pi = \{3\}$.

Рассмотрим случай, когда $\pi = \{3\}$. Тогда B – простая $3'$-группа. Такой может быть только группа Судзуки $Sz(d)$, $d = 2^{2m+1}$, $m \in \mathbf{N}$, порядка

$$d^2(d-1)(d^2+1) = d^2(d-1)(d+2r+1)(d-2r+1), \quad r = 2^m.$$

Согласно [13] группа B может иметь только следующие максимальные подгруппы:

$H(d)$ – группа Фробениуса порядка $d^2(d-1)$;

группа диэдра порядка $2(d-1)$;

группа Фробениуса $[a]b$, где $a^{d+2r+1} = b^4 = 1, r = 2^m$;

группа Фробениуса $[a]b$, где $a^{d-2r+1} = b^4 = 1, r = 2^m$;

$Sz(s)$, где $s^t = q$, t – простой делитель числа $2m+1$.

Учитывая, что A – TI-подгруппа в группе Γ и подгруппы порядков $d-1$, $d+2r+1$ и $d-2r+$ $+1$ являются холловыми в B и циклическими (теоремы XI.3.9 и XI.3.10 [13]), то мы дословным повторением леммы 2 [14] убеждаемся в том, что в группе Γ существуют подгруппы $AY_{2'}$ порядков $|A|(d-1)$; $|A|(d+2r+1)$ и $|A|(d-2r+1)$ и, следовательно, существуют подгруппы AY порядков $2|A|(d-1)$; $4|A|(d+2r+1)$ и $4|A|(d-2r+1)$. Поскольку мы рассматриваем случаи, когда $\theta'(1) \in$

$\in \{|A|-1; |A|+1; 2(|A|-1)\}$, то по утверждению (1_{iii}) леммы 3.2 [1] $Y_{2'} \subseteq C$ для каждой из указанных подгрупп. Тогда мы замечаем, что $|B:C|$ – степень 2. Из леммы 2.9 [1] вытекает, что $B = O_2(B)C$. Это противоречит простоте группы B.

Теперь следует заметить, что случай, когда $\eta' = 0$ невозможен. В самом деле, ибо в противном случае $\theta' = \theta_1 + \theta_2$, где по лемме 3.2 [1] θ_1 и θ_2 – неприводимые характеры степеней $|A|$ и $|A|+1$ соответственно. При этом $|A|+1$ – степень 2 и по утверждению 2_i этой леммы $|\pi > 1|$, что не так.

Предположим теперь, что $\theta'(1) = 2|A|-1$ и $2|A|-1 = r^{\alpha_1}$, где r – некоторое простое число и $\alpha_1 \in \mathbf{N}$.

В этом случае мы имеем три последовательные натуральные числа: $2|A|-1 = r^{\alpha_1}$; $2|A| = 2p^m$ и $2|A|+1 = q^{\alpha}$, которые делят $|\Gamma|$. Одно из них делится на 3. Если 3 делит $|A|$, т. е. $p = 3$, то B – простая $3'$-группа. Чуть ранее мы показали, что этот случай невозможен. Поэтому считаем, что $r = 3$ или $q = 3$, т. е. $B_3 = 3^5$.

Рассмотрим случай, когда $2|A|+1 = |B_3| = 3^5 = 243$. Тогда $|A| = 11^2$ и $2|A|-1 = r^{\alpha_1} = 241$ – простое число. По теореме 1 [15] $B = O_{241}(B)C$. Это противоречит тому, что группа B простая.

Остается случай, когда $r = 3$. Значит, $2|A|-1 = 3^{\alpha_1}$, т. е.

$$2|A|-1 = 2p^m - 1 = 3^{\alpha_1}.$$

Из теоремы [16] вытекает, что либо $m = 1$, либо $m = \alpha_1 = 2$. Поскольку мы рассматриваем случай, когда $m > 1$, то остается случай, когда $m = \alpha_1 = 2$. Значит, $2p^2 - 1 = 3^2 = 9$, т. е. $2p^2 = 10$. Отсюда получаем, что $p^2 = 5$, что не так.

Следовательно, $\theta'(1) \neq 2|A|-1$.

Утверждение 3.15.2 доказано.

3.15.3. *Выполняется: если* $\theta'(1) = |A|$, *то* $B_q \not\subseteq Y$.

Допустим, что $\theta'(1) = |A|$. Тогда $\eta'(1) = |A|+1$ и по утверждению (1_{iv}) леммы 3.2 [1]

$$Y = [Y,A] \times C_Y(A), \quad [Y,A] \subseteq Z(Y).$$

Отсюда следует, что

$$\Gamma_1 = AY = A[Y,A] \times C_Y(A).$$

Предположим, что простое число q делит $|Y|$. Тогда существует подгруппа AY_q, $Y_q \subseteq B_q$. Поскольку подгруппа B_q минимальная нормальная в AB_q, то мы замечаем, что $Y_q = B_q$ и $B_q \subseteq C_Y(A)$ или $B_q \subseteq [Y,A] \subseteq Z(Y)$. В первом случае мы получаем противоречие с леммой 3.3 [1], а во втором – $B_q \triangleleft \Gamma_1$, т. е. $\Gamma_1 \subseteq N$. Так как $N_{\pi'} \subseteq Y$, то $N_{\pi'} = Y = B_q \times C_Y(A)$, т. е. $B_q \subseteq Z(N_{\pi'})$. По теореме Бернсайда группа B содержит нормальное q-дополнение. Это противоречит ее простоте. Наше предположение о том, что q делит $|Y|$ неверно.

Следовательно, Y – q'-группа.

Утверждение 3.15.3 доказано.

3.15.4. *Выполняется: если* $\theta'(1) = 2|A|$, *то* $Y = N_{\pi'}$, *т. е.* $|Y| = 2|B_q|$.

Предположим, что $\theta'(1) = 2|A|$. Тогда $\eta'(1) = 1$ и $N' \subseteq \ker\eta'$.

Допустим, что характер θ' не является неприводимым. Как и в случае, когда $\theta(1) = 2|A|$ в доказательстве утверждения 3.15.1, мы убеждаемся в том, что подгруппа Y абелева. Но $N_{\pi'} \subseteq$ $\subseteq Y$. Следовательно, подгруппа B содержит нормальное q-дополнение. Это противоречит тому, что она простая.

Поэтому характер θ' неприводим.

По теореме Клиффорда

$$(\theta')_Y = \sum_{a \in A} (\beta')^a,$$

где β' – неприводимый характер степени 2 подгруппы Y.

Предположим, что неприводимая компонента β' характера $(\theta')_Y$ подгруппы Y примитивна.

Тогда из теоремы 14.23 [5] вытекает, что

$$|Y : Z(\beta')| = 2^2 \cdot 3; \quad 2^3 \cdot 3; \quad 2^2 \cdot 3 \cdot 5.$$

Отсюда следует, что для $5 < r \in \pi(B)$ подгруппа $B_r \subseteq Z(\beta')$. Поскольку подгруппа B_r – A-инвариантна, то $B_r = (B_r)^a \subseteq (Z(\beta'))^a$ для каждого элемента $a \in A$. Поэтому

$$B_r \subseteq \cap_{a\in A}((Z(\beta'))^a).$$

Но $B_r \subseteq Z(\eta')$, ибо $\eta'(1) = 1$. Поэтому

$$B_r \subseteq Z(\widehat{\chi}_Y) = Z(Y)$$

по лемме 5 [7]. Следовательно, $B_r \subseteq Z(N^{(r)})_{\pi'}$, и поэтому группа B содержит нормальное r-дополнение, что не так. Поэтому $\pi(Y) = \{2;3;5\}$.

Отсюда следует, что в рассматриваемом случае $B_q \not\subseteq Y$, если $|B_q| = q$.

Поэтому $q = 3$, $Y_3 = B_3 = 3^5$ и $B_3 \cap Z(\beta') = 3^4$.

Рассмотрим этот случай. Поскольку $(\beta')_{Z(\beta')} = 2\xi$ для линейного характера ξ подгруппы $Z(\beta')$, то из разложения характера $(\theta')_Y$ выше вытекает, что для каждого неединичного элемента $b \in Z(\beta')$ мы получаем, что

$$(\theta')_Y(b) = \sum_{a\in A}(\beta')^a(b) = 2\sum_{a\in A}\xi^a(b).$$

Поскольку $B_3 \subseteq N'$, что вытекает из теоремы 7.4.4 [4], то $\eta'(b) = 1$. Далее, $\widehat{\chi}(b) = 0$, ибо характер $\widehat{\chi}_{B_q}$ регулярный. Поэтому $\theta'(b) = \widehat{\chi}(b) - \eta'(b) = -1$. Следовательно,

$$2\sum_{a\in A}\xi^a(b) = -1.$$

Отсюда мы получаем, что

$$\sum_{a\in A}\xi^a(b) = -1/2.$$

Это противоречит тому, что алгебраическое рациональное число является целым числом (см. следствие 3.6 и лемму 3.2 [5]).

Предположим теперь, что неприводимая компонента β' характера $(\theta')_Y$ подгруппы Y импримитивна.

Тогда $\beta' = \lambda^Y$ для линейного характера λ подгруппы $X \subseteq Y$, $|Y : X| = 2$, и $\theta' = \lambda^{\Gamma_1}$, $\Gamma_1 = AY$.

Как в аналогичном случае в доказательстве утверждения 3.15.1 мы доказали, что подгруппа $N_{\pi'}$ разрешима, убеждаемся в том, что подгруппа Y разрешима.

Пусть $S = Y_{2'}$. Поскольку фактор группа $\Gamma_1/\ker\eta'$ циклическая, то $S\ker\eta'/\ker\eta'$ также циклическая и нормальна в $\Gamma_1/\ker\eta'$. А по утверждениям 2 и 2_i теоремы [17] факторгруппа $S\ker\theta'/\ker\theta'$ абелева и нормальна в $\Gamma_1/\ker\theta'$. Поскольку $\ker\eta' \cap \ker\theta' = 1$, то группа Γ_1 изоморфна некоторой подгруппе из прямого произведения факторгрупп $\Gamma_1/\ker\eta'$ и $\Gamma_1/\ker\theta'$ с абелевыми и нормальными холловыми подгруппами нечетного порядка. Тогда мы видим, что подгруппа $S \lhd \Gamma_1$ и также абелева. Следовательно, $S \lhd Y$ и $|Y : S|$ – степень 2. Понятно, что $B_q \subseteq S$ и что $B_q \lhd Y$, т. е. $Y \subseteq N_{\pi'}$. Так как по утверждению 3.15.1 $|N_{\pi'}| = 2|B_q|$, то $Y = N_{\pi'}$. Мы делаем вывод, что в случае $\theta'(1) = 2|A|$ максимальная A-инвариантная подгруппа $Y \subseteq B$ имеет порядок

$$|Y| = 2|B_q|.$$

Утверждение 3.15.4 доказано.

3.15.5. *Выполняется: если* $\theta'(1) = |A|$, *то* $B = Y^{(r)}B_q$ *для простого числа* $q \neq r \in \pi(B)$.

Продолжим исследование случая, когда $\theta'(1) = |A|$.

Из теоремы 4.21 [5] и из второй выделенной формулы в доказательстве утверждения 3.15.3 вытекает, что

$$\theta' = (\theta')_{A[Y,A]} \times \mu$$

для некоторого линейного характера μ подгруппы $C_Y(A)$. Ранее отмечено, что характер θ' неприводим. Поэтому и характер $(\theta')_{A[Y,A]}$ неприводим.

Аналогично,

$$\eta' = \varepsilon \times (\eta')_{C_Y(A)}$$

для некоторого линейного характера ε подгруппы $A[Y,A]$.

Поскольку $C_{[Y,A]}(A)=1$, то из условия (*) вытекает, что $C_{[Y,A]}(a)=1$ для каждого элемента $a\in A$. Тогда $A[Y,A]$ – группа Фробениуса с ядром $[Y,A]$. Любой ее линейный характер в своем ядре содержит подгруппу $[Y,A]$. Поэтому $[Y,A]\subseteq \ker\varepsilon$.

Далее, по следствию 13.4 [5] характер $\widehat{\chi}_A$ является рационально значным. Поскольку

$$\widehat{\chi}_A=(\eta')_A+(\theta')_A=\eta'(1)\varepsilon_A+(\theta')_A$$

и $(\theta')_A=\rho_A$ – регулярный характер подгруппы A, то характер ε_A – также рационально значен. Так как $2\notin\pi$ и $\varepsilon(1)=1$, то мы замечаем, что $A\subseteq\ker\varepsilon$. Поэтому $\varepsilon=1_{A[Y,A]}$.

Отсюда и из леммы 2.6 [1] в ее обозначениях нетрудно заметить, что

$$\widehat{\chi}_A=k\rho_A+\varepsilon\beta(1)1_A=(\theta')_A+\eta'(1)\varepsilon_A=\rho_A+(|A|+1)1_A,$$

т. е. $k=\varepsilon=1$ и $\beta(1)=|A|+1$.

Поскольку $\eta'(1)=|A|+1=\beta(1)$, β – неприводимый характер подгруппы C и $C_Y(A)\subseteq C$, то $(\eta')_{C_Y(A)}=\beta_{C_Y(A)}$.

Следовательно,

$$\eta'=1_{A[Y,A]}\times\beta_{C_Y(A)}.$$

Напомним, что $B\supset Y$ – такая любая A-инвариантная подгруппа, что $Y\nsubseteq C$, в том числе и максимальная. Это означает, что $B_r,(N^{(r)})_{\pi'}\subseteq Y^{(r)}$ для $r\in\pi(B)$.

Поэтому два выделенных равенства в доказательстве утверждения 3.15.3 мы можем записать для каждого простого числа $q\neq r\in\pi(B)$, $Y^{(r)}\nsubseteq C$, в виде

$$Y^{(r)}=[Y^{(r)},A]\times C_{Y^{(r)}}(A),\quad [Y^{(r)},A]\subseteq Z(Y^{(r)})$$

и

$$AY^{(r)}=A[Y^{(r)},A]\times C_{Y^{(r)}}(A).$$

Поскольку $Y^{(r)}\nsubseteq C$, то $[Y^{(r)},A]\neq 1$ для таких простых чисел r.

Покажем, что $C\subseteq Y^{(r)}$.

Пусть

$$1\neq M=[Y^{(r)},A]\cap([(Y^{(r)}),A]^c=[(Y^{(r)})^c,A])$$

для некоторого элемента $c\in C\setminus C_{Y^{(r)}}(A)$. Из формулы выше и из теоремы 6.2.2 [4] вытекает, что

$$(Y^{(r)})^c=[(Y^{(r)})^c,A]\times C_{(Y^{(r)})^c}(A).$$

Так как $M\subseteq Z(Y^{(r)})$ и $M\subseteq Z((Y^{(r)})^c)$, то подгруппа

$$L=\langle Y^{(r)},(Y^{(r)})^c\rangle\subseteq C_B(M).$$

Поскольку подгруппа $Y^{(r)}$ максимальна в группе B, то могут быть две возможности: $L=B$ или $L=Y^{(r)}$.

Если $L=B$, то $1\neq M\subseteq Z(B)$, что противоречит простоте группы B.

Если $L=Y^{(r)}$, то мы получаем, что $(Y^{(r)})^c=Y^{(r)}$ для элемента $c\in C$. Это означает, что $c\in N_B(Y^{(r)})$. Так как подгруппа $N_B(Y^{(r)})$ A-инвариантна и отлична от B, то $N_B(Y^{(r)})=Y^{(r)}$ в силу максимальности подгруппы $Y^{(r)}$. Это означает, что $c\in Y^{(r)}$, т. е. $c\in C_{Y^{(r)}}(A)$. Мы получили противоречие с выбором элемента c. Значит, в случае, когда $M\neq 1$ выполняется утверждение $C\setminus C_Y(A)=\varnothing$. Это означает, что $C=C_{Y^{(r)}}(A)$.

Пусть теперь $M=1$ для некоторого элемента $c\in C\setminus C_{Y^{(r)}}(A)$. Из ранее приведенных рассуждений вытекает, что

$$\begin{aligned}\widehat{\chi}_{A[Y^{(r)},A]}&=(\eta')_{A[Y^{(r)},A]}+(\theta')_{A[Y^{(r)},A]}=\\&=(|A|+1)1_{A[Y^{(r)},A]}+(\theta')_{A[Y^{(r)},A]}.\end{aligned}$$

Понятно, что имеет место и такая формула:

$$\widehat{\chi}_{A[(Y^{(r)})^c,A]}=(|A|+1)1_{A[(Y^{(r)})^c,A]}+(\theta')_{A[(Y^{(r)})^c,A]}$$

для таких элементов c.

Мы тогда видим, что

$$(\widehat{\chi}_{A[Y^{(r)},A]}, 1_{A[Y^{(r)},A]})_{A[Y^{(r)},A]} + (\widehat{\chi}_{A[(Y^{(r)})^c,A]}, 1_{A[(Y^{(r)})^c,A]})_{A[(Y^{(r)})^c,A]} = 2(|A|+1) >$$
$$> (\widehat{\chi}_{A[Y^{(r)},A]\cap A[(Y^{(r)})^c,A]}, 1_{A[Y^{(r)},A]\cap A[(Y^{(r)})^c,A]})_{A[Y^{(r)},A]\cap A[(Y^{(r)})^c,A]} = (\widehat{\chi}_A, 1_A)_A = |A|+2.$$

Применим теорему 5.19 [5], по которой подгруппа

$$L = \langle A[Y^{(r)},A], A[(Y^{(r)})^c,A]\rangle$$

собственная.

Пусть $L_{\pi'} \subseteq Y^{(L_{\pi'})}$ – максимальная A-инвариантная подгруппа. Нетрудно видеть, что $A \subseteq L$, $L_{\pi'} \nsubseteq C$ и что компонента θ' характера $\widehat{\chi}_{AY^{(L_{\pi'})}}$ может иметь только степень, равную $|A|$. Тогда

$$Y^{(L_{\pi'})} = [Y^{(L_{\pi'})},A] \times C_{Y^{(L_{\pi'})}}(A), \quad [Y^{(L_{\pi'})},A] \subseteq Z(Y^{(L_{\pi'})}).$$

Так как

$$[Y^{(r)},A], [(Y^{(r)})^c,A] \subseteq L_{\pi'} \subseteq Y^{(L_{\pi'})},$$

то

$$Y^{(r)} \subseteq N_B([Y^{(r)},A]) \subseteq Y^{(L_{\pi'})}$$

и

$$(Y^{(r)})^c \subseteq N_B([(Y^{(r)})^c,A]) \subseteq Y^{(L_{\pi'})}.$$

Поскольку подгруппы $Y^{(r)}$ и $(Y^{(r)})^c$ максимальные A-инвариантные, то

$$Y^{(L_{\pi'})} = Y^{(r)} = (Y^{(r)})^c.$$

Это означает, что $c \in N_B(Y^{(r)})$. Но $N_B(Y^{(r)}) = Y^{(r)}$, ввиду того, что подгруппа $Y^{(r)}$ максимальная A-инвариантная. Поэтому $c \in Y^{(r)}$, т. е. $c \in C_{Y^{(r)}}(A)$. Это противоречит выбору элемента c. Стало быть, и в этом случае $C \setminus C_{Y^{(r)}}(A) = \varnothing$, т. е. $C = C_{Y^{(r)}}(A)$.

Как видим, в обоих случаях $C = C_{Y^{(r)}}(A)$ и, значит,

$$Y^{(r)} = [Y^{(r)},A] \times C, \quad [Y^{(r)},A] \subseteq Z(Y^{(r)}).$$

Пусть $q, r \neq r_i \in \pi(|B|)$, $i = \overline{3, |\pi(|B|)|}$. Понятно, что

$$Y^{(r_i)} = [Y^{(r_i)},A] \times C, \quad [Y^{(r_i)},A] \subseteq Z(Y^{(r_i)})$$

для каждого i.

Обозначим также

$$M = [Y^{(r)},A] \cap [Y^{r_i},A]$$

и таким же образом исследуем случаи, когда $M \neq 1$ и когда $M = 1$. Мы также получим, что $Y^{(r)} = Y^{r_i}$ для каждого $i = \overline{3, |\pi(|B|)|}$. Это означает, что $|B : Y^{(r)}| = |B_q|$. Тогда $B = Y^{(r)}B_q$ и $\Gamma = \Gamma_1 B_q$, где $\Gamma_1 = AY^{(r)}$.

Утверждение 3.15.5 доказано.

Рассмотрим характер $\psi = (1_{\Gamma_1})^{\Gamma}$ степени $|\Gamma : \Gamma_1| = |B_q| = 2|A|+1$. Поскольку $(\psi_\Gamma, 1_\Gamma)_\Gamma = 1$, то для каждой неприводимой компоненты ν характера ψ получаем, что $\nu(1) \leqslant 2|A|$. Так как $A \ntriangleleft \Gamma$, то мы можем считать, что $A\ker\nu/\ker\nu \ntriangleleft \Gamma/\ker\nu$. Тогда по лемме 2.10 [1]

$$\nu(1) \in \{|A|-1; |A|; |A|+1; 2(|A|-1); 2|A|-1\}.$$

По теореме 1 [15] при всех указанных случаях, за исключением значений $\nu(1) = |A|$, $\nu(1) = 2|A|-1 = 17$ и $\nu(1) = 2|A|$ группа $B = O_r(B)C$, где $\nu(1) = r^{\alpha_1}$, $\alpha_1 \in \mathbf{N}$. Это противоречит тому, что группа B простая.

Если $\nu(1) = |A|$, то из теоремы 2 [15] вытекает, что $B = [B,A] \times C_B(A)$. Мы видим, что группа B не является простой.

Если $\nu(1) = 2|A|-1 = 17$, то $|A| = 3^2$. Тогда B – простая $3'$-группа. Ранее мы убедились, что это невозможно.

Остался случай, когда $\nu(1) = 2|A|$. Если характер ν приводим, то группа B абелева, ибо все неприводимые компоненты характера $\widehat{\chi}_B$ линейные. Поэтому остается случай, когда характер ν имеет степень, равную $2|A|$ и неприводим. Применим теорему [17]. Из нее мы узнаем, что группа B разрешима, если неприводимая компонента β^* степени 2 характера ν_B импримитивна. Поэтому предположим, что β^* является примитивным характером. Применим теорему 14.23 [5]. По ней

$$|B : Z(\beta^*)| = 2^2 \cdot 3; 2^3 \cdot 3; 2^2 \cdot 3 \cdot 5.$$

Но $Z(\beta^*) \triangleleft B$, что противоречит тому, что группа B простая. Поэтому $|B| = 2^2 \cdot 3 \cdot 5$. Так как $\widehat{\chi}(1) = = |B_q| = 2|A| + 1 > 7$ делит $|B|$, то мы видим, что $B \subseteq C$, т. е. $A \triangleleft \Gamma$. Это противоречит условию $(^*)$.

Лемма 3.15 доказана. □

Литература

1. *Ядченко А. А.* О разрешимости и факторизации некоторых П-разрешимых неприводимых линейных групп примарной степени. Часть I // Труды Института математики. 2022. Т. 30, № 1–2. С. 84–98.

2. *Ядченко А. А.* О разрешимости и факторизации некоторых П-разрешимых неприводимых линейных групп примарной степени. Часть II // Труды Института математики. 2023. Т. 31, № 1. С. 77–89.

3. *Ядченко А. А.* О разрешимости и факторизации некоторых П-разрешимых неприводимых линейных групп примарной степени. Часть III // Труды Института математики. 2023. Т. 31, № 2. С. 91–102.

4. *Gorenstein D.* Finite groups. New York: Harper and Row, 1968. 527 p.

5. *Isaacs I. M.* Character theory of finite groups. New York: Academic Press, 1976. 303 p.

6. *Ядченко А. А., Романовский А. В.* К проблемме Айзекса о конечных p-разрешимых линейных группах // Математические заметки. 2001. Т. 69, Вып. 1. С. 144–152.

7. *Ядченко А. А.* О П-разрешимых неприводимых линейных группах с холловой TI-подгруппой нечетного порядка I // Труды Института математики. 2008. Т. 16, № 2. С. 118–130.

8. *Glauberman G.* Correspodences of characters for relatively prime operator groups // Canad. J. Math. 1968. N 20. P. 1465–1488.

9. *Isaacs I. M., Robinson G. R.* The Number of Distinct Eigenvalues of Elements in Finite Linear Constituents of certain Characrer Restrictions Groups // Proc. American Math. Soc. 1998. Vol. 126, N 9. P. 2615–2617.

10. *Winter D. L.* On the Structure of Certain p-Solvable Linear Groups II // J. of Algebra. 1975. Vol. 33. P. 170–190.

11. *Ядченко А. А.* О П-разрешимых неприводимых линейных группах с холловой TI-подгруппой нечетного порядка II // Труды Института математики. 2009. Т. 17, № 2. С. 94–104.

12. *Романовский А. В.* Исключительные характеры конечных групп. Минск: Наука и техника, 1985. 148 с.

13. *Huppert B., Blackburn N.* Finite Groups III. Berlin, New York, Heidelberg: Academic Press, 1982. 454 p.

14. *Ядченко А. А.* Автоморфизмы и нормальные подгруппы линейных групп // Математические заметки. 2007. Т. 82, Вып. 3. С. 469–476.

15. *Ядченко А. А.* О факторизации некоторых П-разрешимых неприводимых линейных групп // Труды Института математики. 2019. Т. 27, № 1–2. С. 79–107.

16. *Crescenzo P.* A Diophantine equation which arises in the theory of finite groups // Advances in Mathematics. 1975. Vol. 17, N 1. P. 25–29.

17. *Ядченко А. А.* О нормальных подгруппах и факторизации некоторых π-разрешимых неприводимых линейных групп // Труды Института математики. 2021. Т. 29, № 1–2. С. 149–164.

References

1. Yadchenko A. A. On the solvability and factorization of some π-solvable irreducible linear groups of primary degree. Part I. *Proceedings of the Institute of Mathematics,* 2022, vol. 30, no. 1–2, pp. 84–98 (in Russian).

2. Yadchenko A. A. On the solvability and factorization of some π-solvable irreducible linear groups of primary degree. Part II. *Proceedings of the Institute of Mathematics,* 2023, vol. 31, no. 1, pp. 77–89 (in Russian).

3. Yadchenko A. A. On the solvability and factorization of some π-solvable irreducible linear groups of primary degree. Part III. *Proceedings of the Institute of Mathematics,* 2023, vol. 31, no. 2, pp. 91–102 (in Russian).

4. Gorenstein D. *Finite groups*. New York, Harper and Row, 1968, 527 p.

5. Isaacs I. M. *Character theory of finite groups*. New York, Academic Press, 1976, 303 p.

6. Yadchenko A. A., Romanovskii A. V. On the Isaacs problem concerning finite *p*-solvable linear groups. *Math. Notes*, 2001, vol. 69, no. 1, pp. 144–152 (in Russian).

7. Yadchenko A. A. On the π-solvable irreducible linear groups with Hall TI-subgrops of odd order I. *Proceedings of the Institute of Mathematics,* 2008, vol. 16, no. 2, pp. 118–130 (in Russian).

8. Glauberman G. Correspodences of characters for relatively prime operator groups. *Canad. J. Math.,* 1968, no. 20, pp. 1465–1488.

9. Isaacs I. M., Robinson G. R. The Number of Distinct Eigenvalues of Elements in Finite Linear Constituents of certain Characrer Restrictions Groups *Proc. American Math. Soc.,* 1998, vol. 126, no. 9, pp. 2615–2617.

10. Winter D. L. On the Structure of Certain p-Solvable Linear Groups II. *J. of Algebra,* 1975, vol. 33, pp. 170–190.

11. Yadchenko A. A. On the π-solvable irreducible linear groups with Hall TI-subgrops of odd order II. *Proceedings of the Institute of Mathematics,* 2009, vol. 17, no. 2, pp. 94–104 (in Russian).

12. Romanovskii A. V. *Exceptional characters of finite grops.* Minsk, Nauka i tekhnika Publ., 1985, 148 p. (in Russian).

13. Huppert B., Blackburn N. *Finite Groups III.* Berlin, New York, Heidelberg, Academic Press, 1982, 454 p.

14. Yadchenko A. A. Automorphisms and normal subgroups of linear groups. *Math. Notes*, 2007, vol. 82, no. 3, pp. 469–476 (in Russian).

15. Yadchenko A. A. On the factorization some π-solvable irreducible linear groups. *Proceedings of the Institute of Mathematics,* 2019, vol. 27, no. 1–2, pp. 79–107 (in Russian).

16. Crescenzo P. A Diophantine equation which arises in the theory of finite groups. *Advances in Mathematics,* 1975, vol. 17, no. 1, pp. 25–29.

17. Yadchenko A. A. On the normal subgroups and factorization some π-solvable irreducible linear groups. *Proceedings of the Institute of Mathematics,* 2021, vol. 29, no. 1–2, pp. 149–164 (in Russian).

Национальная академия наук Беларуси
Труды Института математики НАН Беларуси. 2024. Том 32. № 2. С. 31–42

ВЕЩЕСТВЕННЫЙ, КОМПЛЕКСНЫЙ И ФУНКЦИОНАЛЬНЫЙ АНАЛИЗ
REAL, COMPLEX AND FUNCTIONAL ANALYSIS

УДК 517.51; 517.53

ПРИМЕНЕНИЕ ДЕЙСТВИТЕЛЬНОГО ПРОСТРАНСТВА ХАРДИ–СОБОЛЕВА НА ПРЯМОЙ ДЛЯ НАХОЖДЕНИЯ НАИЛУЧШИХ РАЦИОНАЛЬНЫХ ПРИБЛИЖЕНИЙ В L_p

Т. С. Мардвилко

Белорусский государственный университет, Минск, Беларусь
e-mail: mardvilko@gmail.com

Поступила: 08.08.2024 **Исправлена:** 25.10.2024 **Принята:** 12.12.2024

Ключевые слова: пространство Харди, пространство Соболева, пространство Харди–Соболева, рациональная аппроксимация, L_p-приближения, функции ограниченной вариации.

Аннотация. Данная работа посвящена разработке методов действительного пространства Харди–Соболева на прямой для нахождения наилучших рациональных приближений в пространстве L_p. В основе рассмотренных методов лежит представление функции данного пространства суммой простых функций и применение интеграла типа Коши. Получены достаточные условия принадлежности функции рассматриваемому пространству и доказаны неравенства для оценки соответствующей σ-нормы. С помощью полученных результатов найдены точные порядковые оценки наилучших рациональных приближений некоторых функций. В частности, из полученных результатов следует известная оценка наилучших рациональных приближений функции ограниченной вариации.

APPLICATION OF THE REAL HARDY–SOBOLEV SPACE ON THE LINE FOR FINDING THE BEST RATIONAL APPROXIMATIONS IN L_p

T. S. Mardvilko

Belarusian State University, Minsk, Belarus
e-mail: mardvilko@gmail.com

Received: 08.08.2024 **Revised:** 25.10.2024 **Accepted:** 12.12.2024

Keywords: Hardy space, Sobolev space, Hardy–Sobolev space, rational approximation, L_p-approximations, functions of bounded variation.

Abstract. This work is dedicated to developing methods of the real Hardy–Sobolev space on the line for finding the best rational approximations in the L_p space. The methods considered are based on representing a function of this space as a sum of simple functions and the application of a Cauchy-type integral. Sufficient conditions for a function's membership in the considered space have been obtained and inequalities for assessing the corresponding σ-norm have been proven. Using the obtained results, exact order estimates of the best rational approximations of certain functions have been found. In particular, from the obtained results, the well-known estimate of the best rational approximations of a function of bounded variation follows.

1. Введение

Пусть $I \neq \varnothing$ измеримое подмножество прямой $\mathbb{R}$. Через $L_p(I)$, $0 < p \leqslant \infty$, обозначим пространство Лебега измеримых функций $f : I \to \mathbb{C}$, суммируемых в p-й степени на I, т. е. для которых конечна норма при $1 \leqslant p \leqslant \infty$, а при $0 < p < 1$ p-норма

$$\|f\|_{L_p(I)} = \left(\int\limits_I |f(x)|^p \, dx \right)^{1/p}, \quad 0 < p < \infty,$$

$$\|f\|_{L_\infty(I)} = \operatorname{ess\,sup}\{|f(x)| : x \in I\}, \quad p = \infty.$$

В случае отрезка $[a,b]$ будем писать $L_p[a,b]$ вместо $L_p([a,b])$.

Нам также понадобятся пространства Харди для верхней полуплоскости $\Pi = \{z \in \mathbb{C} : \operatorname{Im} z > 0\}$. Аналитическая в Π функция f принадлежит пространству Харди H_p, $0 < p \leqslant \infty$, если

$$\|f\|_{H_p} := \sup_{y>0} \|f(\cdot + iy)\|_{L_p(\mathbb{R})} < \infty.$$

Далее рассмотрим ядро Пуассона и сопряженное ядро Пуассона соответственно [1] для верхней полуплоскости

$$P_y(t) = \frac{y}{t^2+y^2} \quad \text{и} \quad Q_y(t) = \frac{t}{t^2+y^2}, \quad \text{где} \quad t \in \mathbb{R}, \quad y > 0.$$

Заметим, что при $z = x + iy$, $x \in \mathbb{R}$,

$$P_y(x-t) + iQ_y(x-t) = \frac{-i}{t-z} \tag{1.1}$$

– ядро Коши.

Введем аналитическую в Π функцию, соответствующую действительной функции $g \in L_p(\mathbb{R})$, $1 < p < \infty$,

$$f(z) = (Cg)(z) = \frac{1}{\pi i} \int_{\mathbb{R}} \frac{g(t)}{t-z}\, dt. \tag{1.2}$$

Как известно [1], f принадлежит пространству Харди H_p аналитических в Π функций. Следовательно, для почти всех $x \in \mathbb{R}$ существует $\lim\limits_{y\to+0} f(x+iy) =: f(x+i0)$. Из (1.1), (1.2) и свойств интеграла Пуассона [1] получаем, что $\operatorname{Re} f(x+i0) = g(x)$ почти всюду на $\mathbb{R}$. При этом справедливы неравенства

$$\|g\|_{L_p(\mathbb{R})} \leqslant \|f\|_{H_p} \leqslant c(p)\|g\|_{L_p(\mathbb{R})}.$$

Здесь левое неравенство очевидно, правое – составляет содержание теоремы М. Рисса (см. [1]).

Данная статья посвящена разработке методов действительного пространства Харди–Соболева на прямой для нахождения наилучших рациональных L_p-приближений. Удобство использования данного пространства для решения задач рациональной аппроксимации основано на представлении функций данного пространства суммой простых функций (см. п. 2 и п. 4), а также в виде интеграла типа Коши (см. п. 2 и п. 5).

2. Действительное пространство Харди–Соболева

Пусть $s \in \mathbb{N}$, $1 < p < \infty$ и $\sigma = \left(s + \frac{1}{p}\right)^{-1}$, действительная функция $g \in L_p(\mathbb{R})$, а $f(z)$ соответствующая g функция вида (1.2). Тогда g принадлежит действительному пространству Харди–Соболева $\mathcal{H}^s_\sigma$, если f принадлежит комплексному пространству Харди–Соболева H^s_σ, т. е. $f^{(s)} \in H_\sigma$. При этом σ-норма функции g вводится следующим образом:

$$\|g\|_{\mathcal{H}^s_\sigma} = \left\|(Cg)^{(s)}\right\|_{H_\sigma} = \left\|f^{(s)}\right\|_{H_\sigma}.$$

Из данного определения и полноты пространства H_σ следует полнота пространства $\mathcal{H}^s_\sigma$ и σ-неравенство треугольника

$$\|g+h\|^\sigma_{\mathcal{H}^s_\sigma} \leqslant \|g\|^\sigma_{\mathcal{H}^s_\sigma} + \|h\|^\sigma_{\mathcal{H}^s_\sigma}, \quad g, h \in \mathcal{H}^s_\sigma.$$

Для пространства $\mathcal{H}^s_\sigma$ введем еще одну эквивалентную σ-норму. Пусть $I \neq \varnothing$ – связное подмножество $\mathbb{R}$. $C(I) = C^0(I)$ – множество функций непрерывных на I. Через $C^s(I)$, $s \in \mathbb{N}$, обозначим множество s раз непрерывно дифференцируемых функций $f : I \to \mathbb{R}$. Через $W^s_p(I)$ $(s \in \mathbb{N}, 1 \leqslant p \leqslant \infty)$ обозначим пространство Соболева функций $f \in C^{s-1}(I)$, таких, что $f^{(s-1)}$ абсолютно непрерывна на I и $f, f^{(s)} \in L_p(I)$.

Действительную функцию $\varphi \in W^s_\infty(\mathbb{R})$ называем s-простой, если она финитна. С s-простой функцией φ будем ассоциировать отрезок $J = J(\varphi)$, называемый опорным, такой, что $\operatorname{supp}\varphi \subset J$. Далее для s-простой функции введем характеристику

$$\mu_{s\sigma}(\varphi) = |J|^{\frac{1}{\sigma}} \|\varphi^{(s)}\|_{L_\infty(J)},$$

где $|J|$ – длина отрезка J.

Теорема 2.1. *Функция $g \in L_p(\mathbb{R})$, $1 < p < \infty$, принадлежит $\mathcal{H}^s_\sigma$, $s \in \mathbb{N}$, $\sigma = \left(s + \frac{1}{p}\right)^{-1}$, в том и только в том случае, если существует последовательность $\{\varphi_k\}_{k=1}^\infty$ s-простых функций, удовлетворяющих условиям*

$$\sum_{k=1}^{\infty} \mu_{s\sigma}(\varphi)^\sigma =: A < \infty, \tag{2.1}$$

$$\sum_{k=1}^{\infty} \varphi_k(x) = g(x), \tag{2.2}$$

где ряд (2.2) сходится по норме пространства $L_p(\mathbb{R})$. При этом $\|g\|^\#_{\mathcal{H}^s_\sigma} = \inf\left\{A^{\frac{1}{\sigma}} : \text{выполняются (2.1) и (2.2)}\right\}$ является σ-нормой в $\mathcal{H}^s_\sigma$, эквивалентной $\|\cdot\|_{\mathcal{H}^s_\sigma}$.

Теорема 2.1 является следствием результата Р. Койфмана [2–4] об атомическом разложении функций класса $\operatorname{Re} H_p$ при $p \leqslant 1$.

Из определения σ-норм $\|\cdot\|_{\mathcal{H}^s_\sigma}$ и $\|\cdot\|^\#_{\mathcal{H}^s_\sigma}$ следует, что они при линейной замене $x = kt + b$, $k \neq 0$, аргумента функции ведут себя аналогично норме $\|\cdot\|_{L_p(\mathbb{R})}$, $1 < p < \infty$. Именно, справедливо

Утверждение 2.1. *Для любых $k, b \in \mathbb{R}$, $k \neq 0$, и $g \in \mathcal{H}^s_\sigma$ имеют место равенства*

$$\|g(k\cdot + b)\|_{\mathcal{H}^s_\sigma} = |k|^{-1/p}\|g\|_{\mathcal{H}^s_\sigma},$$

$$\|g(k\cdot + b)\|^\#_{\mathcal{H}^s_\sigma} = |k|^{-1/p}\|g\|^\#_{\mathcal{H}^s_\sigma}.$$

В силу эквивалентности σ-норм $\|\cdot\|_{\mathcal{H}^s_\sigma}$ и $\|\cdot\|^\#_{\mathcal{H}^s_\sigma}$ далее будет использоваться обозначение $\|\cdot\|_{\mathcal{H}^s_\sigma}$. Из контекста будет видно, какая из σ-норм применяется.

3. Прямая и обратная теоремы рациональной аппроксимации для функций из действительного пространства Харди–Соболева

Пусть $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. Через $\mathcal{P}_n$ и $\mathcal{R}_n$, $n \in \mathbb{N}_0$, обозначим соответственно множества алгебраических полиномов и рациональных функций степени не выше n с действительными коэффициентами.

Введем $R_n(f)_p = R_n(f; L_p(I))$ – наилучшее приближение функции $f \in L_p(I)$ посредством множества $\mathcal{R}_n$:

$$R_n(f)_p = R_n(f; L_p(I)) = \inf\left\{\|f - r\|_{L_p(I)} : r \in \mathcal{R}_n\right\}, \quad n \in \mathbb{N}_0.$$

Для функций из пространства $\mathcal{H}^s_\sigma$ справедливы следующая прямая (теорема 3.1) и обратная (теорема 3.2) теоремы рациональной аппроксимации. Эти теоремы вытекают из их аналогов для пространства H^s_σ (см. Е. И. Стельмах [5]) и теоремы М. Рисса [1].

Через $c, c_1, c_2, \ldots$ будем обозначать положительные величины (постоянные). В скобках будем указывать параметры, от которых зависят эти величины. Различные постоянные будем нумеровать различными индексами, которые будут меняться в рамках одного раздела.

Теорема 3.1. *Пусть $s \in \mathbb{N}$, $1 < p < \infty$ и $\sigma = \left(s + \frac{1}{p}\right)^{-1}$. Если $g \in \mathcal{H}^s_\sigma$, то*

$$R_n(g)_p \leqslant \frac{c_1(s,p)}{n^s}\|g\|_{\mathcal{H}^s_\sigma}, \quad n \in \mathbb{N}.$$

Теорема 3.2. *Пусть $s \in \mathbb{N}$, $1 < p < \infty$, $\sigma = \left(s + \frac{1}{p}\right)^{-1}$ и $g \in L_p(\mathbb{R})$. Если*

$$\sum_{n=1}^{\infty} \frac{1}{n}\left(n^s R_n(g)_p\right)^\sigma =: B < \infty,$$

то $g \in \mathcal{H}^s_\sigma$ и $\|g\|_{\mathcal{H}^s_\sigma} \leqslant c_2(s,p) B^{\frac{1}{\sigma}}$.

4. Применение простых функций

Для краткости будем полагать $\|f\|_I := \|f\|_{L_\infty(I)}$.

Ранее нами была доказана

Лемма 4.1 [6]. *Пусть* $-\infty < a < b < +\infty$, $s \in \mathbb{N}$, *и* $p \in \mathcal{P}_{2s-1}$. *Тогда для* $l = 0,1,\ldots,2s-1$ *выполняется неравенство*

$$\|p^{(l)}\|_{[a,b]} \leqslant c_1(s) \sum_{j=0}^{s-1} (b-a)^{j-l} \left(\left|p^{(j)}(a)\right| + \left|p^{(j)}(b)\right| \right).$$

Следующая лемма аналогична лемме 2 из [6]. Здесь, в отличие от [6], функция ψ может иметь разрывы в точках a или b, а также рассматривается другая характеристика $\mu_{s\sigma}$ простых функций.

Лемма 4.2. *Пусть* $\psi \in L_\infty(\mathbb{R})$, $\operatorname{supp}\psi \subset [a,b]$, $-\infty < a < b < +\infty$, *и* $\psi\big|_{[a,b]} \in W^s_\infty$, $s \in \mathbb{N}$. *Тогда при* $p \in (1,+\infty)$ *и* $\frac{1}{\sigma} = s + \frac{1}{p}$ *имеем* $\psi \in \mathcal{H}^s_\sigma$ *и*

$$\|\psi\|_{\mathcal{H}^s_\sigma} \leqslant c_2(p,s) \sum_{j=0}^{s} (b-a)^{j+\frac{1}{p}} \|\psi^{(j)}\|_{[a,b]}. \tag{4.1}$$

Доказательство. Рассмотрим вначале случай $[a,b] = [0,1]$. Для каждого $k \in \mathbb{N}_0$ найдем полиномы p_{k0} и p_{k1} степени не выше $2s-1$, являющиеся решением интерполяционной задачи Эрмита, а именно для $j = 0,1,\ldots,s-1$ выполнены равенства

$$p_{k0}^{(j)}(0) = \psi^{(j)}(+0), \quad p_{k0}^{(j)}\left(-2^{-k}\right) = 0;$$

$$p_{k1}^{(j)}(1) = \psi^{(j)}(1-0), \quad p_{k1}^{(j)}\left(1+2^{-k}\right) = 0.$$

Пусть $J_{k0} := [-2^{-k},0]$, $J_{k1} := [1,1+2^{-k}]$ и функции

$$\psi_k(x) = \begin{cases} p_{k0}(x), & x \in J_{k0}, \\ p_{k1}(x), & x \in J_{k1}, \\ \psi(x), & x \in \mathbb{R}\setminus(J_{k0}\cup J_{k1}). \end{cases}$$

Согласно определению $\psi_k \in W^s_\infty(\mathbb{R})$ для всех $k \in \mathbb{N}_0$.

Тогда ψ можно представить в виде

$$\psi(x) = \psi_0(x) + \sum_{k=0}^{\infty}\sum_{i=0}^{1} \varphi_{ki}(x), \quad x \in \mathbb{R}, \tag{4.2}$$

где

$$\varphi_{ki}(x) = \begin{cases} \psi_{k+1}(x) - \psi_k(x), & x \in J_{ki}, \\ 0, & x \in \mathbb{R}\setminus J_{ki}. \end{cases}$$

Если $\psi(0) = \psi(1) = 0$, то ряд (4.2) сходится равномерно на $\mathbb{R}$. Если же $\psi(0) \neq 0$ или $\psi(1) \neq 0$, то сходимость будет точечной на $\mathbb{R}$ из-за того, что $\psi(x)$ разрывна при $x = 0$ или $x = 1$.

Заметим, что ψ_0 и φ_{ki} являются s-простыми функциями с опорными интервалами

$$J(\psi_0) = [-1,2] \quad \text{и} \quad J(\varphi_{ki}) = J_{ki}$$

соответственно.

Для ψ_0 имеем

$$\mu_{s\sigma}(\psi_0) = 3^{1/\sigma}\left\|\psi_0^{(s)}\right\|_{[-1,2]} = 3^{1/\sigma}\max\left\{\left\|\psi_0^{(s)}\right\|_{[-1,0]}, \left\|\psi_0^{(s)}\right\|_{[0,1]}, \left\|\psi_0^{(s)}\right\|_{[1,2]}\right\}. \tag{4.3}$$

На $[-1,0]$ функция ψ_0 является многочленом степени $2s-1$. С учетом условий

$$\psi_0^{(j)}(-1) = 0, \quad \psi_0^{(j)}(0) = \psi^{(j)}(+0), \quad j = 0,1,\ldots,s-1,$$

из леммы 4.1 получим

$$\left\|\psi_0^{(s)}\right\|_{[0,1]} \leqslant c_3(s)\sum_{j=0}^{s-1}\left|\psi_0^{(j)}(0)\right| \leqslant c_3(s)\sum_{j=0}^{s-1}\left\|\psi^{(j)}\right\|_{[0,1]}. \tag{4.4}$$

Рассуждая аналогично, имеем

$$\left\|\psi_0^{(s)}\right\|_{[1,2]} \leqslant c_4(s)\sum_{j=0}^{s-1}\left\|\psi^{(j)}\right\|_{[0,1]}. \tag{4.5}$$

Согласно соотношениям (4.3)–(4.5) справедлива оценка

$$\mu_{s\sigma}(\psi_0) \leqslant c_5(s)\sum_{j=0}^{s}\left\|\psi^{(j)}\right\|_{[0,1]}.$$

Покажем, что

$$\mu_{s\sigma}(\varphi_{ki}) \leqslant c_6(s)2^{-k/p}\sum_{j=0}^{s}\left\|\psi^{(j)}\right\|_{[0,1]}.$$

Проведем рассуждения для $i=0$.

Заметим, что на каждом из отрезков $[-2^{-k},-2^{-k-1}]$ и $[-2^{-k-1},0]$ функция φ_{k0} является многочленом степени $2s-1$, причем

$$\varphi_{k0}(x) = -\psi_k(x) \quad \text{при} \quad x\in[-2^{-k},-2^{-k-1}].$$

Следовательно, согласно лемме 4.1, справедливы неравенства

$$\left\|\varphi_{k0}^{(j)}\right\|_{[-2^{-k},-2^{-k-1}]} = \left\|\psi_k^{(j)}\right\|_{[-2^{-k},-2^{-k-1}]} \leqslant \left\|\psi_k^{(j)}\right\|_{[-2^{-k},0]} \leqslant$$
$$\leqslant c_1(s)\sum_{m=0}^{s-1}2^{-k(m-j)}\left|\psi^{(m)}(+0)\right| \leqslant c_1(s)\sum_{m=0}^{s-1}2^{-k(m-j)}\left\|\psi^{(m)}\right\|_{[0,1]}. \tag{4.6}$$

Учитывая, что

$$\psi_{k0}^{(j)}(0) = \psi_{k+1}^{(j)}(0) - \psi_k^{(j)}(0) = \psi^{(j)}(+0) - \psi^{(j)}(+0) = 0$$

для всех $j=0,1,\ldots,s-1$, и неравенства (4.6), получим

$$\left\|\varphi_{k0}^{(s)}\right\|_{[-2^{-k-1},0]} \leqslant c_7(s)\sum_{j=0}^{s-1}2^{-k(j-s)}\left|\varphi_{k0}^{(j)}(-2^{-k-1})\right| \leqslant$$
$$\leqslant c_7(s)\sum_{j=0}^{s-1}2^{-k(j-s)}\left\|\varphi_{k0}^{(j)}\right\|_{[-2^{-k},-2^{-k-1}]} \leqslant c_8(s)\sum_{j=0}^{s-1}2^{-k(j-s)}\left\|\psi^{(j)}\right\|_{[0,1]}. \tag{4.7}$$

Применяя (4.6) с $j=s$ и (4.7), получим

$$\mu_{s\sigma}(\varphi_{k0}) = 2^{-k/\sigma}\|\varphi^{(s)}\|_{[-2^{-k},0]} \leqslant 2^{-k/\sigma}\left(\|\varphi^{(s)}\|_{[-2^{-k},-2^{-k-1}]} + \|\varphi^{(s)}\|_{[-2^{-k-1},0]}\right) \leqslant$$
$$\leqslant 2^{-k/\sigma}c_9(s)\sum_{j=0}^{s-1}2^{-k(j-s)}\left\|\psi^{(j)}\right\|_{[0,1]} \leqslant 2^{-k/p}c_6(s)\sum_{j=0}^{s-1}\left\|\psi^{(j)}\right\|_{[0,1]}.$$

Согласно теореме 2.1

$$\|\psi\|^{\sigma}_{\mathcal{H}^s_\sigma} \leqslant \mu^{\sigma}_{s\sigma}(\psi_0) + \sum_{k=0}^{\infty}\sum_{i=0}^{1}\mu^{\sigma}_{s\sigma}(\varphi_{ki}) \leqslant$$
$$\leqslant \left(c_5(s)\sum_{i=0}^{s}\left\|\psi^{(j)}\right\|_{[0,1]}\right)^{\sigma} + \sum_{k=0}^{\infty}2^{-(1-s\sigma)k}\left(2c_6(s)\left(\sum_{j=0}^{s-1}\left\|\psi^{(j)}\right\|_{[0,1]}\right)\right)^{\sigma} \leqslant$$
$$\leqslant c_{10}(p,s)\left(\sum_{j=0}^{s}\left\|\psi^{(j)}\right\|_{[0,1]}\right)^{\sigma}.$$

Таким образом, мы получили неравенство (4.1) для $[a,b]=[0,1]$. Для произвольного отрезка $[a,b]$ лемма вытекает из доказанного случая с помощью замены $x=(b-a)t+a$, $t\in[0,1]$, и утверждения 2.1. □

Пример 4.1. *Пусть* $1<p<\infty$, $\alpha>0$ *и*

$$\lambda(x)=\begin{cases}\dfrac{1}{x^{\frac{1}{p}}(\ln x)^{\frac{1}{p}+\alpha}} & \text{при} \quad x\geqslant e,\\[2ex] 0 & \text{при} \quad x<e.\end{cases}$$

Тогда

$$R_n(\lambda,L_p(\mathbb{R}))\asymp n^{-\alpha},\quad n\in\mathbb{N}, \tag{4.8}$$

где постоянные, спрятанные символом $\asymp$, *зависят лишь от* α *и* p.

При получении верхней и нижней оценок из (4.8) нам понадобится следующее свойство функции $\lambda(x)$. Для любых $j\in\mathbb{N}_0$ и $x\geqslant e$ имеет место равенство

$$\lambda^{(j)}(x)=\frac{(-1)^j}{x^{\frac{1}{p}+j}}\sum_{l=0}^{j}\frac{A_{jl}}{(\ln x)^{\frac{1}{p}+\alpha+l}}, \tag{4.9}$$

где $A_{jl}>0$ и зависят лишь от j,l,p и α.

Равенство (4.9) легко доказать методом математической индукции.

Доказательство верхней оценки из (4.8)**.** Пусть $I_k=\left[e^k,e^{k+1}\right]$, $k\in\mathbb{N}$. Положим

$$\psi_k(x)=\begin{cases}\dfrac{1}{x^{\frac{1}{p}}(\ln x)^{\frac{1}{p}+\alpha}}, & x\in I_k,\\[2ex] 0, & x\in\mathbb{R}\setminus I_k.\end{cases}$$

Функции ψ_k удовлетворяют лемме 4.2 для любого $s\in\mathbb{N}$ и $\sigma=\left(s+\frac{1}{p}\right)^{-1}$. Возьмем $s=[\alpha]+1$. Тогда

$$\|\psi_k\|_{\mathcal{H}^s_\sigma}\leqslant c_{11}(p,s)\sum_{j=0}^{s}\exp\left(k\left(\frac{1}{p}+j\right)\right)\left\|\psi_k^{(j)}\right\|_{I_k}.$$

Применяя (4.9), находим

$$\left\|\psi_k^{(j)}\right\|_{I_k}\leqslant\frac{c_{12}(p,\alpha,j)}{k^{\frac{1}{p}+j}}\exp\left(-k\left(\frac{1}{p}+j\right)\right).$$

Таким образом, используя лемму 4.2, получим

$$\|\psi_k\|_{\mathcal{H}^s_\sigma}\leqslant\frac{c_{13}(p,\alpha)}{k^{\frac{1}{p}+\alpha}}.$$

Далее рассмотрим функцию

$$\Lambda_n=\psi_1+\psi_2+\ldots+\psi_n.$$

Для выбранного выше $s\in\mathbb{N}$ и $\sigma=\left(s+\frac{1}{p}\right)^{-1}$ имеем $\Lambda_n\in\mathcal{H}^s_\sigma$ и

$$\|\Lambda_n\|^\sigma_{\mathcal{H}^s_\sigma}=\left\|\sum_{k=1}^{n}\psi_k\right\|^\sigma_{\mathcal{H}^s_\sigma}\leqslant\sum_{k=1}^{n}\|\psi_k\|^\sigma_{\mathcal{H}^s_\sigma}\leqslant\sum_{k=1}^{n}\frac{c_{14}(p,s)}{k^{\left(\frac{1}{p}+\alpha\right)\sigma}}\leqslant\frac{c_{15}(p,s)}{n^{\left(\frac{1}{p}+\alpha\right)\sigma-1}}.$$

Итак, согласно теореме 3.1 получаем

$$R_n\left(\Lambda_n,L_p(\mathbb{R})\right)\leqslant\frac{c_{16}(p,s)}{n^s}\|\Lambda_n\|_{\mathcal{H}^s_\sigma}\leqslant\frac{c_{17}(p,s)}{n^s}\frac{1}{n^{\frac{1}{p}+\alpha-\frac{1}{\sigma}}}=\frac{c_{18}(p,s)}{n^\alpha}.$$

Для доказательства верхней оценки осталось заметить, что

$$R_n(\lambda, L_p(\mathbb{R})) \leqslant R_n(\Lambda_n, L_p(\mathbb{R})) + \|\lambda - \Lambda_n\|_{L_p(\mathbb{R})}.$$

Последнее слагаемое, в свою очередь, легко вычисляется

$$\|\lambda - \Lambda_n\|_{L_p(\mathbb{R})} = \left(\int\limits_{e^{n+1}}^{+\infty} \frac{dx}{x(\ln x)^{1+\alpha p}} \right)^{\frac{1}{p}} = \frac{1}{(\alpha p)^{\frac{1}{p}}(n+1)^{\alpha}}. \qquad \square$$

Для доказательства нижней оценки из (4.8) нам понадобится теорема 4.1. Введем необходимые обозначения.

Через $E_n(f)_p = E_n(f; L_p(I))$ обозначим наилучшее приближение функции $f \in L_p(I)$ множеством $\mathcal{P}_n$, $n \in \mathbb{N}_0$:

$$E_n(f)_p = E_n(f; L_p(I)) = \inf\left\{ \|f - r\|_{L_p(I)} : r \in \mathcal{P}_n \right\}.$$

Пусть $\sigma = \left(s + \frac{1}{p}\right)^{-1}$, $n \in \mathbb{N}$. Модулем изменения функции $f \in L_p[a,b]$, $(a, b \in \mathbb{R},\ a < b)$, порядка $s \in \mathbb{N}$ называется

$$\varkappa_{sp}(n, f) = \sup\left\{ \sum_{k=1}^{n} \left(E_{s-1}(f, L_p(I_k))\right)^{\sigma} \right\}^{1/\sigma},$$

где верхняя грань берется по всем наборам отрезков $I_1, I_2, \ldots, I_n \subset [a,b]$, пересекающихся, разве лишь, по концевым точкам.

Теорема 4.1 [7; 8]. *Пусть* $1 < p \leqslant \infty$, $s, n \in \mathbb{N}$, $\sigma = \left(s + \frac{1}{p}\right)^{-1}$ *и* $f \in L_p[a,b]$, $(a, b \in \mathbb{R}, a < b)$. *Тогда справедливо неравенство*

$$\varkappa_{sp}(n, f) \leqslant c_{19}(s,p) \left\{ \sum_{k=1}^{n} \frac{1}{k} \left(k^s R_k(f; L_p[a,b])\right)^{\sigma} \right\}^{1/\sigma}.$$

Теорема 4.1 была получена для отрезка $[-1,1]$. Для произвольного отрезка $[a,b]$ она следует из этого случая с помощью линейной замены.

В свою очередь, для оценки модуля изменения функции мы воспользуемся следующей теоремой G. M. Phillips.

Теорема 4.2 [9; 10, p. 222]. *Для каждой функции* $f \in C^s[a,b]$, $s \in \mathbb{N}$, *и* $1 \leqslant p \leqslant \infty$ *существует* $\xi \in [a,b]$ *такое, что*

$$E_{s-1}(f, L_p[a,b]) = c_{20}(p,s)(b-a)^{s+\frac{1}{p}} \left| f^{(s)}(\xi) \right|.$$

Нам также понадобится следующая числовая лемма 4.3. Отметим, что подобные рассуждения, используемые при доказательстве этой леммы, применялись ранее в работе С. Б. Стечкина [11].

Лемма 4.3. *Пусть* $a, b, \alpha, \beta, \gamma$ *– положительные числа, причем* $\beta > \alpha$, *и* $\{\delta_n\}_{n=1}^{\infty}$ *– убывающая положительная числовая последовательность. Если при всех* $n \in \mathbb{N}$ *выполняются неравенства*

(i) $\quad \delta_n \leqslant a n^{-\alpha}$,

(ii) $\quad \displaystyle\sum_{k=1}^{n} \frac{1}{k}\left(k^{\beta}\delta_k\right)^{\gamma} \geqslant b n^{\gamma(\beta-\alpha)}$,

то $\delta_n \geqslant c n^{-\alpha}$ *при всех* $n \in \mathbb{N}$. *Здесь* $c > 0$ *и зависит лишь от* $a, b, \alpha, \beta, \gamma$.

Доказательство. Ввиду условия (i) для каждого $n \geqslant 2$ имеем

$$\sum_{k=1}^{n-1} \frac{1}{k}\left(k^{\beta}\delta_k\right)^{\gamma} \leqslant a^{\gamma} \sum_{k=1}^{n-1} k^{\gamma(\beta-\alpha)-1} \leqslant a^{\gamma} \int\limits_{0}^{n} x^{\gamma(\beta-\alpha)-1}\, dx = \frac{a^{\gamma}}{\gamma(\beta-\alpha)} n^{\gamma(\beta-\alpha)}.$$

Следовательно, при $m>n$, с учетом (ii), получим

$$\sum_{k=n}^{m}\frac{1}{k}\left(k^{\beta}\delta_k\right)^{\gamma}\geqslant bm^{\gamma(\beta-\alpha)}-\frac{a^{\gamma}}{\gamma(\beta-\alpha)}n^{\gamma(\beta-\alpha)}.$$

Заметим, что

$$\sum_{k=n}^{m}\frac{1}{k}\left(k^{\beta}\delta_k\right)^{\gamma}\leqslant\delta_n^{\gamma}\sum_{k=n}^{m}k^{\beta\gamma-1}\leqslant\delta_n^{\gamma}\int\limits_{n-1}^{m+1}x^{\beta\gamma-1}\,dx=\frac{\delta_n^{\gamma}}{\beta\gamma}\left[(m+1)^{\beta\gamma}-(n-1)^{\beta\gamma}\right].$$

Следовательно,

$$\delta_n^{\gamma}\geqslant\frac{\beta\gamma}{(m+1)^{\beta\gamma}-(n-1)^{\beta\gamma}}\left[bm^{\gamma(\beta-\alpha)}-\frac{a^{\gamma}}{\gamma(\beta-\alpha)}n^{\gamma(\beta-\alpha)}\right].$$

Остается положить здесь $m\in\mathbb{N}$ наименьшим, для которого выполняются неравенства $m\geqslant 2n$ и $bm^{\gamma(\beta-\alpha)}\geqslant\frac{2a^{\gamma}}{\gamma(\beta-\alpha)}n^{\gamma(\beta-\alpha)}$. □

Доказательство нижней оценки из (4.8)**.** Пусть $J_n=\left[e,e^{n+1}\right]$, $n\in\mathbb{N}$, а $I_k=\left[e^{k},e^{k+1}\right]$, $k==1,2,\ldots,n$.

Заметим, что согласно (4.9)

$$|\lambda^{(s)}(x)|>\frac{1}{x^{\frac{1}{p}+s}}\frac{A_{s0}}{(\ln x)^{\frac{1}{p}+\alpha}}\quad\text{для всех}\quad x\in J_n.$$

Пользуясь последним неравенством и теоремой 4.2, для каждого I_k получим

$$E_{s-1}\left(f,L_p(I_k)\right)\geqslant\frac{c_{21}(p,s)}{k^{\frac{1}{p}+\alpha}}.$$

Тогда, согласно определению модуля изменения f, для $s=[\alpha]+1$, имеем

$$\varkappa_{sp}(n,f)\geqslant\left\{\sum_{k=1}^{n}\left(E_{s-1}\left(f,L_p(I_k)\right)\right)^{\sigma}\right\}^{1/\sigma}\geqslant c_{22}(p,\alpha)n^{s-\alpha}.$$

Таким образом, учитывая последнее неравенство, из теоремы 4.1 и леммы 4.3 следует, что для наилучших рациональных L_p приближений рассматриваемой функции справедлива оценка снизу

$$R_n\left(f;L_p(\mathbb{R})\right)\geqslant R_n\left(f;L_p(J_n)\right)\geqslant\frac{c_{23}(p,\alpha)}{n^{\alpha}}.$$

Тем самым асимптотика (4.8) доказана. □

Пример 4.2. *Пусть λ – функция из примера* 4.1. *Введем функции λ^{+} и λ^{-} на $\mathbb{R}$. Именно,*

$$\lambda^{+}(x)=\lambda(x+e)+\lambda(-x+e)\quad\textit{при}\quad x\in\mathbb{R}\setminus\{0\},\quad\lambda^{+}(0)=\lambda(e);$$

$$\lambda^{-}(x)=\lambda(x+e)-\lambda(-x+e),\quad x\in\mathbb{R}.$$

Функция λ^{+} – четная и непрерывная на $\mathbb{R}$, а λ^{-} – нечетная и в точке $x=0$ имеет разрыв первого рода. При этом

$$R_n\left(\lambda^{\pm};L_p(\mathbb{R})\right)\asymp n^{-\alpha},\quad n\in\mathbb{N}. \tag{4.10}$$

Доказательство. Для $n\geqslant 2$ положим $m=[n/2]$. Поскольку $\lambda^{\pm}(x)=\lambda(x+e)\pm\lambda(-x+e)$ при $x\in\mathbb{R}\setminus\{0\}$, то, используя верхнюю оценку из (4.8), получаем

$$R_n\left(\lambda^{\pm}\right)_p\leqslant R_m\left(\lambda(e+\cdot)\right)_p+R_m\left(\lambda(e-\cdot)\right)_p=2R_m\left(\lambda\right)_p\leqslant c_{24}m^{-\alpha}\leqslant c_{25}n^{-\alpha}.$$

Для получения нижней оценки заметим, что

$$R_n\left(\lambda^{\pm};L_p(\mathbb{R})\right)\geqslant R_n\left(\lambda^{\pm};L_p([0,+\infty))\right)=R_n\left(\lambda;L_p([e,+\infty))\right).$$

Поскольку при получении оценки снизу для $R_n\left(\lambda;L_p(\mathbb{R})\right)$ (см. (4.8)) мы использовали лишь поведение λ на $[e,+\infty)$, то этим нижняя оценка из (4.10) доказана. □

5. Применение интеграла типа Коши

Через $V = V[a,b]$, $-\infty < a < b < +\infty$, обозначим множество функций $f : [a,b] \to \mathbb{R}$, имеющих ограниченное полное изменение, обозначаемое через $v = v(f,[a,b])$. Считаем, что $f \in V[a,b]$, непрерывна слева в точке b и справа в точке a. Если же $x_0 \in (a,b)$ и является точкой разрыва f, то полагаем $f(x_0) = (f(x_0-0)+f(x_0+0))/2$.

Лемма 5.1. *Пусть* $\psi \in L_\infty(\mathbb{R})$, $\operatorname{supp}\psi \subset [a,b]$, $\psi\Big|_{[a,b]} \in V$, $p \in (1,\infty)$ *и* $\dfrac{1}{\sigma} = 1 + \dfrac{1}{p}$. *Тогда* $\psi \in \mathcal{H}^1_\sigma$ *и*

$$\|\psi\|_{\mathcal{H}^1_\sigma} \leqslant c_1(p)(b-a)^{\frac{1}{p}} \left(\|\psi\|_{[a,b]} + v(\psi,[a,b])\right).$$

Доказательство. Согласно утверждению 2.1 можем считать $\operatorname{supp}\psi \subset [-h,h]$, где $h = (b-a)/2$. Тогда

$$(C\psi)(z) = \frac{1}{\pi i}\int_{-h}^{h} \frac{\psi(t)\,dt}{t-z}, \quad z \in \mathbb{C}\setminus[-h,h].$$

Следовательно,

$$(C\psi)'(z) = \frac{1}{\pi i}\int_{-h}^{h} \frac{\psi(t)\,dt}{(t-z)^2}, \quad z \in \mathbb{C}\setminus[-h,h]. \tag{5.1}$$

Для $x \in \mathbb{R}\setminus[-2h,2h]$ имеем

$$\left|(C\psi)'(x)\right| \leqslant \frac{1}{\pi}\|\psi\|_{[-h,h]}\int_{-h}^{h}\frac{dt}{(t-x)^2} \leqslant \frac{3h}{|x|^2}\|\psi\|_{[-h,h]}. \tag{5.2}$$

Выполнив интегрирование по частям, из (5.1) для $z \in \mathbb{C}\setminus[-h,h]$ получим

$$(C\psi)'(z) = \frac{1}{\pi i}\int_{-h}^{h}\psi(t)d\left(\frac{1}{z-t}\right) = \frac{1}{\pi i}\left[\frac{\psi(h)}{z-h} - \frac{\psi(-h)}{z+h}\right] + \frac{1}{\pi i}\int_{-h}^{h}\frac{d\psi(t)}{t-z}. \tag{5.3}$$

Из (5.1) видно, что функция $(C\psi)'(z)$ аналитична в $\mathbb{C}\setminus[-h,h]$ и бесконечность для нее является нулем не ниже второго порядка. С учетом (5.3) и свойств интеграла Коши–Стилтьеса [12, p. 231] следует, что $(C\psi)'(z)$ принадлежит пространству Харди $H_q\left(\hat{\mathbb{C}}\setminus[-h,h]\right)$ при $1/2 < q < 1$. Следовательно, $(C\psi)'(z)$ принадлежит пространству $H_q(\Pi)$. Под $(C\psi)'(x)$ для $x \in [-h,h]$ будем подразумевать $\lim\limits_{y\to+0}\left((C\psi)'(x+iy)\right)$. Указанный предел существует почти для всех $x \in [-h,h]$.

Для измеримой функции $\omega : [-c,c] \to [0,+\infty)$ через ω^* будем обозначать ее симметричную относительно точки 0, убывающую на $[0,c]$ перестановку.

Из свойств интеграла Коши–Стилтьеса [12] и (5.3) получаем, что

$$\left|(C\psi)'(x)\right|^* \leqslant \frac{c_2}{|x|}\left(\|\psi\|_{[-h,h]} + v(\psi,[-h,h])\right), \quad |x| < 2h. \tag{5.4}$$

Из (5.2) и (5.4) находим, что

$$\int_{\mathbb{R}}\left|(C\psi)'(x)\right|^\sigma dx = \int_{\mathbb{R}\setminus[-2h,2h]}\left|(C\psi)'(x)\right|^\sigma dx + 2\int_0^{2h}\left(\left|(C\psi)'(x)\right|^*\right)^\sigma dx \leqslant$$

$$\leqslant c_3(p)(2h)^{\frac{\sigma}{p}}\left(\|\psi\|_{[-h,h]} + v(\psi,[-h,h])\right)^\sigma.$$

Этим лемма 5.1 доказана. □

Следствие 5.2. *Из леммы* 5.1 *и теоремы* 3.1 *следует, что для* $f \in V[a,b]$ *справедливо неравенство*

$$R_n(f;L_p[a,b]) \leqslant c_4\frac{(b-a)^{\frac{1}{p}}}{n}v(f,[a,b]).$$

Ранее этот результат был получен другим методом [13; 14].

В следующей теореме изучаются наилучшие рациональные приближения функции ограниченной вариации на полуоси. Поскольку любая функция ограниченной вариации представима в виде разности двух убывающих функций, то мы ограничимся рассмотрением убывающих функций.

Теорема 5.3. *Пусть функция $f = 0$ при $x < e$ убывает на $[e, +\infty)$ и*

$$0 \leqslant f(x) \leqslant \lambda(x), \quad \text{при} \quad x \geqslant e,$$

где λ – функция из примера 4.1.

Тогда при $n \geqslant 2$

$$R_n(f)_p \leqslant \begin{cases} \dfrac{c_5}{n^\alpha}, & 0 < \alpha < 1; \\ c_6 \dfrac{(\ln n)^{1+\frac{1}{p}}}{n}, & \alpha = 1; \\ \dfrac{c_7}{n}, & \alpha > 1. \end{cases} \tag{5.5}$$

Доказательство. Пусть $I_k = \left[e^k, e^{k+1}\right]$, $k \in \mathbb{N}$. Положим

$$f_k(x) = \begin{cases} f(x), & x \in I_k, \\ 0, & x \in \mathbb{R} \setminus I_k. \end{cases}$$

Функции f_k удовлетворяют условиям леммы 5.1. Следовательно,

$$\|f_k\|_{\mathcal{H}^1_\sigma} \leqslant c_1 |I_k|^{1/p} \left(\|f_k\|_{I_k} + v\left(f_k, I_k\right)\right) \leqslant c_8 e^{\frac{k}{p}} \lambda(e^k) = \frac{c_9}{k^{\frac{1}{p}+\alpha}}.$$

Далее рассмотрим функцию

$$F_n = f_1 + f_2 + \ldots + f_n.$$

Применяя σ-неравенство треугольника, имеем

$$\|F_n\|^\sigma_{\mathcal{H}^1_\sigma} \leqslant \sum_{k=1}^n \|f_k\|^\sigma_{\mathcal{H}^1_\sigma} \leqslant c_{10} \sum_{k=1}^n \frac{1}{k^{\sigma\left(\frac{1}{p}+\alpha\right)}} \leqslant \begin{cases} \dfrac{c_{11}}{n^{\sigma(\alpha-1)}}, & \alpha < 1; \\ c_{12} \ln(n+1), & \alpha = 1; \\ c_{13}, & \alpha > 1. \end{cases}$$

Согласно теореме 3.1, с учетом последнего неравенства имеем

$$R_n\left(F_n\right)_p \leqslant \frac{c_{14}}{n} \|F_n\|_{\mathcal{H}^1_\sigma} \leqslant \begin{cases} \dfrac{c_{15}}{n^\alpha}, & \alpha < 1; \\ c_{16} \dfrac{(\ln(n+1))^{\frac{1}{\sigma}}}{n}, & \alpha = 1 \\ \dfrac{c_{17}}{n}, & \alpha > 1. \end{cases} \tag{5.6}$$

Из определения функций F_n и f следует, что

$$\|f - F_n\|_{L_p(\mathbb{R})} = \|f\|_{L_p([e^{n+1}, +\infty))} \leqslant \left(\int\limits_{e^{n+1}}^{+\infty} \frac{dx}{x(\ln x)^{1+p\alpha}}\right)^{\frac{1}{p}} = \frac{1}{(p\alpha)^{1/p}(n+1)^\alpha}. \tag{5.7}$$

Заметим далее, что

$$R_n\left(f\right)_p \leqslant R_n\left(f - F_n\right)_p + R_n\left(F_n\right)_p \leqslant \|f - F_n\|_{L_p(\mathbb{R})} + R_n\left(F_n\right)_p.$$

Отсюда и оценок (5.6) и (5.7) следует утверждение теоремы. $\square$

Отметим, что верхняя строка в (5.5) является точной в силу оценки (5.7), а точность нижней строки в (5.5) следует из известных результатов для наилучших рациональных приближений функций ограниченной вариации на отрезке [13; 14].

Работа выполнена при финансовой поддержке ГПНИ «Конвергенция–2025».

Литература

1. *Гарнетт Дж.* Ограниченные аналитические функции. М.: Мир, 1984.

2. *Coifman R. R., Weiss G.* Extensions of Hardy spaces and their use in analysis // Bulletin of the American Mathematical Society. 1977. Vol. 83, N 4. P. 569–645.

3. *Grafakos L.* Modern Fourier analysis // Graduate Texts in Mathematics. 2nd ed. New York: Springer, 2009.

4. *Кротов В. Г.* Дифференциальные свойства граничных функций из пространств Харди // Mathematische Nachrichten. 1986. Т. 126, № 1. С. 241–253.

5. *Стельмах Е. И.* Прямая и обратная теоремы рациональной аппроксимации для пространств Харди в полуплоскости // Докл. НАН Беларуси. 2008. Т. 52, № 6. С. 36–41.

6. *Мардвилко Т. С., Пекарский А. А.* Применение действительного пространства Харди–Соболева на прямой для исследования скорости равномерных рациональных приближений функций // Журнал Белорусского государственного университета. Математика. Информатика. 2022. № 3. С. 16–36.

7. *Пекарский А. А.* Оценки производных рациональных функций в $L_p[-1,1]$ // Математические заметки. 1986. Т. 39, № 3. С. 388–394.

8. *Lorentz G. G., Golitschek M. V., Makovoz Y.* Constructive Approximation. Advanced Problems. New York, Berlin, Heidelberg: Springer-Verlag, 1996.

9. *Phillips G. M.* Error estimates for best polynomial approximation // Approximation theory. London: Academic Press, 1979. P. 1–6.

10. *DeVore R. A., Lorentz G. G.* Constructive Approximation: polinomials and splines approximation. New York, Berlin, Heidelberg: Springer-Verlag, 1993.

11. *Стечкин С. Б.* О порядке наилучших приближений непрерывных функций // Изв. АН СССР. Сер. матем. 1951. Т. 15, № 3. С. 219–242.

12. *King F. W.* Hilbert transform. Volume 1. Cambridge: University Press, 2009.

13. *Petrushev P. P., Popov V. A.* Rational approximation of real functions. Cambridge: University Press, 1987.

14. *Пекарский А. А.* Рациональная аппроксимация сингулярных функций // Изв. АН БССР, Сер. физ.-мат. наук. 1980. № 3. С. 32–40.

References

1. Garnett J. B. *Bounded analytic function. Revised first edition.* New York: Springer, 2007, XIV.

2. Coifman R. R., Weiss G. Extensions of Hardy spaces and theiruse in analysis. *Bulletin of the American Mathematical Society,* 1977, vol. 83, no. 4, pp. 569–645.

3. Grafakos L. *Modern Fourier analysis. Graduate Texts in Mathematics.* 2nd ed. New York, Springer, 2009.

4. Krotov V. G. Differential properties of boundary functions of Hardy spaces. *Mathematische Nachrichten*, 1986, vol. 126, no. 1, pp. 241–253 (in Russian).

5. Stelmakh A. I. Direct and inverse theorems of rational approximation for Hardy spaces in the half-plane. *Doklady of the National Academy of Sciences of Belarus*, 2008, vol. 52, no. 6, pp. 36–41 (in Russian).

6. Mardvilko T. S., Pekarskii A. A. Application of the real Hardy–Sobolev space on the line to study the order of uniform rational approximations of functions. *Journal of the Belarusian State University. Mathematics and Informatics*, 2022, no. 3, pp. 16–36 (in Russian).

7. Pekarskii A. A. Estimates of the derivatives of rational functions in $L_p[-1,1]$. *Math. Notes*, 1986, vol. 39, no. 3, pp. 388–394 (in Russian).

8. Lorentz G. G., Golitschek M. V., Makovoz Y. *Constructive Approximation. Advanced Problems.* New York, Berlin, Heidelberg, Springer-Verlag, 1996.

9. Phillips G. M. Error estimates for best polynomial approximation. *Approximation theory*. London, Academic Press, 1979, pp. 1–6.

10. DeVore R. A., Lorentz G. G. *Constructive Approximation: polinomials and splines approximation.* New York, Berlin, Heidelberg, Springer-Verlag, 1993.

11. Stechkin S. B. On the order of the best approximations of continuous functions. *Izv. Akad. Nauk SSSR Ser. Mat.*, 1951, vol. 15, no. 3, pp. 219–242 (in Russian).

12. King F .W. *Hilbert transform.* Volume 1. Cambridge, University Press, 2009.

13. Petrushev P. P., Popov V. A. *Rational approximation of real functions.* Cambridge, University Press, 1987.

14. Pekarskii A. A. Rational approximation of singular functions. *Izv. NAS BSSR. Ser. fiz.-mat. navuk*, 1980, no. 3, pp. 32–40 (in Russian).

Национальная академия наук Беларуси
Труды Института математики НАН Беларуси. 2024. Том 32. № 2. С. 43–55

ВЫЧИСЛИТЕЛЬНАЯ МАТЕМАТИКА
COMPUTATIONAL MATHEMATICS

УДК 519.63

РАЗРУШЕНИЕ РЕШЕНИЯ РАЗНОСТНЫХ СХЕМ, АППРОКСИМИРУЮЩИХ ЗАДАЧИ НЕЙМАНА ДЛЯ НЕЛИНЕЙНЫХ ПАРАБОЛИЧЕСКИХ УРАВНЕНИЙ

Д. А. Щадинский

Институт математики НАН Беларуси, Минск, Беларусь
e-mail: schadinskii@gmail.com

Поступила: 23.10.2024 **Исправлена:** 02.12.2024 **Принята:** 12.12.2024

Ключевые слова: теоремы сравнения, разрушение решения, дискретные аналоги теорем сравнения, задача Неймана, нелинейные параболические уравнения.

Аннотация. На основе дискретного аналога теорем сравнения и неравенства Йенсена в статье получены условия разрушения решения и верхние оценки времени разрушения решения неявных разностных схем, которые аппроксимируют задачи Неймана для различных нелинейных параболических уравнений. Приводятся условия разрушения решения и верхняя оценка времени разрушения решения для аппроксимируемых дифференциальных задач, полученные на основе теорем сравнения и неравенства Йенсена.

BLOW-UP IN DIFFERENCE SCHEMES THAT APPROXIMATE NEUMANN PROBLEMS FOR NONLINEAR PARABOLIC EQUATIONS

D. A. Schadinskii

Institute of Mathematics of the National Academy of Sciences of Belarus, Minsk, Belarus
e-mail: schadinskii@gmail.com

Received: 23.10.2024 **Revised:** 02.12.2024 **Accepted:** 12.12.2024

Keywords: comparison theorems, blow-up, discrete analogue of comparison theorems, nonlinear parabolic equations.

Abstract. In this paper, based on the discrete analogue of comparison theorems and Jensen's inequality, blow-up conditions and upper bound of blow-up time of the solution of implicit finite-different problem which approximates Neumann problems for various nonlinear parabolic equations are obtained. Blow-up conditions and upper bound of blow-up time of approximated differential problems are given, which are obtained and based on comparison theorems and Jensen's inequality.

1. Введение

Теория разрушения решения имеет прикладное значение во многих областях науки. Примером может служить моделирование процессов, таких как безударное сжатие газа, самофокусировка световых пучков в нелинейных средах, термоядерное горение плазмы и других процессов в физике, химии и биологии. В области дифференциальных уравнений теории разрушения решения посвящено большое количество работ. Стоит отметить фундаментальные работы H. Fujita [1] и H. Levine [2]. Однако теория разрушения решения в разностных схемах является мало исследованной областью. Из существующих работ в области разностных схем можно выделить ключевые работы T. Nakagawa [3], Y. Chen [4].

В данной статье приводится новый подход к изучению разрушения решения разностных схем. Он заключается в повторении шагов доказательства разрушения решения дифференциальной задачи, но с использованием дискретных аналогов дифференциальных законов. В пункте 2, для задачи Неймана для параболического уравнения было приведено доказательство разрушения решения с использованием теоремы сравнения Чаплыгина и неравенство Йенсена. В пункте 3 для доказательства разрушения решения разностной схемы, аппроксимирующей задачу из пункта 2, была использована та же техника доказательства, но с применением дискретного аналога теоремы сравнения и неравенства Йенсена для дискретного случая. Стоит отметить, что при таком подходе к доказательствам условия разрушения одинаковы для дифференциального и разностного случаев. С аналогичным подходом в пунктах 4 и 5 были получены условия на разрушения решения задачи Неймана для уравнения с градиентной нелинейностью и для схемы ее аппроксимирующей.

2. Эффект разрушения решения в дифференциальных и разностных задачах

Для изучения разрушения решения рассмотрим следующую задачу:

$$\frac{du}{dt} = g(t)f(u), \quad 0 \leqslant t < \infty, \quad u(0) = u_0, \tag{2.1}$$

где $g(t)$ – непрерывная положительная неубывающая функция для всех $t \in [0,\infty)$; $f(v)$ – непрерывно дифференцируемая положительная возрастающая функция при всех $v \in [u_0,\infty)$.

Определение 2.1. Будем говорить, что решение задачи (2.1) разрушается, если существует такое t_b, для которого

$$\lim_{t\to t_b-0} u(t) = \infty.$$

В этом случае t_b называется временем разрушения решения.

Лемма 2.2. *Для задачи* (2.1) *верны следующие равенства:*

$$\Psi(u(t)) = \Phi(t), \quad t \in [0,\infty), \tag{2.2}$$

где

$$\Psi(v) = \int_{u_0}^{v} \frac{dw}{f(w)}, \quad \Phi(t) = \int_0^t g(\zeta)d\zeta. \tag{2.3}$$

Доказательство. Для доказательства запишем задачу (2.2) в следующем виде:

$$\frac{du(t)}{f(u(t))} = g(t)dt, \quad t \in [0,\infty), \quad u(0) = u_0. \tag{2.4}$$

С учетом равенств (2.4) для $t \in [0,\infty)$ верна следующая цепочка равенств

$$\Psi(u(t)) = \int_{u_0}^{u(t)} \frac{dw}{f(w)} = \int_0^t \frac{du(\zeta)}{f(u(\zeta))} = \int_0^t g(\zeta)dt = \Phi(t).$$

□

Предложение 2.3. *Пусть для задачи* (2.1) *выполнено неравенство*

$$\lim_{v\to\infty} \Psi(v) = \Psi(\infty) < \infty.$$

Тогда решение задачи разрушается за конечное время t_b

$$t_b = \Phi^{-1}\left[\Psi(\infty)\right], \tag{2.5}$$

где $\Psi(v)$, $\Phi(t)$ определяются равенствами (2.3).

Доказательство. Так как $\Phi(t)$ – возрастающая неограниченная функция на $[0,\infty)$ и выполнены условия утверждения $\Psi(\infty) < \infty$, то существует $t_b < \infty$, для которого верно $\Psi(\infty) = \Phi(t_b)$. Тогда в силу строгой монотонности функций $\Psi(v)$, $\Phi(t)$ можно записать соотношение (2.5).

Теперь покажем что решение задачи (2.1) разрушается в точке t_b. Переходя к приделу $t \to \to t_b$ в равенстве (2.2), получим следующие соотношения $\Psi(\lim_{t\to t_b} u(t)) = \Phi(t_b)$, откуда следует, что $\Psi(\lim_{t\to t_b} u(t)) = \Psi(\infty)$, а значит $\lim_{t\to t_b} u(t) = \infty$. □

Введем общее определение разрушения решения. Рассмотрим нестационарную дифференциальную задачу $\mathscr{A}$ в $\Omega \times [t_0, \infty)$, где $\Omega \subset \mathbb{R}^m$, $m \in \mathbb{N}$. Обозначим через $u(x,t)$ решение дифференциальной задачи $\mathscr{A}$. Пусть $||\cdot||_\Omega$ – некоторая норма в области Ω.

Определение 2.4. Решение $u(x,t)$ дифференциальной задачи $\mathscr{A}$ разрушается в норме $||\cdot||_\Omega$, если существует T_b ($t_0 < T_b < \infty$), для которого верно следующее соотношение:

$$\lim_{t \to T_b - 0} ||u(t)||_\Omega = \infty.$$

В этом случае T_b называется временем разрушения решения задачи $\mathscr{A}$.

Предложение 2.5. *Предположим, что $w(t)$ – непрерывная функция, для которой верно следующее соотношение:*

$$\lim_{t \to t_b - 0} w(t) = +\infty,$$

и также предположим, что для решения задачи $\mathscr{A}$ выполнено неравенство

$$||u(t)||_\Omega \geqslant w(t), \quad t \geqslant t_0.$$

Тогда решение задачи $\mathscr{A}$ разрушается в норме $||\cdot||_\Omega$, кроме того $T_b \leqslant t_b$.

В дальнейшем важную роль в изучении задач с разрушением решения будет играть следующее следствие из теоремы сравнения [5, с. 26].

Следствие 2.6. *Пусть $\beta(t)$ верхнее решение задачи* (2.1)*, удовлетворяющее соотношениям*

$$\frac{d\beta}{dt} \geqslant g(t) f(\beta(t)), \quad t \in [0, T_e), \quad \beta(0) = u_0,$$

где $[0, T_e)$ – интервал существования функции $\beta(t)$.

Тогда верно неравенство

$$u(t) \leqslant \beta(t)$$

для всех t, принадлежащих их общему интервалу существования.

Доказательство. Доказательство следует из теоремы 4.1 [5, с. 26] и теоремы Пикара–Линделефа [5, с. 8]. □

Стоит отметить важную роль при исследовании задач с разрушением решения, которая отводится неравенству Йенсена. Неравенство будет широко использоваться в дальнейшем в статье.

Теорема 2.7 (Неравенство Йенсена)**.** *Пусть функция $\varphi(v)$ выпуклая на произвольном интервале χ, тогда для произвольной функции $u(x)$, интегрируемой на χ, следующее неравенство выполнено*

$$\varphi\left(\int_\chi u(x)dx\right) \leqslant \frac{1}{mes(\chi)} \int_\chi \varphi\left(mes(\chi) u(x)\right) dx. \tag{2.6}$$

2.1. Разрушения решения для разностных схем

Так как нельзя говорить о разрушении, основываясь на конечном наборе значений сеточной функции, то для определения разрушения решения понадобится последовательность узлов, на которой будет определено решение разной схемы или сетка со счетным количеством узлов $\omega_\tau = \{t_n | t_{n+1} > t_n, n \in \mathbb{N}\}$.

Предположим, что имеется разностная схема $\mathscr{A}_h$, которая аппроксимирует дифференциальную задачу $\mathscr{A}$ и определена на сетке $\omega = \omega_\tau \times \omega_h$. Обозначим через y_n решение разностной схемы $\mathscr{A}_h$ на сетке ω_h во временном узле t_n. Пусть $||\cdot||_{\omega_h}$ – некоторая сеточная норма по сетке ω_h.

Определение 2.8. Разностная схема $\mathscr{A}_h$, определенная на ω, допускает разрушение решения в норме $||\cdot||_{\omega_h}$, если существует такая сетка ω_τ, для которой последовательность $\{t_n\}_{n=1}^\infty$ стремится к некоторому конечному числу, т. е. $\lim\limits_{n\to\infty} t_n = T_{hb} < \infty$, а решение $\mathscr{A}_h$ стремится к бесконечности в данной норме, т. е.

$$\lim_{n \to \infty} ||y_n||_{\omega_h} = \infty.$$

Замечание 2.9. *В отличие от разрушения решения для дифференциальной задачи в определении* 2.8 *разрушения решения для разностной схемы* $\mathscr{A}_h$ *зависят от выбора сетки* ω_τ. *Всегда можно выбрать сетку* ω_τ *таким образом, чтобы* $\{t_n\}_{n=1}^{\infty}$ *стремилось к некоторому конечному числу, а решение* $\mathscr{A}_h$ *также стремилось к другому конечному числу. Таким образом, если одно решение разностной схемы разрушается, то найдется решение, которое не разрушается. Чтобы отразить эту особенность, будем писать, что разностная схема* $\mathscr{A}_h$ *допускает разрушения решения.*

Предложение 2.10. *Предположим, что* $w(t)$ *– непрерывная функция, которая не ограничена на конечном интервале, т. е.*

$$\lim_{t \to t_b - 0} w(t) = +\infty,$$

и также предположим, что для любого решения задачи $\mathscr{A}_h$ *выполнено неравенство*

$$||y^n||_{\omega_h} \geqslant w(t_n), \quad t_n \in \omega_\tau.$$

Тогда $\mathscr{A}_h$ *допускает разрушение решения в норме* $||\cdot||_{\omega_h}$, *кроме того* $T_{hb} \leqslant t_b$.

Доказательство. Доказательство от противного. Предположим, что для любой сетки ω_τ, для которой последовательность $(t_n)_{n=0}^{\infty}$ стремится к пределу $T_{hb} \leqslant t_b$, решение задачи $\mathscr{A}_h$ в некоторой норме стремится к константе. Построим сетку ω_τ, такую, что выполнено $\lim\limits_{n\to\infty} t_n = t_b$. Тогда для решения $\mathscr{A}_h$ на этой сетке верны следующие соотношения:

$$\infty \neq \lim_{n\to\infty} ||y^n||_{\omega_h} \geqslant \lim_{n\to\infty} w(t_n) = \infty.$$

Противоречие. Следовательно утверждение доказано. □

В дальнейшем важную роль для обнаружения эффекта разрушения решения в разностных схемах будет играть следующий дискретный аналог теорем сравнения [7].

Теорема 2.11. *Пусть на сетке* $\bar{\omega}_\tau = \{\, t_n \mid t_{n+1} = t_n + \tau_n, \tau_n > 0,\ n = 0, 1, \ldots, N_0 - 1, t_0 = 0,\ t_{N_0} = T\}$ *сеточная функция* $\beta_\tau^n \in [u_0, \infty)$, $n = 0, 1, \ldots, N_0$ *удовлетворяет неравенствам*

$$\frac{\beta_\tau^{n+1} - \beta_\tau^n}{\tau_n} \geqslant g(t_{n+1}) f(\beta_\tau^{n+1}), \quad n = 0, \ldots, N_0 - 1, \quad \beta_\tau^0 \geqslant u_0.$$

Тогда

$$\beta_\tau^m \geqslant u(t_n), \quad m = 0, 1, \ldots, N_0.$$

Также стоит отметить важную роль при доказательстве разрушения решения разностных схем дискретного аналога неравенства Йенсена.

Теорема 2.12 (Неравенство Йенсена)**.** *Пусть функция* $\varphi(v)$ *выпуклая на произвольном интервале* U *и* $\{q_i | q_i > 0, i \in \eta\}$ *– множество точек, где* $\eta \subseteq \mathbb{N}$ *и* $\sum\limits_{i\in\eta} q_i = 1$, *тогда для произвольного множества* $\{v_i | v_i \in U, i \in \eta\}$ *выполнено следующее неравенство:*

$$\varphi\left(\sum_{i\in\eta} q_i v_i\right) \leqslant \sum_{i\in\eta} q_i \varphi(v_i).$$

3. Задача Неймана для параболического уравнения

Рассмотрим задачу Неймана для одномерного случая

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(k(u)\frac{\partial u}{\partial x}\right) + f(u), \quad (x,t) \in \Omega \times (0, \infty), \tag{3.1}$$

$$k(u)\frac{\partial u}{\partial x}\Big|_{x=0} = k(u)\frac{\partial u}{\partial x}\Big|_{x=l} = 0, \qquad t \in (0, \infty), \tag{3.2}$$

$$u(x, 0) = u_0(x), \quad x \in \Omega, \tag{3.3}$$

где $\Omega = \{x : 0 < x < l\}$, $l = \text{const} > 0$. Функция k положительная в $\mathbb{R}^+$ и $k \in C^1(\mathbb{R}^+)$, функция f неотрицательная и $f \in C^1(\mathbb{R}^+)$. Функция $u_0 \in C(\bar{\Omega})$ и $u_0(x) > 0, x \in \Omega$.

Теорема 3.1. *Предположим, что следующие условия выполнены:*

1. *Функция $f(v)$ непрерывно-дифференцируемая, монотонно возрастающая и выпуклая при $v \geqslant v_0$, где*

$$v_0 = \frac{1}{l}\int\limits_0^l u_0(x)dx.$$

2. *Существует $T_1 < \infty$, которое удовлетворяет следующему соотношению:*

$$T_1 = \int\limits_{v_0}^{\infty} \frac{dw}{f(w)}.$$

Тогда решение задачи (3.1)–(3.3) *разрушается за конечное время, т. е.*

$$\lim_{t\to T_b-0} \sup_{0\leqslant x\leqslant l} u(x,t) = \infty, \tag{3.4}$$

и, более того, $T_b \leqslant T_1$.

Доказательство. В соответствии с математическим аппаратом, приведенным в книге [6, с. 27], проинтегрируем уравнение (3.1) по $x \in [0,l]$

$$\int\limits_0^l \frac{\partial u}{\partial t}dx = \left(k(u)\frac{\partial u}{\partial x}\right)\Bigg|_0^l + \int\limits_0^l f(u)dx, \tag{3.5}$$

и применим условие (3.2) к (3.5), имеем

$$\frac{\partial}{\partial t}\int\limits_0^l udx = \int\limits_0^l f(u)dx.$$

Учитывая, что функция $f(v)$ выпуклая, используя неравенство Йенсена (2.6) получаем

$$\frac{dv}{dt} \geqslant f(v), \quad v(0) = v_0, \quad v(t) = \frac{1}{l}\int\limits_0^l u(x,t)dx.$$

Тогда из следствия 2.6, следует неравенство

$$v(t) \geqslant w(t), \quad t \in [0,T_2), \tag{3.6}$$

где $w(t)$ – решение задачи (2.1), в которой функция $g = 1$ и $[0,T_2)$ – общий интервал существования функций $v(t)$ и $w(t)$.

Поскольку условие 2 теоремы выполнено, то из предложения 2.3 следует, что $\lim\limits_{t\to T_1-0} w(t) = \infty$. Следовательно, из (3.6) имеем

$$\lim_{t\to T_b-0} v(t) = \infty,$$

более того, $T_b \leqslant T_1$. Тогда

$$\infty = \lim_{t\to T_b-0} v(t) = \frac{1}{l}\lim_{t\to T_b-0}\int\limits_0^l u(x,t)dx \leqslant \lim_{t\to T_b-0}\sup_{0\leqslant x\leqslant l} u(x,t). \tag{3.7}$$

Из (3.7) следует утверждение теоремы (3.4).

□

4. Разностная схема для задачи Неймана для параболического уравнения

В этом пункте рассматривается неявная разностная схема и указываются условия, при которых решение разностной схемы допускает разрушение решения. Введем следующие сетки:

$$\omega_\tau = \{t_n | t_{n+1} = t_n + \tau_n, \quad \tau_n > 0, \quad n = \overline{0, N_0 - 1}, \quad t_0 = 0, \quad t_{N_0} = T\},$$

$$\omega_h = \{x_i = ih, \quad h = l/N, \quad i = \overline{1, N-1}\},$$

$$\bar{\omega}_h = \omega_h \cup \{0\} \cup \{l\}.$$

Рассмотрим теперь разностную схему

$$y_t = (ay_{\bar{x}})_x^{(\sigma)} + f(y^{n+1}), \quad (x,t) \in \omega, \quad \omega = \omega_h \times \omega_\tau, \tag{4.1}$$

$$y(x,0) = u_0(x), \quad x \in \omega_h, \quad a_i^n = 0,5(k(y_{i-1}^n) + k(y_i^n)), \qquad (x,t) \in \omega, \tag{4.2}$$

$$(a_1 y_{x,0})^{(\sigma)} - \frac{h}{2}(y_{t,0} - f(y_0^{n+1})) = (a_N y_{\bar{x},N})^{(\sigma)} + \frac{h}{2}(y_{t,N} - f(y_N^{n+1})) = 0, \tag{4.3}$$

которая аппроксимирует задачу (3.1)–(3.3) с порядком $O(h^2+\tau)$, где $\sigma \in [0,1]$ – значение весового оператора $\mu^{(\sigma)} = \frac{1}{2}\left(\sigma\mu^{n+1} + (1-\sigma)\mu^n\right)$.

Теорема 4.1. *Предположим, что следующие условия выполнены:*

1. *Функция $f(v)$ непрерывно-дифференцируемая, монотонно возрастающая и выпуклая при $v \geqslant v_h^0$, где*

$$v_h^0 = \frac{1}{l}\left(\frac{h}{2}y_0^0 + \sum_{i=1}^{N-1} hy_i^0 + \frac{h}{2}y_N^0\right).$$

2. *Существует $T_1 < \infty$, которое удовлетворяет следующему соотношению:*

$$T_1 = \int\limits_{v_h^0}^{\infty} \frac{dw}{f(w)}. \tag{4.4}$$

Тогда разностная схема (4.1)–(4.3) *допускает разрушение решения, т. е.*

$$\lim_{n\to\infty} \max_{\omega_h} y^n = \infty, \tag{4.5}$$

и кроме того, $\lim\limits_{n\to\infty} t_n = T_{hb} \leqslant T_1$.

Доказательство. Просуммируем (4.1) по ω_h

$$\sum_{i=1}^{N-1} y_t = \sum_{i=1}^{N-1} (ay_{\bar{x}})_x^{(\sigma)} + \sum_{i=1}^{N-1} f(y^{n+1}).$$

После упрощения равенство имеет следующий вид:

$$\sum_{i=1}^{N-1} y_t = \frac{1}{h}\left((a_N y_{\bar{x},N})^{(\sigma)} - (a_1 y_{x,0})^{(\sigma)}\right) + \sum_{i=1}^{N-1} f(y_i^{n+1}).$$

Используя краевые условия (4.3), умножая на h и деля на l, имеем

$$\frac{1}{l}\left(\frac{h}{2}y_{t,0} + \sum_{i=1}^{N-1} hy_t + \frac{h}{2}y_{t,N}\right) = \frac{1}{l}\left(\frac{h}{2}f(y_0^{n+1}) + \sum_{i=1}^{N-1} hf(y_i^{n+1}) + \frac{h}{2}f(y_N^{n+1})\right).$$

Пусть $v_h^n = \frac{1}{l}\left(\frac{h}{2}y_0^n + \sum\limits_{i=1}^{N-1} hy_i^n + \frac{h}{2}y_N^n\right)$ и применяя теорему 2.12 к правой части уравнения, получаем следующую задачу:

$$(v_h)_t \geqslant f(v_h^{n+1}),$$

$$v_h^0 = \frac{1}{l}\left(\frac{h}{2}y_0^0 + \sum_{i=1}^{N-1} hy_i^0 + \frac{h}{2}y_N^0\right).$$

Из теоремы 2.11 следует неравенство

$$v_h^n \geqslant w(t_n), \quad t_n \in \omega_\tau. \tag{4.6}$$

Поскольку условие 2 теоремы выполнено, то $\lim\limits_{t\to T_1} w(t) = \infty$. Следовательно, из предложения 2.10 задача (4.1)–(4.3) допускает разрушение решения, т. е.

$$\lim_{n\to\infty} v_h^n = \infty,$$

более того, $\lim\limits_{n\to\infty} t_n = T_{hb} \leqslant T_1$. Следовательно

$$\infty = \lim_{n\to\infty} v_h^n = \frac{1}{l}\lim_{n\to\infty}\left(\frac{h}{2}y_0^n + \sum_{i=1}^{N-1} hy_i^n + \frac{h}{2}y_N^n\right) \leqslant \lim_{n\to\infty}\max_{0\leqslant i\leqslant N} y_i^n. \tag{4.7}$$

Из (4.5) следует утверждение теоремы (4.7). □

Для вычислительного эксперимента рассматривалась следующая начально-краевая задача:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(u\frac{\partial u}{\partial x}\right) + u^2, \quad (x,t)\in\left(-\frac{L_s}{2},\frac{L_s}{2}\right)\times(0,\infty),$$

$$\left(u\frac{\partial u}{\partial x}\right)\left(-\frac{L_s}{2},t\right) = \left(u\frac{\partial u}{\partial x}\right)\left(\frac{L_s}{2},t\right) = 0, \quad t\in(0,\infty),$$

$$u_0(x) = \frac{4}{3}\cos^2\frac{\pi x}{L_s} + 10^{-20}, \quad x\in\left(-\frac{L_s}{2},\frac{L_s}{2}\right),$$

где $L_s = 2\sqrt{2}\pi$.

Для численного эксперимента использовалась консервативная разностная схема (4.1)–(4.3) при $\sigma = 0{,}5$. Для данной постановки разностной задачи величина $T_1 = \frac{3}{2}$ в соответствии с формулой (4.4). Тогда по теореме 4.1 решение схемы допускает разрушение решения и $T_{hb} < T_1$. Шаг выбирался автоматически из соображений сходимости метода простой итерации и устойчивости вычислительного процесса $\tau_n = \frac{0{,}03}{\max\limits_{\omega_h} y^n}$. Критерием остановки вычислительного процесса было получение приближенного решения $\max\limits_{\omega_h} y^N > 10^{150}$.

Проводились такие же вычислительные эксперименты при $\sigma = 0{,}5$ для аналогичной (4.1)–(4.3) разностной схемы, которая отличается лишь тем, что источник f аппроксимируется на нижнем слое вместо верхнего. Получили подтверждение общеизвестного факта, что в неявных схемах (схемы с опережением) время локализации наступает раньше $T_b \approx 1$, а в явных схемах (схема с запаздыванием) – позже.

Результаты вычислительного эксперимента:

Параметры аппроксимации	h	N	τ_N	y_N	t_N
Неявная	$0,27$	46211	$1{,}01\cdot 10^{-152}$	$1{,}00\cdot 10^{150}$	$1{,}017597$
Явная	$0,27$	45516	$1{,}00\cdot 10^{-152}$	$1{,}01\cdot 10^{150}$	$0{,}976533$

Из результатов вычислительного эксперимента видно, что решение неявной схемы стремится к бесконечности, что соответствует полученным ранее результатам. Приблизительное время разрушения решения, полученное для неявной разностной схемы, меньше T_1, что соответствует полученной оценке (4.6). Так как $v_h^0 = v_0$ в этом вычислительном эксперименте, то время разрушения решения исходной дифференциальной задачи T_b должно быть меньше T_1, что также выполнено.

5. Задача Неймана для уравнения с градиентной нелинейностью

В этом пункте изучается эффект разрушения решения следующей дифференциальной задачи:

$$\frac{\partial g(u)}{\partial t} = \sum_{\alpha=1}^{m}\frac{\partial}{\partial x_\alpha}\left(p\left(\left(\frac{\partial u}{\partial x_\alpha}\right)^2\right)\frac{\partial u}{\partial x_\alpha}\right) + k(t)f(u), \quad (x,t)\in\Omega\times(0,\infty), \tag{5.1}$$

$$(\bar{F},\bar{n}) = 0, \quad (x,t)\in\partial\Omega\times(0,\infty), \tag{5.2}$$

$$\bar{F} = \left(p\left(\left(\frac{\partial u}{\partial x_1} \right)^2 \right) \frac{\partial u}{\partial x_1}, \ldots, p\left(\left(\frac{\partial u}{\partial x_m} \right)^2 \right) \frac{\partial u}{\partial x_m} \right),$$

$$u(x,0) = u_0(x), \quad x \in \Omega, \tag{5.3}$$

где $\bar{\Omega}$ – ограниченное связное множество в $\mathbb{R}^m$ с гладкой границей $\partial\Omega$. Функция $g \in C^1(\mathbb{R}^+)$ и $g'(s) > 0$ для всех $s > 0$, функция p положительная в $\mathbb{R}^+$ и $p \in C^1(\mathbb{R}^+)$, функция k положительная в $\mathbb{R}^+$ и $k \in C(\mathbb{R}^+)$, функция f неотрицательная и $f \in C^1(\mathbb{R}^+)$, $\bar{n}$ – внешняя единичная нормаль к границе области Ω. Функция $u_0 \in C(\bar{\Omega})$ и $u_0(x) > 0, x \in \bar{\Omega}$.

Теорема 5.1. *Предположим, что следующие условия выполнены:*

1. *Функция*

$$\varphi(v) = mes\Omega f\left(g^{-1}\left(\frac{1}{mes\Omega} v \right) \right),$$

$mes\Omega$ – мера области Ω, непрерывно-дифференцируемая, монотонно возрастающая и выпуклая при $v \geqslant v_0$, где

$$v_0 = v(0) = \int_{\Omega} g(u(x,0))dx.$$

2. *Существует $T_1 < \infty$, которое удовлетворяет уравнению*

$$\int_{v_0}^{\infty} \frac{dw}{\varphi(w)} = \int_0^{T_1} k(t)dt.$$

Тогда решение задачи (5.1)–(5.3) *разрушается за конечное время, т. е.*

$$\lim_{t \to T_b - 0} \sup_{x \in \Omega} u(x,t) = \infty,$$

и, более того, $T_b \leqslant T_1$.

Доказательство. Проинтегрируем (5.1) по $x \in \Omega$ и применим формулу Гаусса–Остроградского

$$\int_{\Omega} \frac{\partial g(u)}{\partial t} dx = (\bar{F}, \bar{n})\Big|_{\partial\Omega} + k(t) \int_{\Omega} f(u)dx, \quad t \in (0,T], \tag{5.4}$$

и применим условие (5.2) к (5.4), имеем

$$\frac{\partial}{\partial t} \int_{\Omega} g(u)dx = k(t) \int_{\Omega} f(u)dx, \quad t \in (0,T].$$

Заметим, что можно записать функцию в виде $f(u) = f\left(g^{-1}\left(g(u)\right)\right)$, следовательно имеет место уравнение

$$\frac{\partial}{\partial t} \int_{\Omega} g(u)dx = k(t) \int_{\Omega} f\left(g^{-1}\left(g(u)\right)\right) dx, \quad t \in (0,T]. \tag{5.5}$$

Учитывая, что функция $\varphi(v)$ выпуклая, приведем правую часть уравнения (5.5) к виду, удобному для применения неравенства Йенсена (2.6), и применим его

$$\begin{gathered} mes\Omega \frac{1}{mes\Omega} k(t) \int_{\Omega} f\left(g^{-1}\left(mes\Omega \frac{1}{mes\Omega} g(u) \right) \right) dx = \\ = \frac{1}{mes\Omega} k(t) \int_{\Omega} \varphi\left(mes\Omega g(u)\right) dx \geqslant k(t)\varphi\left(\int_{\Omega} g(u)dx \right). \end{gathered} \tag{5.6}$$

Тогда из (5.5) при учете (5.3) и неравенства (5.6) можно получить следующую задачу:

$$\frac{dv}{dt} \geqslant k(t)\varphi(v), \quad v(0) = v_0, \quad v(t) = \int_{\Omega} g(u)dx. \tag{5.7}$$

Тогда из следствия 2.6, следует неравенство

$$v(t) \geqslant w(t), \quad t \in [0, T_2), \tag{5.8}$$

где $w(t)$ – решение задачи (2.1), в которой функция $g = k$ и $f = \varphi$ из (5.7) и $[0, T_2)$ – общий интервал существования функций $v(t)$ и $w(t)$.

Поскольку условие 2 теоремы выполнено, то из предложения 2.3 следует, что $\lim\limits_{t\to T_1} w(t) = \infty$. Следовательно, из (5.8) имеем

$$\lim_{t\to T_b - 0} v(t) = \infty,$$

более того, $T_b \leqslant T_1$. Тогда в силу того, что функция $g(u)$ непрерывная и возрастающая, получаем следующие равенства:

$$\infty = \lim_{t\to T_b-0} v(t) = \lim_{t\to T_b-0} \int\limits_{\Omega} g(u(x,t))dx \leqslant$$

$$\leqslant mes\Omega \lim_{t\to T_b-0} \sup_{\Omega} g(u(x,t)) = mes\Omega \lim_{t\to T_b-0} g(\sup_{\Omega} u(x,t)).$$

Отсюда следует равенство

$$\lim_{t\to T_b-0} \sup_{\Omega} u(x,t) = \infty.$$

□

6. Разностная схема для задачи Неймана для уравнения с градиентной нелинейностью

В этом пункте рассматривается неявная разностная схема и указывается условие, при котором решение разностной схемы допускает разрушение решения.

Предполается, что Ω является областью с гладкой границей, которая достаточно близка к m-мерному параллелепипеду $[0,l_1]\times ... \times[0,l_m]$, чтобы не влиять существенно на погрешность решения, не менять порядок погрешности аппроксимации разностной схемы и начальной функции в граничных точках сетки при использовании равномерной прямоугольной сетки. В таком случае считаем, что углы между внешней нормалью к поверхности в точках соприкасающихся граней параллелепипеда и положительными направлениями оси равны $\pm\frac{\pi}{4}$ для правой границы и $\pm\frac{3\pi}{4}$ для левой границы.

Примером такой области с гладкой границей может быть следующая двумерная область с крайне малой величиной параметра ε:

$$z_1(x_1) \leqslant x_2(x_1) \leqslant z_2(x_1) \quad x_1 \in [0, l_1], \tag{6.1}$$

где

$$z_1(x_1) = \begin{cases} \varepsilon - \sqrt{\varepsilon^2 - (x_1-\varepsilon)^2}, & x_1 \in [0,\varepsilon]; \\ 0, & x_1 \in (\varepsilon, l_1-\varepsilon); \\ \varepsilon - \sqrt{\varepsilon^2 - (x_1 - l_1+\varepsilon)^2}, & x_1 \in [l_1-\varepsilon, l_1]; \end{cases}$$

$$z_2(x_1) = \begin{cases} l_2 - \varepsilon + \sqrt{\varepsilon^2 - (x_1-\varepsilon)^2}, & x_1 \in [0,\varepsilon]; \\ l_2, & x_1 \in (\varepsilon, l_1-\varepsilon); \\ l_2 - \varepsilon + \sqrt{\varepsilon^2 - (x_1 - l_1+\varepsilon)^2}, & x_1 \in [l_1-\varepsilon, l_1]. \end{cases}$$

Определение 6.1. Многомерную область, построенную по примеру области (6.1), будем называть ε-сглаженной прямоугольной областью $[0,l_1]\times ... \times[0,l_m]$.

На основе множества Ω вводим сетку

$$\bar\omega = \bar\omega_h \times \bar\omega_\tau, \quad \omega = \omega_h\times\omega_\tau, \quad \gamma_h = \bar\omega_h/\omega_h, \quad \bar\omega_\tau = \omega_\tau \cup \{0\},$$

где

$$\bar\omega_\tau = \{t_n | t_{n+1} = t_n + \tau_n, \quad \tau_n > 0, \quad n = \overline{0, N_0 - 1}, \quad t_{N_0} = T\},$$

$$\bar{\omega}_h = \left\{ x_i = \left(x_1^{(i_1)}, \ldots, x_m^{(i_m)} \right) \middle| x_\alpha^{(i_\alpha)} = i_\alpha h_\alpha, \quad i_\alpha = \overline{0, N_\alpha}, \quad \alpha = \overline{1,m} \right\},$$

$$\omega_h = \left\{ x_i = \left(x_1^{(i_1)}, \ldots, x_m^{(i_m)} \right) \middle| x_\alpha^{(i_\alpha)} = i_\alpha h_\alpha, \quad i_\alpha = \overline{1, N_\alpha - 1}, \quad \alpha = \overline{1,m} \right\},$$

$$\gamma_\alpha^0 = \left\{ x_i = \left(x_1^{(i_1)}, \ldots, x_m^{(i_m)} \right) \middle| x \in \bar{\omega}, \quad x_\alpha^{(i_\alpha)} = 0 \right\},$$

$$\gamma_\alpha^N = \left\{ x_i = \left(x_1^{(i_1)}, \ldots, x_m^{(i_m)} \right) \middle| x \in \bar{\omega}, \quad x_\alpha^{(i_\alpha)} = (N_\alpha - 1) h_\alpha \right\},$$

с постоянными шагами по пространству $h_1, h_2, \ldots, h_m$ ($h_\alpha = \frac{l_\alpha}{N_\alpha}$) и адаптивным шагом по времени τ_n. Мера множества $mes\Omega = \prod\limits_{\alpha=1}^{m} l_\alpha$.

Рассмотрим разностную схему, которая аппроксимирует задачу (5.1)–(5.3)

$$g(y)_t = \sum_{\alpha=1}^{m} A_\alpha(y) + k(t_{n+1}) f(y^{n+1}), \quad (x,t) \in \bar{\omega}_h \times \omega_\tau, \tag{6.2}$$

$$A_\alpha(y) = \begin{cases} \frac{2}{h_\alpha} D(y_{x_\alpha}^{n+1}), & x \in \gamma_\alpha^0; \\ D(y_{x_\alpha}^{n+1})_{\bar{x}_\alpha}, & x \in \omega_h; \quad \alpha = \overline{1,m}, \\ -\frac{2}{h_\alpha} D(y_{\bar{x}_\alpha}^{n+1}), & x \in \gamma_\alpha^N; \end{cases} \tag{6.3}$$

$$y^0 = y_0(x), \quad x \in \bar{\omega}_h, \tag{6.4}$$

где

$$D(u) = p\left(u^2\right) u.$$

Теорема 6.2. *Предположим, что следующие условия выполнены:*

1\. *Функция*

$$\varphi(v) = mes\Omega f\left(g^{-1}\left(\frac{1}{mes\Omega} v \right) \right)$$

непрерывно-дифференцируемая, монотонно возрастающая и выпуклая при $v \geqslant v_h^0$, *где*

$$v_h^0 = s \sum_{\omega_h} g(y_0) + s \frac{1}{2} \sum_{\gamma_h} g(y_0), \quad s = \prod_{\alpha=1}^{m} h_\alpha.$$

2\. *Существует* $T_1 < \infty$, *которая удовлетворяет уравнению*

$$\int\limits_{v_h^0}^{\infty} \frac{dw}{\varphi(w)} = \int\limits_0^{T_1} k(t) dt.$$

Тогда разностная схема (6.2)–(6.4) *допускает разрушение решения, т. е.*

$$\lim_{n\to\infty} \max_{\omega_h} y^n = \infty,$$

и кроме того, $\lim\limits_{n\to\infty} t_n = T_{hb} \leqslant T_1$.

Доказательство. Просуммируем (6.2) по ω_h

$$\sum_{\omega_h} g(y)_t = \sum_{\omega_h} \sum_{\alpha=1}^{m} A_\alpha(y) + k(t_{n+1}) \sum_{\omega_h} f(y^{n+1}).$$

Используя (6.3), для случая $x \in \gamma_h$ получаем

$$\sum_{\omega_h} g(y)_t + \frac{1}{2} \sum_{\gamma_h} g(y)_t = k(t_{n+1}) \left(\sum_{\omega_h} f(y^{n+1}) + \frac{1}{2} \sum_{\gamma_h} f(y^{n+1}) \right).$$

Умножив на $s = \prod\limits_{\alpha=1}^{m} h_\alpha$ и введя оператор $\Lambda = \sum\limits_{\omega_h} + \frac{1}{2} \sum\limits_{\gamma_h}$, имеем

$$s\, \Lambda g(y)_t = mes\Omega\, k(t_{n+1})\, \Lambda \frac{s}{mes\Omega} f\left(g^{-1}\left(g(y^{n+1}) \right) \right).$$

Пусть $v_h^n = s\,\Lambda g(y^n)$ и применяя теорему 2.12 к уравнению, получаем следующую задачу:

$$(v_h)_t \geqslant k(t_{n+1})\varphi(v^{n+1}),$$

$$v_h^0 = s\,\Lambda g(y_0) = s\sum_{\omega_h} g(y_0) + s\frac{1}{2}\sum_{\gamma_h} g(y_0).$$

Из теоремы 2.11 следует неравенство

$$v_h^n \geqslant w(t_n), \quad t_n \in \omega_\tau.$$

Поскольку условие 2 теоремы выполнено, то $\lim\limits_{t\to T_1} w(t) = \infty$. Следовательно, из предложения 2.10 задача (6.2)–(6.4) допускает разрушение решения, т. е.

$$\lim_{n\to\infty} v_h^n = \infty,$$

более того, $\lim\limits_{n\to\infty} t_n = T_{hb} \leqslant T_1$. Следовательно, в силу того, что $g(u)$ непрерывная возрастающая, имеем следующее равенство:

$$\infty = \lim_{n\to\infty} v_h^n = \lim_{n\to\infty} s\Lambda g(y^n) \leqslant mes\Omega \lim_{n\to\infty} \max_{\bar\omega_h} g(y^n) = mes\Omega \lim_{n\to\infty} g\left(\max_{\bar\omega_h} y^n\right).$$

Отсюда следует равенство

$$\lim_{n\to\infty} \max_{\bar\omega_h} y^n = \infty.$$

□

6.1. Вычислительный эксперимент

Для вычислительного эксперимента использовалась задача (5.1)–(5.3) при $g(u) = u$, $p(v) = v$, $f(u) = u^5$, $k(t) = 1$,

$$\frac{\partial u}{\partial t} = \sum_{\alpha=1}^{2} \frac{\partial}{\partial x_\alpha}\left(\left(\frac{\partial u}{\partial x_\alpha}\right)^3\right) + f(u), \quad (x,t) \in \Omega\times(0,\infty),$$

$$(\bar F, \bar n) = 0, \quad (x,t) \in \partial\Omega\times(0,\infty), \quad \bar F = \left(\left(\frac{\partial u}{\partial x_1}\right)^3, \left(\frac{\partial u}{\partial x_2}\right)^3\right),$$

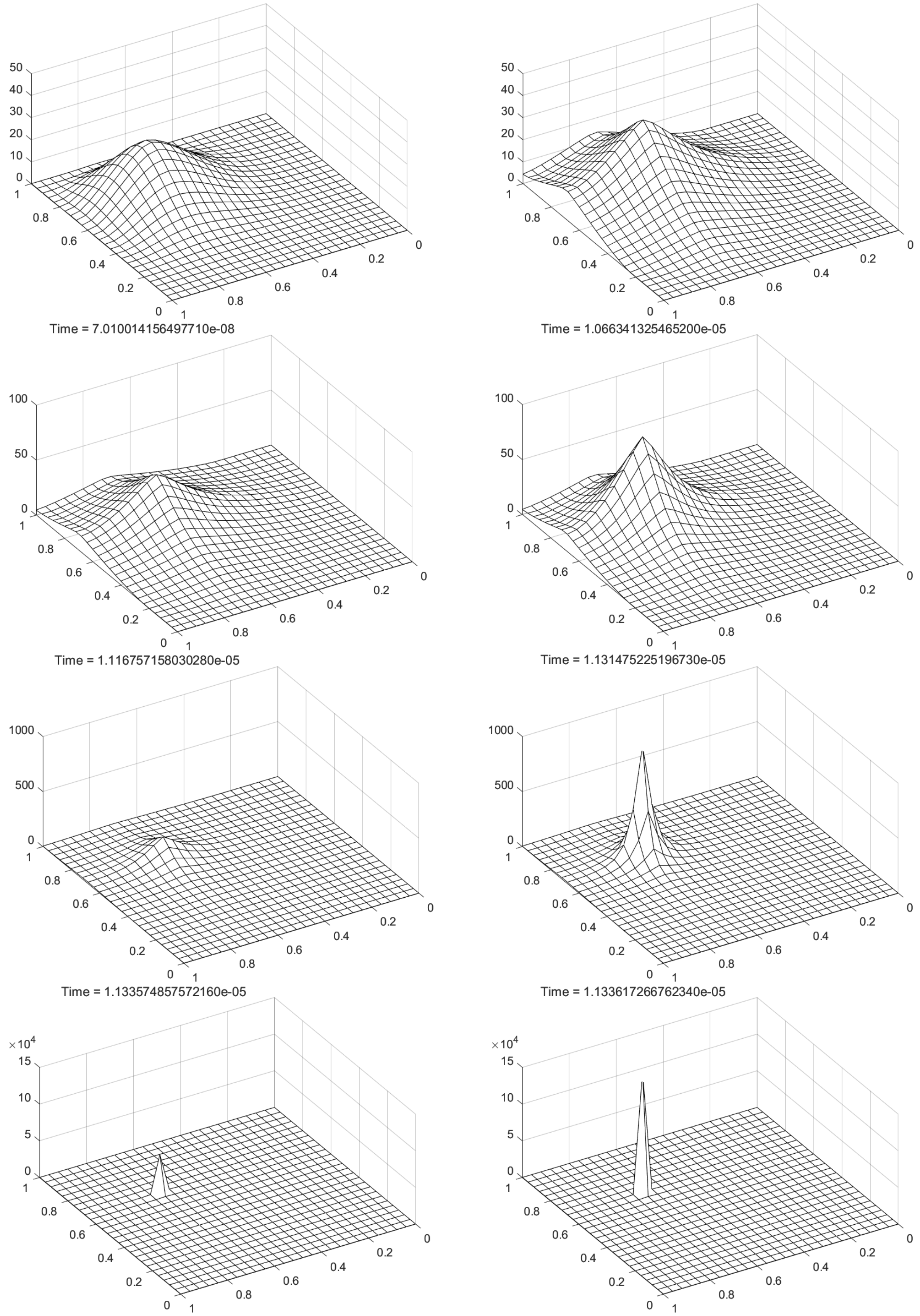
$$u_0(x) = \frac{25}{4}\left(\cos(-\pi + 2\pi x_1^2) + 1\right)\left(\cos(-\pi + 2\pi x_2^2) + 1\right) + 10^{-20}, \quad x \in \Omega,$$

где $\bar\Omega$ является ε-сглаженной прямоугольной областью $[0,1]\times[0,1]$ со значением $\varepsilon = 10^{-20}$.

Для аппроксимации использовалась разностная схема (6.2)–(6.4). Численно было найдено время разрушения решения $T_b \approx 1,13361735232337\cdot 10^{-5}$. Из теоремы 6.2 получим время $T_1 = = \int\limits_{v_0}^{\infty} \frac{dw}{\varphi(w)} = 1,53755\cdot 10^{-3}$, которое оценивает время разрушения T_b сверху. Графики решения при некоторых фиксированных t приведены на рисунке. Заметим, что разница во времени между слоями на последних двух графиках меньше, чем 10^{-20}.

Таким образом, решение задачи разрушается, что соответствует теоремам 5.1, 6.2 и время $T_1 = 1,53755\cdot 10^{-3}$ оценивает сверху время разрушения $T_b \approx 1,13361735232337\cdot 10^{-5}$.

Данный эксперимент проводился на сетке размером 101 на 101 узел. Условием окончания вычислительного эксперимента было достижения нормы (max-норма) по пространству решения значения, большего 10^9. Программа использовала параллельные алгоритмы для решения задачи на восьмиядерном Ryzen 7 5700x. Программа решила свою задачу за 1 час 16 минут. Были вычислены решения на более чем 900000 временных слоев. Время выполнения программы может быть значительно уменьшено за счет выбора более совершенного алгоритма численного поиска минимума функции.



Численное решение на разных временных слоях

7. Заключение

В статье был продемонстрирован новый подход к получению условий разрушения решения в разностных схемах, который может быть альтернативой для представленных в научных трудах методов изучения разрушения решения разностных схем. Основным элементом доказательства является дискретный аналог теорем сравнения Чаплыгина, использование которого позволяет доказывать разрушение решения в разностных схемах подходами, более характерными для доказательства разрушения решения дифференциальных задач. Это было продемонстрировано в статье на примере двух задач Неймана для нелинейных параболических уравнений. Как результат, были получены условия разрушения решения и верхние оценки времени разрушения решения рассмотренных разностных схем, которые согласуются с соответствующими условиями и оценками аппроксимируемых дифференциальных задач.

Работа поддержана Институтом математики НАН Беларуси в рамках государственной программы «Конвергенция–2025» (проект 2021025).

Литература

1. *Fujita H.* On the blowing up of solutions of the Cauchy problem for $u_t = \Delta u + u^{1+\alpha}$ // J. Fac. Sci., Univ. Tokyo, Sect. I. 1966. Vol. 13. P. 109–124.

2. *Levine H. A.* Some nonexistence and instability theorems for solutions of formally parabolic equations of the form $Pu_t = -Au + \mathfrak{F}(u)$ // Arch. Ration. Mech. Anal. 1973. Vol. 51. P. 371–386.

3. *Nakagawa T.* Blowing up of a finite difference solution to $u_t = u_{xx} + u^2$ // Appl. Math. Optim. 1976. Vol. 2. P. 337–350.

4. *Chen Y.* Asymptotic behaviours of blowing-up solutions for finite difference analogue of $u_t = u_{xx} + u^{1+\alpha}$ // J. Fac. Sci., Univ. Tokyo, Sect. I A. 1986. Vol. 33. P. 541–574.

5. *Hartman P.* Ordinary differential equations. New York; London; Sydney: John Wiley and Sons, Inc. 1964. 612 p.

6. *Самарский А. А., Галактионов В. А., Курдюмов С. П., Михайлов А. П.* Режимы с обострением в задачах для квазилинейных параболических уравнений. Москва: Наука, 1987. 480 с.

7. *Матус П. П., Парадинска А., Щадинский Д. А.* Дискретные аналоги теорем сравнения и их применение // Докл. Нац. акад. наук Беларуси. 2013. Т. 57, № 4. С. 16–20.

References

1. Fujita H. On the blowing up of solutions of the Cauchy problem for $u_t = \Delta u + u^{1+\alpha}$. *J. Fac. Sci., Univ. Tokyo, Sect. I.*, 1966, vol. 13, pp. 109–124.

2. Levine H. A. Some nonexistence and instability theorems for solutions of formally parabolic equations of the form $Pu_t = -Au + \mathfrak{F}(u)$. *Arch. Ration. Mech. Anal.*, 1973, vol. 51, pp. 371–386.

3. Nakagawa T. Blowing up of a finite difference solution to $u_t = u_{xx} + u^2$. *Appl. Math. Optim.*, 1976, vol. 2, pp. 337–350.

4. Chen Y. Asymptotic behaviours of blowing-up solutions for finite difference analogue of $u_t = u_{xx} + u^{1+\alpha}$. *J. Fac. Sci., Univ. Tokyo, Sect. I A.*, 1986, vol. 33, pp. 541–574.

5. Hartman P. *Ordinary differential equations*. New York, London, Sydney, John Wiley and Sons, Inc. 1964, 612 p.

6. Samarskii A. A., Galaktionov V. A., Kurdyumov S. P., Mikhailov A. P. *Blow-Up in Quasilinear Parabolic Equations*. Berlin, New York, De Gruyter, 1995, 554 p.

7. Matus P. P., Paradzinska A., Schadinsky D. A. Discrete analogs of the theorems comparison and their applications. *Doklady of the National Academy of Sciences of Belarus*, 2013, vol. 57, no. 4, pp. 16–20. (in Russian).

Национальная академия наук Беларуси
Труды Института математики НАН Беларуси. 2024. Том 32. № 2. С. 56–68

ДИСКРЕТНАЯ МАТЕМАТИКА И МАТЕМАТИЧЕСКАЯ КИБЕРНЕТИКА
DISCRETE MATHEMATICS AND MATHEMATICAL CYBERNETICS

UDC 519.85

NON-EXPOSED FACES OF THE CONE OF COMPLETELY POSITIVE MATRICES

O. I. Kostyukova

Institute of Mathematics of the National Academy of Sciences of Belarus, Minsk, Belarus
e-mail: kostyukova@im-bas-net.by

Received: 15.11.2024 **Revised:** 26.11.2024 **Accepted:** 12.12.2024

Keywords: conic optimization, completely positive matrices, K-semidefinite matrices, a face of a cone, exposed and non-exposed faces of a cone.

Abstract. In this paper, we consider the cone of completely positive matrices. Currently, some families of non-exposed polyhedral faces of this cone were constructed. Inspired by these results, in this paper, we continue the study of the existence and properties of non-exposed faces of the cone of completely positive matrices. We prove a criterion for a face of this cone to be non-exposed. We also provide sufficient conditions that can be easily checked numerically. We show that for any $p \geqslant 6$, there exist non-exposed non-polyhedral faces of the cone of $p \times p$ completely positive matrices. Illustrative examples are given.

НЕВЫСТУПАЮЩИЕ ФАСАДЫ КОНУСА ПОЛНОСТЬЮ ПОЛОЖИТЕЛЬНЫХ МАТРИЦ

О. И. Костюкова

Институт математики НАН Беларуси, Минск, Беларусь
e-mail: kostyukova@im-bas-net.by

Поступила: 15.11.2024 **Исправлена:** 26.11.2024 **Принята:** 12.12.2024

Ключевые слова: коническая оптимизация, полностью положительные матрицы, K-полуопределенные матрицы, фасад конуса, выступающие и невыступающие фасады конуса.

Аннотация. В данной работе мы рассматриваем конус полностью положительных матриц. К настоящему время в литературе были построены некоторые семейства невыступающих полиэдральных фасадов этого конуса. Мотивированные этими результатами, в данной работе мы продолжаем изучение свойств невыступающих фасадов конуса полностью положительных матриц. Доказаны условия, выполнение которых необходимо и достаточно для того, чтобы фасад этого конуса был невыступающим. Также получены достаточные условия, которые можно легко проверить численно. Показано, что для любого $p \geqslant 6$ существуют невыступающие неполиэдральные фасады конуса $p \times p$ полностью положительных матриц. Приведены иллюстративные примеры.

1. Introduction

The optimization of linear functions in feasible sets determined by the intersection of an affine space and a convex cone is known as linear conic optimization. Conic optimization has significant applications and is a pertinent subfield of convex optimization because the conical structure of the feasible set allows one to model many nonlinearities that arise in practice. The number of such applications is increasing, and software is developing quickly (see e.g. [1; 2], and the references therein).

Since the feasible set of a linear conic program is given by the intersection of an affine space and a convex cone, the cone under consideration plays an important role and imposes the main difficulties because the objective function and all the other constraints are linear. The relevant conic problems can be solved more efficiently the better we understand a cone's structure or local description. Facial exposedness is fundamental in understanding the boundary structure of convex cones.

The copositive cone and its dual cone, namely the completely positive cone, have many applications and are actively studied for many years (see [3–6]). However, due to their complicated structures, knowledge about the geometric aspects (especially facial structures) of the copositive and completely positive cones is very limited.

In 2015, Berman et al. in [7] described the main open problems which are currently of interest in the theory of copositive and completely positive matrices. One of the stated open problems was: Is the $p \times p$ completely positive cone facially exposed?

In papers [8; 9], the author answered this question. He gave two concrete parametric classes of non-exposed polyhedral faces of the 5×5 completely positive cone constructing on the base of the Horn matrix and a semidefinite matrix in the form $T(\theta) = a(\theta)\, a'(\theta) + b(\theta)\, b'(\theta)$ with some special vectors $a(\theta),\, b(\theta) \in \mathbb{R}^5$ depending on a parameter $\theta \in \mathbb{R}^5$. At the end of the paper [8], the author asked "Whether there is another type of non-exposed faces in the $p \times p$ completely positive cones, or specifically in the 5×5 completely positive cones, will be an interesting research topic".

In paper [10], for any integer $p \geqslant 5$, non-exposed polyhedral faces of the cone of completely positive $p \times p$ matrices were studied. Criterion and several more explicit (easily checked) sufficient conditions for a polyhedral sub-cone of completely positive matrices to be a non-exposed face were proved. Also for any $p \geqslant 5$, the author presented new families of non-exposed faces of the cone of completely positive $p \times p$ matrices that are different from ones considered in [8; 9] for $p = 5$.

Note that all non-exposed faces considered in examples in [8–10] are polyhedral. As far as we know, there are no examples or studies in the literature of the existence of non-exposed non-polyhedral faces of the cone of completely positive matrices.

In this paper, we explore non-exposed non-polyhedral faces of the cone of completely positive matrices.

The paper is organized as follows. In Section 2 we provide some basic definitions and properties associated with a convex cone. In Section 3, necessary and sufficient conditions for a face of completely positive cone to be non-exposed are proved. Also necessary and sufficient conditions for the face to be non-polyhedral are given. In Section 4, it is shown that for any $p \geqslant 6$ there exist non-exposed non-polyhedral faces of the cone of $p \times p$ completely positive matrices. Section 5 contains some illustrative examples. The paper ends with some conclusions.

2. Definitions and notation

Given a finite dimensional space $\mathfrak{X}$, let us first remind some general definitions.

A set $C \subset \mathfrak{X}$ is convex if for any $x, y \in C$ and any $\alpha \in [0,1]$ it holds $\alpha x + (1-\alpha)y \in C$. Given a set $\mathcal{B} \subset \mathfrak{X}$, denote by $\mathrm{conv}\mathcal{B}$ its *convex hull*, i. e., the minimal (by inclusion) convex set, containing this set, by $\mathrm{span}(\mathcal{B})$ its span, i. e., the smallest linear subspace that contains $\mathcal{B}$, and by $\mathrm{cone}(\mathcal{B})$ its *conic hull*. A set $K \subset \mathfrak{X}$ is a cone if for any $x \in K$ and any $\alpha > 0$, it holds $\alpha x \in K$.

A nonempty convex subset F of a convex closed set $C \subset \mathfrak{X}$ is called a *face* of C if $\alpha x + (1-\alpha)y \in F$ with $x, y \in C$ and $\alpha \in (0,1)$ imply $x, y \in F$. We say that a face F of C is *proper* if $F \neq C$ and $F \neq 0$. A face F of a closed convex set $C \subset \mathfrak{X}$ is called *exposed* if it can be represented as the intersection of C with a supporting hyperplane, i. e. there exist $y \in \mathfrak{X}$ and $d \in \mathbb{R}$ such that for all $x \in C$ it holds $\langle y, x\rangle \geqslant d$ and $\langle y, x\rangle = d$ iff $x \in F$. Every exposed face should also be a face. Given a face F of a set C, the minimal (by inclusion) exposed face containing F will be called here the *minimal exposed face* for this face.

Given a cone $K \subset \mathfrak{X}$, its dual cone K^* is given by

$$K^* = \{x \in \mathfrak{X} : \langle x, y\rangle \geqslant 0 \ \forall y \in K\}.$$

Given an integer $p > 1$, denote by $\mathbb{R}^p_+$ the set of all p vectors with non-negative components (the non-negative orthant in $\mathbb{R}^p$); let $\mathbb{S}(p)$ and $\mathbb{S}_+(p)$ be the space of real symmetric $p \times p$ matrices and the cone of symmetric positive semidefinite $p \times p$ matrices, respectively.

For a given closed subset $K \subset \mathbb{R}^p$ denote by $\mathcal{COP}(K)$ the cone of symmetric K-semidefinite matrices [11]

$$\mathcal{COP}(K) := \{D \in \mathbb{S}(p) : t^\top Dt \geqslant 0 \; \forall t \in K\}.$$

It is evident that $\mathcal{COP}(K) = \mathcal{COP}(K_{norm})$ where $K_{norm} = \{t \in K : ||t||_1 = 1\}$, $||t||_1 = |t_1| + ... + |t_p|$ for $t = (t_k, k = 1,...,p)'$.

For $K = \mathbb{R}^p_+$, we obtain the cone of copositive matrices [3]

$$\mathcal{COP}(\mathbb{R}^p_+) := \{D \in \mathbb{S}(p) : t^\top Dt \geqslant 0 \; \forall t \in \mathbb{R}^p_+\} = \mathcal{COP}(T(p)),$$

where

$$T(p) := \{t \in \mathbb{R}^p_+ : ||t||_1 = 1\}$$

is the simplex in $\mathbb{R}^p$.

For cone $\mathcal{COP}(T(p))$, the dual cone is the cone of completely positive matrices

$$\mathcal{CP}(T(p)) := (\mathcal{COP}(T(p)))^* = \text{cone}\{tt', \; t \in T(p)\}.$$

Recall that a completely positive matrix $U \in \mathcal{CP}(T(p))$ can be written as $\sum\limits_{i \in I} t(i)t'(i)$ with some $t(i) \in \mathbb{R}^p_+, \; i \in I, \; |I| \leqslant p(p+1)/2$.

For given closed subset $K \subset \mathbb{R}^p$ and $X \in \mathcal{COP}(K)$, denote by $Z_0(X, K_{norm})$ the set of normalized zeroes of X in K:

$$Z_0(X, K_{norm}) := \{t \in K_{norm} : t'Xt = 0\}.$$

For a vector $t = (t_k, k = 1,...p)' \in T(p)$, denote by $\text{supp}(t)$ its support: $\text{supp}(t) = \{k \in \{1,...,p\} : t_k > 0\}$. For $X \in \mathcal{COP}(T(p))$, a normalized zero $\tau \in Z_0(X, T(p))$ is called a minimal zero of X if there does not exist zero $\bar{\tau} \in Z_0(X, T(p))$ such that the inclusion $\text{supp}(\bar{\tau}) \subset \text{supp}(\tau)$ holds true strictly.

In what follows we will use the following proposition.

Proposition 2.1. *Let M, $N(m)$, $m \in M$, be given finite index sets and $\xi(j) \in \mathbb{R}^p$, $j \in N(m), m \in M$, be given vectors. Set*

$$\mathcal{T} := \bigcup_{m \in M} \mathcal{T}(m) \textit{ where } \mathcal{T}(m) := \text{conv}\{\xi(j), j \in N(m)\}$$

and consider $X \in \mathcal{COP}(\mathcal{T})$. Then the set

$$Z_0(X, \mathcal{T}) := \{t \in \mathcal{T} : t'Xt = 0\}$$

of all normalized zeros of X in $\mathcal{T}$ is empty or a union of a finite number of polytopes.

Proof. It is evident that $Z_0(X, \mathcal{T}) = \bigcup\limits_{m \in M} Z_0(X, \mathcal{T}(m))$. For $m \in M$, denote

$$p_m = |N(m)|, \; W(m) = (\xi(j), j \in N(m)), \; B(m) = W(m)'XW(m).$$

It is easy to see that

$$X \in \mathcal{COP}(\mathcal{T}(m)) \iff B(m) \in \mathcal{COP}(T(p_m)).$$

Let $Z_0(B(m), T(p_m)) := \{t \in T(p_m) : t'B(m)t = 0\}$ be the set of normalized zeros of $B(m)$ in $T(p_m)$ and $\{\tau(m, j), j \in J(m)\}$ be the set of normalized minimal zeros of $B(m)$ in $T(p_m)$. It is known (see Proposition 2.3 and its proof with $A_* = B(m)$ in [12]) that there exists the set $\{J(s,m), s \in S(m)\}$ of subsets of $J(m)$ such that

$$Z_0(B(m), T(p_m)) = \bigcup_{s \in S(m)} \text{conv}\{\tau(m, j), j \in J(s,m)\}.$$

Then it is easy to see that

$$Z_0(X, \mathcal{T}(m)) = \bigcup_{s \in S(m)} \text{conv}\{\bar{\tau}(m, j), j \in J(s,m)\},$$

where $\bar{\tau}(m,j) = W(m)\tau(m,j)$ for all $j \in J(m)$. Consequently,

$$Z_0(X,\mathcal{T}) = \bigcup_{m\in M}\bigcup_{s\in S(m)} Z_{ms}(X,\mathcal{T}),$$

where $Z_{ms}(X,\mathcal{T}) := \operatorname{conv}\{\bar{\tau}(m,j), j \in J(s,m)\}$ for $s \in S(m)$, $m \in M$, are polytopes. □

It follows from Proposition 2.1 that if K_{norm} is a union of a finite number m_* of polytopes, then $Z_0(X,K)$ is empty or a union of a finite number m_0 of polytopes. Note that, as a rule, $m_0 \neq m_*$.

3. Non-exposed faces

In this section, we formulate and prove a criterion for a proper face of $\mathcal{CP}(T(p))$ to be non-exposed without assumption that the face is polyhedral.

Let $\mathcal{F}$ be a proper face of $\mathcal{CP}(T(p))$. Then applying Algorithm 2 from [13], we conclude that there exists a finite integer $k_0 \geqslant 0$ such that $\mathcal{F} = \mathcal{F}_{k_0+1}$, where

$$\mathcal{F}_0 = \mathcal{CP}(T(p)),\ X_k \in \mathcal{F}_k^*,\ X_k \notin \mathcal{F}_k^{\perp},\ \mathcal{F}_{k+1} = \mathcal{F}_k \cap X_k^{\perp},\ k = 0,\dots,k_0.$$

It follows from these relations that

$$\begin{gathered} X_0 \in \mathcal{COP}(T(p)),\ Z_0(0) := Z_0(X_0,T(p)) \neq \varnothing,\ Z_0(0) \neq T(p),\\ X_{k+1} \in \mathcal{COP}(Z_0(k)),\ Z_0(k+1) := Z_0(X_{k+1},Z_0(k)) \neq \varnothing,\ Z_0(k+1) \neq Z_0(k),\\ k = 0,\dots,k_0-1; \end{gathered} \tag{3.1}$$

$$\mathcal{F} = \{U = \sum_{i=1}^{p(p+1)/2} \alpha_i t(i)t'(i) : \alpha_i \geqslant 0,\ t(i) \in Z_0(k_0)\ \forall i = 1,\dots,p(p+1)/2\}. \tag{3.2}$$

On the other hand, let $\{X_k, k = 0,\dots,k_0\}$ be a set of matrices and $\{Z_0(k), k = 0,\dots,k_0\}$ be a set of subsets of $T(p)$ satisfying (3.1), then the set $\mathcal{F}$ defined by the rule (3.2) is a proper face of $\mathcal{CP}(T(p))$. This gives us a way for constructing proper faces of $\mathcal{CP}(T(p))$.

It is evident that if $k_0 = 0$, then the face (3.2) is an exposed face of $\mathcal{CP}(T(p))$.

Corollary 3.1. *Let $\{X_k, k = 0,\dots,k_0\}$ be a set of matrices and $\{Z_0(k), k = 0,\dots,k_0\}$ be a set of subsets of $T(p)$ satisfying* (3.1). *Then for any $k = 0,\dots,k_0$, the set $Z_0(k)$ is non-empty and is a union of a finite number of polytopes. Hence the set $Z_0(k_0)$ in* (3.2) *can be represented in the form*

$$Z_0(k_0) =: \mathcal{Z} = \bigcup_{s\in\mathcal{S}} \mathcal{Z}(s),\ \mathcal{Z}(s) = \operatorname{conv}\{\xi(j), j \in \mathcal{J}(s)\} \tag{3.3}$$

with certain finite index sets $\mathcal{J}(s), s \in \mathcal{S}$, and vectors $\xi(j) \in T(p)$, $j \in \mathcal{J}(s)$, $s \in \mathcal{S}$.

Proof. By construction, $Z_0(k_0) \subset Z_0(k_0-1) \subset \dots \subset Z_0(1) \subset Z_0(0) \subset T(p)$, and for any $k = 0,\dots,k_0$, the set $Z_0(k)$ is non-empty.

Applying Proposition 2.1 with $m = 1$ and $\mathcal{T} = \mathcal{T}(1) = T(p)$, we conclude that the set $Z_0(0) := Z_0(X_0,T(p))$ is a union of a finite number of polytopes. After applying Proposition 2.1 successively with $\mathcal{T} = Z_0(k)$ for $k = 0,\dots,k_0$, we obtain that for any $k = 0,\dots,k_0$, the set $Z_0(k)$ is a union of a finite number of polytopes. Consequently, the set $Z_0(k_0)$ in (3.2) can be represented in the form (3.3) with certain finite index sets $\mathcal{J}(s), s \in \mathcal{S}$, and vectors $\xi(j) \in T(p)$, $j \in \mathcal{J}(s)$, $s \in \mathcal{S}$. □

In what follows, we will use the following lemma.

Lemma 3.2. *Let $t \in \mathbb{R}^p_+ \setminus \{\mathbf{0}\}$, $\mu(i) \in T(p)$, $i \in I$, $|I| < \infty$, be given and*

$$t\,t' = \sum_{i\in I} \alpha_i \mu(i)(\mu(i))' \text{ with some } \alpha_i \geqslant 0\ \forall i \in I. \tag{3.4}$$

Then there exist $\beta > 0$ and $i_* \in I$ such that $t = \beta\mu(i_*)$.

Proof. Denoting $B = (\sqrt{\alpha_i}\mu(i), i \in I)$, we can rewrite the equality from (3.4) in the form $t\,t' = BB'$ whereform we conclude that

$$1 = \operatorname{rank}(t\,t') = \operatorname{rank}(BB').$$

It follows from the equality $\mathrm{rank}(BB') = 1$ that $\mathrm{rank}(B) = 1$ and hence there exist an index $i_* \in I$ and numbers $\beta_i \geqslant 0$, $i \in I$, such that $\sqrt{\alpha_i}\mu(i) = \beta_i\mu(i_*)$ for all $i \in I$. Then equality in (3.4) takes the form

$$tt' = \tilde{\beta}\mu(i_*))\,\mu'(i_*),$$

where $\tilde{\beta} = \sum_{i\in I} \beta_i^2 > 0$. This implies that $t = \beta\mu(i_*)$ with $\beta = \sqrt{\tilde{\beta}} > 0$. □

Proposition 3.3. *Let $k_0 \geqslant 1$ and $\{X_k, k = 0,...,k_0\}$ be a set of matrices satisfying* (3.1)*, then the set $\mathcal{F}$ defined by the rule* (3.2) *is a non-exposed face of $\mathcal{CP}(T(p))$ iff there exists $t^* \in Z_0(X_0,T(p)) \setminus Z_0(k_0)$ such that for any $D \in \mathcal{COP}(T(p))$*

$$\text{the equalities } t'Dt = 0\ \forall t \in Z_0(k_0) \text{ imply the equality } t^{*\prime}Dt^* = 0. \tag{3.5}$$

Proof. Let $\mathcal{F}$ be a face of $\mathcal{CP}(T(p))$ having representation (3.2) and suppose that there exists $t^* \in Z_0(X_0,T(p)) \setminus Z_0(k_0)$ such that for any $D \in \mathcal{COP}(T(p))$, condition (3.5) holds true. Suppose the contrary: the face $\mathcal{F}$ is an exposed face of $\mathcal{CP}(T(p))$. Then there exists a matrix $A \in \mathcal{COP}(T(p))$ such that

$$\mathcal{F} = \mathcal{CP}(T(p)) \cap A^{\perp} =$$
$$= \{U = \sum_{i=1}^{p_*} \bar{\alpha}_i\mu(i)\mu(i)',\ \bar{\alpha}_i \geqslant 0,\ \mu(i) \in Z_0(A,T(p))\ \forall i = 1,...,p_*\}, p_* = p(p+1)/2. \tag{3.6}$$

Let us show that

$$Z_0(A,T(p)) = Z_0(k_0). \tag{3.7}$$

Consider any $t \in Z_0(A,T(p))$. Then it follows from (3.6) that $tt' \in \mathcal{F}$ and it follows from the representation (3.2) that

$$tt' = \sum_{i\in I_*} \alpha_i t(i)t(i)' \text{ with some } \alpha_i > 0,\ t(i) \in Z_0(k_0)\ \forall i \in I_* \subset \{1,...,p_*\}.$$

Then it follows from Lemma 3.2 that $t = \beta t(i_0)$ with some $i_0 \in I_*$ and $\beta > 0$ and taking into account that $t \in Z_0(A,T(p))$ and $t(i_0) \in Z_0(k_0)$, we obtain that $t = t(i_0) \in Z_0(k_0)$. Consequently, the inclusion $Z_0(A,T(p)) \subset Z_0(k_0)$ holds true.

Now consider any $t \in Z_0(k_0)$. This inclusion implies that $tt' \in \mathcal{F}$, and it follows from (3.6) that

$$tt' = \sum_{i\in \bar{I}_*} \bar{\alpha}_i\mu(i)\mu(i)',\ \bar{\alpha}_i > 0,\ \mu(i) \in Z_0(A,T(p))\ \forall i \in \bar{I}_* \subset \{i = 1,...,p_*\}.$$

Then it follows from Lemma 3.2 that $t = \beta\mu(i_0)$ with some $i_0 \in \bar{I}_*$ and $\beta > 0$ and taking into account that $t \in Z_0(k_0)$ and $\mu(i_0) \in Z_0(A,T(p))$, we obtain that $t = \mu(i_0) \in Z_0(A,T(p))$. Consequently, the inclusion $Z_0(k_0) \subset Z_0(A,T(p))$ holds true. Hence the equality (3.7) takes place.

It follows from (3.7) that for $A \in \mathcal{COP}(T(p))$ equalities $t'At = 0\ \forall t \in Z_0(k_0)$ hold true. Taking into account that for any $D \in \mathcal{COP}(T(p))$, condition (3.5) holds true, we obtain that $t^{*\prime}At^* = 0$, and consequently, $t^* \in Z_0(A,T(p))$. This inclusion and (3.7) imply that $t^* \in Z_0(k_0)$. But by construction, $t^* \in Z_0(X_0,T(p)) \setminus Z_0(k_0)$. Hence our assumption that $\mathcal{F}$ is an exposed face is wrong.

Now suppose that a face $\mathcal{F}$ represented in the form (3.2) is a non-exposed face of $\mathcal{CP}(T(p))$. Let $\mathcal{F}_{min} := \mathcal{CP}(T(p)) \cap X_{min}^{\perp}$, $X_{min} \in \mathcal{COP}(T(p))$, be the minimal exposed face of $\mathcal{CP}(T(p))$ containing the face $\mathcal{F}$. Since $\mathcal{F} \subset \mathcal{F}_{min}$ and $\mathcal{F} \neq \mathcal{F}_{min}$, there exists $U^* \in \mathcal{F}_{min}$ such that $U^* \notin \mathcal{F}$.

It follows form the inclusion $U^* \in \mathcal{F}_{min}$ that U^* admits a representation

$$U^* = \sum_{i\in I_*} \alpha_i^* t^*(i)t^*(i)' \text{ with some } \alpha_i^* > 0,\ t^*(i) \in Z_0(X_{min},T(p))\ \forall i \in I_*.$$

Then the condition $U^* \notin \mathcal{F}$ implies that there exists $i_0 \in I_*$ such that

$$t^*(i_0) \in Z_0(X_{min},T(p)),\ t^*(i_0) \notin Z_0(k_0).$$

Let us show that the condition (3.5) holds true with $t^* = t^*(i_0)$. Suppose the contrary: there exists $B \in \mathcal{COP}(T(p))$ such that

$$t'Bt = 0\ \forall t \in Z_0(k_0),\ t^{*\prime}Bt^* > 0. \tag{3.8}$$

Consider a matrix $\bar{X} := X_{min} + B \in \mathcal{COP}(T(p))$. It follows from (3.8) that

$$Z_0(k_0) \subset Z_0(\bar{X}, T(p)) \subset Z_0(X_{min}, T(p)),\ t^* \in Z_0(X_{min}, T(p)),\ t^* \notin Z_0(\bar{X}, T(p)).$$

But these relations contradict the assumption that $\mathcal{F}_{min} = \mathcal{CP}(T(p)) \cap X_{min}^{\perp}$, is the minimal exposed face of $\mathcal{CP}(T(p))$ containing the face $\mathcal{F}$. Thus we have shown that condition (3.5) holds true with $t^* = t^*(i_0)$.

Now let us show that $t^* = t^*(i_0) \in Z_0(X_0, T(p))$. Suppose the contrary: $t^*(i_0) \notin Z_0(X_0, T(p))$. Consider a matrix $\widetilde{X} := X_{min} + X_0 \in \mathcal{COP}(T(p))$. Then, taking into account that by construction, $\mathcal{F} \subset \subset \mathcal{CP}(T(p)) \cap X_{min}^{\perp}$ and $\mathcal{F} \subset \mathcal{CP}(T(p)) \cap X_0^{\perp}$, we obtain

$$\mathcal{F} \subset \mathcal{CP}(T(p)) \cap \widetilde{X}^{\perp} \subset \mathcal{CP}(T(p)) \cap X_{min}^{\perp},\ t^*(i_*) \in Z_0(X_{min}, T(p)),\ t^*(i_*) \notin Z_0(\widetilde{X}, T(p)).$$

But these relations contradict the assumption that $\mathcal{F}_{min} = \mathcal{CP}(T(p)) \cap X_{min}^{\perp}$, is the minimal exposed face of $\mathcal{CP}(T(p))$ containing the face $\mathcal{F}$. Thus we have shown that $t^* = t^*(i_0) \in Z_0(X_0, T(p))$.

Hence, we have proved that if $\mathcal{F}$ is a non-exposed face of $\mathcal{CP}(T(p))$, then there exists $t^* \in \in Z_0(X_0, T(p)) \setminus Z_0(k_0)$ such that for any $D \in \mathcal{COP}(T(p))$ condition (3.5) holds true. □

A cone $C \subset \mathbb{R}^n$ is said to be a polyhedral if there exists a finite set $\{c(i), i \in I\}$ of vectors $c(i) \in \mathbb{R}^n$, $||c(i)||_1 = 1$, $i \in I$, such that $C = \text{cone}\{c(i)c(i)', i \in I\}$.

Proposition 3.4. *For a given proper face $\mathcal{F}$ of $\mathcal{CP}(T(p))$, let $\mathcal{Z} := Z_0(k_0)$ be a subset of $T(p)$ such that $\mathcal{F}$ admits the representation* (3.2). *Then the face $\mathcal{F}$ is polyhedral iff the set $\mathcal{Z}$ consists of a finite number of elements, i. e., $\mathcal{Z} = \{\tau(j),\ j \in J\}$ with certain index set J, $|J| < \infty$, and vectors $\tau(j) \in T(p)$, $j \in J$.*

Proof. Suppose that the set $\mathcal{Z}$ consists of a finite number of elements, i. e., $\mathcal{Z} = \{\tau(j),\ j \in J\}$. Then it is evident that

$$\mathcal{F} = \{U = \sum_{i=1}^{p(p+1)/2} \alpha_i t(i)t(i)' : \alpha_i \geqslant 0,\ t(i) \in \mathcal{Z}\ \forall i = 1, ..., p(p+1)/2\} = \text{cone}\{\tau(j)\tau(j)', j \in J\},$$

and consequently, the face $\mathcal{F}$ is polyhedral.

Now suppose that a face $\mathcal{F}$ of $\mathcal{CP}(T(p))$ having representation (3.2) is polyhedral. Hence there exist a finite set J and vectors $\tau(j) \in T(p)$, $j \in J$, such that

$$\mathcal{F} = \text{cone}\{\tau(j)\tau(j)',\ j \in J\}. \tag{3.9}$$

Hence, for $j \in J$, we have $\tau(j)\tau(j)' \in \mathcal{F}$, and it follows from (3.2) that

$$\tau(j)\tau(j)' = \sum_{i \in I(j)} \alpha_{ij} t(i)t(i)' \text{ with some } \alpha_{ij} > 0,\ t(i) \in \mathcal{Z}\ \forall i \in I(j).$$

Then applying Lemma 3.2 we conclude that there exists $i_j \in I(j)$ and $\beta(j) > 0$ such that $\tau(j) = \beta(j)t(i_j)$. Taking into account that $\tau(j) \in T(p)$ and $t(i_j) \in T(p)$, we obtain that $\tau(j) = t(i_j)$ and consequently $\tau(j) = t(i_j) \in \mathcal{Z}$. Thus, $\{\tau(j),\ j \in J\} \subset \mathcal{Z}$.

Now consider $t \in \mathcal{Z}$. There it follows from (3.2) that $tt' \in \mathcal{F}$, and it follows from (3.9) that

$$tt' = \sum_{j \in J_*} \alpha_j^* \tau(j)\tau(j)' \text{ with some } \alpha_j > 0,\ j \in J_* \subset J.$$

Then applying Lemma 3.2 and taking into account that $t \in T(p)$ and $\tau(j) \in T(p)$ for all $j \in J$, we conclude that there exists $j_0 \in J_*$ such that $t = \tau(j_0)$. This implies that $t \in \{\tau(j), j \in J\}$ for any $t \in \mathcal{Z}$. Hence $\mathcal{Z} \subset \{\tau(j), j \in J\}$. It follows from this inclusion and the inclusion $\{\tau(j), j \in J\} \subset \mathcal{Z}$ proved above that $\mathcal{Z} = \{\tau(j), j \in J\}$. □

Corollary 3.5. *For a given proper face $\mathcal{F}$ of $\mathcal{CP}(T(p))$, let $\mathcal{Z} := Z_0(k_0)$ be a subset of $T(p)$ such that $\mathcal{F}$ admits the representation* (3.2). *Then the face $\mathcal{F}$ is non-polyhedral iff there exist $\tau \in \mathcal{Z}$ and $\mu \in \mathcal{Z}$ such that $\tau \neq \mu$, $\alpha\tau + (1-\alpha)\mu \in \mathcal{Z}$ for all $\alpha \in [0,1]$.*

Proof. Suppose that there exist $\tau \in \mathcal{Z}$ and $\mu \in \mathcal{Z}$ such that $\tau \neq \mu$, $\alpha\tau + (1-\alpha)\mu \in \mathcal{Z}$ for all $\alpha \in \in [0,1]$. Then, it is evident that the set $\mathcal{Z}$ consists of an infinite number of elements, and it follows from Proposition 3.4 that the face $\mathcal{F}$ is non-polyhedral.

Now suppose that the face $\mathcal{F}$ is non-polyhedral. Then it follows from Proposition 3.4 that the set $\mathcal{Z}$ consists of an infinite number of elements. Moreover, due to Proposition 2.1, we know that the set $\mathcal{Z}$ is a union of a finite number of polytopes. Taking into account these facts, one conclude that there exist $\tau \in \mathcal{Z}$ and $\mu \in \mathcal{Z}$ such that $\tau \neq \mu$, $\alpha\tau + (1-\alpha)\mu \in \mathcal{Z}$ for all $\alpha \in [0,1]$. $\square$

3.1. Sufficient condition for a proper face to be non-exposed

In Proposition 3.3, for a given $t^* \in T$, it is not easy to test if for any $D \in \mathcal{COP}(T(p))$, condition (3.5) holds true. In this subsection, we give easily tested sufficient conditions for fulfillment of this condition.

For a given set $\{X_k, k = 0,...,k_0\}$ of matrices satisfying (3.1), consider the set $\mathcal{F}$ defined by the rule (3.2). It was shown above that the set $\mathcal{F}$ is a proper face of $\mathcal{CP}(T(p))$. It was shown above that the set $Z_0(k_0) =: \mathcal{Z}$ is a union of a finite number of polytopes.

Knowing the set of matrices $\{X_k, k = 0,...,k_0\}$ and consistently using Lemma 2.1 and the algorithms from [14], we can find the sets $\mathcal{J}(s)$, $s \in \mathcal{S}$, and vectors $\xi(j) \in T(p)$, $j \in \mathcal{J}(s)$, $s \in \mathcal{S}$, such that the set $Z_0(k_0)$ admits the representation

$$Z_0(k_0) = \bigcup_{s\in\mathcal{S}} \mathcal{Z}(s),\ \mathcal{Z}(s) = \operatorname{conv}\{\xi(j), j \in \mathcal{J}(s)\}. \tag{3.10}$$

Denote $\mathcal{J} := \bigcup_{s\in\mathcal{S}} \mathcal{J}(s)$, $\mathcal{S}(j) = \{s \in \mathcal{S} : j \in \mathcal{J}(s)\}$ for all $j \in \mathcal{J}$,

$$P_*(s) := \bigcup_{j\in\mathcal{J}(s)} \operatorname{supp}(\xi(j))\ \forall s \in \mathcal{S};\ M_*(j) := \bigcup_{s\in\mathcal{S}(j)} P_*(s)\ \forall j \in \mathcal{J}.$$

Proposition 3.6. *Consider a set of matrices $\{X_k, k = 0,...,k_0\}$ and the corresponding set $\{Z_0(k), k = 0,...,k_0\}$ of subsets of $T(p)$ satisfying (3.1). Suppose that the representation (3.10) of the set $Z_0(k_0)$ is known and let $M_*(j)$, $\xi(j)$, $j \in J$, be the corresponding sets and vectors defined above.*

Suppose that there exists $t^* \in Z_0(0) \setminus Z_0(k_0)$ such that for any $D \in \mathbb{S}(p)$, the qualities

$$e_k' D\xi(j) = 0\ \forall k \in M_*(j),\ \forall j \in \mathcal{J}, \tag{3.11}$$

imply the equalities

$$e_k' Dt^* = 0\ \forall k \in \operatorname{supp}(t^*). \tag{3.12}$$

Then the set $\mathcal{F}$ defined in (3.2) is a non-exposed face of $\mathcal{CP}(T(p))$. Here $\{e_k, k = 1,...,p\}$ is the standard basis in $\mathbb{R}^p$.

Proof. Consider the sets of matrices

$$\mathcal{A} := \{D \in \mathcal{COP}(T(p)) : t'Dt = 0\ \forall t \in Z_0(k_0)\},$$
$$\mathcal{B} := \{D \in \mathbb{S}(p) : e_k' D\xi(j) = 0\ \forall k \in M_*(j),\ \forall j \in J\}.$$

Let us show that

$$\mathcal{A} \subset \mathcal{B}. \tag{3.13}$$

It is evident that it follows from the inclusion $D \in \mathcal{A}$ and representation (3.10) of the set $Z_0(k_0)$ that

$$\xi(j)' D\xi(j) = 0\ \forall j \in \mathcal{J}. \tag{3.14}$$

It is easy to see that these equalities and condition $D \in \mathcal{COP}(T(p))$ imply the inequalities

$$D\xi(j) \geqslant\ \forall j \in \mathcal{J}.$$

Let us show that

$$\xi(i)' D\xi(j) = 0\ \forall i \in \mathcal{J}(s), \forall j \in \mathcal{J}(s),\ \forall s \in \mathcal{S},\ \forall D \in \mathcal{A}. \tag{3.15}$$

In fact, it follows from the inclusion $D \in \mathcal{A}$ and (3.10) that

$$t'Dt = 0\ \forall t \in \mathcal{Z}(s) = \operatorname{conv}\{\xi(j), j \in \mathcal{J}(s)\},\ \forall s \in \mathcal{S}.$$

Hence, taking into account (3.14) and inclusions $0.5(\xi(i) + \xi(j)) \in \mathcal{Z}(s)$ for all $i \in \mathcal{J}(s)$, $j \in \mathcal{J}(s)$ and $s \in \mathcal{S}$, we obtain $0 = (\xi(i) + \xi(j))' D(\xi(i) + \xi(j)) = 2\xi(i)' D\xi(j)$ for all $i \in \mathcal{J}(s)$, $j \in \mathcal{J}(s)$ and $s \in \mathcal{S}$. Thus, we have shown that equalities (3.15) hold true.

Now we will show that

$$e_k' D\xi(j) = 0\ \forall k \in P_*(s),\ \forall j \in \mathcal{J}(s),\ \forall s \in \mathcal{S},\ \forall D \in \mathcal{A}. \tag{3.16}$$

Suppose contrary: there exist $D \in \mathcal{A}$, $s_0 \in \mathcal{S}$, $k_0 \in P_*(s_0)$ and $j_0 \in \mathcal{J}(s_0)$ such that

$$e_{k_0}' D\xi(j_0) > 0.$$

Since $k_0 \in P_*(s_0)$, there exists an index $i_0 \in \mathcal{J}(s_0)$ such that $k_0 \in \operatorname{supp}(\xi(i_0))$. Taking into account the inequalities $\xi(i_0) \geqslant \mathbf{0}$, $D\xi(j_0) \geqslant \mathbf{0}$ and $\xi_{k_0}(i_0) > 0$, let us calculate

$$\xi(i_0)' D\xi(j_0) = \sum_{k \in P} \xi_k(i_0) e_k' D\xi(j_0) \geqslant \xi_{k_0}(i_0) e_{k_0}' D\xi(j_0) > 0.$$

But this contradicts equalities (3.15). Thus, we have shown that that equalities (3.16) hold true.

Now let us consider any $j \in \mathcal{J}$ and the corresponding set $\mathcal{S}(j)$. It follows from (3.16) and the definition of the set $\mathcal{S}(j)$ that for any $D \in \mathcal{A}$ we have

$$e_k' D\xi(j) = 0\ \forall k \in P_*(s),\ \forall s \in \mathcal{S}(j) \iff e_k' D\xi(j) = 0\ \forall k \in \bigcup_{s \in \mathcal{S}(j)} P_*(s).$$

Taking into account the latter equalities and the definition of the set $M_*(j)$, we conclude that $e_k' D\xi(j) = 0\ \forall k \in M_*(j),\ \forall j \in \mathcal{J},\ \forall D \in \mathcal{A}$. This implies inclusion (3.13).

It follows from inclusion (3.13) that the conditions of this proposition guarantee the fulfillment of conditions of Proposition 3.3. Wherefrom we conclude that the face $\mathcal{F}$ under consideration is a non-exposed face of $\mathcal{CP}(T(p))$. $\square$

In contrast to conditions of Proposition 3.3, the conditions of Proposition 3.6 can be easily verified.

In fact, for any matrix $D \in \mathbb{S}(p)$ with elements $d_{ij}, i = 1,...,p,\ j = i,...,p$, we define the vector $\operatorname{svec}(D) \in \mathbb{R}^{p(p+1)/2}$ by the rule (see [15])

$$\operatorname{svec}(D) = (d_{11}, \sqrt{2}d_{12}, ..., \sqrt{2}d_{1p}, d_{22}, \sqrt{2}d_{23}, ..., \sqrt{2}d_{2p}, ..., d_{pp})^\top.$$

Suppose that $\xi(j)$, $M_*(j)$, $j \in J$, and $t^* \in Z_0(0) \setminus Z_0(k_0)$ are given. Then it is easy to construct matrices $\mathbb{B} \in \mathbb{R}^{p_a \times p(p+1)/2}$ with $p_a = \sum\limits_{j \in J} |M_*(j)|$ and $\mathbb{W} \in \mathbb{R}^{p_b \times p(p+1)/2}$ with $p_b = |\operatorname{supp}(t^*)|$ such that

$$\mathcal{B} = \{D \in \mathbb{S}(p) : \mathbb{B}\operatorname{svec}(D) = \mathbf{0}\},$$

$$\{D \in \mathbb{S}(p) : e_k' D t^* = 0\ \forall k \in \operatorname{supp}(t^*)\} = \{D \in \mathbb{S}(p) : \mathbb{W}\operatorname{svec}(D) = \mathbf{0}\}.$$

Then it is evident that for any $D \in \mathbb{S}(p)$, the qualities (3.11) imply the equalities (3.12) if and only if

$$\operatorname{rank}\mathbb{B} = \operatorname{rank}\begin{pmatrix} \mathbb{B} \\ \mathbb{W} \end{pmatrix}. \tag{3.17}$$

The latter condition can be easily tested by any available algorithm calculating the rank of a given matrix.

4. Non-exposed non-polyhedral faces generated by a non-exposed polyhedral face

Theorem 4.1. *Let $\mathcal{F} := \operatorname{cone}\{\tau(s)\tau(s)', s \in \mathcal{S}\}$ with $|\mathcal{S}| < \infty$, $\tau(s) \in T(p)$ for all $s \in \mathcal{S}$, be a non-exposed polyhedral face of $\mathcal{CP}(T(p))$. Then for any integer $q > 0$, the set*

$$\mathcal{F}_* := \operatorname{cone}\{\bar{t}\bar{t}', \bar{t} \in \bigcup_{s \in \mathcal{S}} T_*(s)\},$$

where $T_*(s) = \operatorname{conv}\{\bar{\tau}(s), \mathbf{e}_{p+k}, k = 1,...,q\}$, $\bar{\tau}(s) = (\tau'(s), 0, ..., 0)' \in \mathbb{R}^{p+q}\ \forall s \in \mathcal{S}$, $\{\mathbf{e}_k, k = 1,...,p+q\}$ is the standard basis in $\mathbb{R}^{p+q}$, is a non-exposed non-polyhedral face of $\mathcal{CP}(T(p+q))$ with $T(p+q) = \{t \in \mathbb{R}^{p+q}_+ : ||t||_1 = 1\}$.

Proof. Since $\mathcal{F}$ is a face of $\mathcal{CP}(T(p))$, the following holds true

$$A \in \mathcal{CP}(T(p)),\ B \in \mathcal{CP}(T(p)),\ A + B \in \mathcal{F} \implies A \in \mathcal{F},\ B \in \mathcal{F}. \tag{4.1}$$

By assumption $\mathcal{F}$ is a non-exposed polyhedral face of $\mathcal{CP}(T(p))$, hence, it follows from [10] that there exists $\tau_* \in T(p) \setminus \{\tau(s), s \in \mathcal{S}\}$ such that, for any $D \in \mathcal{COP}(T(p))$, the equalities

$$\tau'(s)D\tau(s) = 0 \; \forall s \in \mathcal{S}, \tag{4.2}$$

imply the equality

$$\tau_*{}'D\tau_* = 0. \tag{4.3}$$

Let us show that $\mathcal{F}_*$ is a face of $\mathcal{CP}(T(p+q))$.

Notice that, for any $s \in \mathcal{S}$,

$$\bar{t} \in T_*(s) \iff \bar{t} = \begin{pmatrix} \alpha\tau(s) \\ t \end{pmatrix} \text{ with some } \alpha \geqslant 0,\ t \in \mathbb{R}^q_+,\ \alpha + ||t||_1 = 1. \tag{4.4}$$

Consequently, if $U_* \in \mathcal{F}_*$ then U_* admits a presentation

$$U_* = \sum_{s\in\mathcal{S}} \begin{pmatrix} \alpha_s\tau(s) \\ t(s) \end{pmatrix} (\alpha_s\tau'(s), t'(s)) \text{ with some } \alpha_s \geqslant 0,\ t(s) \in \mathbb{R}^q_+ \; \forall s \in \mathcal{S}. \tag{4.5}$$

Suppose that

$$A_* \in \mathcal{CP}(T(p+q)),\ B_* \in \mathcal{CP}(T(p+q)), \text{ and } U_* := A_* + B_* \in \mathcal{F}_*.$$

Since $A_* \in \mathcal{CP}(T(p+q))$, $B_* \in \mathcal{CP}(T(p+q))$, then the matrices A_* and B_* can be represented in the forms

$$A_* = \sum_{j=1}^{\bar{p}_*} \begin{pmatrix} \xi_1(j) \\ \xi_2(j) \end{pmatrix} (\xi_1(j)', \xi_2(j)'),\ B_* = \sum_{j=1}^{\bar{p}_*} \begin{pmatrix} \mu_1(j) \\ \mu_2(j) \end{pmatrix} (\mu_1(j)', \mu_2(j)'), \tag{4.6}$$

where $\bar{p}_* = (p+q)(p+q+1)/2$, $\xi_1(j) \in \mathbb{R}^p_+$, $\xi_2(j) \in \mathbb{R}^q_+$, $\mu_1(j) \in \mathbb{R}^p_+$, $\mu_2(j) \in \mathbb{R}^q_+$, for all $j = 1, ..., \bar{p}_*$. It follows from (4.5), (4.6), and equality $U_* = A_* + B_*$ that

$$\sum_{s\in\mathcal{S}} (\alpha_s\tau(s))(\alpha_s\tau(s))' = \sum_{j=1}^{\bar{p}_*} (\xi_1(j)\xi_1(j)' + \mu_1(j)\mu_1(j)'). \tag{4.7}$$

It is evident that

$$\sum_{s\in\mathcal{S}} (\alpha_s\tau(s))(\alpha_s\tau(s))' \in \mathcal{F},\ \sum_{j=1}^{\bar{p}_*} \xi_1(j)\xi_1(j)' \in \mathcal{CP}(T(p)),\ \sum_{j=1}^{\bar{p}_*} \mu_1(j)\mu_1(j)' \in \mathcal{CP}(T(p)).$$

Then it follows from (4.1) and (4.7) that

$$\xi_1(j)\xi_1(j)' \in \mathcal{F},\ \mu_1(j)\mu_1(j)' \in \mathcal{F} \; \forall j = 1, ..., \bar{p}_*.$$

Due to Lemma 3.2 these inclusions imply the equalities

$$\xi_1(j) = \beta_j\tau(s_j),\ \mu_1(j) = \bar{\beta}_j\tau(\bar{s}_j) \text{ with some } \beta_j \geqslant 0,\ s_j \in \mathcal{S},\ \bar{\beta}_j \geqslant 0,\ \bar{s}_j \in \mathcal{S},$$
$$\forall j = 1, ..., \bar{p}_*.$$

Consequently, for all $j = 1, ..., \bar{p}_*$,

$$\begin{pmatrix} \xi_1(j) \\ \xi_2(j) \end{pmatrix} = \begin{pmatrix} \beta_j\tau(s_j) \\ \xi_2(j) \end{pmatrix} = (\beta_j + ||\xi_2(j)||_1)\bar{t}(s_j),\ \bar{t}(s_j) \in T_*(s_j),$$

$$\begin{pmatrix} \mu_1(j) \\ \mu_2(j) \end{pmatrix} = \begin{pmatrix} \bar{\beta}_j\tau(\bar{s}_j) \\ \mu_2(j) \end{pmatrix} = (\bar{\beta}_j + ||\mu_2(j)||_1)\bar{t}(\bar{s}_j),\ \bar{t}(\bar{s}_j) \in T_*(\bar{s}_j).$$

These equalities and relations (4.4), (4.6) imply that $A_* \in \mathcal{F}_*$, $B_* \in \mathcal{F}_*$, and consequently $\mathcal{F}_*$ is a face of $\mathcal{CP}(T(p+q))$.

Now we will show that $\mathcal{F}_*$ is a non-exposed face of $\mathcal{CP}(T(p+q))$.

Let $\mathcal{F}_*^{exp} = \mathcal{CP}(T(p+q)) \cap X_*^{\perp}$ with some $X_* \in \mathcal{COP}(T(p+q))$ be the minimal exposed face of $\mathcal{CP}(T(p+q))$ containing $\mathcal{F}_*$. Hence

$$\bigcup_{s\in\mathcal{S}} T_*(s) \subset Z_0(X_*, T(p+q)) \implies \bar{\tau}(s) \in Z_0(X_*, T(p+q)) \; \forall s \in \mathcal{S}.$$

Let us rewrite the matrix X_* in the form

$$X_* = \begin{pmatrix} \tilde{X} & \tilde{X}_1 \\ \tilde{X}_1' & \tilde{X}_2 \end{pmatrix}, \text{ where } \tilde{X} \in \mathcal{COP}(T(p)).$$

Then

$$0 = (\bar{\tau}(s))' X_* \bar{\tau}(s) = (\tau(s))' \tilde{X} \tau(s) \ \forall s \in \mathcal{S}.$$

Taking into account these equalities, the inclusion $\tilde{X} \in \mathcal{COP}(T(p))$ and conditions (4.2), (4.3), we conclude that $\tau_*' \tilde{X} \tau_* = 0$. It follows from the latter equality that

$$\bar{\tau}_*' X_* \bar{\tau}_* = 0 \text{ with } \bar{\tau}_* = (\tau_*', 0, ..., 0)' \in \mathbb{R}^{p+q},$$

and consequently $\bar{\tau}_* \bar{\tau}_*' \in \mathcal{F}_*^{exp}$.

Let us show that $\bar{\tau}_* \bar{\tau}_*' \notin \mathcal{F}_*$. In fact, suppose the contrary: $\bar{\tau}_* \bar{\tau}_*' \in \mathcal{F}_*$. Then it follows from the latter inclusion that $\tau_* \tau_*' \in \mathcal{F}$. Hence taking into account this inclusion, Lemma 3.2 and equalities $||\tau_*||_1 = ||\tau(s)||_1 = 1$ for all $s \in \mathcal{S}$, we obtain that $\tau_* = \tau(s_*)$ with some $s_* \in \mathcal{S}$. But this contradicts the assumption that $\tau_* \in T(p) \setminus \{\tau(s), s \in \mathcal{S}\}$. Thus, we have proved that $\bar{\tau}_* \bar{\tau}_*' \notin \mathcal{F}_*$.

Since $\bar{\tau}_* \bar{\tau}_*' \notin \mathcal{F}_*$, $\bar{\tau}_* \bar{\tau}_*' \in \mathcal{F}_*^{exp}$, and $\mathcal{F}_*^{exp}$ is the minimal exposed face of $\mathcal{CP}(T(p+q))$ containing $\mathcal{F}_*$, we conclude that $\mathcal{F}_*$ is a non-exposed face of $\mathcal{CP}(T(p+q))$.

Now we will show that $\mathcal{F}_*$ is a non-polyhedral face. By construction, for any $s \in \mathcal{S}$, the set $T_*(s)$ contains a continuum of elements. Hence, the set $\mathcal{Z} := \bigcup\limits_{s \in S} T_*(s)$ does not consist of a finite number of elements. Then it follows from Proposition 3.4 that the face $\mathcal{F}_*$ is not polyhedral. □

Corollary 4.1. *For any $p \geqslant 6$, there exist non-exposed non-polyhedral faces of $\mathcal{CP}(T(p))$.*

Proof. It was shown in [10] that for any $p \geqslant 5$, there exist non-exposed polyhedral faces of $\mathcal{CP}(T(p))$. Thus, it follows from this result and Theorem 4.1 that for any $p \geqslant 6$, there exist non-exposed non-polyhedral faces of $\mathcal{CP}(T(p))$. □

Remark 4.2. *It follows from the proof of Theorem* 4.1 *that statement of this theorem holds true, if $\mathcal{F}$ is a non-exposed (not necessary polyhedral) face of $\mathcal{CP}(T(p))$ and for $s \in \mathcal{S}$, the set $T_*(s)$ is as follows: $T_*(s) = \text{conv}\{\bar{\xi}(j), j \in \mathcal{J}(s),\ \mathbf{e}_{p+k}, k = 1, ..., q\}$, $\bar{\xi}(j) = (\xi'(j), 0, ..., 0)' \in \mathbb{R}^{p+q}\ \forall j \in \mathcal{J}(s), \forall s \in \mathcal{S}$. Here we consider that the face $\mathcal{F}$ is represented in the form* (3.2) *where the set $Z_0(k_0)$ is as in* (3.10). *Note that such representation exists for any proper face of $\mathcal{CP}(T(p))$.*

5. Examples

In this section, we illustrate the application of Proposition 3.6 and Theorem 4.1 by examples.

We start with an example which illustrates Proposition 3.6.

Let us set $p = 6$ and consider the set $T(6) = \{t \in \mathbb{R}^6_+,\ ||t||_1 = 1\}$ and matrix

$$X_0 = \begin{pmatrix} H & \mathbf{0} \\ \mathbf{0}' & 0 \end{pmatrix} \in \mathbb{R}^{6\times 6}, \text{ where } H = \begin{pmatrix} 1 & -1 & 1 & 1 & -1 \\ -1 & 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & 1 & -1 \\ -1 & 1 & 1 & -1 & 1 \end{pmatrix}.$$

The matrix H is the Horn matrix and it is known (see, for example, [7]) that it is copositive, hence, $X_0 \in \mathcal{COP}(T(6))$.

Let us denote

$$\tau(1) = 0.5(1, 1, 0, 0, 0, 0)',\ \tau(2) = 0.5(0, 1, 1, 0, 0, 0)',\ \tau(3) = 0.5(0, 0, 1, 1, 0, 0)',$$
$$\tau(4) = 0.5(0, 0, 0, 1, 1, 0)',\ \tau(5) = 0.5(1, 0, 0, 0, 1, 0)',\ \tau(6) = (0, 0, 0, 0, 0, 1)'.$$

One can easily check that

$$Z_0(0) := Z_0(X_0, T(6)) = \text{conv}\{\tau(1), \tau(2), \tau(6)\} \cup \text{conv}\{\tau(2), \tau(3), \tau(6)\} \cup$$
$$\cup \text{conv}\{\tau(3), \tau(4), \tau(6)\} \cup \text{conv}\{\tau(4), \tau(5), \tau(6)\} \cup \text{conv}\{\tau(1), \tau(5), \tau(6)\}.$$

Let us consider a matrix

$$X_1 = \begin{pmatrix} D_0 & \mathbf{0} \\ \mathbf{0}' & 0 \end{pmatrix} \in \mathbb{R}^{6\times 6}, \text{ where } D_0 = \begin{pmatrix} 6.0 & -11.5 & 5.5 & 5.5 & -5.5 \\ -11.5 & 21.0 & -19.0 & 13.0 & 7.0 \\ 5.5 & -19.0 & 21.0 & -19.0 & 7.0 \\ 5.5 & 13.0 & -19.0 & 21.0 & -13.0 \\ -5.5 & 7.0 & 7.0 & -13.0 & 9.0 \end{pmatrix}.$$

Since for $t_* = (0.27, 0.46, 0.27, 0, 0, 0)' \in T(6)$ we have $t_*'X_1t_* = -0.3624$, it is evident that $X_1 \notin \mathcal{COP}(T(6))$.

Let us show that $X_1 \in \mathcal{COP}(Z_0(0))$. To do this we will show that $X_1 \in \mathcal{COP}(\text{conv}(Z_0(0))$ which is equivalent to $\tilde{D} := \mathcal{T}'X_1\mathcal{T} \in \mathcal{COP}(T(6))$, where

$$\mathcal{T} := (\tau(j), j = 1, ..., 6) \in \mathbb{R}^{6\times 6}.$$

In fact, the matrix $\tilde{D}$ takes the form

$$\tilde{D} = 4(X_0 + C), \text{ where } C = \begin{pmatrix} 0 & 0 & 0.25 & 4 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0.25 & 0 & 0 & 0 & 0.25 & 0 \\ 4 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0.25 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Since all elements of the matrix C are non-negative and $X_0 \in \mathcal{COP}(T(6))$, we conclude that $\tilde{D} \in \mathcal{COP}(T(6))$ and hence $X_1 \in \mathcal{COP}(\text{conv}(Z_0(0))$ that implies $X_1 \in \mathcal{COP}(Z_0(0))$.

Now let us find the set $Z_0(1) := Z_0(X_1, Z_0(0))$. It is easily to do this via the matrix $\tilde{D}$. It is easy to see that

$$Z_0(\tilde{D}, T(6)) = \text{conv}\{\tau(1), \tau(6)\} \cup \text{conv}\{\tau(2), \tau(6)\} \cup \text{conv}\{\tau(3), \tau(6)\} \cup$$
$$\cup \text{conv}\{\tau(4), \tau(6)\} \cup \text{conv}\{\tau(5), \tau(6)\}.$$

This implies that

$$Z_0(X_1, \text{conv}(Z_0(0))) = \text{conv}\{\xi(1), \xi(6)\} \cup \text{conv}\{\xi(2), \xi(6)\} \cup \text{conv}\{\xi(3), \xi(6)\} \cup$$
$$\cup \text{conv}\{\xi(4), \xi(6)\} \cup \text{conv}\{\xi(5), \xi(6)\},$$

where

$$\xi(1) = 0.5(\tau(1) + \tau(2)),\ \xi(2) = 0.5(\tau(2) + \tau(3)),\ \xi(3) = 0.5(\tau(3) + \tau(4)),$$

$$\xi(4) = 0.5(\tau(4) + \tau(5)),\ \xi(5) = 0.5(\tau(1) + \tau(5)),\ \xi(6) = \tau(6).$$

It is easy to see that if $T_1 \subset T_2 \subset T$ then $D \in \mathcal{COP}(T_2)$ implies that $D \in \mathcal{COP}(T_1)$, and $Z_0(D, T_1) \subset \subset Z_0(D, T_2)$. Hence, if $Z_0(D, T_2) \subset T_1$, then $Z_0(D, T_1) = Z_0(D, T_2)$.

In our case, we have $T_2 := \text{conv}(Z_0(0))$, $T_1 := Z_0(0)$, $X_1 \in \mathcal{COP}(T_2)$ and $Z_0(X_1, T_2) \subset T_1$. Hence, the equality $Z_0(X_1, T_1) = Z_0(X_1, T_2)$ holds true, and we obtain

$$Z_0(1) := Z_0(X_1, Z_0(0)) = Z_0(X_1, \text{conv}(Z_0(0))).$$

Consequently,

$$\mathcal{S} = \{1, ..., 5\},\ \mathcal{J}(s) = \{s, 6\}\ \forall s \in \mathcal{S};,\ \mathcal{J} = \{1, ..., 6\},$$
$$M_*(1) = \{1,2,3,6\},\ M_*(2) = \{2,3,4,6\},\ M_*(3) = \{3,4,5,6\},$$
$$M_*(4) = \{1,4,5,6\},\ M_*(5) = \{1,2,5,6\},\ M_*(6) = \{1, ..., 6\}.$$

Consider the set

$$\mathcal{F} := \{U = \sum_{i=1}^{p_*} \alpha_i t(i) t(i)',\ \alpha_i \geqslant 0,\ t(i) \in Z_0(1)\}, p_* = p(p+1)/2.$$

Since matrices X_0, X_1 and the sets $Z_0(0)$, $Z_0(1)$ satisfy conditions (3.1) with $k_0 = 1$, we conclude that $\mathcal{F}$ is a face of $\mathcal{CP}(T(6))$.

Let us show that the face $\mathcal{F}$ is non-exposed. To do this we will use Proposition 3.6.

Set $t^* = \tau(1) \in Z_0(0) \setminus Z_0(1)$ and consider the sets of matrices

$$\mathcal{B} := \{D \in \mathbb{S}(6) : e_k' D\xi(j) = 0\ \forall k \in M_*(j),\ \forall j \in \mathcal{J}\},$$
$$\mathcal{W} := \{D \in \mathbb{S}(6) : e_k' Dt^* = 0\ \forall k \in \mathrm{supp}(t^*)\} \text{ with } \mathrm{supp}(t^*) = \{1,2\},$$

where $\{e_k, k = 1,...,6\}$ is the standard basis in $\mathbb{R}^6$.

For any matrix $D \in \mathbb{S}(6)$ with elements $d_{ij}, i = 1,...,6,\ j = i,...,6$, we define the vector $\mathrm{svec}(D) \in \mathbb{R}^{21}$ by the rule (see [15])

$$\mathrm{svec}(D) = (d_{11}, \sqrt{2}d_{12},...,\sqrt{2}d_{16}, d_{22}, \sqrt{2}d_{23},...,\sqrt{2}d_{26},...,d_{66})^\top.$$

It is easy to construct matrices $\mathbb{B} \in \mathbb{R}^{26\times 21}$, $\mathbb{W} \in \mathbb{R}^{2\times 21}$ such that

$$\mathcal{B} = \{D \in \mathbb{S}(6) : \mathbb{B}\,\mathrm{svec}(D) = \mathbf{0}\}, \quad \mathcal{W} = \{D \in \mathbb{S}(6) : \mathbb{W}\,\mathrm{svec}(D) = \mathbf{0}\}.$$

In our example, we have $\mathrm{rank}\,\mathbb{B} = 20$ and $\mathrm{rank}\begin{pmatrix}\mathbb{B}\\ \mathbb{W}\end{pmatrix} = 20$. Hence it follows from (3.17) and Proposition 3.6 that the face $\mathcal{F}$ is non-exposed.

Since the set $Z_0(k_0) = Z_0(1) = Z_0(X_1, \mathrm{conv}(Z_0(0))$ contains a continuum of elements, it follows from Proposition 3.4 that $\mathcal{F}$ is non-polyhedral.

In [10], for any $p \geqslant 5$, examples of non-exposed polyhedral faces of $\mathcal{CP}(T(p))$ were given. Hence, to illustrate Theorem 4.1 we can choose any non-exposed face form [10]. To simplify the presentation, let us choose $p = 5$. Set $\mathcal{S} = \{1,...,5\}$,

$$\tau(1) = (1,2,1,0,0)'/4,\ \tau(2) = (0,1,2,1,0)'/4,\ \tau(3) = (0,0,1,2,1)'/4,$$
$$\tau(4) = (1,0,0,1,2)'/4,\ \tau(5) = (2,1,0,0,1)'/4.$$

It was shown in [9; 10] that the set $\mathcal{F} := \mathrm{cone}\{\tau(s), i \in \mathcal{S}\}$ is a non-exposed face of $\mathcal{CP}(5)$. Let us choose $q = 2$ and denote

$$\bar\tau(1) = (1,2,1,0,0,0,0)'/4,\ \bar\tau(2) = (0,1,2,1,0,0,0)'/4,\ \bar\tau(3) = (0,0,1,2,1,0,0)'/4,$$
$$\bar\tau(4) = (1,0,0,1,2,0,0)'/4,\ \bar\tau(5) = (2,1,0,0,1,0,0)'/4,$$

$$T_*(s) = \mathrm{conv}\{\bar\tau(s), e_6, e_7\} = \{(\alpha_1\tau'(s), \alpha_2, \alpha_3)', \alpha_i \geqslant 0, i = 1,2,3; \sum_{i=1}^{3}\alpha_i = 1\} \subset \mathbb{R}^7_+,\ s \in \mathcal{S}.$$

Then it follows from Theorem 4.1 that the set $\mathcal{F}_* := \mathrm{cone}\{\bar t\bar t', \bar t \in \bigcup_{s\in\mathcal{S}} T_*(s)\}$ is a non-exposed non-polyhedral face of $\mathcal{CP}(7)$.

Conclusion. In this paper, we studied properties of non-exposed faces of the cone of completely positive matrices. A special attention was paid on non-polyhedral faces. Based on results obtained, we proved easily tested sufficient conditions for a face to be non-exposed one. The novelty of the results obtained lies in the fact that it was proved and illustrated that for any $p \geqslant 6$, there exist non-polyhedral non-exposed faces of the cone of completely positive $p \times p$ matrices. This makes a significant contribution to the study of the facial structure of the completely positive cone.

This work was supported by the Institute of Mathematics of the National Academy of Sciences of Belarus within the framework of the state programme "Convergence–2025", tasks 1.3.01 and 1.3.04.

References

1. Anjos M. F., Lasserre J. B. (eds). Handbook on Semi-definite, Conic and Polynomial Optimization, *International Series in OR/MS,* 2012, vol. 166, Springer, New York. https://doi.org/10.1007/978-1-4614-0769-0

2. Letchford A. N., Parkes A. J. A guide to conic optimisation and its applications. *RAIRO – Oper. Res.,* 2018, vol. 52, iss. 4–5, pp. 1087–1106. https://doi.org/10.1051/ro/2018034

3. Bomze I. M., Schachinger W., Uchida G. Think co(mpletely)positive! Matrix properties, examples and a clustered bibliography on copositive optimization. *Journal of Global Optimization*, 2012, vol. 52, iss. 3, pp. 423–445. https://doi.org/10.1007/s10898-011-9749-3

4. Dickinson P. J. Geometry of the copositive and completely positive cones. *Journal of Mathematical Analysis and Applications*, 2011, vol. 380, iss. 1, pp. 377–395. https://doi.org/10.1016/j.jmaa.2011.03.005

5. Kostyukova O. I., Tchemisova T. V. On equivalent representations and properties of faces of the cone of copositive matrices. *Optimization,* 2022, vol. 71, iss. 11, pp. 3211–3239. https://doi.org/10.1080/02331934.2022.2027939

6. Hoffman A. J., Pereira F. On copositive matrices with -1, 0, 1 entries. *Journal of Combinatorial Theory,* 1973, Series A, vol. 14, iss. 3, pp. 302–309. https://doi.org/10.1016/0097-3165(73)90006-X

7. Berman A., Dür M., Shaked-Monderer N. Open problems in the theory of completely positive and copositive matrices. *The Electronic Journal of Linear Algebra*, 2015, vol. 29, pp. 46–58. https://doi.org/10.13001/1081-3810.2943

8. Zhang Q. Completely positive cones: are they facially exposed? *Linear Algebra and its Applications*, 2018, vol. 558, pp. 195–204. https://doi.org/10.1016/j.laa.2018.08.028

9. Zhang Q. Faces of the 5×5 completely positive cone. *Linear and Multilinear Algebra*, 2020, vol. 68, iss. 12, pp. 2523–2540. https://doi.org/10.1080/03081087.2019.1586827

10. Kostyukova O. I. Non-exposed polyhedral faces of the completely positive cone. *Linear and Multilinear Algebra*, 2024, pp. 1–28. https://doi.org/10.1080/03081087.2024.2346313

11. Eichfelder G., Jahn J. Set-Semidefinite Optimization. *Journal of Convex Analysis*, 2008, vol. 15, iss. 4, pp. 767–801.

12. Kostyukova O. I., Tchemisova T. V., Dudina O. S. On the Uniform Duality in Copositive Optimization. *Journal of Optimization Theory and Applications,* 2024. https://doi.org/10.1007/s10957-024-02515-1

13. Wang F., Wolkowicz H. Singularity degree of non-facially exposed faces. 2022, arXiv preprint. arXiv: 2211.00834.

14. Kostyukova O. I., Tchemisova T. V. Representation of Zeros of a Copositive Matrix via Maximal Cliques of a Graph. 2024. arXiv preprint. arXiv: 2410.08066.

15. Alizadeh F., Schmieta S. H. Optimization with Semi-definite, Quadratic and Linear Constraints, Report 23-97. 1997. Rutgers Center for Operations Research, Rutgers University.

Национальная академия наук Беларуси
Труды Института математики НАН Беларуси. 2024. Том 32. № 2. С. 69–72

УДК 517.962.2+517.929.2

ЛИНЕЙНЫЕ РЕКУРРЕНТНЫЕ УРАВНЕНИЯ В ПРОСТРАНСТВЕ ВЫПУКЛЫХ МНОГОУГОЛЬНИКОВ С НЕПЕРЕСЕКАЮЩИМИСЯ РЕШЕНИЯМИ

А. С. Войделевич

Институт математики НАН Беларуси, Минск, Беларусь
e-mail: aliaksei.vaidzelevich@gmail.com

Поступила: 24.07.2024 **Исправлена:** 02.10.2024 **Принята:** 12.12.2024

Ключевые слова: линейные рекуррентные уравнения, выпуклые многоугольники.

Аннотация. Получено необходимое и достаточное условие на матрицу коэффициентов линейного рекуррентного уравнения в пространстве выпуклых многоугольников, любые два различных решения которого не пересекаются, т. е. значения решений при каждом аргументе различны.

LINEAR RECURRENCE EQUATIONS IN THE SPACE OF CONVEX POLYGONS WITH NON-INTERSECTING SOLUTIONS

A. S. Vaidzelevich

Institute of Mathematics of the National Academy of Sciences of Belarus, Minsk, Belarus
e-mail: aliaksei.vaidzelevich@gmail.com

Received: 24.07.2024 **Revised:** 02.10.2024 **Accepted:** 12.12.2024

Keywords: linear recurrence equations, convex polygons.

Abstract. A necessary and sufficient condition is obtained for the coefficient matrix of a linear recurrence equation in the space of convex polygons, any two different solutions of which do not intersect, i. e. the values of the solutions for each argument are different.

В работе рассматривается частный случай линейных рекуррентных уравнений в пространстве выпуклых компактов [1], решения которых представляют собой последовательности выпуклых многоугольников. Прежде чем сформулировать полученный результат, введем необходимые обозначения и приведем определения.

Суммой Минковского $Z = X + Y$ двух непустых множеств $X, Y \subset \mathbb{R}^2$ на плоскости называется множество $Z \stackrel{\text{def}}{=} \{x+y\colon x \in X,\, y \in Y\}$. Для действительной (2×2)-матрицы A и множества $X \subset \mathbb{R}^2$ через AX обозначим множество $\{Ax\colon x \in X\}$. В том частном случае, когда $A = \operatorname{diag}[\alpha, \alpha]$, вместо AX пишем αX. Непосредственно из определения следует, что для произвольных множеств X, $Y \subset \mathbb{R}^2$ и (2×2)-матриц A верно равенство $A(X+Y) = AX + AY$. Отметим, что для произвольных действительных (2×2)-матриц A, B и множества $X \subset \mathbb{R}^2$, вообще говоря, $(A+B)X \neq AX + BX$. В то же время если числа α и β одного знака, то для любого выпуклого множества X выполнено равенство $(\alpha+\beta)X = \alpha X + \beta X$.

Через $\mathcal{P}_2$ обозначим множество всех выпуклых многоугольников на плоскости, где под выпуклым многоугольником будем понимать выпуклую оболочку конечного множества точек.

В частности, точки и отрезки также считаем выпуклыми многоугольниками. Будем говорить, что два многоугольника P и Q различные, если они отличаются как множества точек на плоскости, а не в том смысле, что не существует движения плоскости, переводящего многоугольник P в многоугольник Q. Нетрудно видеть, что семейство $\mathcal{P}_2$ замкнуто относительно операции сложения по Минковскому и умножения на матрицу, а значит, мы можем рассмотреть линейное рекуррентное уравнение

$$X(t+1) = X(t) + AX(t), \quad X(t) \in \mathcal{P}_2, \quad t \in \mathbb{N} \cup \{0\}, \tag{1}$$

где A – действительная (2×2)-матрица коэффициентов. Два решения $X_1(\cdot)$ и $X_2(\cdot)$ уравнения (1) называются *непересекающимися*, если $X_1(t) \neq X_2(t)$ при всех $t \in \mathbb{N} \cup \{0\}$.

Уравнение (1) является дискретным аналогом однородного линейного дифференциального уравнения

$$\dot{x} = ax, \quad x(t) \in \mathbb{R}, \quad t \geqslant 0, \tag{2}$$

где $a \in \mathbb{R}$. Действительно, дискретизируя уравнение (1), получаем $x_{t+1} - x_t = ax_t$, т. е. $x_{t+1} = x_t + ax_t$. Хорошо известно, что графики любых двух различных решений уравнения (2) не пересекаются. Возникает естественный вопрос о том, обладает ли уравнение (1) аналогичным свойством. В общем случае ответ на этот вопрос отрицательный. Действительно, если $A = -E$, где E – единичная матрица, то для любого выпуклого многоугольника $M \in \mathcal{P}_2$ решения $X_1(\cdot)$ и $X_2(\cdot)$ такие, что $X_1(0) = M$ и $X_2(0) = -M$, принимают одно и тоже значение $M + (-M)$ при $t = 1$. В работе доказано следующее утверждение.

Теорема. *Любые два решения уравнения* (1) *либо совпадают, либо принимают различные значения при всех* $t \in \mathbb{N} \cup \{0\}$, *если и только если среди собственных значений матрицы* A *нет корней натуральной степени из* -1.

Для доказательства сформулированной теоремы установим справедливость нескольких вспомогательных утверждений.

Лемма 1. *Пусть хотя бы одно из собственных значений матрицы* A *равно корню натуральной степени из* -1. *Тогда найдутся два различных выпуклых многоугольника* P *и* Q, *удовлетворяющих равенству*

$$P + AP = Q + AQ. \tag{3}$$

Доказательство. Пусть n – наименьшее натуральное число такое, что у матрицы A^n одно из собственных значений равно -1. Пусть v – собственный вектор матрицы A^n, соответствующий собственному значению -1, т. е. $A^n v = -v$ и $v \neq 0$. Через m обозначим наименьшее натуральное число такое, что $A^m v = v$. Тогда $A^i v \neq A^j v$ для любых целых $0 \leqslant i \leqslant j < m$ (в противном случае было бы верно равенство $A^{m-j+i}v = v$, но $m - j + i < m$). Так как $A^{2n}v = v$, то $2n$ делится на m, при этом $n < m$, а значит, $m = 2n$. Положим

$$P = \operatorname{conv}\{v, A^2v, \ldots, A^{2n-2}v\} \quad \text{и} \quad Q = \operatorname{conv}\{Av, A^3v, \ldots, A^{2n-1}v\}.$$

Выпуклые многоугольники P и Q различные и выполнены равенства $AP = Q$, $AQ = P$, а значит, $P + AP = P + Q = Q + AQ$. □

Определение. *Опорной функцией* произвольного множества $X \subset \mathbb{R}^2$ называется функция $s(X,\cdot)\colon \mathbb{R}^2 \to \mathbb{R}$, определяемая равенством $s(X,v) \stackrel{\text{def}}{=} \sup\limits_{x\in X} v^\top x$, $v \in \mathbb{R}^2$.

Если множество X представимо в виде выпуклой оболочки конечного числа точек $p_1, p_2, \ldots, p_n$, т. е. является выпуклым многоугольником, то $s(X,v) = \max\limits_{1\leqslant i\leqslant n} v^\top p_i$, $v \in \mathbb{R}^2$.

Из определений опорной функции и суммы Минковского следует, что для любых двух множеств $X, Y \subset \mathbb{R}^2$ выполнено равенство $s(X+Y,\cdot) \equiv s(X,\cdot) + s(Y,\cdot)$. Из равенства $P + AP = Q + AQ$ следует, что $s(P+AP, v) = s(Q+AQ, v)$ для любого $v \in \mathbb{R}^2$, а значит,

$$s(P,v) + s(AP,v) = s(Q,v) + s(AQ,v). \tag{4}$$

Если $X \subset Y$, то $s(X,v) \leqslant s(Y,v)$ для любого $v \in \mathbb{R}^2$. Обратно, если Y – выпуклое компактное множество и $s(X,v) \leqslant s(Y,v)$ при всех $v \in \mathbb{R}^2$, то из теоремы об отделимости выпуклых множеств следует, что $X \subset Y$. Таким образом, два выпуклых компактных множества X и Y совпадают, если и только если равны их опорные функции $s(X,\cdot) \equiv s(Y,\cdot)$. В частности, выпуклые компактные множества X, Y и Z удовлетворяют равенству $X+Z=Y+Z$ тогда и только тогда, когда $X=Y$.

Лемма 2. *Если два выпуклых многоугольника P и Q удовлетворяют равенству* (3)*, то для любого $n \in \mathbb{N}$ также выполнены равенства*

$$P+A^{2n+1}P = Q+A^{2n+1}Q \quad \text{и} \quad P+A^{2n}Q = Q+A^{2n}P. \tag{5}$$

Доказательство. Воспользуемся методом математической индукции. Предположим, что верно равенство $P+A^{2n-1}P = Q+A^{2n-1}Q$, тогда для любого вектора $v \in \mathbb{R}^2$ имеем

$$s(P,v)+s(A^{2n-1}P,v) = s(Q,v)+s(A^{2n-1}Q,v). \tag{6}$$

Подставляя $A^\top v$ вместо v в равенство (6) и пользуясь тем, что $s(X,A^\top v) = s(AX,v)$ для любого подмножества $X \subset \mathbb{R}^2$, получаем

$$s(AP,v)+s(A^{2n}P,v) = s(AQ,v)+s(A^{2n}Q,v). \tag{7}$$

Вычитая из равенства (4) равенство (7), получаем $s(P,v)-s(A^{2n}P,v) = s(Q,v)-s(A^{2n}Q,v)$, т. е. $s(P+A^{2n}Q,v) = s(Q+A^{2n}P,v)$, а значит, $P+A^{2n}Q = Q+A^{2n}P$.

Предположим теперь, что верно равенство $P+A^{2n}Q = Q+A^{2n}P$. Поступая аналогично, получаем, что для любого вектора $v \in \mathbb{R}^2$ выполнено равенство

$$s(AP,v)+s(A^{2n+1}Q,v) = s(AQ,v)+s(A^{2n+1}P,v). \tag{8}$$

Вычитая из равенства (4) равенство (8), получаем

$$s(P,v)-s(A^{2n+1}Q,v) = s(Q,v)-s(A^{2n+1}P,v),$$

т. е. $s(P+A^{2n+1}P,v) = s(Q+A^{2n+1}Q,v)$, а значит, $P+A^{2n+1}P = Q+A^{2n+1}Q$. □

Лемма 3. *Различные выпуклые многоугольники P и Q удовлетворяют равенству* (3)*, тогда при некотором $k \in \mathbb{N}$ хотя бы одно из собственных значений матрицы A^k равно* 1.

Доказательство. Через $p_1, p_2, \ldots, p_n$ и $q_1, q_2, \ldots, q_m$ обозначим вершины многоугольников P и Q соответственно. Без нарушения общности будем считать, что $P \setminus Q \neq \varnothing$. Выберем вершину p_{i_0}, не принадлежащую многоугольнику Q. Для любого $k \in \mathbb{N}$ найдется индекс $i_1 = i_1(k)$ такой, что $p_{i_0}+A^{2k+1}p_{i_1}$ – вершина многоугольника $P+A^{2k+1}P$. Из леммы 2 следует, что для любого $k \in \mathbb{N}$ найдутся индексы $j_0 = j_0(k)$ и $j_1 = j_1(k)$ такие, что $p_{i_0}+A^{2k+1}p_{i_1} = q_{j_0}+A^{2k+1}q_{j_1}$. Так как существует лишь конечное число способов выбрать тройку индексов (i_1, j_0, j_1), то для некоторых $k_1 < k_2 \in \mathbb{N}$ будут выбраны одни и те же индексы, т. е. $p_{i_0}+A^{2k_1+1}p_{i_1} = q_{j_0}+A^{2k_1+1}q_{j_1}$ и $p_{i_0}+ +A^{2k_2+1}p_{i_1} = q_{j_0}+A^{2k_2+1}q_{j_1}$. Поэтому

$$p_{i_0}-q_{j_0} = A^{2k_2+1}(q_{j_1}-p_{i_1}) = A^{2(k_2-k_1)}A^{2k_1+1}(q_{j_1}-p_{i_1}) = A^{2(k_2-k_1)}(p_{i_0}-q_{j_0}).$$

Так как $p_{i_0} \neq q_{j_0}$, то $(p_{i_0}-q_{j_0})$ – собственный вектор матрицы $A^{2(k_2-k_1)}$, соответствующий собственному значению 1. □

Лемма 4. *Пусть $A = \operatorname{diag}[\lambda, 1]$, где $\lambda \neq -1$. Тогда если выпуклые многоугольники P и Q удовлетворяют равенству* (3)*, то они совпадают.*

Доказательство. Если $\lambda = 1$, то из равенства (3) получаем, что $P+P = Q+Q$, т. е. $2P = 2Q$, а значит, $P = Q$.

Если $\lambda = 0$, то умножим обе части равенства (3) на A. Так как $A^2 = A$, то получаем, что $AP+AP = AQ+AQ$, т. е. $AP = AQ$, а значит, из равенства (3) следует, что $P = Q$.

Предположим, что $\lambda \in (-1,1)$. Согласно лемме 1 при всех $n \in \mathbb{N}$ выполнены равенства (5). Следовательно, $s(P,v)+s(A^{2n+1}P,v) = s(Q,v)+s(A^{2n+1}Q,v)$ при всех $v \in \mathbb{R}^2$. Так как $s(A^{2n+1}P,v) = = s\big(P,(A^{2n+1})^\top v\big) \to s(P,\operatorname{diag}[0,1]v) = s(\operatorname{diag}[0,1]P,v)$ при $n \to \infty$. Аналогично, $s(A^{2n+1}Q,v) \to \to s(\operatorname{diag}[0,1]Q,v)$ при $n \to \infty$. Поэтому

$$P+\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} P = Q+\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} Q,$$

а значит, как доказано выше, $P = Q$.

Наконец, если $|\lambda| > 1$, то домножение обеих частей равенства (3) на A^{-1} сводит этот случай к уже рассмотренному, так как $\lambda^{-1} \in (-1,1)$. □

Лемма 5. *Пусть* $A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. *Тогда если выпуклые многоугольники* P *и* Q *удовлетворяют равенству* (3), *то они совпадают.*

Доказательство. Методом математической индукции доказывается, что $A^k = \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}$ при всех $k \in \mathbb{N}$. Для любых векторов u, $v \in \mathbb{R}^2$ функция $f(t) = v^\top \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} u$ линейная по t, а значит, найдется такое $k_0 \in \mathbb{N}$, что либо $v^\top A^k u \leqslant 0$ при всех целых $k \geqslant k_0$, либо $v^\top A^k u \geqslant 0$ при всех целых $k \geqslant k_0$.

Пусть $p_1, p_2, \ldots, p_n$ – вершины многоугольника P, а $q_1, q_2, \ldots, q_m$ – вершины многоугольника Q. Так как матрица A невырожденная, то $A^k p_1, A^k p_2, \ldots, A^k p_n$ – различные вершины многоугольника $A^k P$, $k \in \mathbb{N}$. Аналогично, $A^k q_1$, $A^k q_2$, $\ldots$, $A^k q_m$ – различные вершины многоугольника $A^k Q$, $k \in \mathbb{N}$.

Произвольным образом выберем вершину p_{i_0}. Через $\mathcal{V} \subset \mathbb{R}^2$ обозначим множество векторов v таких, что $v^\top p_{i_0} > v^\top p_i$ при всех $i \neq i_0$. Так как множество $\mathcal{V}$ имеет мощность континуум, то найдется вектор $v \in \mathcal{V}$, неортогональный ни одному из векторов вида $q_{j_2} - q_{j_1}$, $A^k(p_{i_2} - p_{i_1})$, $A^k(q_{j_2} - q_{j_1})$ при всех $i_1 \neq i_2$, $j_1 \neq j_2$, $k \in \mathbb{N}$. Однозначно определены индексы

$$j_0 = \arg\max_{1 \leqslant j \leqslant m} v^\top q_j, \quad i_1 = \arg\max_{1 \leqslant i \leqslant n} v^\top A^k p_i \quad \text{и} \quad j_1 = \arg\max_{1 \leqslant j \leqslant m} v^\top A^k p_j$$

(при всех достаточно больших k индексы i_1 и j_1 одни и те же). Из леммы 2 следует, что

$$p_{i_0} + A^{2k+1} p_{i_1} = q_{j_0} + A^{2k+1} q_{j_1}, \quad p_{i_0} + A^{2k} q_{i_1} = q_{i_0} + A^{2k} p_{i_1}.$$

Умножим второе равенство на A и сложим с первым. После приведения подобных слагаемых, получаем, что $(A+E)p_{i_0} = (A+E)q_{j_0}$, а значит, $p_{i_0} = q_{i_0}$. Так как вершина p_{i_0} выбрана произвольным образом, то $P \subset Q$. Аналогично, получаем, что $Q \subset P$, а значит, многоугольники P и Q совпадают. □

Доказательство теоремы. Предположим, что хотя бы одно из собственных значений матрицы A равно корню натуральной степени из -1. Согласно лемме 1 найдутся два выпуклых многоугольника $P \neq Q$, удовлетворяющих равенству (3). Рассмотрим два решения $X_1(\cdot) \neq X_2(\cdot)$ уравнения (1) такие, что $X_1(0) = P$ и $X_2(0) = Q$. Тогда

$$X_1(1) = P + AP = Q + AQ = X_2(1).$$

Пусть теперь два решения $X_1(\cdot) \neq X_2(\cdot)$ приняли одно и тоже значение. Выберем наибольшее $t \in \mathbb{N} \cup \{0\}$ такое, что $X_1(t) \neq X_2(t)$. Обозначим $P = X_1(t)$ и $Q = X_2(t)$. Так как $X_1(t+1) = X_2(t+1)$, то для выпуклых многоугольников P и Q выполнено равенство (3). Согласно лемме 3 одно из собственных значений λ матрицы A равно корню степени $m \in \mathbb{N}$ из 1. Без нарушения общности будем считать, что λ – примитивный корень из 1 порядка m. Если m четно, т. е. $m = 2n$, то $\lambda^n = -1$. Если же m нечетно, т. е. $m = 2n+1$, то рассмотрим доказанное в лемме (1) равенство $P + A^{2n+1}P = Q + A^{2n+1}Q$. Оба собственных значения матрицы A^{2n+1} действительные. Пусть $J = S^{-1}A^{2n+1}S$ – жорданова нормальная форма матрицы A^{2n+1}, где S – действительная обратимая матрица. Без нарушения общности будем считать, что в правом нижнем углу матрицы J стоит 1. Выпуклые многоугольники $P' = S^{-1}PS$ и $Q' = S^{-1}QS$ различные и удовлетворяют равенству $P' + JP' = Q' + JQ'$. Из лемм 4 и 5 следует, что это может быть лишь только в одном случае, а именно, когда $J = \mathrm{diag}[-1, 1]$. □

Работа выполнена в Институте математики НАН Беларуси по заданию 1.2.01 «Развитие конструктивных и асимптотических методов исследования сложных управляемых дифференциальных и дискретных систем» ГПНИ «Конвергенция–2025» (подпрограмма «Математические модели и методы»).

Литература

1. *Войделевич А. С.* Линейные рекуррентные уравнения в пространстве выпуклых компактов и диаметры их решений // Дифференц. уравнения. 2023. Т. 59, № 8. С. 1084–1088.

References

1. Voidelevich A. S. Linear Recurrent Equations in the Space of Convex Compact Sets and the Diameters of Their Solutions *Differential Equations*, 2023, vol. 59, pp. 1090–1094.

Национальная академия наук Беларуси
Труды Института математики НАН Беларуси. 2024. Том 32. № 2. С. 73–81

УДК 517.968.7

ИНТЕГРО-ДИФФЕРЕНЦИАЛЬНОЕ УРАВНЕНИЕ, СВЯЗАННОЕ С КРАЕВОЙ ЗАДАЧЕЙ РИМАНА–КАРЛЕМАНА

А. П. Шилин

Белорусский государственный университет, Минск, Республика Беларусь
e-mail: a.p.shilin@gmail.com

Поступила: 09.10.2024 **Исправлена:** 05.12.2024 **Принята:** 12.12.2024

Ключевые слова: интегро-дифференциальное уравнение, гиперсингулярный интеграл, обобщенные формулы Сохоцкого, краевая задача Римана–Карлемана, линейное дифференциальное уравнение.

Аннотация. Рассматривается линейное интегро-дифференциальное уравнение на замкнутой кривой, расположенной на комплексной плоскости. Коэффициенты уравнения имеют специальную структуру. Уравнение содержит регулярные и гиперсингулярные интегралы и сводится вначале к смешанной краевой задаче Римана–Карлемана для аналитических функций. Далее решаются два дифференциальных уравнения в областях комплексной плоскости с дополнительными условиями. Указываются в явном виде условия разрешимости исходного уравнения. При их выполнении решение дается в замкнутой форме. Приводится пример.

INTEGRO-DIFFERENTIAL EQUATION ASSOCIATED WITH THE RIEMANN–CARLEMAN BOUNDARY VALUE PROBLEM

A. P. Shilin

Belarusian State University, Minsk, Belarus
e-mail: a.p.shilin@gmail.com

Received: 09.10.2024 **Revised:** 05.12.2024 **Accepted:** 12.12.2024

Keywords: integro-differential equation, hypersingular integral, generalized Sokhotsky formulas, Riemann–Carleman boundary problem, linear differential equation.

Abstract. We consider a linear integro-differential equation on a closed curve located on the complex plane. The coefficients of the equation have a special structure. The equation is first reduced to the mixed Riemann–Carleman boundary value problem for analytic functions. Next, two differential equations are solved in areas of the complex plane with additional conditions. The conditions for the solvability of the original equation are indicated explicitly. When they executed, the solution is given in closed form. An example is given.

1. Введение

Пусть L – простая гладкая замкнутая положительно ориентированная кривая на комплексной плоскости. Под гладкостью кривой понимается наличие у нее непрерывно меняющейся касательной без точек заострения. Обозначим D_+ и D_- соответственно внутренность и внешность этой кривой. Зададим комплексные числа a_k, b_k, $k=\overline{0,n}$, $n \in \mathbb{N}$, $a_n \neq 0$, $b_n \neq 0$. Зададим также H-непрерывную (т. е. удовлетворяющую условию Гельдера) функцию $h(t)$, $t \in L$. Искомой будет в дальнейшем функция $\varphi(t)$, $t \in L$, H-непрерывная вместе со своими производными, входящими в уравнение. Производные этой функции находятся по ее комплексному аргументу $t \in L$.

Для предельных значений на кривой L интеграла типа Коши

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

и его производных справедливы полученные в [1] обобщенные формулы Сохоцкого

$$\Phi_\pm^{(k)}(t) = \pm\frac{1}{2}\varphi^{(k)}(t) + \frac{k!}{2\pi i}\int_L \frac{\varphi(\tau)d\tau}{(\tau - t)^{k+1}}, \quad k=\overline{0,n}, \quad t \in L, \tag{1.1}$$

частным случаем которых при $k=0$ являются классические формулы Сохоцкого. Гиперсингулярные интегралы в формулах (1.1) понимаются в смысле конечной части по Адамару, что согласно [1] приводит для их вычисления к формулам

$$\int_L \frac{\varphi(\tau)d\tau}{(\tau-t)^{k+1}} = \frac{\pi i \varphi^{(k)}(t)}{k!} + \int_L \frac{\varphi(\tau)-\sum_{j=0}^{k}\frac{\varphi^{(j)}(t)}{j!}(\tau-t)^j}{(\tau-t)^{k+1}}d\tau,$$

в правых частях которых интегралы сходятся в обычном смысле. В [2] с использованием формул (1.1) решено в замкнутой форме уравнение

$$\sum_{k=0}^{n}\left(a_k\varphi^{(k)}(t) + \frac{b_k k!}{\pi i}\int_L \frac{\varphi(\tau)d\tau}{(\tau-t)^{k+1}}\right) = h(t), \quad t\in L. \tag{1.2}$$

В [3–5] и некоторых других работах автора уравнение вида (1.2) решено для частных случаев переменных коэффициентов. В настоящей работе будет решено новое гиперсингулярное интегро-дифференциальное уравнение, важной особенностью которого является добавление в его левую часть также регулярных интегралов с искомой функцией. В интегральных уравнениях с сингулярными интегралами такие добавки (с которыми уравнения называются полными) совсем в немногих случаях позволяют провести исчерпывающее конструктивное исследование уравнения. Подобные же «полные гиперсингулярные уравнения» ранее, видимо, нигде не изучались.

2. Исходное уравнение и его связь с краевой задачей

Зададим также H-непрерывные функции $a(t)\neq 0$, $b(t)\neq 0$, $t\in L$. Предположим, что кривая L лежит целиком по одну из сторон относительно какой-либо прямой, проходящей через точку $z=0$. Будем решать уравнение

$$\sum_{k=0}^{n}\left[(a(t)a_k+b(t)b_k)\,\varphi^{(2k)}(t) + \frac{(a(t)a_k-b(t)b_k)(2k)!}{\pi i}\left(\int_L \frac{\varphi(\tau)d\tau}{(\tau-t)^{2k+1}} + \int_L \frac{\varphi(\tau)d\tau}{(\tau+t)^{2k+1}}\right)\right] = h(t), \quad t\in L. \tag{2.1}$$

Обозначим D_+^* и L^* область и кривую, симметричные относительно точки $z=0$ соответственно области D_+ и кривой L, а $D_* = D_-\setminus(L^*\bigcup D_+^*)$ – область с симметрией относительно точки $z=0$. Введем функции

$$\frac{1}{2\pi i}\int_L \frac{\varphi(\tau)d\tau}{\tau-z} + \frac{1}{2\pi i}\int_L \frac{\varphi(\tau)d\tau}{\tau+z} = \begin{cases} \Psi_+(z), & z\in D_+,\\ \Psi_*(z), & z\in D_*.\end{cases}$$

Функция $\Psi_+(z)$ будет аналитической в области D_+, а функция $\Psi_*(z)$ будет четной аналитической функцией в области D_*, причем $\Psi_*(\infty)=0$. Для предельных значений на кривой L этих функций и их производных можно записать формулы

$$\Psi_+^{(2k)}(t) = \frac{1}{2}\varphi^{(2k)}(t) + \frac{(2k)!}{2\pi i}\int_L \frac{\varphi(\tau)d\tau}{(\tau-t)^{2k+1}} + \frac{(2k)!}{2\pi i}\int_L \frac{\varphi(\tau)d\tau}{(\tau+t)^{2k+1}}, \tag{2.2}$$

$$\Psi_*^{(2k)}(t) = -\frac{1}{2}\varphi^{(2k)}(t) + \frac{(2k)!}{2\pi i}\int_L \frac{\varphi(\tau)d\tau}{(\tau-t)^{2k+1}} + \frac{(2k)!}{2\pi i}\int_L \frac{\varphi(\tau)d\tau}{(\tau+t)^{2k+1}}, \quad k=\overline{0,n}, \quad t\in L, \tag{2.3}$$

которые получаются после использования обобщенных формул Сохоцкого (1.1) для интеграла с $\tau-z$ и дифференцирования под знаком интеграла для интеграла с $\tau+z$. Вычитая и складывая формулы (2.2) и (2.3), получим

$$\varphi^{(2k)}(t) = \Psi_+^{(2k)}(t) - \Psi_*^{(2k)}(t), \tag{2.4}$$

$$\frac{(2k)!}{\pi i}\left(\int_L \frac{\varphi(\tau)d\tau}{(\tau-t)^{2k+1}} + \int_L \frac{\varphi(\tau)d\tau}{(\tau+t)^{2k+1}}\right) = \Psi_+^{(2k)}(t) + \Psi_*^{(2k)}(t), \quad k=\overline{0,n},$$

что позволяет придать уравнению (2.1) вид краевой задачи для аналитических функций

$$\sum_{k=0}^{n}\left[(a(t)a_k+b(t)b_k)\left(\Psi_+^{(2k)}(t)-\Psi_*^{(2k)}(t)\right)+(a(t)a_k-b(t)b_k)\left(\Psi_+^{(2k)}(t)+\Psi_*^{(2k)}(t)\right)\right] \;=\; h(t),$$

$t\in L$, или после очевидных упрощений

$$2a(t)\sum_{k=0}^{n}a_k\Psi_+^{(2k)}(t)-2b(t)\sum_{k=0}^{n}b_k\Psi_*^{(2k)}(t)=h(t),\quad t\in L. \tag{2.5}$$

Первая сумма в левой части равенства (2.5) является предельным значением на кривой L функции $Y_+(z)=\sum_{k=0}^{n}a_k\Psi_+^{(2k)}(z)$, аналитической в D_+. С учетом того, что производная четного порядка от четной аналитической функции сама является четной аналитической функцией, вторая сумма будет предельным значением четной функции $Y_*(z)=\sum\limits_{k=0}^{n}b_k\Psi_*^{(2k)}(z)$, аналитической в D_*. Краевую задачу (2.5) теперь можно записать в виде

$$Y_+(t)=\frac{b(t)}{a(t)}Y_*(t)+\frac{h(t)}{2a(t)},\quad t\in L. \tag{2.6}$$

Такая задача относится к типу смешанных краевых задач Римана–Карлемана. Решать ее следует с условием $Y_*(\infty)=0$, вытекающим из условия $\Psi_*(\infty)=0$. Укажем работу [6], где возникла и была решена близкая краевая задача: роль кривой L там играла прямая, параллельная действительной оси, а условие четности одной из искомых функций заменялось на некоторое аналогичное условие. Мы решаем задачу (2.6), следуя классической схеме Ф. Д. Гахова [7] решения задачи Римана и используя при этом подходящую факторизацию коэффициента $b(t)/a(t)$ задачи. В результате получится

$$Y_+(z)=X_+(z)\left(T_+(z)+P(z)\right);\quad Y_*(z)=X_*(z)\left(T_*(z)+P(z)\right),$$

где

$$X_+(z)=\exp\Gamma_+(z),\quad X_*(z)=\left(z^2-z_0^2\right)^{-\alpha}\exp\Gamma_*(z),\quad z_0\in D_+,\quad \alpha=\mathrm{Ind}_L\frac{b(t)}{a(t)},$$

$$\frac{1}{2\pi i}\int_L\frac{\ln\left[(\tau^2-z_0^2)^{-\alpha}b(\tau)/a(\tau)\right]d\tau}{\tau-z}+\frac{1}{2\pi i}\int_L\frac{\ln\left[(\tau^2-z_0^2)^{-\alpha}b(\tau)/a(\tau)\right]d\tau}{\tau+z}=\begin{cases}\Gamma_+(z), & z\in D_+,\\ \Gamma_*(z), & z\in D_*,\end{cases}$$

$$\frac{1}{4\pi i}\int_L\frac{h(\tau)d\tau}{a(\tau)X_+(\tau)(\tau-z)}+\frac{1}{4\pi i}\int_L\frac{h(\tau)d\tau}{a(\tau)X_+(\tau)(\tau+z)}=\begin{cases}T_+(z), & z\in D_+,\\ T_*(z), & z\in D_*,\end{cases}$$

$$P(z)=\begin{cases}\sum\limits_{k=0}^{\alpha-1}c_kz^{2k},\quad c_k\in\mathbb{C},\quad \text{если}\quad \alpha>0,\\ 0,\quad \text{если}\quad \alpha\leqslant 0.\end{cases}$$

При $\alpha\geqslant 0$ задача (2.6) разрешима безусловно, а при $\alpha<0$ для ее разрешимости необходимо и достаточно выполнение условий

$$\int_L\frac{h(\tau)\tau^{2k-1}d\tau}{a(\tau)X_+(\tau)}=0,\quad k=\overline{1,-\alpha}. \tag{2.7}$$

Предположим, что задача (2.6) разрешима, а ее решение найдено. Далее следует решать дифференциальные уравнения с постоянными коэффициентами

$$\sum_{k=0}^{n}a_k\Psi_+^{(2k)}(z)=Y_+(z),\quad z\in D_+, \tag{2.8}$$

$$\sum_{k=0}^{n}b_k\Psi_*^{(2k)}(z)=Y_*(z),\quad z\in D_*,\quad \Psi_*(\infty)=0, \tag{2.9}$$

и в случае нахождения их решений воспользоваться формулой (2.4) при $k = 0$:

$$\varphi(t) = \Psi_+(t) - \Psi_*(t), \quad t \in L. \tag{2.10}$$

3. Решение дифференциальных уравнений

Применяя к уравнению (2.8) метод вариации произвольных постоянных (напр., [8, с. 94]), получим его общее решение

$$\Psi_+(z) = \sum_{j=1}^{2n} f_j(z) \left(C_j^+ + \int_{z_1}^{z} \frac{W_j(\zeta)d\zeta}{W(\zeta)} \right). \tag{3.1}$$

В формуле (3.1) функции $f_j(z)$, явный вид которых известен и здесь не приводится, образуют фундаментальную систему решений соответствующего однородного уравнения, $C_j^+ \in \mathbb{C}$, $z_1 \in D_+$, $W(\zeta)$ – вронскиан функций $f_j(\zeta)$, $W_j(\zeta)$ – определители, полученные из $W(\zeta)$ заменой элементов j-го столбца на $0, 0, ..., 0, Y_+(\zeta)/a_n$, $j = \overline{1, 2n}$. Отметим, что в частности, когда корни $\lambda_1, \lambda_2, ..., \lambda_{2n}$ соответствующего характеристического уравнения являются однократными, формулу (3.1) можно упростить [3]:

$$\Psi_+(z) = \sum_{j=1}^{2n} e^{\lambda_j z} \left(C_j^+ - \frac{\int_{z_1}^{z} e^{-\lambda_j \zeta} Y_+(\zeta) d\zeta}{a_n \prod\limits_{\substack{m=1 \\ m \neq j}}^{2n} (\lambda_m - \lambda_j)} \right). \tag{3.2}$$

Интегрирование в формуле (3.1) производится по любой кривой в области D_+, соединяющей фиксированную точку z_1 с произвольной точкой z, и в силу конечности и односвязности области D_+ приводит к однозначной аналитической функции.

Уравнение (2.9) сложнее. Будем решать это уравнение также методом вариации произвольных постоянных. Так как такого отдельного метода для искомых четных функций не существует, то мы вначале найдем все решения и затем выберем из них четные. Очевидно, что корнями соответствующего характеристического уравнения

$$\sum_{k=0}^{n} b_k \mu^{2k} = 0 \tag{3.3}$$

будут пары $\pm\mu_m$ противоположных комплексных чисел некоторых кратностей k_m, $m = \overline{1, s}$, $\sum\limits_{m=1}^{s} k_m = n$. Знаки «$+$» и «$-$» в каждой паре выбираются произвольно. Фундаментальную систему решений однородного уравнения (2.9) возьмем в виде

$$z^l \operatorname{ch}(\mu_m z), \quad z^l \operatorname{sh}(\mu_m z), \quad l = \overline{0, k_m - 1}, \quad m = \overline{1, s}. \tag{3.4}$$

Если для некоторого m корнем характеристического уравнения будет число 0, то соответствующие функции в совокупности (3.4) следует заменить на z^l, $l = \overline{0, 2k_m - 1}$, где k_m – кратность каждого нуля в паре ± 0. Так как функции (3.4) линейно независимы и имеют на бесконечности существенную особую точку, то никакая их линейная комбинация с ненулевым набором коэффициентов, в том числе дающая четную функцию, не приведет к аналитическому на бесконечности решению с условием $\Psi_*(\infty) = 0$. Случай $\mu_m = 0$ не изменит этой ситуации, поскольку будет добавлять в упомянутой линейной комбинации полюсы разных порядков. Следовательно, однородное уравнение (2.9) имеет лишь решение $\Psi_*(z) \equiv 0$. Ненулевым решением уравнения (2.9) может быть только некоторое частное решение неоднородного уравнения. Это решение записывается по формуле

$$\Psi_*(z) = \sum_{j=1}^{2n} g_j(z) \left(C_j^* + \int_{0}^{z} \frac{U_j(\zeta)d\zeta}{U(\zeta)} \right), \tag{3.5}$$

аналогичной формуле (3.1). В формуле (3.5) $g_j(z)$ – это функции (3.4), как-либо занумерованные, C_j^* – подлежащие нахождению комплексные числа, $U(\zeta)$ – вронскиан функций $g_j(\zeta)$, $U_j(\zeta)$ – определители, полученные из $U(\zeta)$ заменой элементов j-го столбца на $0,0,...,0,Y_*(\zeta)/b_n$, $j=\overline{1,2n}$. Интегрирование в формуле (3.5) производится по любой кривой в области D_*, соединяющей точки 0 и z. Так как D_* является бесконечной двусвязной областью, то следует наложить условия, приводящие к однозначности функции $\Psi_*(z)$. Такими необходимыми и достаточными условиями будут, очевидно, условия выполнения равенств

$$\int_L \frac{U_j(t)dt}{U(t)} = 0, \quad j=\overline{1,2n}, \tag{3.6}$$

$$\int_{L^*} \frac{U_j(t)dt}{U(t)} = 0, \quad j=\overline{1,2n}. \tag{3.7}$$

Обоснуем, что для дальнейшего можно ограничиться условиями (3.6).

Лемма. *Условия* (3.7) *являются следствием условий* (3.6).

Доказательство. Для доказательства, а также и для последующих рассуждений, нумерацию функций (3.4) сделаем более определенной. Для четных значений l функции $z^l\operatorname{ch}(\mu_m z)$ будут четными, а функции $z^l\operatorname{sh}(\mu_m z)$ нечетными. Для нечетных значений l, наоборот, функции $z^l\operatorname{ch}(\mu_m z)$ будут нечетными, а функции $z^l\operatorname{sh}(\mu_m z)$ четными. При $\mu_m=0$ при четном l функции z^l будут четными, а при нечетном l нечетными. Всего в совокупности (3.4) n четных и n нечетных функций. Будем считать, что при нечетном j функции $g_j(z)$ являются четными, а при четном j нечетными.

Обозначим $U(z)=\det(a_{pq}(z))_{p,q=1}^{2n}$. Первая строка вронскиана $U(z)$ будет состоять из чередующихся четных и нечетных функций, начиная с четной функции $a_{11}(z)=g_1(z)$. Так как дифференцирование меняет характер четности функций, то вторая строка вронскиана $U(z)$ будет также состоять из чередующихся четных и нечетных функций, начиная с нечетной функции $a_{21}(z)= = g_1'(z)$. Аналогично чередуются четные и нечетные функции в последующих строках. Расположение четных и нечетных функций во вронскиане $U(z)$ будет похоже на расположение черных и белых клеток на шахматной доске размером $2n\times 2n$. Очевидно, что сумма $p+q$ индексов для четной функции $a_{pq}(z)$ будет четным числом, а для нечетной функции $a_{pq}(z)$ – нечетным числом. Вронскиан $U(z)$ есть сумма слагаемых, имеющих с точностью до знака вид

$$a_{1,m_1}(z)\, a_{2,m_2}(z)\cdot\ldots\cdot a_{2n,m_{2n}}(z), \tag{3.8}$$

где $m_1,m_2,...,m_{2n}$ – перестановка чисел $1,2,...,2n$. Если в произведении (3.8) было бы нечетное число нечетных функций, то сумма всех индексов всех множителей также была бы нечетным числом. Но это невозможно, так как $1+m_1+2+m_2+...+2n+m_{2n}=2(1+2+...+2n)$ – четное число. Следовательно, в произведении (3.8) четное число нечетных функций (а также некоторое количество четных функций), поэтому произведение является четной функцией. Отсюда вытекает четность $U(z)$.

Определитель $U_j(z)$ лишь j-м столбцом отличается от определителя $U(z)$. Если j – четное число, то в j-м столбце в качестве последнего элемента $a_{2n,j}(z)$ будет четная функция $Y_*(z)/b_n$. Нули, которыми являются остальные элементы этого столбца, можно считать чередующимися четными и нечетными функциями, начиная с нечетной функции $a_{1j}(z)$. В результате расположение четных и нечетных функций в определителе $U_j(z)$ будет таким же, как и в $U(z)$, поэтому определитель $U_j(z)$ будет четной функцией.

Если j – нечетное число, то можно считать, что лишь в j-м столбце определителя $U_j(z)$ характер четности элементов иной, чем в $U(z)$. Тогда один из множителей в каждом из слагаемых, в сумму которых раскладывается определитель $U_j(z)$, изменит характер четности по сравнению с аналогичным множителем в $U(z)$. Вместе с тем изменится характер четности и самого определителя $U_j(z)$ по сравнению с $U(z)$ – определитель $U_j(z)$ будет нечетным.

Теперь понятно, что в интегралах из равенств (3.6) все подынтегральные функции являются четными или нечетными. Сделав в этих интегралах замену $t=-t_1$, придем (с точностью, возможно, до знака) к интегралам из равенств (3.7), поэтому из равенств (3.6) вытекают равенства (3.7). □

Предполагая в дальнейшем условия (3.6) выполненными, вернемся к формуле (3.5), придав ей вид

$$\Psi_*(z)=\sum_{j=1}^{n}g_{2j-1}(z)\left(C^*_{2j-1}+\int\limits_0^z\frac{U_{2j-1}(\zeta)d\zeta}{U(\zeta)}\right)+$$
$$+\sum_{j=1}^{n}g_{2j}(z)\left(C^*_{2j}+\int\limits_0^z\frac{U_{2j}(\zeta)d\zeta}{U(\zeta)}\right). \tag{3.9}$$

Поскольку $\frac{U_{2j-1}(\zeta)}{U(\zeta)}$ являются нечетными функциями, то первообразные $\int\limits_0^z\frac{U_{2j-1}(\zeta)d\zeta}{U(\zeta)}$ будут четными функциями, и тогда первая сумма в правой части формулы (3.9) даст четную функцию. Функции $\frac{U_{2j}(\zeta)}{U(\zeta)}$ являются четными, поэтому первообразные $\int\limits_0^z\frac{U_{2j}(\zeta)d\zeta}{U(\zeta)}$ являются нечетными функциями. Отсюда вытекает, что для четности второй суммы в (3.9) (а вместе с ней для четности функции $\Psi_*(z)$) следует в дальнейшем брать $C^*_{2j}=0,\ j=\overline{1,n}$.

Осталось добиться аналитичности функции $\Psi_*(z)$ на бесконечности вместе с условием $\Psi_*(\infty)=0$. Для этого разложим функцию $\Psi_*(z)$ на бесконечности в ряд Лорана и приравняем к нулю соответствующие коэффициенты. В результате придем к следующей бесконечной линейной алгебраической системе для нахождения постоянных $C^*_{2j-1},\ j=\overline{1,n}$:

$$\sum_{j=1}^{n}\alpha_{\gamma j}C^*_{2j-1}=\beta_\gamma,\quad \gamma=1,2,3,..., \tag{3.10}$$

где

$$\alpha_{\gamma j}=\frac{1}{2\pi i}\int\limits_{|t|=\rho}\frac{g_{2j-1}(t)dt}{t^{2\gamma-1}},$$

$$\beta_\gamma=-\frac{1}{2\pi i}\sum_{j=1}^{2n}\int\limits_{|t|=\rho}\frac{g_j(t)}{t^{2\gamma-1}}\int\limits_0^t\frac{U_j(\zeta)d\zeta}{U(\zeta)},$$

ρ – достаточно большое положительное число.

4. Формулировка результата. Пример

Проведенные рассуждения и формула (2.10) позволяют сформулировать результат.

Теорема. *Для разрешимости уравнения* (2.1) *необходимо и достаточно, чтобы были верны равенства* (2.7) *при* $\alpha<0$, *равенства* (3.6) *и была совместна система* (3.10). *Если эти условия выполняются, то искомая функция находится по формуле*

$$\varphi(t)=\sum_{j=1}^{2n}f_j(t)\left(C^+_j+\int\limits_{z_1}^t\frac{W_j(\zeta)d\zeta}{W(\zeta)}\right)-$$
$$-\sum_{j=1}^{n}\left(g_{2j-1}(t)\left(C^*_{2j-1}+\int\limits_0^t\frac{U_{2j-1}(\zeta d\zeta)}{U(\zeta)}\right)+g_{2j}(t)\int\limits_0^t\frac{U_{2j}(\zeta)d\zeta}{U(\zeta)}\right),$$

где C^+_j *– произвольные комплексные постоянные,* $j=\overline{1,2n}$, *а комплексные постоянные* C^*_{2j-1} *являются решением системы* (3.10), $j=\overline{1,n}$.

Рассмотрим пример:

$$(-1+4i)\varphi(t)+3\varphi''(t)+\frac{1+4i}{\pi i}\left(\int\limits_{|\tau-1|=0,5}\frac{\varphi(\tau)d\tau}{\tau-t}+\int\limits_{|\tau-1|=0,5}\frac{\varphi(\tau)d\tau}{\tau+t}\right)+$$

$$+\frac{2}{\pi i}\left(\int\limits_{|\tau-1|=0,5}\frac{\varphi(\tau)d\tau}{(\tau-t)^3}+\int\limits_{|\tau-1|=0,5}\frac{\varphi(\tau)d\tau}{(\tau+t)^3}\right)=\frac{1+2t-t^2}{(t-1)^3},\quad |t-1|=0,5.$$

Так выглядит уравнение (2.1) при $n=1$, $a(t)=b(t)=1$, $a_0=4i$, $b_0=-1$, $a_1=2$, $b_1=1$, $h(t)=$ $=\frac{1+2t-t^2}{(t-1)^3}$. Окружность $|t-1|=0,5$ расположена по одну из сторон, например, мнимой оси. Краевая задача (2.6) принимает вид задачи о скачке

$$Y_+(t)-Y_*(t)=\frac{1+2t-t^2}{2(t-1)^3},\quad |t-1|=0,5,$$

которая безусловно разрешима и имеет единственное решение. Несложно найти представление

$$\frac{1+2t-t^2}{2(t-1)^3}=\frac{1-2t-t^2}{2(t+1)^3}-\frac{t^4-8t^2-1}{(t^2-1)^3},$$

так что, очевидно, будет получаться

$$Y_+(z)=\frac{1-2z-z^2}{2(z+1)^3},\quad Y_*(z)=\frac{z^4-8z^2-1}{(z^2-1)^3}.$$

Уравнение (2.8) принимает вид

$$2\Psi''_+(z)+4i\Psi_+(z)=\frac{1-2z-z^2}{2(z+1)^3},\quad |z-1|<0,5.$$

Взяв функции $e^{(1-i)z}$, $e^{(-1+i)z}$ в качестве фундаментальной системы решений соответствующего однородного уравнения, по формуле (3.2) запишем общее решение

$$\Psi_+(z)=e^{(1-i)z}\left(C_1^+ +\frac{1+i}{16}\int\limits_1^z\frac{e^{(-1+i)\zeta}(1-2\zeta-\zeta^2)}{(\zeta+1)^3}d\zeta\right)+$$

$$+e^{(-1+i)z}\left(C_2^+ -\frac{1+i}{16}\int\limits_1^z\frac{e^{(1-i)\zeta}(1-2\zeta-\zeta^2)}{(\zeta+1)^3}d\zeta\right).$$

Уравнение (2.9) в примере принимает вид

$$\Psi''_*(z)-\Psi_*(z)=\frac{z^4-8z^2-1}{(z^2-1)^3},\quad z\in\{z:\ |z-1|>0,5\}\bigcap\{z:\ |z+1|>0,5\}.$$

Взяв функции $\operatorname{ch}z$, $\operatorname{sh}z$ в качестве фундаментальной системы решений однородного уравнения, получим по формуле (3.9)

$$\Psi_*(z)=\operatorname{ch}z\left(C_1^*-\int\limits_0^z\frac{\zeta^4-8\zeta^2-1}{(\zeta^2-1)^3}\operatorname{sh}\zeta d\zeta\right)+\operatorname{sh}z\left(C_2^*+\int\limits_0^z\frac{\zeta^4-8\zeta^2-1}{(\zeta^2-1)^3}\operatorname{ch}\zeta d\zeta\right),$$

где в дальнейшем следует взять $C_2^*=0$. Последние два интеграла удается вычислить:

$$\int\limits_0^z\frac{\zeta^4-8\zeta^2-1}{(\zeta^2-1)^3}\operatorname{sh}\zeta d\zeta=\frac{1}{z^2-1}\operatorname{ch}z+\frac{2z}{(z^2-1)^2}\operatorname{sh}z+1,$$

$$\int\limits_0^z\frac{\zeta^4-8\zeta^2-1}{(\zeta^2-1)^3}\operatorname{ch}\zeta d\zeta=\frac{1}{z^2-1}\operatorname{sh}z+\frac{2z}{(z^2-1)^2}\operatorname{ch}z.$$

В результате получится

$$\Psi_*(z)=(C_1^*-1)\operatorname{ch}z+\frac{1}{1-z^2}.$$

Очевидно, что лишь при $C_1^*=1$ функция $\Psi_*(z)=\frac{1}{1-z^2}$ будет аналитической на бесконечности с условием $\Psi_*(z)=0$. Благодаря простому выражению для функции $\Psi_*(z)$ записывать и анализировать систему (3.10) в примере не требуется. Окончательная формула решения примера имеет вид

$$\varphi(t)=e^{(1-i)t}\left(C_1^+ + \frac{1+i}{16}\int\limits_1^t \frac{e^{(-1+i)\zeta}(1-2\zeta-\zeta^2)}{(\zeta+1)^3}d\zeta\right)+$$

$$+e^{(-1+i)t}\left(C_2^+ - \frac{1+i}{16}\int\limits_1^t \frac{e^{(1-i)\zeta}(1-2\zeta-\zeta^2)}{(\zeta+1)^3}d\zeta\right)-\frac{1}{1-t^2}.$$

5. Заключение

Подсчитаем еще число произвольных комплексных постоянных, входящих в решение исходного уравнения (2.1) в случае его разрешимости. Краевая задача (2.6) даст $\max(0,\alpha)$ постоянных c_k. Решение уравнения (2.8) добавит $2n$ постоянных C_j^+. При $\alpha>0$ условия (3.6) в развернутом виде будут представлять собой систему линейных алгебраических уравнений для нахождения постоянных c_k, входящих в решение задачи (2.6). Вид этой системы может быть легко записан и здесь не приводится. После решения системы число произвольных постоянных уменьшится на ранг r ее матрицы. Будем считать $r=0$ при $\alpha\leqslant 0$. Укажем матрицу системы (3.10) в случае однократных корней уравнения (3.3). При этом $g_{2j-1}(z)=\operatorname{ch}(\mu_j z),\ \ j=\overline{1,n}$, так что матрица примет вид

$$\begin{pmatrix} 1 & 1 & \dots & 1\\ \frac{\mu_1^2}{2!} & \frac{\mu_2^2}{2!} & \dots & \frac{\mu_n^2}{2!}\\ \frac{\mu_1^4}{4!} & \frac{\mu_2^4}{4!} & \dots & \frac{\mu_n^4}{4!}\\ \dots & \dots & \dots & \dots\\ \frac{\mu_1^{2k}}{(2k)!} & \frac{\mu_2^{2k}}{(2k)!} & \dots & \frac{\mu_n^{2k}}{(2k)!}\\ \dots & \dots & \dots & \dots \end{pmatrix},$$

т. е. будет бесконечной матрицей типа матрицы Вандермонда для попарно различных чисел $\mu_1^2,\mu_2^2,\dots,\mu_n^2$. Ранг такой матрицы равен n, поэтому однородная система (3.10) имеет лишь нулевое решение. Следовательно, при решении системы не появится новых произвольных постоянных. Произвольные постоянные в правых частях системы могут использоваться для достижения ее совместности, из-за чего их количество уменьшится на некоторое число r_1. Всего при однократных корнях уравнения (3.3) будет, таким образом, $\max(0,\alpha)+2n-r-r_1$ произвольных постоянных. Указать ранг матрицы системы (3.10) при корнях уравнения (3.3) любой кратности представляется автору несколько затруднительным. По-видимому, он снова будет равен n, поэтому число произвольных постоянных в решении исходного уравнения останется тем же.

Отметим, что H-непрерывность функций $\Psi_+(t)$ и $\Psi_*(t)$ вместе с их производными до порядка $2n$ (а тогда согласно формуле (2.10) и искомой функции $\varphi(t)$) обосновывается несложно и вполне аналогично [4].

Литература

1. *Зверович Э. И.* Обобщение формул Сохоцкого // Весці Нац. акад. навук Беларусі. Сер. фіз.-мат. навук. 2012. № 2. С. 24–28.

2. *Зверович Э. И.* Решение гиперсингулярного интегро-дифференциального уравнения с постоянными коэффициентами // Докл. Нац. акад. наук Беларуси. 2010. Т. 54, № 6. С. 5–8.

3. *Шилин А. П.* Гиперсингулярные интегро-дифференциальные уравнения со степенными множителями в коэффициентах // Журн. Белорус. гос. ун-та. Математика. Информатика. 2019. № 3. С. 48–56. https://doi.org/10.33581/2520-6508-2019-3-48-56

4. *Шилин А. П.* Гиперсингулярное интегро-дифференциальное уравнение с линейными функциями в коэффициентах // Весці Нац. акад. навук Беларусі. Сер. фіз.-мат. навук. 2022. Т. 58,

№ 4. С. 358–369. https://doi.org/10.29235/1561-2430-2022-58-4-358-369

5. *Шилин А. П.* Гиперсингулярное интегро-дифференциальное уравнение с квадратичными функциями в коэффициентах // Весцi Нац. акад. навук Беларусi. Сер. фiз.-мат. навук. 2024. Т. 60, № 2. С. 117–131. https://doi.org/10.29235/1561-2430-2024-60-2-117-131

6. *Черский Ю. И.* Интегральное уравнение типа свертки с экспонентой в ядре // Дифференц. уравнения. 1997. Т. 33, № 11. С. 1566–1567.

7. *Гахов Ф. Д.* Краевые задачи. М.: Наука, 1977.

8. *Камке Э.* Справочник по обыкновенным дифференциальным уравнениям, 6-е изд. СПб.: Лань, 2003.

References

1. Zverovich E. I. Generalization of Sokhotsky formulas. *Proceedings of the National Academy of Sciences of Belarus. Physics and Mathematics Series*, 2012, no. 2, pp. 24–28 (in Russian).

2. Zverovich E. I. Solution of the hypersingular integro-differential equation with constant coefficients. *Doklady of the National Academy of Sciences of Belarus*, 2010, vol. 54, no. 6, pp. 5–8 (in Russian).

3. Shilin A. P. Hypersingular integro-differential equations with power factors in coefficients. *Journal of the Belarusian State University. Mathematics and Informatics*, 2019, no. 3, pp. 48–56 (in Russian). https://doi.org/10.33581/2520-6508-2019-3-48-56

4. Shilin A. P. A hypersingular integro-differential equation with linear functions in coefficients. *Proceedings of the National Academy of Sciences of Belarus. Physics and Mathematics Series*, 2022, vol. 58, no. 4, pp. 358–369 (in Russian). https://doi.org/10.29235/1561-2430-2022-58-4-358-369

5. Shilin A. P. Hypersingular integro-differential equation with quadratic functions in coefficients. *Proceedings of the National Academy of Sciences of Belarus. Physics and Mathematics Series*, 2024, vol. 60, no. 2, pp. 117–131 (in Russian). https://doi.org/10.29235/1561-2430-2024-60-2-117-131

6. Chersky Yu. I. An integral equation of convolution type with an exponent in the kernel. *Differential Equations*, 1997, vol. 33, no. 11, pp. 1566–1567 (in Russian).

7. Gakhov F. D. *Boundary value problems*. Moscow, Nauka, 1977 (in Russian).

8. Kamke E. *Handbook of ordinary differential equations*. 6^{th} ed. Saint Peterburg, Lan’, 2003 (in Russian).

Национальная академия наук Беларуси
Труды Института математики НАН Беларуси. 2024. Том 32. № 2. С. 82–92

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ И ЧИСЛЕННЫЕ МЕТОДЫ
MATHEMATICAL MODELING AND NUMERICAL METHODS

УДК 619.63

О РЕАЛИЗАЦИИ СПЕКТРАЛЬНОГО МЕТОДА ЧЕБЫШЁВА ДЛЯ ДВУМЕРНЫХ ЭЛЛИПТИЧЕСКИХ УРАВНЕНИЙ СО СМЕШАННЫМИ ПРОИЗВОДНЫМИ

В. М. Волков, Дун Цзинхуэй

Белорусский государственный университет, Минск, Беларусь
e-mail: v.volkov@tut.by

Поступила: 13.08.2024 **Исправлена:** 15.11.2024 **Принята:** 12.12.2024

Ключевые слова: спектральный метод Чебышёва, эллиптические уравнения со смешанными производными, стабилизированный метод би-сопряженных градиентов, метод переменных направлений, алгоритм Бартелса–Стюарта.

Аннотация. Рассмотрены вопросы построения численных алгоритмов на основе спектрального метода Чебышёва для приближенного решения эллиптических уравнений со смешанными производными в прямоугольной области с однородными краевыми условиями Дирихле. Для реализации спектрального метода использован стабилизированный метод би-сопряженных градиентов с переобусловливателями в виде разностных или спектральных аналогов оператора Лапласа. Проведено сравнение эффективности обработки переобусловлевателя с применением итерационного метода переменных направлений и алгоритма Бартелса–Стюарта. Представленные результаты показывают, что рассмотренные алгоритмы демонстрируют вычислительные характеристики, сопоставимые по времени вычислений на сетках одинаковой размерности с характеристиками разностных методов, однако многократно превосходят последние по точности в случае достаточно гладких решений.

ON THE IMPLEMENTATION OF THE CHEBYSHEV SPECTRAL METHOD FOR TWO-DIMENSIONAL ELLIPTIC EQUATIONS WITH MIXED DERIVATIVES

V. M. Volkov, Dong JingHui

Belarusian State University, Minsk, Belarus
e-mail: v.volkov@tut.by

Received: 13.08.2024 **Revised:** 15.11.2024 **Accepted:** 12.12.2024

Keywords: Chebyshev spectral method, Elliptic equations with mixed derivatives, biconjugate gradient stabilized method, alternating directions implicit method, Bartels–Stewart algorithm.

Abstract. The issues of constructing numerical algorithms based on the Chebyshev spectral method for approximate solution of elliptic equations with mixed derivatives in a rectangular domain with homogeneous Dirichlet boundary conditions are considered. To implement the spectral method, the biconjugate gradients stabilized method with preconditioners in the form of finite difference or spectral analogs of the Laplace operator is used. A comparison of the efficiency of processing the preconditioner using the iterative method of alternating directions and the Bartels–Stewart algorithm is carried out. The presented results show that the considered algorithms demonstrate computational characteristics comparable in computation time on grids of the same dimension with the characteristics of difference methods, but they are many times superior to the latter in accuracy in the case of sufficiently smooth solutions.

1. Введение

Методы конечных разностей и конечных элементов характеризуются полиномиальной скоростью сходимости погрешности δ относительно шагов дискретизации h: $\delta(h) = O(h^p)$, где, как правило, $p \leqslant 2$ [1; 2]. При этом, как в одномерном, так и в многомерных случаях, существуют прямые и итерационные алгоритмы реализации этих методов, обеспечивающие оптимальную вычислительную сложность, для которой характерен линейный (или близкий к линейному) рост вычислительных затрат в зависимости от общего количества узлов сетки N [3–5]. Для спектральных методов на основе полиномов Чебышёва [5; 6] при достаточной гладкости решения достигается экспоненциальная скорость сходимости $\delta(h) = O(h^N)$ [6, p. 25; 7, p. 48], однако их реализация более трудоемка. Атуальной задачей является разработка численных алгоритмов реализации спектральных методов, вычислительная сложность которых была бы в определенном смысле близка к оптимальной, как в методах конечных разностей и конечных элементов.

К настоящему времени достигнуты определенные успехи в разработке алгоритмов реализации спектральных методов для многомерных уравнений Пуассона. В случае двумерных задач в прямоугольной области дискретная спектральная модель, как и в разностном случае, сводится к системе матричных уравнений Ляпунова. Среди методов численного решения данного класса задач (см., например, обзор [8]) можно отметить методы переменных направлений [8–11], алгоритм Бателса–Стюарта [12], а также итерационные методы в подпространствах Крылова с переобуславливателем на основе метода переменных направлений [13; 14]. Сопоставимые по вычислительным затратам алгоритмы для двумерных эллиптических задач с переменными коэффициентами предложены в работе [15].

В настоящей работе дано обобщение итерационной техники реализации спектрального метода Чебышёва, развитой в работе [15], на случай уравнений со смешанными производными, что имеет важное значение при решении эллиптических задач в случаях неоднородных анизотропных сред и в областях сложной геометрии путем их конформного отображения в прямоугольник [16]. На примере модельной двумерной задачи с известными аналитическими решениями различной степени гладкости продемостририровано, что, несмотря на более высокую вычислительную сложность, реальные вычислительные затраты предложенного итерационного алгоритма реализации спектрального метода Чебышёва на сетках одинакового размера сопоставимы по времени с аналогичным алгоритмом реализации разностного метода. При этом точность спектрального метода в случае решений, имеющих четыре и более непрерывных производных, превосходит точность разностного метода при равных временных затратах на реализацию данных методов.

2. Постановка задачи. Спектральный метод коллокации

Рассмотрим задачу Дирихле для эллиптического уравнения со смешанными производными в двумерной прямоугольной области:

$$\frac{\partial}{\partial x}\sigma_{xx}\frac{\partial u}{\partial x} + \frac{\partial}{\partial y}\sigma_{yy}\frac{\partial u}{\partial y} + \frac{\partial}{\partial y}\sigma_{yx}\frac{\partial u}{\partial x} + \frac{\partial}{\partial x}\sigma_{xy}\frac{\partial u}{\partial y} = f(x,y), \tag{2.1}$$

$$(x,y) \in \Omega = [-1,1]\times[-1,1], \;\; u(\pm 1, y) = u(x, \pm 1) = 0. \tag{2.2}$$

Здесь коэффициенты $\sigma_{xx}, \sigma_{yy}, \sigma_{xy} = \sigma_{yx}$ – достаточно гладкие функции переменных x, y, удовлетворяющие условию эллиптичности [1, c. 34]:

$$\sigma_{xy}^2(x,y) = \sigma_{yx}^2(x,y) < \sigma_{xx}(x,y)\sigma_{yy}(x,y), \quad (x,y) \in \Omega. \tag{2.3}$$

Разностные методы для решения задачи (2.1)–(2.3) на квадратной сетке размерности $n \times n$ приводят к системе линейных алгебраических уравнений с семидиагональной матрицей, размера $N \times N$, где $N = (n-2)^2$ – количество внутрених узлов сетки [1]. Для реализации систем разностных уравнений могут эффективно использоваться многосеточные итерационные методы [4] или итерационные методы в пространствах Крылова со спектрально оптимальным переобусловливателем [5].

Спектральные методы коллокации на основе полиномов Чебышёва требуют соответствующую неравномерную сетку [6]:

$$\omega = \left\{ (x_k, y_m),\; x_k = \cos\left(\frac{\pi(k-1)}{n_x - 1}\right),\; y_m = \cos\left(\frac{\pi(m-1)}{n_y - 1}\right), k = \overline{1, n_x},\; m = \overline{1, n_y} \right\}. \tag{2.4}$$

Далее, для простоты, полагаем $n_x = n_y = n$.

Для перехода от дифференциальной задачи (2.1)–(2.3) к дискретной спектральной модели достаточно заменить производные в уравнении (2.1) на соответствующие матрицы спектрального дифференцирования с учетом краевых условий задачи [6]. В итоге задача сводится к системе алгебраических уравнений

$$A_s\overline{U} = F, \tag{2.5}$$

матрица которой имеет вид

$$A_s = A_{xx} + A_{yy} + A_{xy} + A_{yx},$$

где

$$A_{xx} = \mathrm{diag}(\sigma_{xx})(I\otimes\overline{D^2}) + \mathrm{diag}(\partial_x\sigma_{xx})(I\otimes\overline{D})],\ \ A_{yy} = \mathrm{diag}(\sigma_{yy})(\overline{D^2}\otimes I) + \mathrm{diag}(\partial_y\sigma_{yy})(\overline{D}\otimes I),$$

$$A_{yx} = \mathrm{diag}(\sigma_{yx})(\overline{D}\otimes I)(I\otimes\overline{D}) + \mathrm{diag}(\partial_y\sigma_{xy})(I\otimes\overline{D}),$$

$$A_{xy} = \mathrm{diag}(\sigma_{xy})(I\otimes\overline{D})(\overline{D}\otimes I) + \mathrm{diag}(\partial_x\sigma_{xy})(\overline{D}\otimes I).$$

Здесь $\overline{D}$, $\overline{D^2} \in R^{(n-2)\times(n-2)}$ – матрицы дифференцирования Чебышёва первого и второго порядка соответственно (см., например, [5, p. 53]), в которых для учета однородных краевых условий удалены первые и последние строки и столбцы, $\mathrm{diag}(\sigma_{xx})$, $\mathrm{diag}(\sigma_{xy})$, $\mathrm{diag}(\partial_x\sigma_{xx})$ и т. д. – диагональные матрицы коэффициентов задачи и их частных производных во внутренних узлах сетки, $\otimes$ – символ кронекеровского произведения матриц, $I \in R^{(n-2)\times(n-2)}$ – единичная матрица. Использование недивергентной формы дифференциальных операторов в данной дискретной модели позволяет упростить учет однородных краевых условий (2.2).

Несложно убедиться, что при наличии в уравнении смешанных производных системная матрица A_s дискретной спектральной модели (2.5) является полной, что практически исключает применение традиционных итерационных алгоритмов решения таких систем. Например, выполняемое на каждой итерации непосредственное умножение такой матрицы $A_s \in R^{N\times N}$ на вектор приближенного решения задачи $\overline{U} \in R^N$, $N = (n-2)^2$, требует $2N^2$ арифмерических операций, что на порядок превосходит оптимальную вычислительную сложность алгоритма. В силу этого, по аналогии с [15], воспользуемся алгоритмом вычисления произведения матрицы A_s на вектор приближенного решения в виде суммы таких произведений для каждой составляющей дискретного аналога дифференциального оператора в отдельности:

$$A_s\overline{U} = \Re\left\{D\left(S_{xx}*(DU)\right) + \left(S_{yy}*\left(UD^T\right)\right)D^T + \left(S_{yx}*(DU)\right)D^T + D\left(S_{xy}*\left(UD^T\right)\right)\right\}. \tag{2.6}$$

Здесь $D \in R^{n\times n}$ – матрица спектрального дифференцирования Чебышёва, S_{xx}, S_{yy}, S_{xy}, $S_{yx} \in R^{n\times n}$ – двумерные массивы (матрицы) значений соответствующих коэффициентов уравнения (2.1) в точках сетки (2.4), символ $*$ обозначает поэлементное умножение массивов, а символ $\Re$ – удаление граничных точек массива и последующую трансформацию его в вектор столбец ($R^{(n-2)\times(n-2)} \to R^{(n-2)^2}$). Вычислительные затраты на реализацию (2.6) составляют $16n^3 + 4n^2 \simeq O\left(N\sqrt{N}\right)$ арифметических операций, что существенно меньше, чем при непосредственном умножении $A_s\overline{U}$. При отсутствии смешанных производных формула (2.6) сохраняет эффективность при вычислении произведения системной матрицы на вектор приближенного решения [15].

Таким образом, спектральный метод Чебышёва для задачи (2.1)–(2.3) при стандартной формулировке дискретной модели приводит к системе линейных алгебраических уравнений (2.5) с полной матрицей $A_s \in R^{N\times N}$, однако при итерационной реализации данной системы нам не требуется явного вычисления самой матрицы A_s, а достаточно реализовать лишь умножение данной матрицы на вектор решения согласно формуле (2.6).

3. Итерационная реализация спектрального метода

Для реализации спектрального метода Чебышёва воспользуемся стабилизированной версией итерационного метода би-сопряженных градиентов (Bi-Conjugate Gradient – biCG) с переобусловливателем [17]. Выбор данного метода обусловлен его высокой эффективностью при решении больших систем линейных алгебраических уравнений с несимметричной матрицей [13; 18], а также наличием доступной программной реализации в виде функции bicgstab математического пакета MATLAB.

Используем двухэтапное переобусловливание системы. На первом этапе задействуем диагональный переобусловливатель, элементы которого для каждого узла сетки (x_k, y_m) вычисляются следующим образом:

$$S = \{\sigma_{j,j}\} = \mathrm{Sp}\begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{pmatrix}, \tag{3.1}$$

или

$$S = \{\sigma_{j,j}\} = \mathrm{Sp}\left(M_{km}^{-1}\begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{pmatrix} M_{km}\right), \tag{3.2}$$

где $M_{km} \in R^{2\times 2}$ – матрицы преобразования подобия, приводящие матрицы коэффициентов в соответствующих узлах сетки к диагональному виду, а компоненты σ_{xx}, σ_{yy}, σ_{xy}, σ_{yx} – значения соответствующих коэффициентов в точках (x_k, y_m), для которых $j = k + m(k-1)$.

Элементы матрицы S согласно (3.2) численно равны собственным значениям матриц коэффициентов, что не требует существенных дополнительных вычислительных затрат по сравнению с (3.1), и в экспериментах показывает несколько большую эффективность в снижении количества итераций метода biCG. Данный переобусловливатель может быть применен непосредственно к системе алгебраических уравнений (2.5) до начала итерационного процесса.

На втором этапе применим переобусловливатель в виде дискретного аналога оператора Лапласса, построенного на основе конечно-разностной или спектральной аппроксимации с учетом краевых условий задачи. Для обработки данного переобусловливателя на каждой итерации biCG требуется решение системы алгебраических уравнений, которую удобно представить в виде матричной системы Ляпунова:

$$PU + UP^T = F, \tag{3.3}$$

где $U, P \in R^{(n-2)\times(n-2)}$ – массив решения и матрица переобусловливателя. Для конечно-разностного переобусловливателя спектрального метода Чебышёва $P = P_{FD}$, P_{FD} – трехдиагональная матрица второй разностной производной на неравномерной сетке вида (2.4), или $P = \overline{D^2}$, $\overline{D^2}$ – спектральная матрица дифференцирования Чебышёва второго порядка, в которой удалены первые и последние строки и столбцы. Для переобусловливателя разностного метода $P = \overline{P}_{FD}$, $\overline{P}_{FD}$ – трехдиагональная матрица второй разностной производной на равномерной сетке. Для решения матричной системы Ляпунова (3.3) можно эффективно использовать итерационный метод переменных направлений с оптимальным набором итерационных параметров [3], а также алгоритм Бартелса–Стюарта [17]. Эффективность данных методов в случае двумерных эллиптических задач с переменными коэффициентами показана в работе [15]. Ниже проанализированы возможности столь же эффективного использования данного подхода для уравнений со смешанными производными.

4. Результаты численных экспериментов

Рассмотрим модельную задачу, на примере которой сравним эффективность изложенного выше спектрального метода Чебышёва с аналогичным подходом, основанным на методе конечных разностей. В качестве разностного метода, использованного в численных экспериментах, взяты разностные схемы [18; 19]. Реализация разностного метода проводилась аналогично, как и для спектрального метода, с помощью стабилизированного итерационного алгоритма би-сопряженных градиентов с переобусловливателем. Критерий остановки итераций определялся значением относительной нормы невязки $||r||/||f|| < \varepsilon = 10^{-10}$.

Вначале рассмотрим случай гладких коэффициентов:

$$\begin{aligned}&\sigma_{xx}(x,y)=\sigma_{yy}(x,y)=\alpha+\cos(\pi x)+\cos(\pi y),\\&\sigma_{xy}(x,y)=\sigma_{yx}(x,y)=\alpha^{-1}\left(\cos(\pi x)-\cos(\pi y)\right),\ \alpha>1.\end{aligned}\tag{4.1}$$

Правая часть уравнения такова, что бесконечно дифференцируемое аналитическое решение задачи имеет вид

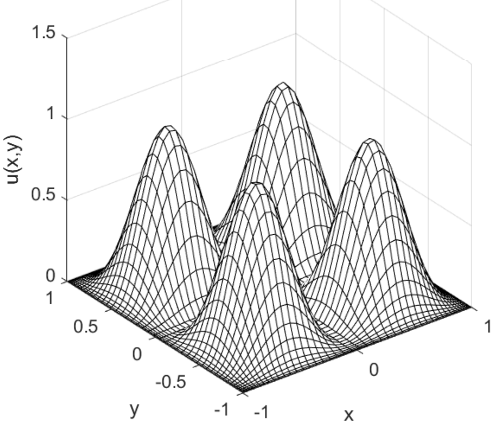
$$u(x,y)=\sin^2(m_x\pi x)\sin^2(m_y\pi y).\tag{4.2}$$

Для определенности полагаем $\alpha=3$, $m_x=m_y=1$. Для рассмотренного случая максимальный коэффициент анизотропии $\xi\simeq 1,57$, а коэффициент неоднородности $\psi=5$, где ξ и ψ определяются отношениями:

$$\xi=\max_{\omega}\frac{\max\left(\lambda(s)\right)}{\min\left(\lambda(s)\right)},\quad \psi=\max_{\omega}\frac{\max\sqrt{\lambda_1^2(s)+\lambda_2^2(s)}}{\min\sqrt{\lambda_1^2(s)+\lambda_2^2(s)}},\quad s=\begin{pmatrix}\sigma_{xx}&\sigma_{xy}\\\sigma_{yx}&\sigma_{yy}\end{pmatrix},$$

где $\lambda(s)=\{\lambda_1(s),\lambda_2(s)\}$ – собственные значения матрицы коэффициентов задачи s.

Вид решения задачи и зависимость относительной среднеквадратичной погрешности спектрального и разностного методов, $\delta=\|U-u\|/\|u\|$, представлены на рис. 1. Как видно из рисунка, для спектрального метода предельная точность, $\delta<10^{-10}$, ограниченная преимущественно вычислительной погрешностью, достигается при наличии примерно 15 узлов сетки на один период волнообразного гладкого решения. В то же время разностный метод второго порядка точности на такой сетке имеет относительную погрешность всего лишь $\delta\simeq 10^{-2}$.



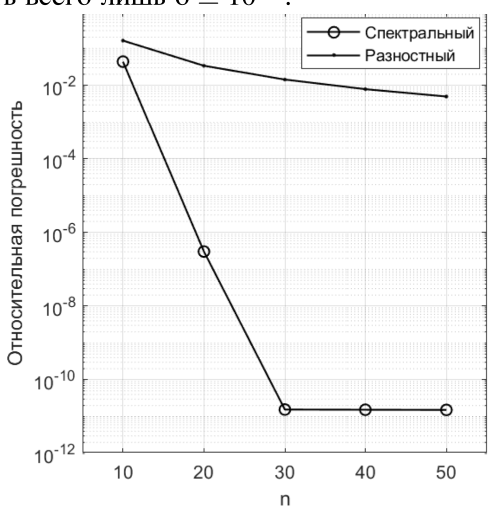


Рис. 1. Вид решения задач (2.1), (2.2), (4.1), (4.2) (*а*); зависимость относительной погрешности приближенного решения спектрального и разностного методов от количества узлов сетки n (*б*)

Далее остановимся более подробно на сравнении эффективности реализации спектрального и разностного методов. Прежде всего отметим некоторые особенности использования для реализации дискретных моделей двух вложенных итерационных методов, в качестве которых роль внешней процедуры играет стабилизированный метод би-сопряженных градиентов, а внутренний итерационный процесс для обработки переобусловливателя основан на методе переменных направлений. За основу алгоритма стабилизированного метода би-сопряженных градиентов взята соответствующая функция цифрового пакета MATLAB bicgstab, в которой для случая спектрального метода модифицирована процедура умножения системной матрицы на вектор приближенного решения задачи согласно (3.1), а также процедура обработки переобусловливателя. В частности, используя матричную форму двумерной задачи с постоянными коэффициентами (3.3), векторизованный алгоритм итерационного метода переменных направлений имеет вид:

$$S_k^- U_{k+1/2} = U_k S_k^{+^T} + F, \quad U_{k+1} S_k^{-^T} = S_k^+ U_{k+1/2} + F, \quad k = 0, 2, \ldots K, \tag{4.3}$$

где $S_k^{\pm} = \omega_{k+1} E \pm P$, $E \in R^{(n-2)\times(n-2)}$ – единичная матрица, P – трехдиагональная матрица второй разностной производной на сетке (2.4) в случае спектрального метода или равномерной сетке для разностного метода, ω_{k+1}, $k = \overline{1, K}$, – оптимальный набор итерационных параметров метода переменных направлений (см., например, [3, с. 438]):

$$\omega_k = \chi_k = \Delta \mathrm{dn}\left(\frac{2k-1}{2K} F(\pi/2, \eta'), \eta'\right). \tag{4.4}$$

Здесь $\mathrm{dn}(x, \eta)$ – эллиптическая функция аргумента x и модуля η, $F(\pi/2, \eta)$ – полный нормальный эллиптический интеграл Лежандра первого рода, $\eta = \delta/\Delta$, $\eta' = \sqrt{1-\eta^2}$, δ и Δ – нижняя и верхняя границы спектра матрицы P соответственно.

Минимальное количество итераций метода переменных направлений (4.3)–(4.4) для достижения точности ε удовлетворяет оценке [3, с. 439] :

$$K \geqslant K_0 = \pi^{-1} \ln\left(4\eta^{-1}\right) \ln\left(4\varepsilon^{-1}\right). \tag{4.5}$$

Согласно (4.5) при $n = 300$ и $\varepsilon = 10^{-10}$ минимальное количество итераций $K_0 = 49$. Тем не менее, как показано в работе [13] на примере численного решения обобщенных матричных уравнений Ляпунова, и при реализации спектрального метода Чебышёва для двумерного эллиптического уравнения с переменными коэффициентами [15], 4-5 внутренних итераций метода переменных направлений оказывается достаточным для сходимости внешнего итерационного процесса в подпространствах Крылова, и при этом обеспечиваются минимальные вычислительные затраты на реализацию алгоритма в целом. Согласно результатам численных экспериментов для задачи (2.1), (2.2), (4.1), (4.2), оптимальное число внутренних итераций метода переменных направлений, обеспечивающее минимизацию вычислительных затрат для разностного метода $K_{\mathrm{opt}} = 4$, а для спектрального метода $K_{\mathrm{opt}} = 6$. Полученные эмпирические оптимальные значения количества внутренних итераций метода переменных направлений практически не зависят ни от размерности сетки в пределах $n \leqslant 300$, ни от характеристик неоднородности и анизотропии коэффициентов задачи ξ и ψ.

Количество итераций метода би-сопряженных градиентов с переобусловливателем на основе разностного аналога оператора Лапласа, обрабатываемого методом переменных направлений (4.3), (4.4), $K = K_{\mathrm{opt}}$, и удельное время итерационной реализации спектрального и разностного методов для задачи (2.1), (2.2), (4.1), (4.2) в зависимости от количества узлов сетки n представлены на рис. 2. Отметим, что вычислительные затраты на одну итерацию имеют порядок $O(n^2)$ для разностного метода, а для спектрального метода – $O(n^3)$. Поскольку рост количества итераций с возрастанием размерности сетки в пределах $n = 50 \div 300$ практически отсутствует, то вычислительная сложность рассматриваемых алгоритмов в целом имеет тот же порядок, что и для одной итерации.

Обработка переобуславливателя на основе разностного аналога оператора Лапласа при фиксированном количестве внутренних итераций имеет вычислительную сложность $O(n^2)$. При размерности сетки в пределах $50 \leqslant n \leqslant 300$, как при реализации разностного, так и в случае спектрального методов, на данную процедуру приходится больше половины всех вычислений. В силу этого, при существенных отличиях в асимптотических оценках вычислительной сложности спектрального и разностного алгоритмов в целом, наблюдаются не столь значительные отличия фактических удельных вычислительных затрат на их реализацию, поскольку наибольшие затраты приходятся на реализацию переобусловливателя и разностный метод выигрывает в скорости за счет меньшего числа итераций.

Анализируя результаты численных экспериментов, представленные на рис. 2, можно отметить, что рост удельных вычислительных затрат заметно отстает от асимптотических оценок вычислительной сложности как для разностного, так и для спектрального методов. Это можно объяснить увеличением производительности процессора при выполнении векторных операций с увеличением размерности массивов данных, а также снижением относительной доли накладных расходов. С этим связано также аномальное падение удельного времени выполнения спектрального алгоритма в при $n = 60 \div 120$ и разностного метода во всем диапазоне рассмотренных размерностей сетки.

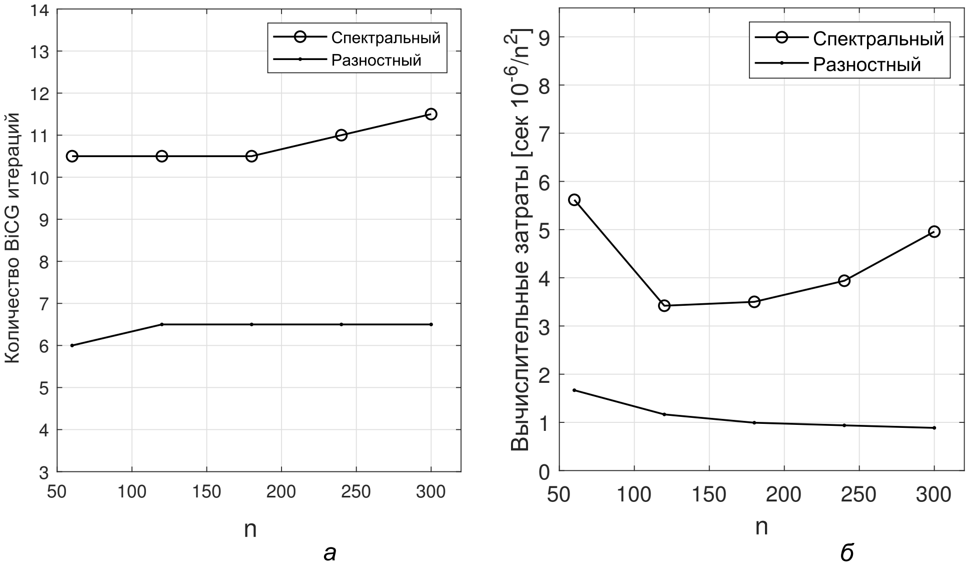


Рис. 2. Зависимость количества итераций метода би-сопряженных градиентов с разностным переобусловливателем на основе МПН от количества узлов сетки n (*а*); удельное время итерационной реализации спектрального и разностного методов в зависимости от количества узлов сетки n (*б*)

Следует отметить также, что стабильная работа рассмотренных алгоритмов наблюдается лишь при размерности сетки $n \leqslant 300$. В случае $n \geqslant 350$ стабилизированный итерационный метод би-сопряженных градиентов при реализации спектральной версии алгоритма не сходится даже при количестве внутренних итераций метода переменных направлений $K \geqslant 50$, что достаточно для достижения точности $\varepsilon = 10^{-10}$ при обработке переобусловливателя.

Рассмотрим альтернативный подход, основанный на обработке переобусловливателя с использованием алгоритма Бартелса–Стюарта [12], реализованного в соответствующей функции lyap пакета MATLAB. Важно отметить, что при решении матричного уравнения вида (3.3) около 75 % вычислительных затрат алгоритма Бартелса–Стюарта приходится на разложение Шура матрицы P (случай не симметричной матрицы P):

$$P = QRQ^T, \tag{4.6}$$

где Q – унитарная матрица, R – верхняя треугольная матрица. В силу этого непосредственное использование функции lyap на каждой итерации метода би-сопряженных градиентов представляется не рациональным. Для повышения эффективности алгоритма матричные компоненты разложения Шура (4.6) Q и R могут быть предвычислены до начала итераций, что позволяет многократно сократить затраты на обработку переобусловливателя.

На рис. 3 представлены результаты численных экспериментов для модельной задачи (2.1), (2.2), (4.1), (4.2), полученные посредством спектрального и разностного методов с описанной выше альтернативной техникой обработки переобусловливателя. Удельные вычислительные затраты для разностного метода практически не отличаются количественно от соответствующих результатов, представленных на рис. 2. Вместе с тем модифицированный разностный алгоритм проигрывает качественно, поскольку использование метода Бартелса–Стюарта повышает его вычислительную сложность, что подтверждается ростом удельных вычислительных затрат при возрастании размерности сетки n (рис. 3, *б*).

Для спектрального метода модифицированный алгоритм имеет в целом ту же вычислительную сложность, что и при использовании разностного переобусловливателя, обрабатываемого методом переменных направлений. При этом уменьшается общее число итераций и полностью устраняется их зависимость от размерности сетки, что позволяет в итоге примерно в полтора раза сократить удельные вычислительные затраты по сравнению со случаем разностного переобусловливателя, обрабатываемого методом переменных направлений (см. рис. 2 и 3) . Количество итераций метода

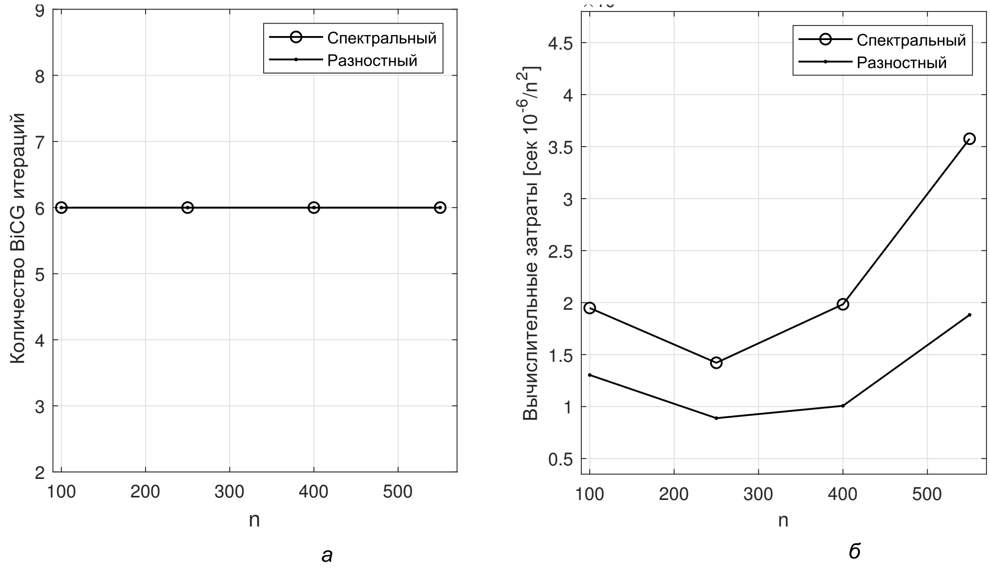


Рис. 3. Зависимость количества итераций метода би-сопряженных градиентов с обработкой переобусловливателя на основе алгоритма Бартелса–Стюарта от количества узлов сетки n (a); удельное время итерационной реализации спектрального и разностного методов в зависимости от количества узлов сетки n ($б$)

би-сопряженных градиентов для спектральной и разностной версий алгоритма с соответствующими переобусловливателями совпадает. Наиболее значимое преимущество модифицированного спектрального алгоритма состоит в стабильной сходимости итераций при достаточно большой размерности сетки. Относительно эффективности спектрального и разностного методов для задач с достаточно гладким решением докажем следующее.

Утверждение. *Рассмотрим два метода решения двумерной эллиптической задачи* (2.1)–(2.3)*, имеющей достаточную гладкость коэффициентов и решения:*

I. Разностный метод второго порядка точности, для которого погрешность $\delta_1 = c_1 n^{-2}$*, а вычислительная сложность* $M_1 = Q_1 n^2$*;*

II. Спектральный метод, имеющий экспоненциальную скорость сходимости, $\delta_2 = \exp(-c_2 n)$*, и вычислительную сложность* $M_2 = Q_2 n^3$*.*

Здесь c_1, c_2, Q_1, Q_2 *– положительные постоянные, не зависящие от* n*.*

Для достижения достаточно малой погрешности приближенного решения, $\delta < \varepsilon_0$*, вычислительные затраты разностного метода I превосходят соответствующие вычислительные затраты спектрального метода II.*

Доказательство. Размерности сетки $n = n_1$ и $n = n_2$, обеспечивающие соответственно одинаковую погрешность методов I и II, удовлетворяют приближенному равенству $n_1 \simeq \sqrt{c_1}\exp(c_2 n_2/2)$. Несложно убедиться, что для положительной, монотонно убывающей функции $\theta(n_2) = M_2(n_2)/M_1(n_1(n_2))$, описывающей отношение вычислительных затрат спектрального и разностного методов для достижения фиксированной погрешности, имеем

$$\theta(n) = \frac{Q_2 n^2}{Q_1\sqrt{c_1^3}\exp(3c_2 n/2)}, \quad \lim_{n\to\infty}\theta(n) = 0. \tag{4.7}$$

В силу (4.7) всегда найдется такое значение n_0, для которого $\theta(n \geqslant n_0) < 1$, при этом $\varepsilon_0 = c_1/n_0^2$, что и требовалось доказать.

Полученные оценки согласуются с данными численных экспериментов. В частности, для дифференциальных задач, решение которых $u \in C^\infty$, спектральный метод Чебышёва превосходит в эффективности метод конечных разностей при любых требованиях точности, т. е. $\forall n > 2$, $\theta(n) < 1$ (см. рис. 4).

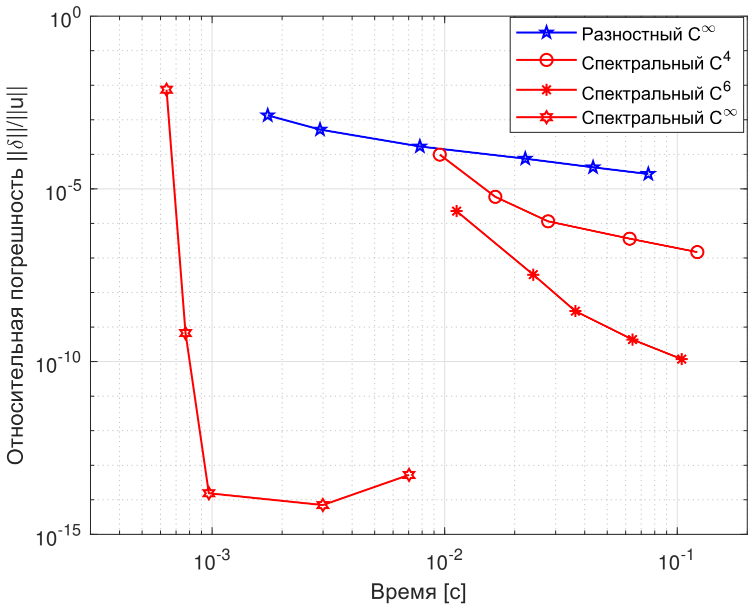


Рис. 4. Зависимости погрешностей спектрального и разностного методов от вычислительных затрат с ростом размерности сетки для решений, обладающих различной степенью гладкости

Для сравнительного анализа эффективности спектрального и разностного методов при пониженных требованиях к гладкости решения рассмотрим пример задачи (2.1), (2.2), в которой коэффициенты заданы кусочно-постоянными функциями:

$$\sigma_{xy}(x,y) = \sigma_{yx}(x,y) = 0, \sigma_{xx}(x,y) = \sigma_{yy}(x,y) = \begin{cases} 1+d, & xy \geqslant 0, \\ 1, & xy < 0, \end{cases}$$

а правая часть уравнения (2.1) такова, что ее решение имеет вид

$$u(x,y) = \begin{cases} (\sin(\pi x)\sin(\pi y))^p, & xy > 0, \\ (\sin(\pi x)\sin(\pi y))^{p+d} & xy \leqslant 0. \end{cases} \tag{4.8}$$

Для $p \geqslant 1$, в случае $d = 0$ и $d = 1$ решение (4.8) $u(x,u) \in C^\infty(\Omega)$ и $u(x,u) \in C^{p-1}(\Omega)$ соответственно.

На рис. 4 приведены результаты численных экспериментов, показывающие скорость сходимости спектрального и разностного методов в зависимости от вычислительных затрат (времени реализации задачи) для решений, обладающих различной гладкостью. Характеристики разностного метода представлены только для решений класса C^∞ в силу того, что при понижении гладкости решения вплоть до класса C^4 скорость сходимости разностного решения практически не меняется. Из рисунка видно, что для рассмотренных классов решений, имеющих четыре и более непрерывных производных спектральный метод демонстрирует заметное преимущество в сравнении с разностным, и данное преимущество прогрессирует с ростом гладкости решения задачи.

5. Заключение

В контексте общей дискуссии о разработке быстрых алгоритмов реализации спектральных методов решения двумерных краевых задач для эллиптических уравнений, проведем сравнительный анализ представленных выше результатов с аналогичными исследованиями других авторов.

За редким исключением, большинство исследований посвящено разработке быстрого спектрального солвера для уравнения Пуассона в прямоугольной области. В этом случае дискретная модель дифференциальной задачи как в разностном, так и в спектральном подходах сводится к решению матричной системы Ляпунова (Сильвестра–Ляпунова). Среди последних результатов можно отметить работу [11], где на основе факторизованной схемы переменных направлений предложен эффективный алгоритм реализации спектрального метода Чебышёва для уравнения Пуассона с близкой к оптимальной вычислительной сложностью порядка $O(n^2 \log n)$. Доступность программной реализации данного алгоритма позволила провести непосредственное сравнение его эффективности с эффективностью представленных выше алгоритмов для двумерных эллиптических задач общего вида на одинаковой вычислительной платформе. При $n = 100 \div 600$ время решения задачи для уравнения Пуассона с использованием методики [11] примерно в $7 \div 8$ раз меньше, нежели рассмотренный выше алгоритм для задач общего вида (см. рис. 3). Однако методика [11] исключает непосредственное ее использование для решения эллиптических задач общего вида (2.1)–(2.3). Если рассматривать методику [11] в качестве алгоритма обработки спектрального переобусловливателя для итерационных методов в подпространствах Крылова, то важно отметить дополнительные вычислительные затраты, возникающие при решении эллиптических задач общего вида. Во-первых, на контуре внешнего итерационного процесса возникнет необходимость вычислять произведение системной матрицы на вектор приближенного решения, что в подобных алгоритмах составляет около 30 % общих вычислительных затрат даже без учета смешанных производных (см., например [15]). Во-вторых, общее время реализации алгоритма возрастает пропорционально количеству внешних итераций по сравнению с временем реализации одной итерации.

Таким образом, представленный в работе итерационный алгоритм реализации спектрального метода Чебышёва для двумерных эллиптических уравнений со смешанными производными имеет вычислительную сложность $O(n^3)$, которая, с одной стороны, представляется субоптимальной в концептуальном смысле [5]. Однако, с другой стороны, возможности улучшения эффективности универсального двумерного спектрального итерационного солвера имеют, по-видимому, принципиальные ограничения, связанные с необходимостью умножения матрицы дифференцирования Чебышёва на матрицу (двумерный массив) приближенного решения задач, что требует порядка $O(n^3)$ арифметических операций. Использование для этих целей алгоритма быстрого дискретного преобразования Фурье позволяет снизить асимптотическую оценку вычислительных затрат до уровня $O(n^2 \log n)$, однако, это дает возможность получить фактическое преимущество лишь при $n \geqslant 1000$.

Авторы выражают благодарность Е. В. Прокониной, предоставившей программную реализацию алгоритма расчета системной матрицы для разностного метода решения двумерного уравнения Пуассона со смешанными производными.

Литература

1. *Самарский А. А., Андреев В. Б.* Разностные методы для эллиптических уравнений. М.: Наука, 1976.

2. *Ciarlet P. G.* The finite element method for elliptic problems // Society for Industrial and Applied Mathematics, 2002.

3. *Самарский А. А., Николаев Е. С.* Методы решения сеточных уравнений. М.: Наука, 1978.

4. *Hackbusch W.* Multi-grid methods and applications. Springer Science. Business Media, 2013. Vol. 4.

5. *D'yakonov E. G.* Optimization in solving elliptic problems. CRC Press, 2018.

6. *Boyd J. P.* Chebyshev and Fourier spectral methods. Courier Corporation, 2001.

7. *Trefethen L. N.* Spectral Methods in MATLAB. Philadelphia: SIAM, 2000.

8. *Simoncini V.* Computational methods for linear matrix equations // SIAM REVIEW. 2016. Vol. 58, N 3. P. 377–441.

9. *Peaceman D. W., Rachford Jr H. H.* The numerical solution of parabolic and elliptic differential equations // Journal of the Society for Industrial and Applied Mathematics. 1955. Vol. 3, N 1. P. 28–41.

10. *Benner P., Li R. C., Truhar N.* On the ADI method for Sylvester equations // Journal of Computational and Applied Mathematics. 2009. Vol. 233, N 4. P. 1035–1045.

11. *Fortunato D., Townsend A.* Fast Poisson solvers for spectral methods // IMA Journal of Numerical Analysis. 2020. Vol. 40, N 3. P. 1994–2018.

12. *Bartels R. H., Stewart G. W.* Algorithm 432 [C2]: Solution of the matrix equation $AX + XB = C$ [F4] // Communications of the ACM. 1972. Vol. 15, N 9. P. 820–826.

13. *Damm T.* Direct methods and ADI preconditioned Krylov subspace methods for generalized Lyapunov equations // Numerical Linear Algebra with Applications. 2008. Vol. 15, N 9. P. 853–871.

14. *Jbilou K.* ADI preconditioned Krylov methods for large Lyapunov matrix equations // Linear algebra and its applications. 2010. Vol. 432, N 10. P. 2473–2485.

15. *Волков В. М., Качаловская Е. И.* Итерационная реализация спектрального метода Че-

бышёва для двумерных эллиптических уравнений с переменными коэффициентами // Журнал Белорусского государственного университета. Математика. Информатика. 2023. № 3. С. 53–62.

16. *Orszag S. A.* Spectral methods for problems in complex geometrics // Numerical methods for partial differential equations. Academic Press, 1979. P. 273–305.

17. *Van der Vorst H. A.* Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems // SIAM Journal on scientific and Statistical Computing. 1992. Vol. 13, N 2. P. 631–644.

18. *Волков В. М., Проконина Е. В.* Разностные схемы и итерационные методы для многомерных эллиптических уравнений со смешанными производными // Весці Нацыянальнай акадэміі навук Беларусі. Серыя фізіка-матэматычных навук. 2019. Т. 54, № 4. С. 454–459.

19. *Самарский А. А., Мажукин В. И., Матус П. П., Шишкин Г. И.* Монотонные разностные схемы для уравнений со смешанными производными // Математическое моделирование. 2001. Т. 13, № 2. С. 17–26.

References

1. Samarski A. A., Andreev V. B. *Finite-difference methods for elliptical equations.* Moscow, Nauka, 1976 (in Russian).

2. Ciarlet P. G. *Finite element method for elliptic problems.* Society for Industrial and Applied Mathematics, 2002.

3. Samarski A. A., Nikolaev E. S. *Methods for Solving Grid Equations.* Moscow, Nauka, 1978 (in Russian).

4. Hackbusch W. *Multi-grid methods and applications*. Springer Science, Business Media, 2013, vol. 4.

5. D’yakonov E. G. *Optimization in solving elliptic problems.* CRC Press, 2018.

6. Boyd J. P. *Chebyshev and Fourier spectral methods*. Courier Corporation, 2001.

7. Trefethen L. N.*Spectral Methods in MATLAB.* Philadelphia, SIAM, 2000.

8. Simoncini V. *Computational methods for linear matrix equations.* SIAM REVIEW, 2016, vol. 58, no. 3, pp. 377–441.

9. Peaceman D. W., Rachford Jr H. H. The numerical solution of parabolic and elliptic differential equations. *Journal of the Society for Industrial and Applied Mathematics,* 1955, vol. 3, no. 1, pp. 28–41.

10. Benner P., Li R. C., Truhar N. On the ADI method for Sylvester equations. *Journal of Computational and Applied Mathematics,* 2009, vol. 233, no. 4, pp. 1035–1045.

11. Fortunato D., Townsend A. Fast Poisson solvers for spectral methods. *IMA Journal of Numerical Analysis,* 2020, vol. 40, no. 3, pp. 1994–2018.

12. Bartels R. H., Stewart G. W. Algorithm 432 [C2]: Solution of the matrix equation $AX + XB = C$ [F4]. *Communications of the ACM,* 1972, vol. 15, no. 9, pp. 820–826.

13. Damm T. Direct methods and ADI preconditioned Krylov subspace methods for generalized Lyapunov equations. *Numerical Linear Algebra with Applications,* 2008, vol. 15, no. 9, pp. 853–871.

14. Jbilou K. ADI preconditioned Krylov methods for large Lyapunov matrix equations. *Linear algebra and its applications,* 2010, vol. 432, no. 10, pp. 2473–2485.

15. Volkov V. M., Kachalouskaya E. I. An iterative Chebyshev spectral solver for two-dimensional elliptic equations with variable coefficients. *Journal of the Belarusian State University. Mathematics and Informatics,* 2023, iss. 3, pp. 53–62 (in Russian).

16. Orszag S. A. Spectral methods for problems in complex geometrics. *Numerical methods for partial differential equations.* Academic Press, 1979, pp. 273–305.

17. Van der Vorst H. A. Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing,* 1992, vol. 13, no. 2, pp. 631–644.

18. Volkov V. M., Prakonina A. U. Finite-difference schemes and iterative methods for multidimensional elliptic equations with mixed derivatives. *Proceedings of the National Academy of Sciences of Belarus. Physics and Mathematics series,* 2019, vol. 54, iss. 4, pp. 454–459 (in Russian).

19. Samarskii A. A., Mazhukin V. I., Matus P. P., Shishkin G. I. Monotone difference schemes for equations with mixed derivatives. *Mathematical Modeling,* 2001, vol. 13, iss. 2, pp. 17–26.

Национальная академия наук Беларуси
Труды Института математики НАН Беларуси. 2024. Том 32. № 2. С. 93–96

КРАТКИЕ СООБЩЕНИЯ
BRIEF COMMUNICATIONS

УДК 511.643

АЛЬТЕРНАТИВНОЕ ПОСТРОЕНИЕ ТЕОРИИ ОПРЕДЕЛИТЕЛЕЙ

С. М. Агеев, Е. С. Агеева

Белорусский государственный университет, Минск, Беларусь
e-mail: ageev_sergei@inbox.ru

Поступила: 05.07.2024 **Исправлена:** 05.09.2024 **Принята:** 12.12.2024

Ключевые слова: теорема о равноправности, мультипликативное свойство определителя, обобщенная теорема Лапласа, теорема Бине–Коши.

Аннотация. Непосредственно, без привлечения четностей перестановок и приведения матриц к ступенчатому виду, устанавливается эквивалентность разложения определителя по любой строчке и любому столбцу. С помощью этого существенно упрощается оставшаяся часть теории определителей: мультипликативное свойство определителя, обобщенная теорема Лапласа, теорема Бине–Коши и др.

ALTERNATIVE CONSTRUCTION OF THE DETERMINANT THEORY

S. M. Ageev, H. S. Ageeva

Belarusian State University, Minsk, Belarus
e-mail: ageev_sergei@inbox.ru

Received: 05.07.2024 **Revised:** 05.09.2024 **Accepted:** 12.12.2024

Keywords: the equality theorem, the multiplicative property of determinants, the generalized Laplace expansion and Cauchy–Binet formula.

Abstract. We establish in a direct way, without involving the sigh function of permutations and matrice reducing to echelon form, the equivalence of the expansion of determinant along any row and any column. On base of this the rest of the theory of determinants is significantly simplified: determinant multiplicativity, the generalized Laplace expansion and Cauchy–Binet formula and so on.

1. Введение

Определитель матрицы входит в число основных инструментов современной математики. Несмотря на то что он представляет собой всего лишь замысловатую[1] формулу от элементов матрицы, язык теории определителей оказался весьма гибким и удобным, в его терминах выражаются многие другие важные понятия и результаты.

С момента возникновения понятия определителя матрицы A было разработано несколько подходов к его обоснованию. Однако каждый из них, строгий и последовательный, доставляет большие трудности для изучающего предмет впервые.

Самое устоявшееся (хотя далеко не самое простое) определение $\det(A)$ состоит в использовании формулы полного развертывания. В качестве основного инструмента здесь выступают *подстановки* и их *четности* – понятия, стоящие несколько в стороне от теории матриц. Кроме того, при обосновании их свойств легко увязнуть в многочисленных деталях.

[1] Удивительно то, что похожая (но более простая) формула для перманента матрицы играет существенно менее значимую роль в математике.

В двух других часто используемых подходах – аксиоматическом и индуктивном, главный акцент делается на приведение матриц к ступенчатому виду. Мало того, что это вновь приводит к рассуждениям технического характера (что в целом затуманивает суть дела), при этом не удается доказать важное мультипликативное свойство для определителей над произвольными коммутативными кольцами.

Вниманию читателя предлагается иное построение теории определителей. С одной стороны, это позволит коротко и прозрачно обосновать все свойства определителей[2]. С другой стороны, удается избежать использования четности перестановок[3] и приведения матриц к ступенчатому виду (что хорошо, по крайней мере, на первых порах). Ключевыми здесь являются следующие два факта:

1) теорема о равноправности, утверждающая, что в некотором естественном смысле разложения D_i и D^j определителя по i-й строчке и по j-му столбцу, соответственно, коммутируют. Этот факт (в пределах одной страницы) непосредственно устанавливает равносильность разложения определителя по любой строчке и любому столбцу;

2) доказательство теоремы о мультипликативном свойстве определителя $|AB| = |A||B|$, сводящее его к случаю, когда один из сомножителей совпадает с матрицей, получающейся перестановкой строчек единичной матрицы. Мало того, что оно кажется проще имеющихся доказательств, фактически теми же самыми рассуждениями можно установить классические теоремы Лапласа и Бине–Коши (последнюю, ввиду ее трудоемкости, обычно изучают факультативно).

После получения этих двух ключевых фактов доказательство всех стандартных свойств определителей становится простым и кратким. Отметим, что при нашем подходе четность подстановки и ее свойства (а также формула полного развертывания определителей) играют второстепенную роль (и могут быть изложены в любой подходящий момент после теоремы об определителе произведения матриц).

2. Теорема о равноправности

Определитель $\det(B)$ матрицы $B = \|b\|$ порядка 1 равен элементу b. Если уже введено понятие определителя $\det(B)$ для всех матриц порядка $k, k < n$, то определитель $\det(A)$ матрицы A порядка $n \geqslant 2$ вводится индуктивно, следующим образом:

Определение 2.1. Определителем $\det(A)$ называется одна из знакочередующихся сумм

$$D_i(A) = \sum_{\beta=1}^{n} (-1)^{i+\beta} a_{i\beta} \det(M_{i\beta}) \ \text{ или } D^j(A) = \sum_{\alpha=1}^{n} (-1)^{\alpha+j} a_{\alpha j} \det(M_{\alpha j}),$$

(которые назовем *разложением* $\det A$ *по i-й строчке* и *разложением* $\det A$ *по j-му столбцу* соответственно). Здесь через M_{ij} обозначена матрица порядка $n-1$, получающаяся из A вычеркиванием ее i-строчки и j-столбца; M_{ij} называется *дополнительным минором элемента a_{ij} матрицы A* (или кратко, *минором A*).

Для установления корректности определения 2.1 достаточно проверить следующее утверждение

Теорема 2.2. *Для любой квадратной матрицы A порядка n и для любой i-й строчки и любого j-го столбца справедливо равенство $D_i(A) = D^j(A)$.*

Отсюда следует центральный результат о равноправности: вычислить определитель $\det(A)$ можно, разлагая его по любой строчке или столбцу.

Доказательство. Для $n = 2$ проверка теоремы осуществляется непосредственно, поэтому далее предполагаем, что $n \geqslant 3$.

Поскольку $D_i(A)$ и $D^j(A)$ содержат общее слагаемое $(-1)^{i+j} a_{ij} \det(M_{ij})$, то все сводится к проверке равенства $D'_i(A) = {}'D^j(A)$, где

$$D'_i(A) = D_i(A) - (-1)^{i+j} a_{ij} \det(M_{ij}) = \sum_{\beta \neq j}^{n} (-1)^{i+\beta} a_{i\beta} \det(M_{i\beta}) \text{ и} \tag{2.1}$$

[2] Недавним примером того, как классический результат получает существенное упрощение, служит теорема Жордана о разложении линейного оператора, см. [1] и [2].

[3] Наоборот, свойства четности будут легко выведены из мультипликативного свойства.

$$'D^j(A) = D^j(A) - (-1)^{i+j} a_{ij} \det(M_{ij}) = \sum_{\alpha\neq i}^{n} (-1)^{\alpha+j} a_{\alpha j} \det(M_{\alpha j}). \tag{2.2}$$

Для того чтобы разложить определитель $\det(M_{i\beta})$ минора[4] $M_{i\beta}$ порядка $n-1$ из (2.1) по его j-му столбцу, введем следующее обозначение: если $a \neq b \in \bar{n} = \{1,2,\ldots,n\}$, то $[a \mid b\rangle$ есть порядковый номер числа b в $\bar{n}$ после удаления из него a. Почти очевидно, что

(*) если целые числа $k \neq l$ взяты из $\bar{n}$, то число $k + l + [k \mid l\rangle + [l \mid k\rangle$ нечетно.

Теперь определитель $\det(M_{i\beta}), \beta \neq j$, из (2.1) разложим[5] по **j-му столбцу** $M_{i\beta}$:

$$\det(M_{i\beta}) = \sum_{\alpha\neq i}^{n} (-1)^{[i|\alpha\rangle + [\beta|j\rangle}\, a_{\alpha j} \det(M_{i\alpha;j\beta}). \tag{2.3}$$

Здесь через $M_{i\alpha;j\beta}$ обозначен минор порядка $n-2 \geqslant 1$, получающийся из A вычеркиванием ее строчек с номерами i и α, а также ее j-го и β-го столбцов; показатель степени в формуле (2.3) есть сумма номера строчки и номера столбца элемента $a_{\alpha j}$ в $M_{i\beta}$, т. е. $[i \mid \alpha\rangle + [\beta \mid j\rangle$.

После подстановки $\det(M_{i\beta})$ из (2.3) в $D_i'(A)$ получим

$$D_i'(A) = \sum_{\beta\neq j}^{n} \Big(\sum_{\alpha\neq i}^{n} (-1)^{i+\beta+[i|\alpha\rangle+[\beta|j\rangle} a_{i\beta} a_{\alpha j} \det(M_{i\alpha;\beta j}) \Big). \tag{2.4}$$

Аналогично определитель $\det(M_{\alpha j}), \alpha \neq i$, из (2.2) разложим **по i-й строчке**[6] $M_{\alpha j}$:

$$\det(M_{\alpha j}) = \sum_{\beta\neq j}^{n} (-1)^{[\alpha|i\rangle + [j|\beta\rangle} a_{i\beta} \det(M_{i\alpha;\beta j}). \tag{2.5}$$

При подстановке $\det(M_{\alpha j})$ из формулы (2.5) в $'D^j(A)$ получим

$$'D^j(A) = \sum_{\alpha\neq i}^{n} \Big(\sum_{\beta\neq j}^{n} (-1)^{\alpha+j+[\alpha|i\rangle+[j|\beta\rangle}\, a_{\alpha j} a_{i\beta} \det(M_{i\alpha;\beta j}) \Big). \tag{2.6}$$

Наконец, непосредственно сравним формулы (2.4) и (2.6), предварительно воспользовавшись заменой порядка суммирования в (2.6). Из свойства (*) легко следует, что

$$\big(\alpha + j + [\alpha \mid i\rangle + [j \mid \beta\rangle\big) + \big(i + \beta + [i \mid \alpha\rangle + [\beta \mid j\rangle\big) = \big(\alpha + i + [\alpha \mid i\rangle + [i \mid \alpha\rangle\big) + \big(j + \beta + [j \mid \beta\rangle + [\beta \mid j\rangle\big)$$

есть число четное. Поэтому $(-1)^{\alpha+j+[\alpha|i\rangle+[j|\beta\rangle} = (-1)^{i+\beta+[i|\alpha\rangle+[\beta|j\rangle}$ и все соответствующие слагаемые в формулах для $'D^j(A)$ и $D_i'(A)$ совпадают – $D_i'(A) = {}'D^j(A)$. □

3. Мультипликативное свойство определителя

Теорема 3.1. $\det(AB) = \det(A)\det(B)$ *для квадратных матриц A и B порядка n.*

Доказательство. Фиксируем B и рассмотрим множество $\mathcal{A} = \mathcal{A}_B$ всех матриц A, для которых теорема верна. Ясно, что

(**) $\mathcal{A}$ содержит любую матрицу с нулевой строчкой или с двумя равными строчками; матрица A при перестановке местами двух ее строчек или при умножении какой-либо ее строчки на скаляр $\alpha \neq 0$ продолжает принадлежать (или не принадлежать) семейству $\mathcal{A}$.

Проведем доказательство теоремы индукцией по числу $\sharp A \geqslant 0$ ненулевых элементов матрицы A. Если $\sharp A < n$, то одна строчка A нулевая и поэтому в силу (**) $A \in \mathcal{A}$.

Если $\sharp A > n$, то возьмем строчку A (например, 1-ю), в которой $\geqslant$2 ненулевых элементов. Ясно, что существуют такие матрицы A' и A'', совпадающие с A вне 1-й строчки, что для их 1-х строчек справедливо равенство $(A_1) = (A_1') + (A_1'')$ и при этом $\sharp A', \sharp A'' < \sharp A$. В силу индуктивного предположения $A', A'' \in \mathcal{A}$. Из линейного свойства определителя следует, что и $A \in \mathcal{A}$.

[4]Так как порядок $n-1 < n$, то понятие определителя для минора $M_{i\beta}$ введено корректным образом.

[5]Так как порядок $n-2 < n$, то понятие определителя для минора $M_{i\alpha;j\beta}$ тоже введено корректным образом.

[6]По-видимому, в этом состоит ключевой прием доказательства – в попеременном разложении определителей по строчке и столбцу.

Наконец, рассмотрим **основной случай** $\sharp A = n$. В силу (**) можно считать, что на каждой строчке матрицы A лежит ровно по одному ненулевому элементу, равному 1, а любые 2 строчки A различны. Поэтому A получается из единичной матрицы E при помощи конечного числа, скажем s, попарных перестановок ее строчек – $|A| = (-1)^s|E| = (-1)^s$.

Но легко видеть, AB получается из матрицы B при помощи тех же самых попарных s перестановок строчек. Поэтому $|AB| = (-1)^s|B|$, что легко влечет $A \in \mathcal{A}$. $\square$

Литература

1. *Botha J. D.* Alternative proofs of the rational canonical form theorem // Int. J. Math. Educ. Sci. Technol. 1994. Vol. 25, № 5. P. 745–749.

2. *Филиппов А. Ф.* Краткое доказательство теоремы о приведении матрицы к жордановой форме // Вестн. МГУ. Сер. матем. 1971. Т. 26, N 1–2. С. 70–71.

References

1. Botha J. D. Alternative proofs of the rational canonical form theorem. *Int. J. Math. Educ. Sci. Technol.*, 1994, vol. 25, no. 5, pp. 745–749.

2. Filippov A. F. A short proof of the theorem on the reduction of a matrix to Jordan form. *Mosc. Univ. Math. Bull.,* 1971, vol. 26, no. 1–2, pp. 70–71.

ПРАВИЛА ДЛЯ АВТОРОВ

1. Для публикации в журнале принимаются ранее не опубликованные в других изданиях научные статьи. Обзорные статьи публикуются по решению редколлегии.

2. Объем статьи не более 20 журнальных страниц (с учетом таблиц и рисунков, а также списка литературы на 2 языках), объем краткого сообщения – до 5 страниц.

3. Статьи в Журнале публикуются на русском или английском языке.

4. Статья должна быть подготовлена в системе LaTeX по образцу, находящемуся на сайте Журнала. Не допускается использование в TeX-файлах «нестандартных» TeX-команд (т. е. команд, не входящих в стандартную поставку LaTeX), а также переопределение стандартных команд. При подаче статьи автору необходимо указывать рубрику Журнала, к которому относится статья.

5. Текст статьи начинается с индекса УДК, затем следуют название статьи, инициалы и фамилии авторов, а также краткая аннотация (не более 15 строк) и ключевые слова (5–10 слов). Аннотация не должна содержать ссылок на формулы и литературу статьи. Сведения об авторе (авторах), название статьи, аннотация и ключевые слова пишутся на русском и английском языках, остальные элементы оформляются на русском (английском) языке.

6. Ссылки на литературу в тексте нумеруются в порядке их упоминания и даются в квадратных скобках. Использование ссылок на неопубликованные работы не допускается. Необходимо приводить два списка ссылок на использованную в статье литературу – «Литература» и «References».

7. Если присутствует информация о финансировании (поддержке грантами проектов и т. п.), ее следует размещать в конце статьи.

GUIDELINES FOR AUTHORS

1. Scientific articles not previously published in other publications are accepted for publication in the journal. Review articles are published by decision of the editorial board.

2. The volume of the article is no more than 20 journal pages, taking into account tables and figures, the volume of a short message is up to 5 pages.

3. Articles in the Journal are published in Russian or English.

4. The article must be prepared in the LaTeX system according to the model located on the Journal's website. It is not allowed to use "non-standard" in TeXfiles TeXcommands (i. e. commands not included in the standard LaTeX distribution), and also override standard commands. When submitting an article, the author must indicate the category of the Journal to which the article belongs.

5. The text of the paper must begin with the UDC, the title, and the name(s) of the author(s) preceded by initials followed by a short abstract (no more than 15 lines) and keywords (5–10 words). The abstract must not contain numbered references to the formulas and bibliography items. Information about the author (authors), title of the article, abstract and keywords are written in Russian and English, other elements are written in Russian (English).

6. The references in the text are numbered in order of their appearance and in square brackets. The use of references to unpublished works is not permitted. It is necessary to provide two lists of references to the literature used in the article – "References" and "Literature". An example of bibliographic descriptions is given in the model located on the Journal's website.

7. If there is information about financial support for the research (grant support for projects, etc.), it should be placed at the end of the article.

ТРУДЫ ИНСТИТУТА МАТЕМАТИКИ НАН БЕЛАРУСИ
2024. ТОМ 32, № 2

Редактор Т. П. Петрович
Компьютерная верстка И. В. Близнец

Подписано в печать __.__.2024. Формат 60×84 $^1/_8$. Усл. печ. л. 14,42. Уч.-изд. л. 15,25.
Государственное научное учреждение «Институт математики Национальной академии наук Беларуси».
220072, г. Минск, ул. Сурганова, д. 11.
Свидетельство о государственной регистрации издателя, изготовителя, распространителя печатных изданий № 1 / 257 от 2 апреля 2014 г.

Отпечатано в Республиканском унитарном предприятии «Издательский дом «Беларуская навука».
Свидетельство о государственной регистрации издателя, изготовителя, распространителя печатных изданий № 1 / 18 от 02.08.2013. ЛП № 02330 / 455 от 30.12.2013. Ул. Ф. Скорины, 40, 220084, г. Минск.
Тираж 30 экз. Заказ . . . .