

Том 63, Номер 8

ISSN 0044-4669

Август 2023



ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ

www.sciencejournals.ru



СОДЕРЖАНИЕ

Том 63, номер 8, 2023 год

ОБЩИЕ ЧИСЛЕННЫЕ МЕТОДЫ

Обобщение быстрого преобразования Фурье с постоянной структурой <i>М. С. Беспалов</i>	1241
Аппроксимация дифференциальных операторов с учетом граничных условий <i>В. П. Варин</i>	1251
Улучшенная разностная схема для задачи Коши в случае уравнения переноса <i>Г. И. Шишкин, Л. П. Шишкина</i>	1272
Об одновременном определении коэффициента теплопроводности и объемной теплоемкости вещества <i>А. Ю. Горчаков, В. И. Зубов</i>	1279

ОБЫКНОВЕННЫЕ ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ

О решениях краевой задачи для одного дифференциального уравнения второго порядка с параметром и разрывной правой частью <i>О. В. Басков, Д. К. Потапов</i>	1296
---	------

УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

О некоторых эллиптических краевых задачах в конических областях <i>В. Б. Васильев</i>	1309
Application of IBSEF Method to Chaffee-Infante Equation in $(1 + 1)$ and $(2 + 1)$ Dimensions <i>U. Demirbilek, Kh. R. Mamedov</i>	1316
О единственности решения систем линейных алгебраических уравнений, к которым редуцируются обратные линейные задачи гравиметрии и магнитометрии: локальный случай <i>И. И. Колотов, Д. В. Лукьяненко, И. Э. Степанова, А. Г. Ягола</i>	1317
Аналог теоремы Келлога для кусочно-ляпуновских областей <i>А. П. Солдатов</i>	1332

МАТЕМАТИЧЕСКАЯ ФИЗИКА

Внутренние гравитационные волны от локализованного источника в потоке стратифицированной среды с модельным распределением частоты плавучести <i>В. В. Булатов, И. Ю. Владимиров</i>	1343
Моделирование эмиссионных процессов в сильных электромагнитных полях <i>Т. А. Кудряшова, С. В. Поляков, Н. И. Тарасов</i>	1354
Оценка погрешности и оптимизация метода прямого статистического моделирования с учетом пространственной регуляризации <i>М. Ю. Плотников, Е. В. Шкарупа</i>	1367

Numerical Simulation of Phase Transfer During Cryosurgery for an Irregular Tumor
Using Hybrid Approach

Surabhi Nishada, Rama Bhargava

1380

ИНФОРМАТИКА

Методы построения оценок множеств достижимости в задаче моделирования потоков людей

М. В. Зайцева, П. А. Точилин

1381

О множестве стабильных матчингов двудольного графа

А. В. Карзанов

1395

ОБЩИЕ ЧИСЛЕННЫЕ
МЕТОДЫ

УДК 519.113

ОБОБЩЕНИЕ БЫСТРОГО ПРЕОБРАЗОВАНИЯ ФУРЬЕ
С ПОСТОЯННОЙ СТРУКТУРОЙ

© 2023 г. М. С. Беспалов^{1,*}

¹ 600000 Владимир, ул. Горького, 87, Владимирский гос. ун-т, Российская Федерация

*e-mail: bespalov@vlsu.ru

Поступила в редакцию 09.02.2023 г.
Переработанный вариант 14.03.2023 г.
Принята к публикации 28.04.2023 г.

Широко популярны знаменитые быстрые алгоритмы Кули–Тьюки для дискретного преобразования Фурье составного основания, представленные в двух видах — классическом и с постоянной структурой. В статье предложено матричное представление этих алгоритмов в обозначениях двух видов тензорного произведения матриц: кронекерова произведения и b -произведения. Предложенное матричное представление указывает на идентичность структуры этих алгоритмов с двумя быстрыми алгоритмами Гуда для кронекеровой степени матрицы. Продемонстрирована методика построения матричной формы быстрых алгоритмов для дискретных преобразований: Фурье и Крестенсона с составным основанием, а также Виленкина. Показана предпочтительность использования алгоритма с постоянной структурой в случаях более сложных конструкций. Библ. 13.

Ключевые слова: дискретное преобразование Фурье, дискретное преобразование Уолша, быстрый алгоритм, кронекерово произведение матриц.

DOI: 10.31857/S0044466923080033, EDN: WRXNSJ

ВВЕДЕНИЕ

Центральное место в теории алгоритмов цифровой обработки сигналов занимают быстрые алгоритмы Кули и Тьюки [1] для реализации дискретного преобразования Фурье (ДПФ) с составным основанием N . При практическом использовании наиболее популярными среди них служат алгоритмы быстрого преобразования Фурье (БПФ), соответствующие факторизации матриц ДПФ порядка $N = 2^n$. Из них выделяют [2] два основных БПФ с прореживанием по частоте и по времени, которые соответствуют одной и той же факторизации матриц ДПФ. Это объясняется тем, что матрица ДПФ $F = F_N$ симметрична; поэтому ДПФ можно считать для столбцов x , y по формуле $y = F \cdot x$ или для строк x , y по формуле $y = x \cdot F$. При переходе от одной формулы к другой меняется на противоположный порядок выполнения операций.

Другой подход к построению БПФ порядка $N = 2^n$ (с названием БПФ с постоянной структурой) не столь популярен, но более удобен при больших n . В данной статье проводится качественный анализ этих двух алгоритмов БПФ с представлением их в матричном виде. В основе построения первого из них лежит лемма факторизации кронекерова произведения $\otimes$ с добавлением диагональных матриц вращения. Формальное описание второго из них возможно с помощью нового тензорного произведения матриц $\oslash$, свойства которого получены в [3]. Естественно, что невозможно обойтись без сомножителей в виде диагональных матриц вращения и в алгоритме БПФ с постоянной структурой.

Если из каждого из этих двух вариантов факторизации матрицы ДПФ порядка $N = 2^n$ удалим все сомножители в виде диагональных матриц вращения, то получим два алгоритма Гуда [4] факторизации матрицы дискретного преобразования Уолша (ДПУ). Так как диагональные матрицы вращения не оказывают влияния на структуру алгоритма в виде блок-схем, то в качестве вывода получаем

Утверждение 1. *Скелет каждой из двух разновидностей БПФ составляют два алгоритма Гуда для ДПУ.*

Другой прием увеличения порядка матрицы ортогонального преобразования состоит в переходе от ДПФ к дискретному преобразованию Крестенсона (ДПК), матрица которого есть кронекерова степень матрицы ДПФ. При этом переходе возникает задача композиции двух быстрых алгоритмов – внутреннего для ДПФ и внешнего для кронекеровой степени. Внешний алгоритм можно построить по одному из алгоритмов Гуда, где в приоритете опять алгоритм с постоянной структурой. А конструкцию сопряжения этих алгоритмов можно провести с помощью свойств двух тензорных произведений матриц, описанных в [3].

1. ПРЕДВАРИТЕЛЬНЫЕ СВЕДЕНИЯ

1.1. В каждом варианте быстрого алгоритма ДПФ присутствует перестановка отсчетов сигнала, встречающаяся под названиями *разрядно-инверсная*, *реверсная* или *идеальная перетасовка*. Будем использовать более короткое название *реверсная* (применяемое в [5]), хотя название *разрядно-инверсная* более информативно. Вариант двоичной реверсной перестановки подробно описан в книге [5].

Рассмотрим общий случай реверсной перестановки относительно системы счисления по смешанному основанию.

Пусть $N = pq$ и $p \neq q$. Числа x от 0 до $N - 1$ представляются в виде $x = bp + a$ или $x = aq + b$, где $a = 0, 1, \dots, p - 1$, $b = 0, 1, \dots, q - 1$. Для представления $x = aq + b$ внутренний цикл по a , а внешний цикл по b приводит к одному реверсному порядку

$$\underbrace{0 \ q \ 2q \ \dots \ (p-1) \ q}_{\text{внутренний цикл}} \underbrace{1 \ q + 1 \ \dots \ (p-1)q + 1}_{\text{внешний цикл}} \dots \underbrace{q-1 \ 2q-1 \ \dots \ pq-1}_{\text{внешний цикл}}, \quad (1)$$

а реверсное прочтение другого представления $x = bp + a$ приводит к другому реверсному порядку того же набора чисел.

Если же $N = pqr$ для различных p, q, r , то получается уже шесть различных реверсных перестановок N чисел в соответствии с перестановкой порядка сомножителей в N .

Поскольку перестановки нам потребуется применять не к числам, а к строкам и столбцам матриц, то предложим описание реверсных перестановок через соответствующую перестановочную матрицу.

1.2. Напомним определения двух разных видов тензорного произведения матриц. Будем обозначать через $M_{n,m}$ множество матриц размера $n \times m$ и через $M_n := M_{n,n}$ множество квадратных матриц порядка n .

Если $A = (a_{kj}) \in M_{n,l}$, $B \in M_{m,s}$, то кронекерово произведение определяется в виде блочной матрицы

$$A \otimes B = (a_{kj} B)_{k,j=0}^{n-1,l-1} \in M_{nm,ls}.$$

Временно обозначим тем же символом с нижним индексом A_k строки матрицы A , а с верхним индексом A^j столбцы матрицы A . Тогда b -произведение $A \in M_{n,l}$ на $B \in M_{m,s}$ определяется в виде блочной матрицы (блок есть поэлементное произведение столбца на строку)

$$A \oslash B = (A^j B_k)_{k,j=0}^{m-1,l-1} \in M_{nm,ls}.$$

Быстрые алгоритмы первого типа ДПФ и ДПУ строятся на основе следующей леммы факторизации (доказательство приведено в [3]).

Лемма 1. *Если $A \in M_{n,l}$, $B \in M_{l,s}$, $C \in M_{m,k}$, $D \in M_{k,t}$, то*

$$(A \cdot B) \otimes (C \cdot D) = (A \otimes C) \cdot (B \otimes D) = (C \oslash A) \cdot (B \oslash D),$$

$$(A \cdot B) \oslash (C \cdot D) = (A \oslash C) \cdot (B \otimes D) = (C \otimes A) \cdot (B \oslash D).$$

Следствие 1. Два вида тензорного произведения матриц связаны через реверсную перестановку строк

$$B \oslash D = (I_l \oslash I_k) \cdot (B \otimes D).$$

Доказывается подстановкой $A = I_l, C = I_k$.

Аналогично в общем случае матрица реверсной перестановки определяется как b -произведение единичных матриц, а в частном случае равных оснований – как b -степень единичной матрицы.

В [6] получено следующее свойство, где I_x – единичная матрица указанного порядка.

Лемма 2. Если $A \in M_m, B \in M_l$, то

$$(A \otimes I_x) \cdot (B \otimes I_{mx}) = (B \otimes A) \otimes I_x.$$

1.3. Матрица ДПУ (уровня n или порядка $N = 2^n$) в нумерации Адамара определяется как кронекерова степень $H_n = H^{\otimes n}$ матрицы

$$H = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}. \quad (2)$$

Матрица ДПУ в нумерации Пэли (уровня n) определяется как b -степень той же матрицы $W_n = H^{\odot n}$.

Матрица (обратного) ДПФ, которое берем в качестве основного,

$$F_N = \left(\omega_N^{jk} \right)_{j,k=0}^{N-1}, \quad \text{где} \quad \omega_N = \exp \frac{2\pi i}{N}. \quad (3)$$

Прямое ДПФ задается матрицей $\bar{F}_N$ или заменой ω_N на $\bar{\omega}_N = \omega_N^{-1}$ в (3).

Матрица дискретного преобразования Крестенсона (ДПК) определяется как кронекерова степень матрицы (3). Часто рассматривают другой вариант ДПК с матрицей в виде b -степени матрицы ДПФ, на который мы в данной статье отвлекаться не будем.

Строки (столбцы) матрицы (3) в [7], [2] названы дискретными экспоненциальными функциями (ДЭФ). Обозначим основную ДЭФ как вектор

$$\mathbf{r} = r_N = (1 \ \omega \ \omega^2 \ \dots \ \omega^{N-1}).$$

Остальные ДЭФ есть степени по Адамару основной ДЭФ:

$$\mathbf{r}^2 = \mathbf{r} \bullet \mathbf{r} = (1 \ \omega^2 \ \omega^4 \ \dots \ \omega^{2N-2}), \quad \mathbf{r}^3 = \mathbf{r} \bullet \mathbf{r} \bullet \mathbf{r} = (1 \ \omega^3 \ \omega^6 \ \dots \ \omega^{3N-3}), \dots$$

Для фиксированного N все ДЭФ образуют группу порядка N :

$$\mathbf{r}^k \bullet \mathbf{r}^m = \mathbf{r}^{k+m}, \quad \mathbf{r}^N = \mathbf{r}^0 = (1 \ 1 \ \dots \ 1) - \text{нейтральный элемент.}$$

Если N составное, т.е. $N = pq$, то введем понятие *срез дискретной экспоненциальной функции* (СДЭФ) с основным представителем

$$\mathbf{s} = s_{(p,q)} = (1 \ \omega \ \omega^2 \ \dots \ \omega^{q-1}), \quad \text{где} \quad \omega = \omega_N = \exp \frac{2\pi i}{N}.$$

Остальные СДЭФ также являются степенями по Адамару основной СДЭФ:

$$\begin{aligned} \mathbf{s}^0 &= (1 \ 1 \ \dots \ 1), \quad \mathbf{s}^2 = \mathbf{s} \bullet \mathbf{s} = (1 \ \omega^2 \ \omega^4 \ \dots \ \omega^{2q-2}), \quad \mathbf{s}^3 = \mathbf{s} \bullet \mathbf{s} \bullet \mathbf{s}, \dots \\ \dots, \mathbf{s}^{p-2}, \mathbf{s}^{p-1} &= (1 \ \omega^{p-1} \ \omega^{2p-2} \ \dots \ \omega^{(p-1)(q-1)}). \end{aligned}$$

Всего различных СДЭФ p штук. Отсчет с номером 1 служит образующим элементом как для ДЭФ, так и для СДЭФ.

Составим N -мерный вектор всех СДЭФ, определяемый по двум параметрам p, q таким, что

$$N = pq: \quad \mathbf{s}_{p,q} = (\mathbf{s}^0 \ \mathbf{s}^1 \ \mathbf{s}^2 \ \dots \ \mathbf{s}^{p-1}) = \left(\underbrace{1 \ 1 \ \dots \ 1}_p \underbrace{1 \ \omega \ \dots \ \omega^{q-1}}_q \underbrace{1 \ \omega^2 \ \dots \ \omega^{2q-2}}_q \dots \underbrace{1 \ \omega^{p-1} \ \dots \ \omega^{(p-1)(q-1)}}_q \right),$$

который является аналогом сигнала Франка [8].

Если N раскладывается на 3 сомножителя $N = pqr$, то обозначим $N_1 = pq$ и каждый блок в виде среза N_1 -мерного вектора СДЭФ повторим r раз

$$\mathbf{S}_{p,q,r} = \left(\underbrace{\mathbf{s}^0 \dots \mathbf{s}^0}_r \underbrace{\mathbf{s}^1 \dots \mathbf{s}^1}_r \dots \underbrace{\mathbf{s}^{p-1} \dots \mathbf{s}^{p-1}}_r \right). \quad (4)$$

2. БЫСТРЫЕ АЛГОРИТМЫ ОРТОГОНАЛЬНЫХ ПРЕОБРАЗОВАНИЙ

2.1. К быстрым алгоритмам для ортогональных преобразований в 1958 г. обратился Гуд, предложивший в [4] два варианта факторизации кронекеровой степени матрицы. Он отметил, что первый из алгоритмов есть следствие леммы факторизации кронекерова произведения. А второй алгоритм как раз и является алгоритмом с постоянной структурой, так как на каждом шаге не меняется вид матрицы сомножителя. Приведем эти алгоритмы в матричном виде применительно к ДПК.

Утверждение 2 (первый алгоритм Гуда). *Матрицу ДПК можно факторизовать в виде произведения слабозаполненных матриц*

$$F_p^{\otimes n} = T_n \cdot T_{n-1} \cdot \dots \cdot T_1, \quad \text{где} \quad T_j = E^{\otimes n-j} \otimes F_p \otimes E^{\otimes j-1},$$

E — единичная матрица p -го порядка. Причем матрицы T_j допускают любую перестановку сомножителей.

Алгоритм с постоянной структурой, выделенный Гудом как основной результат статьи [4], приведем в форме, полученной в [6].

Утверждение 3 (второй алгоритм Гуда). *Матрицу ДПК можно представить в виде степени слабозаполненной матрицы*

$$F_p^{\otimes n} = \left(F_p \otimes I_{p^{n-1}} \right)^n.$$

Замечание. В случае $p = 2$ в представленных алгоритмах приведены быстрые алгоритмы для ДПУ-Адамара, где $F_2 = H$ из (2).

Другой подход к этим алгоритмам, с построением соответствующих базисов, предложен в статьях [9], [10], а матричный вид первого алгоритма в [11].

Введем для случая порядка матрицы $N = pqr$ диагональные матрицы вращений $D_{p,q,r}$ с вектором $\mathbf{S}_{p,q,r}$ из (4) на диагонали: p групп по q элементов в группе с r повторами.

В частности,

$$D_{2,2^{j-1},1} = \text{diag} \left(\underbrace{1, 1, \dots, 1}_{2^{j-1}}, \omega^0, \omega^1, \dots, \omega^{2^{j-1}-1} \right)$$

для меняющихся порядков матриц $N_j = 2^j = pq$ с соответствующим $\omega = \omega_{pq} = \omega_{2^j}$. Это обозначение плюс запись матрицы реверсной перестановки в виде b -степени позволяют получить следующий (уточняющий предложенный в [12], [13]) матричный вариант алгоритма Кули-Тьюки для ДПФ порядка $N = p^n$.

Утверждение 4 (основной случай алгоритма Кули-Тьюки). *Пусть $N = p^n$. Допускается факторизация*

$$F_N = R_n \cdot R_{n-1} \cdot \dots \cdot R_1 \cdot L,$$

где $R_j = E^{\otimes n-j} \otimes \left((F_p \otimes E^{\otimes j-1}) \cdot D_{p,p^{j-1},1} \right)$ при $j > 1$, $R_1 = E^{\otimes n-1} \otimes F_p$, E — единичная матрица p -го порядка, $L = E^{\otimes n}$ — матрица реверсной перестановки.

Так как повторы здесь внешние, а не внутренние, то для их записи применен прием с помощью левого кронекерова произведения.

Утверждение 5 (алгоритм Кули-Тьюки с постоянной структурой). Пусть $N = p^n$. Допускается факторизация

$$F_N = L \cdot T \cdot D_{p,p,p^{n-2}} \cdot T \cdot D_{p^2,p,p^{n-3}} \cdot T \cdot D_{p^3,p,p^{n-4}} \cdot \dots \cdot T \cdot D_{p^{n-1},p,1} \cdot T,$$

где $T = F_p \otimes I_{p^{n-1}}$, $L = E^{\otimes n}$, E — единичная p -го порядка.

Вывод. Скелет стандартных видов алгоритма Кули-Тьюки составляют первый и второй алгоритмы Гуда. Добавлением к алгоритмам Гуда множителя в виде матрицы реверсной перестановки $L = E^{\otimes n}$ в начало алгоритма (умножение справа) для первого или в конец алгоритма (умножение слева) для второго алгоритма Гуда превращает алгоритм ДПУ-Адамара в алгоритм ДПУ-Пэли. Следующее добавление между каждой парой сомножителей алгоритма Гуда специально подобранных диагональных матриц вращений приводит к факторизации матрицы ДПФ порядка $N = p^n$.

Число сложений (вычитаний) в каждом из этих четырех алгоритмов одинаковое и равно $N \log_p N$. Для ДПУ это сложения действительных чисел, а для ДПФ комплексные сложения. Умножений для ДПУ нет. Для ДПФ число умножений вычисляется по числу комплексных множителей в диагональных матрицах вращений.

2.2. Наибольший практический интерес представляют случаи $p = 3$, как простейший модельный случай, и $p = 4$, как оптимальный по числу операций вариант основного случая $N = 2^n$. Замечание про случай $p = 4$ проиллюстрируем примерами.

Для вычисления $y = F_{16} \cdot x$ предложим две факторизации с постоянной структурой по утверждению 5. Первая факторизация

$$F_{16} = E^{\otimes 4} \cdot T \cdot D_{2,2,4} \cdot T \cdot D_{4,2,2} \cdot T \cdot D_{8,2,1} \cdot T,$$

где $T = H \otimes I_8$, $E = I_2$. Вторая факторизация

$$F_{16} = (I_4 \otimes I_4) \cdot (F_4 \otimes I_4) \cdot D_{4,4,1} \cdot (F_4 \otimes I_4)$$

дополняется факторизацией $F_4 = (H \otimes E) \cdot D_{2,2,1} \cdot (H \otimes E)$. Далее удобнее воспользоваться приемами параллельных вычислений, представив входной столбец $x \in \mathbb{C}^{16}$ в виде блочного столбца

$$\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} \text{ с элементами } a, b, c, d \in \mathbb{C}^4. \text{ Как первый шаг } (H \otimes E) \cdot \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = \begin{pmatrix} a+c \\ a-c \\ b+d \\ b-d \end{pmatrix}, \text{ так и второй (после}$$

$$\text{умножения на i) шаг } (H \otimes E) \cdot \begin{pmatrix} a+c \\ a-c \\ b+d \\ i(b-d) \end{pmatrix} = \begin{pmatrix} a+b+c+d \\ a+ib-c-id \\ a-b+c-d \\ a-ib-c+id \end{pmatrix} \text{ состоит из 16 комплексных слагаемых}$$

(12 сложений). Умножение на числа i и $-i$ выполняется в виде перестановки элементов с изменением одного знака, что операцией не считается.

Из этих двух вариантов при первом варианте факторизации имеем 10 умножений: 4 в матрице $D_{4,2,2} = \text{diag}(11111u1u1i1i1u^31u^3)$, где $u = \omega_8$, и 6 в матрице $D_{8,2,1} = \text{diag}(111\omega1\omega^21\omega^31i1\omega^51\omega^61\omega^7)$, где $\omega = \omega_{16}$.

При втором варианте факторизации имеем 8 умножений ($\omega = \omega_{16}$): $D_{4,4,1} = \text{diag}(11111\omega\omega^2\omega^31\omega^2i\omega^61\omega^3\omega^6\omega^9)$.

3. ОСНОВНЫЕ ТЕОРЕМЫ О БЫСТРЫХ АЛГОРИТМАХ

Доказательство сформулированных выше утверждений 4 и 5, в которых приведены наиболее удобные и чаще применяемые БПФ, вытекает из следующего общего случая ДПФ с произвольным составным основанием $N = pqr$. Далее покажем переход к основанию $N = pqrs$.

Пусть $N = pqr$. Назовем *матрицей дискретного преобразования* Виленкина (ДПВ) кронекеро-произведение матриц ДПФ $F = F_p \otimes F_q \otimes F_r$.

Теорема 1. Данное ДПВ допускает факторизации Гуда

- 1) $F_p \otimes F_q \otimes F_r = (F_p \otimes I_q \otimes I_r) \cdot (I_p \otimes F_q \otimes I_r) \cdot (I_p \otimes I_q \otimes F_r),$
- 2) $F_p \otimes F_q \otimes F_r = (F_r \otimes I_{pq}) \cdot (F_q \otimes I_{pr}) \cdot (F_p \otimes I_{qr}).$

Доказательство. По лемме 1 получаем известную первую формулу

$$\begin{aligned} F_p \otimes F_q &= (F_p \cdot I_p) \otimes (I_q \cdot F_q) = (F_p \otimes I_q) \cdot (I_p \otimes F_q). \\ F &= ((F_p \otimes F_q) \cdot I_{pq}) \otimes (I_r \cdot F_r) = (((F_p \otimes I_q) \cdot (I_p \otimes F_q)) \otimes I_r) \cdot (I_{pq} \otimes F_r) = \\ &= (F_p \otimes I_q \otimes I_r) \cdot (I_p \otimes F_q \otimes I_r) \cdot (I_{pq} \otimes F_r). \end{aligned}$$

Для второй формулы по лемме 2 имеем

$$(F_p \otimes F_q) \otimes I_r = (F_q \otimes I_{pr}) \cdot (F_p \otimes I_{qr}),$$

при обозначении $B = F_p \otimes F_q \in M_{pq}$ по лемме 1 и этой формуле получим

$$F = B \otimes F_r = (F_r \otimes I_{pq}) \cdot (B \otimes I_r) = (F_r \otimes I_{pq}) \cdot (F_q \otimes I_{pr}) \cdot (F_p \otimes I_{qr}).$$

Замечание по терминологии. В технической литературе [2], следуя предложенному в [7], закрепился термин Виленкина-Крестенсона функции (ВКФ) как для ДПК, так и для их непрерывного аналога. Но Виленкин рассматривал мультипликативную систему функций в общем виде для смешанных оснований. Поэтому термин ВКФ воспринимался как те из функций Виленкина, которые изучались Крестенсоном (т.е. с постоянным основанием). Поэтому и предлагается вариант терминов с конкретизацией: ДПУ (основание 2), ДПК (основание $p > 2$) и ДПВ (случай смешанного основания). Если быть исторически точными, то дискретный вариант в виде кронекеровых произведений ДПФ нельзя приписывать Виленкину. Более правильно термином ДПВ для $N = pqr$ было бы называть преобразование с матрицей (соответствующее функциям Виленкина непрерывного аргумента)

$$F = F_p \otimes F_q \otimes F_r.$$

Утверждение 6. Два вида ДПВ связаны через матрицу реверсной перестановки

$$F_p \otimes F_q \otimes F_r = (I_p \otimes I_q \otimes I_r) \cdot (F_p \otimes F_q \otimes F_r).$$

Доказательство. По следствию 1 имеем $F_p \otimes F_q = (I_p \otimes I_q) \cdot (F_p \otimes F_q)$. Отсюда вновь по лемме 1 получим

$$F_p \otimes F_q \otimes F_r = ((I_p \otimes I_q) \cdot (F_p \otimes F_q)) \otimes (I_r \cdot F_r) = (I_p \otimes I_q \otimes I_r) \cdot (F_p \otimes F_q \otimes F_r).$$

Теорема 2. Две факторизации матрицы ДПФ основания $N = pqr$:

- 1) $F_N = (F_p \otimes I_{qr}) \cdot D_{p,qr,1} \cdot (I_p \otimes ((F_q \otimes I_r) \cdot D_{q,r,1})) \cdot (I_{pq} \otimes F_r) \cdot (I_r \otimes I_q \otimes I_p),$
- 2) $F_N = (I_r \otimes I_q \otimes I_p) \cdot (F_p \otimes I_{qr}) \cdot D_{p,q,r} \cdot (F_q \otimes I_{pr}) \cdot D_{pq,r,1} \cdot (F_r \otimes I_{pq}).$

Доказательство. Первый случай для двух сомножителей прямой проверкой

$$F_{pq} = (F_p \otimes I_q) \cdot D_{p,q,1} \cdot (I_p \otimes F_q) \cdot (I_q \otimes I_p) = (F_p \otimes I_q) \cdot D_{p,q,1} \cdot (F_q \otimes I_p).$$

Сначала дано представление с выделенной реверсной перестановкой, а потом последние два сомножителя объединены по лемме 1.

Теперь первый случай для трех сомножителей

$$F_{p(qr)} = (F_p \otimes I_{qr}) \cdot D_{p,qr,1} \cdot (I_p \otimes F_{qr}) \cdot (I_{qr} \otimes I_p),$$

где по лемме 1 имеем

$$\begin{aligned} I_p \otimes F_{qr} &= (I_p \cdot I_p \cdot I_p) \otimes ((F_q \otimes I_r) \cdot D_{q,r,1} \cdot (I_q \otimes F_r) \cdot (I_r \otimes I_q)) = \\ &= (I_p \otimes ((F_q \otimes I_r) \cdot D_{q,r,1})) \cdot (I_p \otimes I_q \otimes F_r) \cdot (I_p \otimes (I_r \otimes I_q)). \end{aligned}$$

Внесем в предыдущее выражение и преобразуем два последних множителя по лемме 1

$$(I_p \otimes (I_r \otimes I_q)) \cdot (I_{qr} \otimes I_p) = ((I_r \otimes I_q) \cdot I_{qr}) \otimes (I_p \cdot I_p) = I_r \otimes I_q \otimes I_p.$$

Второй случай для двух сомножителей имеет вид

$$F_{pq} = (I_q \otimes I_p) \cdot (F_p \otimes I_q) \cdot D_{p,q,1} \cdot (F_q \otimes I_p) \quad (5)$$

Обозначим через $\tilde{F}_{p,q}$ реверсную перестановку строк матрицы F_{pq} , равную произведению последних трех сомножителей в (5). Уточним вид этой реверсной перестановки на примере $p = 2, q = 3$:

$$F_2 \otimes I_3 = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ & & -1 & & & \\ & & & 1 & & \\ & & & & 1 & \\ & & & & & -1 \end{pmatrix},$$

$$S_{2,3,1} = (1 \ 1 \ 1 \ \omega^0 \ \omega^1 \ \omega^2),$$

$$(F_2 \otimes I_3) \cdot D_{2,3,1} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & \omega & 0 \\ 0 & 1 & 0 & 0 & -\omega & 0 \\ 0 & 0 & 1 & 0 & 0 & \omega^2 \\ 0 & 0 & 1 & 0 & 0 & -\omega^2 \end{pmatrix}, \quad F_3 \otimes I_2 = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & q & 0 & q^2 & 0 \\ 1 & 0 & q^2 & 0 & q & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & q & 0 & q^2 \\ 0 & 1 & 0 & q^2 & 0 & q \end{pmatrix} = \begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ & & q & & & \\ & & & q^2 & & \\ & & & & q & \\ & & & & & 1 \end{pmatrix},$$

где $\omega = \omega_6 = e^{\frac{2\pi i}{6}}, q = \omega_3 = e^{\frac{2\pi i}{3}}, q = \omega^2$. Получили

$$\tilde{F}_{2,3} = (F_2 \otimes I_3) \cdot D_{2,3,1} \cdot (F_3 \otimes I_2) = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & \omega & q & q\omega & q^2 & q^2\omega \\ 1 & -\omega & q & -q\omega & q^2 & -q^2\omega \\ 1 & \omega^2 & q^2 & q^2\omega^2 & q & q\omega^2 \\ 1 & -\omega^2 & q^2 & -q^2\omega^2 & q & -q\omega^2 \end{pmatrix}.$$

Набор строк составляет полный набор различных ДЭФ порядка 6. В столбце с номером 1 приведены образующие этих ДЭФ в реверсном порядке:

$$\omega^0, \ \omega^3, \ \omega^1, \ \omega^4, \ \omega^2, \ \omega^5.$$

Умножение слева на перестановочную матрицу $I_3 \otimes I_2$ возвращает столбцу из этих элементов прямой порядок.

Аналогично в общем случае: матрицу $(F_p \otimes I_q)$ рассматриваем как блочную — в начальной блочной строке приведен прореженный нулями носитель в виде матрицы F_p , записанной по столбцам в начальном столбце каждого блока. Блочная строка с номером $j + 1$ получается циклическим сдвигом вправо j -й блочной строки. Поэтому в каждом блочном столбце идет повтор со сдвигом вправо носителя. Этот процесс лучше виден, если нули в матрицах убрать, оставив носитель и пустые клетки. Произведение трех матриц будем вычислять по столбцам по правилу:

j -й столбец произведения двух матриц есть линейная комбинация столбцов первого сомножителя с коэффициентами в j -м столбце второго. Поэтому предлагается элементы матрицы $D_{p,q,1}$ внести в столбцы первого сомножителя, а далее каждый j -й столбец последнего сомножителя выбирает (наложением носителя) и суммирует с указанными коэффициентами все столбцы j -го блока полученного первого сомножителя. При этом начальный столбец получается состоящим из единиц. А столбец с номером 1 (в нем образующие элементы) имеет вид $(r_p \ u r_p \ u^2 r_p \dots u^{q-1} r_p)^T$, где $r_p = (1 \ v \ v^2 \dots v^{p-1})$, $v = \omega_p$, $u = \omega_{pq}$. Если представим по правилу $v = u^q$ степени v как степени u , то относительно суммарных степеней u элементы этого столбца в реверсном порядке (1). Домножением слева на матрицу реверсной перестановки $I_q \otimes I_p$ восстанавливаем стандартный порядок элементов первого столбца.

Покажем, что если произвольный k -й элемент первого столбца (т.е. с номером $k, 1$) имеет вид $v^a u^b$, то элемент той же строки с номером k, j имеет вид $v^{aj} u^{bj}$. Для этого будем рассматривать u , v как не взаимосвязанные символы. Подставим вместо $D_{p,q,1} \cdot (F_q \otimes I_p)$ матрицу $K \otimes I_p$ той же структуры, где $K \in M_q$ и все элементы K единицы: $(F_p \otimes I_q) \cdot (K \otimes I_p) = K \otimes F_p$ – блочная матрица с одинаковыми блоками F_p . Для всех элементов этой матрицы степени v такие же, как и для $\tilde{F}_{p,q}$. Доказали, что $\tilde{f}_{k,j} = v^{aj} u^j$.

Повторим прием с заменой матрицы $F_p \otimes I_q$ на матрицу той же структуры $K \otimes I_q$, а множители диагональной матрицы $D_{p,q,1}$ внесем в строки матрицы $F_q \otimes I_p$. Если $k = a + bp$, то вносимые множители для k -й строки $1, u^a, u^{2a}, \dots, u^{(p-1)a}$ в строки с номерами $a, q + a, 2q + a, \dots, (p-1)q + a$. Умножаем $K \otimes I_q$ на $D_{p,q,1} \cdot (F_q \otimes I_p)$ по правилу: k -я строка есть линейная комбинация строк второго сомножителя с коэффициентами из k -строки первого сомножителя. Так как носители u слагаемых строк дизъюнктивны, то получаем строку

$$\begin{aligned} & (1 \ u^a \ u^{2a} \dots u^{(p-1)a} \quad w^a (u^a w^a) (u^{2a} w^a) \dots (u^{(p-1)a} w^a) \dots \\ & \dots w^{(q-1)a} (u^a w^{(q-1)a}) (u^{2a} w^{(q-1)a}) \dots (u^{(p-1)a} w^{(q-1)a})). \end{aligned} \quad (6)$$

В каждом блоке правой матрицы выбирали строку с одним и тем же номером a , поэтому начальные множители каждой группы есть элементы ДЭФ $\mathbf{r}^a = (1 \ w^a \ w^{2a} \dots w^{(q-1)a})$ (a -й строки матрицы F_q). Выразим все элементы через u из равенства $w = u^p$ и получим в (6) прямую нумерацию по степеням u^a , т.е. ДЭФ $\mathbf{r}^a$ порядка $N = pq$.

Перейдем к второму случаю для трех сомножителей $N = pqr$.

Из доказываемой формулы выделим три сомножителя:

$$(F_p \otimes I_{qr}) \cdot D_{p,q,r} \cdot (F_q \otimes I_{pr}) \quad (7)$$

и сравним со случаем $r = 1$: $\tilde{F}_{p,q} = (F_p \otimes I_q) \cdot D_{p,q,1} \cdot (F_q \otimes I_p)$. Простой структурный анализ проведем по формуле леммы 2 при $r = 1$ и в общем случае:

$$(F_p \otimes I_{qr}) \cdot (F_q \otimes I_{pr}) = (F_q \otimes F_p) \otimes I_r.$$

Чтобы в блочных столбцах матрицы (7) сохранялся вид столбца, каждый блок диагонали матрицы $D_{p,q,1}$ дублируем r раз, что приводит к $D_{p,q,r}$. Это приводит к формуле

$$\tilde{F}_{p,q} \otimes I_r = (F_p \otimes I_{qr}) \cdot D_{p,q,r} \cdot (F_q \otimes I_{pr}),$$

что аккуратно проверяется непосредственным просчетом.

При подстановке во вторую формулу утверждения теоремы получим

$$\tilde{F}_{p,q,r} = (\tilde{F}_{p,q} \otimes I_r) \cdot D_{p,q,r,1} \cdot (F_r \otimes I_{pq}),$$

что проверим далее.

Так как $\tilde{F}_{p,q}$ есть перестановка строк матрицы F_{pq} , то и $\tilde{F}_{p,q} \otimes I_r$ есть перестановка строк $F_{pq} \otimes I_r$. Следовательно, $(\tilde{F}_{p,q} \otimes I_r) \cdot D_{p,q,r,1} \cdot (F_r \otimes I_{pq})$ есть перестановка строк

$(F_{pq} \otimes I_r) \cdot D_{pq,r,1} \cdot (F_r \otimes I_{pq})$. Поэтому достаточно проанализировать первый (образующий) столбец этой матрицы.

В первом столбце матрицы $\tilde{F}_{p,q}$ приведены все различные степени (но в реверсном порядке) от 0 до $pq - 1$ элемента $\omega = \omega_{pq}$. В матрице $(\tilde{F}_{p,q} \otimes I_r)$ эти столбцы повторяются со смещением, а в произведении матриц $(\tilde{F}_{p,q} \otimes I_r) \cdot D_{pq,r,1}$ входят с множителями, собранными в векторе $s = (1 \ \omega \ \omega^2 \ \dots \ \omega^{r-1})$, где $\omega = \omega_{pqr}$. Таким образом, образующий столбец проверяемой матрицы содержит все степени ω в реверсном порядке. Теорема доказана.

Обобщение теоремы 1 для большого числа множителей очевидно. При обобщении теоремы 2 для четырех множителей остановимся только на формулировке результата.

Следствие 2. Для ДПФ порядка $N = pqr$ верны факторизации

$$F_N = (F_p \otimes I_{qrs}) \cdot D_{p,qrs,1} \cdot (I_p \otimes ((F_q \otimes I_{rs}) \cdot D_{q,rs,1})) \cdot (I_{pq} \otimes ((F_r \otimes I_s) \cdot D_{r,s,1})) \cdot L,$$

$$F_N = L \cdot (F_p \otimes I_{qrs}) \cdot D_{p,q,rs} \cdot (F_q \otimes I_{prs}) \cdot D_{pq,r,s} \cdot (F_r \otimes I_{pqs}) \cdot D_{pqr,s,1} \cdot (F_s \otimes I_{pqr}),$$

где $L = I_s \otimes I_r \otimes I_q \otimes I_p$.

4. БЫСТРОЕ ПРЕОБРАЗОВАНИЕ КРЕСТЕНСОНА С ОСНОВАНИЕМ p^n

В прикладных задачах часто исследуются сигналы с очень большим числом отчетов с помощью ДПФ такого же большого основания. Это вынуждает увеличивать разрядность чисел для сохранения точности вычислений. Альтернативой является замена ДПФ на дискретное преобразование Крестенсона (ДПК).

Рассмотрим возникающий при этом случай ДПК с матрицей $F_N^{\otimes m}$ при $N = p^n$, когда ставится вопрос о возможности применения суперпозиции рассмотренных быстрых алгоритмов из утверждений 5 и 3.

Теорема 3. *Предлагается следующая факторизация матрицы:*

$$F_{p^n}^{\otimes m} = \left((I_p^{\otimes n} \otimes I_{N^{m-1}}) \cdot (T \otimes I_{N^{m-1}}) \prod_{k=1}^{n-1} ((D_{p^k, p, p^{n-k-1}} \otimes I_{N^{m-1}}) \cdot (T \otimes I_{N^{m-1}})) \right)^m,$$

где $T = F_p \otimes I_{p^{n-1}}$, $N = p^n$.

Доказательство. Пусть

$$\tilde{F}_N = T \cdot D_{p,p,p^{n-2}} \cdot T \cdot D_{p^2,p,p^{n-3}} \cdot T \cdot D_{p^3,p,p^{n-4}} \dots \cdot T \cdot D_{p^{n-1},p,1} \cdot T$$

в утверждении 5, которое примет вид $F_{p^n} = I_p^{\otimes n} \cdot \tilde{F}_N$.

По утверждению 3 имеем $F_{p^n}^{\otimes m} = (F_{p^n} \otimes I_{N^{m-1}})^m$. Преобразуем выражение в скобках по второй формуле леммы 1

$$F_{p^n} \otimes I_{N^{m-1}} = (I_p^{\otimes n} \cdot \tilde{F}_N) \otimes (I_{N^{m-1}} \cdot I_{N^{m-1}}) = (I_p^{\otimes n} \otimes I_{N^{m-1}}) \cdot (\tilde{F}_N \otimes I_{N^{m-1}}).$$

По первой формуле леммы 1 к каждому сомножителю $\tilde{F}_N$ добавляется справа одинаковый кронекеров множитель

$$\tilde{F}_N \otimes I_{N^{m-1}} = (T \otimes I_{N^{m-1}}) \cdot (D_{p,p,p^{n-2}} \otimes I_{N^{m-1}}) \cdot (T \otimes I_{N^{m-1}}) \cdot \dots \cdot (D_{p^{n-1},p,1} \otimes I_{N^{m-1}}) \cdot (T \otimes I_{N^{m-1}}).$$

Далее согласно утверждению 3 проводится m итераций полученной формулы для $F_{p^n} \otimes I_{N^{m-1}}$. Теорема доказана.

Возможны и другие, более громоздкие, факторизации матрицы $F_{p^n}^{\otimes m}$. Именно постоянство структуры выбранного варианта алгоритма выделяет его как оптимальный вариант. Рекомендуется для сокращения числа операций при очень больших N брать $p = 4$, n равным 2 или 3 и варьировать m .

СПИСОК ЛИТЕРАТУРЫ

1. *Cooley J.W., Tukey J.W.* An algorithm for the machine calculation of complex Fourier series // *Math. Comput.* 1965. V. 19 (90). P. 297–301.
2. *Залманзон Л.А.* Преобразования Фурье, Уолша, Хаара и их приложения в управлении, связи и других областях. М.: Наука, 1989. 496 с.
3. *Беспалов М.С.* О свойствах тензорного произведения матриц // *Ж. вычисл. матем. и матем. физ.* 2014. Т. 54. № 4. С. 547–561.
<https://doi.org/10.1134/S0965542514040046>
4. *Good I.J.* The interaction algorithm and practical Fourier analysis // *J. Royal Stat. Soc.* 1958. Ser. B. V. 20. P. 361–372.
5. *Малоземов В.Н., Машарский С.М.* Основы дискретного гармонического анализа. СПб.: Лань, 2012. 304 с.
6. *Беспалов М.С.* Новые разложения кронекеровой степени по Гуду // *Проблемы передачи информации.* 2018. Т. 54. № 3. С. 62–66.
<https://doi.org/10.1134/S0032946018030043>
7. *Трахтман А.М., Трахтман В.А.* Основы теории дискретных сигналов на конечных интервалах. М.: Сов. радио, 1975.
8. *Малоземов В.Н., Машарский С.М., Цветков К.Ю.* Сигнал Франка и его обобщения // *Проблемы передачи информации.* 2001. Т. 37. № 2. С. 18–26.
9. *Малоземов В.Н., Машарский С.М.* Обобщенные вейвлетные базисы, связанные с дискретным преобразованием Виленкина–Крестенсона // *Алгебра и анализ.* 2001. Т. 13. Вып. 1. С. 111–157.
10. *Машарский С.М.* Быстрое преобразование Виленкуина – Крестенсона на основе факторизации Гуда // *Ж. вычисл. матем. и матем. физ.* 2002. Т. 42, № 6. с. 784–790.
11. *Беспалов М.С.* Дискретные преобразования Крестенсона // *Проблемы передачи информации.* 2010. Т. 46. № 4. С. 91–115.
<https://doi.org/10.1134/S003294601004006X>
12. *Johnson J., Johnson R.W., Rodriguez D., Tolimieri R.* A methodology for designing, modifying and implementing Fourier transform algorithms on various architectures // *Circuits, Systems and Signal Processing.* 1990. V. 9. № 4. P. 449–500.
13. *Малоземов В.Н., Просеков О.В.* Факторизация Кули–Тьюки матрицы Фурье // *Избранные главы дискретного гармонического анализа и геометрического моделирования.* Ч. I. Изд. 2-е. Под ред. проф. В. Н. Малоземова. СПб.: Изд-во ВВМ. 2014. С. 20–29.

ОБЩИЕ ЧИСЛЕННЫЕ МЕТОДЫ

УДК 519.16

АППРОКСИМАЦИЯ ДИФФЕРЕНЦИАЛЬНЫХ ОПЕРАТОРОВ С УЧЕТОМ ГРАНИЧНЫХ УСЛОВИЙ

© 2023 г. В. П. Варин^{1,*}

¹ 125047 Москва, Миусская пл., 4, Институт прикладной математики им. М.В. Келдыша РАН, Россия

*e-mail: varin@keldysh.ru

Поступила в редакцию 11.01.2023 г.
Переработанный вариант 19.02.2023 г.
Принята к публикации 30.03.2023 г.

Применение спектральных методов для решения краевых задач является весьма эффективным, но сопряженным с большими техническими трудностями, связанными с учетом граничных условий. Существует несколько способов такого учета, но все они либо весьма трудоемки, либо требуют предварительного анализа задачи и записи ее в интегральной форме. Мы предлагаем универсальный способ учета граничных условий для линейных дифференциальных операторов на конечном отрезке, который весьма прост в реализации. Применение рациональной арифметики позволяет оценить эффективность метода без учета ошибок округления. Мы применили наш подход к вычислению рациональных приближений для некоторых фундаментальных констант. Получены приближения, которые в ряде случаев лучше, чем те, которые дают обыкновенные цепные дроби этих констант. Библ. 22. Фиг. 2.

Ключевые слова: полиномы Лежандра, краевые задачи, доказательства иррациональности, голономные функции, высокоточные вычисления.

DOI: 10.31857/S004446692308015X, **EDN:** IIGDWE

1. ВВЕДЕНИЕ

Численные методы решения краевых задач на конечном интервале можно условно разделить на сеточные и спектральные. В сеточных методах дифференциальные операторы аппроксимируются с помощью конечных разностей, а функции приближаются с помощью таблиц их значений на выбранной сетке. В спектральных методах функции приближаются отрезками разложений этих функций по некоторым наборам пробных функций, в качестве которых чаще всего используются полиномы, ортогональные на данном интервале с некоторым весом.

Среди спектральных методов выделяются два, которые обладают уникальными свойствами, выделяющими их из всех прочих. Это разложение периодических функций в ряды Фурье и разложение функций, аналитических в некоторой окрестности отрезка, по (смещенным) полиномам Чебышёва (что можно рассматривать как модификацию метода Фурье).

Принципиальное отличие этих двух методов от всех прочих спектральных состоит в том, что они одновременно являются также и сеточными, что гарантируется дискретным преобразованием Фурье или Фурье–Чебышёва.

Эти преобразования, которые также можно записать в матричной форме в явном виде, всегда хорошо обусловлены, являются (почти) инволютивными, и не добавляют погрешности к результату при вычислениях с конечной разрядной сеткой (см. [1]).

Второе важное отличие этих двух методов от (почти) всех остальных спектральных — это сходимость к решению не только в среднем, что гарантировано изначально, но также и равномерно на выбранном интервале при весьма слабых ограничениях на решение задачи.

Эти свойства позволяют по таблице полученного решения сделать заключение об аналитических свойствах решения, которые редко заранее известны. Иными словами, скорость убывания коэффициентов Фурье–(Чебышёва) находится в хорошо известной зависимости от гладкости решения. Например, для аналитических функций коэффициенты Фурье–(Чебышёва) убывают экспоненциально быстро.

Для решений, сингулярных в концах интервала, коэффициенты Фурье–(Чебышёва) убывают обычно все же достаточно быстро, чтобы метод был эффективным.

Поэтому, получив таблицу коэффициентов Фурье решения, можно сделать ряд важных заключений о свойствах полученного решения, которые невозможно сделать в случае применения сеточных методов. Помимо выводов о гладкости решения, можно достоверно оценить достигнутую глобальную точность, а также оценить достаточность (или нет) выбранной размерности аппроксимации при данной разрядной сетке.

Изложенные соображения были положены К.И. Бабенко в основу т.н. *методов без насыщения* (см. [2]).

К сожалению, Бабенко не дал в своей книге простого и понятного определения этих методов, а явление насыщения излагал в рамках общей теории аппроксимаций. Это послужило поводом к большому количеству спекуляций на тему методов без насыщения в русскоязычной литературе.

Применение спектральных методов часто сопряжено с большими техническими трудностями, связанными с учетом краевых условий задачи. Например, книга [1] дает практически все о полиномах Чебышёва, что нужно при решении практических задач. Однако в примерах решения задач учет краевых условий там осуществляется путем подбора вручную уравнений для нужных коэффициентов. Похожий подход применялся в [3], р. 119.

Эта же проблема возникает в т.н. τ -методе Ланцоша (см. [4]), который также является по сути спектральным. Учет краевых условий в τ -методе проводится путем добавления уравнений для некоторых малых параметров τ_i , $i = 1, 2, \dots$, а также выбрасывания некоторых уравнений для коэффициентов разложения решения из полученной переопределенной системы, но каждый раз в индивидуальном порядке.

Наиболее общий метод учета краевых условий — это обращение дифференциального оператора и запись задачи в интегральной форме. Однако обращение дифференциального оператора с учетом граничных условий — это весьма нетривиальная и очень трудоемкая задача, которая каждый раз требует индивидуального подхода.

Другой (и весьма малоизвестный) способ учета краевых условий был предложен Бабенко (см. [2]).

Бабенко предложил способ аппроксимации некоторых типовых дифференциальных операторов с учетом граничных условий так, как если бы аппроксимировались интегральные. Далее нужный дифференциальный оператор собирается из имеющихся прототипов так, что его граничные условия достаточно хорошо аппроксимируются (подробности см. в [5]).

Однако и этот способ учета краевых условий задачи является весьма трудоемким и имеющим ограниченное применение (см. ниже).

Таким образом, проблема учета краевых условий в спектральных методах остается (с практической точки зрения) открытой и пока не имеет общего и эффективного решения.

В этой статье мы даем, в определенном смысле, полное решение этой задачи для конкретного спектрального метода — разложения функций на интервале $[0, 1]$ по смещенным полиномам Лежандра (т.е. с единичным весом).

Выбор этих полиномов обусловлен тем, что разложения по ним также сходятся равномерно (при определенных условиях), т.е. полученные аппроксимации будут давать методы без насыщения.

Вторая причина, по которой мы выбрали эти полиномы — это наличие приложений данного метода к аналитической теории чисел, что возможно при выборе рациональной арифметики в расчетах и использовании компьютерной алгебры (CAS).

Учет краевых условий в разложениях по полиномам Чебышёва, на самом деле, осуществляется таким же образом. Однако эти разложения больше подходят для численных расчетов, и им будет посвящена отдельная статья.

В следующем разделе мы приведем готовые формулы для некоторых типовых дифференциальных операторов, а также покажем, как моделируются (в принципе) любые алгебраические линейные дифференциальные операторы с любыми краевыми условиями для них. В этом смысле наш подход напоминает (чисто внешне) метод Бабенко.

В последующих разделах мы дадим ряд примеров приложений нашего метода, эффективность которого особенно видна при использовании рациональной арифметики.

2. АРИФМЕТИКА РАЗЛОЖЕНИЙ ПО ПОЛИНОМАМ ЛЕЖАНДРА

Везде в этой статье рассматриваются аналитические функции на отрезке $[0, 1]$, которые могут иметь особенности на концах интервала. Эти функции раскладываются в ряды по смещенным полиномам Лежандра, и эти разложения сходятся (вообще говоря) равномерно на любом подынтервале отрезка $[0, 1]$.

Так же как и при вычислениях с обычными числами, мы всегда имеем дело лишь с конечным представлением функций в виде отрезков рядов. Поэтому необходимо определить операции на данном множестве функций так, чтобы разложения автоматически обрывались нужным образом. Для сложения и вычитания функций это очевидно, однако умножение, дифференцирование и интегрирование функций представляют некоторую проблему, которой мы и займемся.

Пусть $\tilde{P}(n, x)$ — это обычные полиномы Лежандра, ортогональные на отрезке $[-1, 1]$ с единичным весом. Мы используем смещенные полиномы Лежандра, $P(n, x) = \tilde{P}(n, 2x - 1)$, ортогональные на отрезке $[0, 1]$,

$$P(n, x) = \frac{1}{n!} \frac{d^n}{dx^n} (x^n (x-1)^n), \quad \int_0^1 P^2(n, x) dx = \frac{1}{2n+1}, \quad n \in \mathbb{N}_0.$$

Пусть N — это размерность аппроксимации, т.е. любая функция $f(x)$ в этом разделе понимается как отрезок ее разложения по полиномам $P(n, x)$ до степени $N-1$, т.е. $f \in \mathcal{P}_N$. Это представление функций аналогично представлению чисел в арифметике с фиксированной разрядной сеткой. Ясно, что дифференцирование и интегрирование таких функций нарушают это представление и требуют переразложения результата операции по полиномам $P(n, x)$ до номера $N-1$ включительно. Это же справедливо для умножения функций. Например,

$$x \sum_{k=0}^{N-1} a_k P(k, x) = \sum_{k=0}^{N-1} b_k P(k, x),$$

где равенство понимается в проективном смысле, т.е. при известных коэффициентах a_n коэффициенты b_n необходимо определить так, чтобы разность этих функций была ортогональна полиномам $P(n, x)$, $n < N$.

Таким образом, любая функция $f(x)$ представлена N -мерным вектором $f = \langle a_0, a_1, \dots, a_{N-1} \rangle^t$, а операции умножения на x , дифференцирование и т.д. — это линейные операторы в N -мерном векторном пространстве.

Мы будем обозначать векторы обычными, а линейные операторы — большими латинскими буквами. Введем следующие обозначения:

e — единичный вектор, т.е. $e(x) = 1$, $e = \langle 1, 0, \dots, 0 \rangle^t$;

E — единичная матрица;

X — оператор умножения на x ;

$X2$ — оператор умножения на x^2 ;

x — вектор, соответствующий функции x , т.е. $x = \langle \frac{1}{2}, \frac{1}{2}, 0, \dots, 0 \rangle^t$;

D — оператор дифференцирования по x ;

I — оператор формального интегрирования по x .

Таким образом,

$$x = X.e, \quad f'(x) = D.f, \quad \int f(x) dx = I.f, \quad \text{и т.д.,}$$

где точка обозначает применение оператора к вектору либо умножение матриц, а интеграл неопределенный, т.е. константа интегрирования неизвестна и может меняться при изменении N .

Заметим также, что $X.X \neq X2$, хотя эти две матрицы отличаются только в одном элементе, $(X2 - X.X)[N, N] = N^2/(4N^2 - 1)/4$. Однако оператор $X2$ все же лучше в численных (но не символьных) расчетах, чем $X.X$, но не настолько, чтобы вводить отдельные операторы $X3$, $X4$ и т.д.

Мы приведем готовые формулы для всех перечисленных операторов. Вывод этих формул весьма громоздок, и мы его опускаем. Сами формулы легко проверяются и могут быть доказаны с помощью математической индукции.

Матрицы X и $X2$ являются соответственно трех- и пятидиагональными. Приведем для примера эти матрицы для размерности $N = 6$:

$$X = \begin{bmatrix} \frac{1}{2} & \frac{1}{6} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{5} & 0 & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{2} & \frac{3}{14} & 0 & 0 \\ 0 & 0 & \frac{3}{10} & \frac{1}{2} & \frac{2}{9} & 0 \\ 0 & 0 & 0 & \frac{2}{7} & \frac{1}{2} & \frac{5}{22} \\ 0 & 0 & 0 & 0 & \frac{5}{18} & \frac{1}{2} \end{bmatrix}, \quad X2 = \begin{bmatrix} \frac{1}{3} & \frac{1}{6} & \frac{1}{30} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{2}{5} & \frac{1}{5} & \frac{3}{70} & 0 & 0 \\ \frac{1}{6} & \frac{1}{3} & \frac{8}{21} & \frac{3}{14} & \frac{1}{21} & 0 \\ 0 & \frac{1}{10} & \frac{3}{10} & \frac{17}{45} & \frac{2}{9} & \frac{5}{99} \\ 0 & 0 & \frac{3}{35} & \frac{2}{7} & \frac{29}{77} & \frac{5}{22} \\ 0 & 0 & 0 & \frac{5}{63} & \frac{5}{18} & \frac{44}{117} \end{bmatrix}.$$

Для матрицы X диагонали образуют последовательности, начиная с нижней, соответственно,

$$\frac{n}{2(2n-1)}, \frac{1}{2}, \frac{n}{2(2n+1)}, \quad n \in \mathbb{N}.$$

Для матрицы $X2$ диагонали образуют последовательности, начиная с нижней, соответственно,

$$\frac{n(n+1)}{4(2n-1)(2n+1)}, \frac{n}{2(2n-1)}, \frac{3n^2-3n-2}{2(2n+1)(2n-3)}, \frac{n}{2(2n+1)}, \frac{n(n+1)}{4(2n+3)(2n+1)}, \quad n \in \mathbb{N}.$$

Проще всего устроена матрица дифференцирования D , а сложнее всего — матрица интегрирования I . Эти матрицы для размерности $N = 6$ имеют вид:

$$D = \begin{bmatrix} 0 & 2 & 0 & 2 & 0 & 2 \\ 0 & 0 & 6 & 0 & 6 & 0 \\ 0 & 0 & 0 & 10 & 0 & 10 \\ 0 & 0 & 0 & 0 & 14 & 0 \\ 0 & 0 & 0 & 0 & 0 & 18 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad I = \begin{bmatrix} \frac{1}{2} & -\frac{1}{6} & -\frac{1}{4} & \frac{5}{16} & \frac{3}{16} & -\frac{7}{16} \\ \frac{1}{2} & 0 & -\frac{1}{10} & 0 & 0 & 0 \\ 0 & \frac{1}{6} & 0 & -\frac{1}{14} & 0 & 0 \\ 0 & 0 & \frac{1}{10} & 0 & -\frac{1}{18} & 0 \\ 0 & 0 & 0 & \frac{1}{14} & 0 & -\frac{1}{22} \\ 0 & 0 & 0 & 0 & \frac{1}{18} & 0 \end{bmatrix}.$$

Элемент с индексами $[m, n]$ в верхнетреугольной матрице D имеет вид

$$D[m, n] = \begin{cases} 2(2m-1), & m+n \text{ нечетно,} \\ 0, & m+n \text{ четно.} \end{cases}$$

Матрица I является трехдиагональной, за исключением первой строки. Элементы главной диагонали нулевые. Ненулевой элемент с индексами $[m, n]$ в матрице I имеет вид

$$I[m, n] = \begin{cases} c(n), & m=1, \\ \frac{1}{2(2m-3)}, & m-n=1, \\ \frac{-1}{2(2m+1)}, & n-m=1, \end{cases}$$

где функция номера столбца $c(n)$ дается формулой

$$c(n) = \begin{cases} -\frac{1}{6}, & n = 2, \\ (-1)^{\frac{n-1}{2}} \Gamma\left(\frac{n}{2}\right) / \Gamma\left(\frac{n+1}{2}\right) / 2/\sqrt{\pi}, & n \text{ нечетно}, \\ (-1)^{\frac{n}{2}} 2\Gamma\left(\frac{n+3}{2}\right) / \Gamma\left(\frac{n}{2}-1\right) / (n-1)/n/\sqrt{\pi}, & n \text{ четно}. \end{cases}$$

Матрица X размерности N обладает важным свойством, которое нам понадобится. Справедливо

Предложение 1. Матрица $P(N, X)$ нулевая, $N = 1, 2, \dots$

Доказательство. Введем вектор $p = \langle P(0, x), P(1, x), \dots, P(N-1, x) \rangle^t$. Для трехдиагональной матрицы Якоби J размерности N , ассоциированной с системой ортогональных полиномов Лежандра $\{P(n, x), n = 0, 1, \dots\}$, справедливо тождество

$$xp = J.p + \left\langle 0, 0, \dots, \frac{N}{2(2N-1)} P(N, x) \right\rangle^t, \quad (1)$$

которое является прямым следствием трехзвенного линейного рекуррентного соотношения для этих полиномов (см. [6]).

Результат умножения любого вектора $f \in \mathcal{P}_N$ на x можно записать как скалярное произведение этого вектора с левой частью (1). Поэтому транспонируя (1), получим $X = J^t$.

Но спектр матрицы J — это корни полинома $P(N, x)$, что следует из (1). Поэтому матрица J , а значит, и X аннулируются своим характеристическим полиномом $P(N, x)$ (теорема Кэли–Гамильтона). Что и требовалось доказать.

Как следствие, получаем,

$$P(n, x) = \frac{4^n \Gamma(n+1/2)}{\sqrt{\pi} n!} \det(xE - X), \quad n \in \mathbb{N}.$$

Также при нечетном N все матрицы $P(M, X)$ для нечетных M вырождены (но ненулевые). Это связано с тем, что у всех полиномов $P(M, x)$ нечетной степени есть общий рациональный корень $x = 1/2$. Для четных N все матрицы $P(M, X)$, $M \neq N$ невырождены.

Обратим внимание, что последняя строка матрицы дифференцирования D всегда нулевая, т.е. этот оператор всегда вырожден, а его матрица имеет ранг $N-1$. Это прямое следствие того, что ОДУ $y'(x) = f(x)$ имеет одномерное многообразие решений. Иными словами, для однозначного определения решения $y(x)$ необходимо задать начальное (либо краевое) условие.

Эти условия можно учесть совершенно естественным образом в последней строке матрицы D . Поскольку

$$P(k-1, 0) = (-1)^{k+1}, \quad P(k-1, 1) = 1, \quad k \in \mathbb{N},$$

то последнюю строку матрицы D надо заполнить этими значениями для учета, соответственно, начального (т.е. при $x = 0$) или краевого (т.е. при $x = 1$) условия (или их линейной комбинации). Затем нужно заменить последний элемент вектора f в правой части на нужное начальное (краевое) условие. Тогда полученная линейная система будет невырожденна, а полученное решение $y(x)$ будет автоматически иметь нужное начальное (краевое) условие.

Последнее утверждение можно сформулировать точно. Справедливо

Предложение 2. Пусть правая часть $f(x)$ ОДУ $y'(x) = f(x)$ имеет вид

$$f(x) = \sum_{k=0}^{N-2} a_k P(k, x), \quad (2)$$

т.е. последний элемент вектора $f = f(x)$ (который заменяется на граничное условие) равен нулю. Тогда решение $y(x)$ находится точно в виде полинома описанным выше способом при любом начальном (краевом) условии или их линейной комбинации.

Доказательство. Для того, чтобы ОДУ $y'(x) = f(x)$ имело решение $y \in \mathcal{P}_N$ необходимо и достаточно, чтобы правая часть f имела вид (2). Любое линейное краевое условие учитывается подбором константы интегрирования, для чего достаточно одного линейного уравнения. Что и требовалось доказать.

Можно дать еще одну интерпретацию описанного способа решения ОДУ $y'(x) = f(x)$ с полиномиальной правой частью.

Как было отмечено, ранг матрицы аппроксимации оператора D равен $N - 1$. Поэтому если решать полученную вырожденную линейную систему методом Гаусса с выбором главного элемента по всей матрице, то на последнем шаге N получится уравнение $0 \times \tilde{y}[N] = \tilde{f}[N]$, где тильда символизирует операции, проделанные в методе Гаусса. Последнее уравнение может быть совместно, только если $\tilde{f}[N] = 0$, т.е. $f(x)$ — это полином вида (2).

Практически можно взять любое значение $\tilde{y}[N]$, а затем скорректировать начальное (краевое) условие полученного решения, как требуется. Эта операция полностью эквивалентна неопределенному интегрированию функции, т.е. третий способ решения ОДУ $y'(x) = f(x)$ — это вычисление решения $y = I.f$, где вектор f вида (2), с последующей корректировкой константы интегрирования.

В операторе $D^2 = D.D$ будут, очевидно, две нулевые последние строки, которые следует заполнить нужными краевыми условиями. Вообще, для учета краевых условий нашим методом необходимо и достаточно знать следующие векторы-строки, линейные комбинации которых заполняют последние строки нужных линейных дифференциальных операторов. Эти векторы размерности N имеют вид

$$i_m = \langle P^{(m)}(k-1, 0) \rangle, \quad b_m = \langle P^{(m)}(k-1, 1) \rangle, \quad k = 1, \dots, N, \quad m \in \mathbb{N}_0,$$

где m обозначает номер производной полинома $P(n, x)$ по x . Векторы i_m отвечают за граничные значения функции при $x = 0$, а b_m — за граничные значения при $x = 1$. Эти векторы устроены весьма просто:

$$\begin{aligned} i_0 &= \langle (-1)^{k+1} \rangle, & b_0 &= \langle 1 \rangle, \\ i_1 &= \langle (-1)^k k(k-1) \rangle, & b_1 &= \langle k(k-1) \rangle, \\ i_2 &= \left\langle \frac{(-1)^{k+1} k(k-2)(k^2-1)}{2} \right\rangle, & b_2 &= \left\langle \frac{k(k-2)(k^2-1)}{2} \right\rangle, \\ &\dots & & \end{aligned}$$

причем этот список может быть легко продолжен (с использованием CAS).

Любые начальные либо краевые линейные условия для функций $f(x)$, $f'(x)$, $f''(x)$ и т.д. и их комбинаций получаются в виде линейных комбинаций перечисленных векторов.

Аналогично можно задать условия и внутри интервала, но у нас пока нет содержательной задачи для демонстрации этого факта.

Приведем численный пример, на котором видны преимущества данного метода. Рассмотрим спектральную задачу на интервале $x \in [0, 1]$:

$$\frac{d^2 y}{dx^2} + a^2 y = 0, \quad y(0) = 0, \quad y(1) + y'(1) = 0, \quad (3)$$

с нестандартными краевыми условиями. Эта задача приводилась в [7] для демонстрации приложения τ -метода.

Спектральная задача (3) решается в явном виде. Несложные вычисления показывают, что все собственные значения являются решениями трансцендентного уравнения $x + \tan x = 0$, т.е.

$$a_1 \approx 2.028757838110, \quad a_2 \approx 4.913180439434, \quad a_3 \approx 7.978665712413, \dots$$

Если решать эту задачу методом Бабенко, то необходимо сперва сконструировать интерполяционный полином Чебышёва с указанными краевыми условиями. Затем этот полином нужно дважды продифференцировать и получить таким образом конечномерную спектральную задачу. Мы проделали эти вычисления, но они весьма громоздки, и мы их опускаем.

Согласно нашему методу, нужно взять матрицу $A = D.D$ и заменить ее две последние строки на векторы i_0 и $b_0 + b_1$ (в любом порядке). Полученная спектральная задача размерности $N = 10$ дает приближенные значения a_1, a_2, a_3 , соответственно, с шестью, тремя и одним десятичными разрядами при вычислениях в обычной арифметике, что соответствует точности, полученной для этой задачи в [7] с помощью τ -метода той же размерности.

При увеличении N количество полученных собственных значений увеличивается, а их точность возрастает, т.е. спектр стабилизируется. Никаких “паразитных” собственных значений в нашем методе не появляется, т.е. данный метод является методом без насыщения.

В модификации τ -метода, которая использовалась в [7], для учета краевых условий к разложению функции добавляются дополнительные члены, уравнения для коэффициентов которых получаются из краевых условий. Это требует весьма громоздких вычислений, предназначенных только для данного конкретного случая. В результате для системы (3) τ -метод в [7] дает спектральную задачу для пучка матриц, т.е. $Au = \lambda Bu$, где матрицы A и B имеют блочную структуру, собранную из отдельных кусков.

Отметим также, что, помимо учета краевых условий, наш подход схож с операторным подходом к τ -методу, предложенному в [8]. Однако в [8] используются бесконечные матрицы, которые обрезаются нужным образом для получения конечномерных аппроксимаций операторов. Затем эти конечномерные матрицы обрамляются дополнительными блоками (примерно как в [7]) для учета краевых условий.

3. КВАДРАТУРЫ ГАУССА И ПАДЕ-АППРОКСИМАЦИИ

Вычислительные эксперименты показали (неожиданно высокую) численную эффективность предложенного алгоритма. При расчетах в обычной плавающей арифметике обнаруженные нами эффекты ненаблюдаемы, и, вероятно, поэтому они (насколько нам известно) никогда ранее не отмечались в спектральных методах.

Вычисления с расширенной разрядной сеткой, а также в рациональной арифметике, проделанные нами в CAS Maple, позволили установить связь нашего метода с квадратурами Гаусса и паде-аппроксимациями функций.

Напомним, что в классической M -точечной квадратуре Гаусса с единичным весом на интервале $[-1, 1]$ требуется вычислить M корней полинома Лежандра $\tilde{P}(M, x)$ и M весов квадратурной формулы (см. [9]). На самом деле достаточно вычислить половину из каждого набора ввиду их симметрии, но проще вычислить их все сразу. Это возможно сделать только в плавающей арифметике. Тогда полученная квадратура будет точна на полиномах степени не выше $2M - 1$.

Поскольку при нашем подходе функция приближается полиномами Лежандра степени не выше $N - 1$, то можно было ожидать, что данный метод будет соответствовать квадратуре Гаусса порядка $[N/2]$. Приведем пример, показывающий, что точность на самом деле значительно выше.

Рассмотрим задачу Коши

$$y'(x) = \frac{1}{1+x^2} = r(x), \quad x \in [0, 1], \quad y(0) = 0, \quad (4)$$

решением которой является функция $y(x) = \arctan x$.

Выберем для этого примера размерность аппроксимации $N = 10$.

Правая часть уравнения (4) может быть сразу записана в виде вектора $r = (E + X^2)^{-1}e$ (см. теорему ниже), т.е. мы получаем приближенное разложение рациональной функции $r(x)$ по сдвинутым полиномам Лежандра, просто обратив нужную матрицу. Обозначим точные коэффициенты разложения через $\tilde{r}_k$:

$$\tilde{r}_k = (2k - 1) \int_0^1 r(x) P(k - 1, x) dx, \quad k = 1, 2, \dots, N.$$

Тогда

$$|r_k - \tilde{r}_k| \approx \{3 \times 10^{-15}, 3 \times 10^{-15}, 1 \times 10^{-13}, 6 \times 10^{-13}, 6 \times 10^{-13}, 2 \times 10^{-11}, 6 \times 10^{-11}, 2 \times 10^{-10}, 2 \times 10^{-9}, 4 \times 10^{-9}\}. \quad (5)$$

Поскольку

$$\tilde{r}_1 = \int_0^1 r(x)P(0, x)dx = \int_0^1 \frac{1}{1+x^2} dx = \frac{\pi}{4}, \quad (6)$$

то мы вычислили интеграл от правой части (4), даже не решая дифференциального уравнения. Имеем

$$r_1 = \frac{281845376409124}{358856678744865}, \quad \frac{\pi}{4} - r_1 \approx 3.057054520150 \times 10^{-15}. \quad (7)$$

Уместно сравнить полученную точность с той, которую можно достичь с помощью квадратур Гаусса при вычислении интеграла (6). Обозначим соответствующие квадратуры G_M . Тогда

$$\frac{\pi}{4} - G_{10} \approx -6.3 \times 10^{-14}, \quad \frac{\pi}{4} - G_{11} \approx 3.056102 \times 10^{-15}, \quad \frac{\pi}{4} - G_{12} \approx -4.8 \times 10^{-17},$$

т.е. наш метод соответствует (в данном примере) 11-точечной квадратуре Гаусса.

Теперь вычислим решение задачи Коши (4) согласно нашему методу. Возьмем матрицу $A = D$ и заменим последнюю (нулевую) строку на вектор i_0 . Далее заменим последний (в данном случае ненулевой) элемент вектора r на ноль, согласно начальному значению решения. Получаем конечномерную аппроксимацию задачи (4), $A \cdot y = r$, откуда находим вектор $y = A^{-1} \cdot r$. Таким образом, получаем приближенное разложение арктангенса

$$y(x) = \arctan x \approx \sum_{k=1}^N y_k P(k-1, x).$$

Подставив $x = 1$ в эту формулу, получаем рациональное приближение числа $\pi/4$, которое оказывается в точности равным значению r_1 в (7), вычисленному ранее.

Вычислим паде-аппроксимации арктангенса различной размерности в рациональной арифметике, подставим туда $x = 1$ и сравним с приближением числа $\pi/4$, полученным нами. Оказалось, что

$$P[19,18](y(1)) = \frac{270807143217152}{344802363740835}, \quad P[19,18](y(1)) - \frac{\pi}{4} \approx 3.672520113 \times 10^{-15}.$$

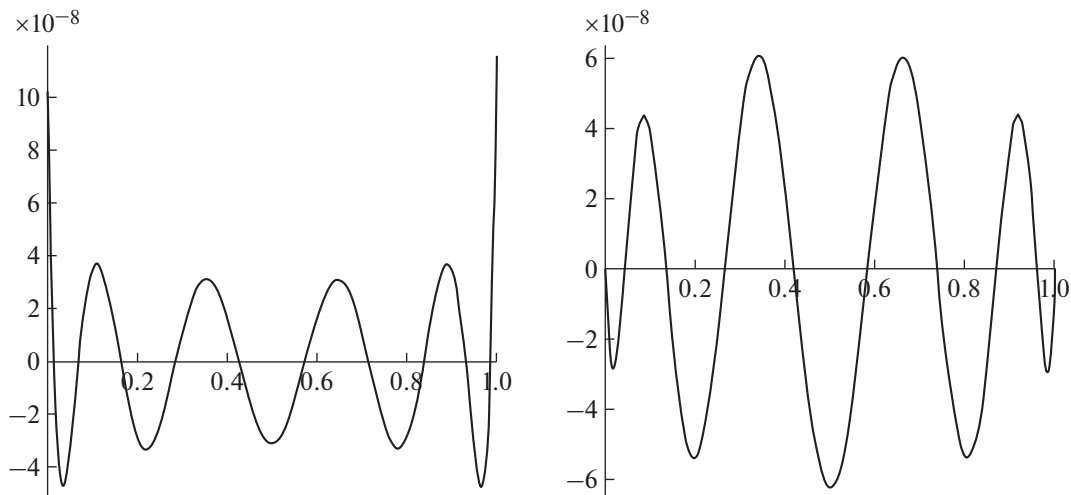
Как мы увидим, близость (и даже схожесть) полученных аппроксимаций не является случайной.

Коэффициенты y_k отличаются от точных значений разложения арктангенса примерно так же, как в (5). На фиг. 1 показаны графики разности $\arctan x$ с их аппроксимациями полиномами Лежандра для $N = 10$. На фиг. 1 слева использовались точные (иррациональные) коэффициенты разложения, а на фиг. 1 справа — полученные нами.

Обратим внимание, что глобальная (или равномерная) погрешность нашей аппроксимации примерно соответствует ожидаемой для данной размерности (и даже лучше). Однако краевое значение в задаче (4) находится с точностью, превышающей равномерную на несколько порядков.

Последнее обстоятельство также не является случайным. Чтобы это увидеть, решим задачу (4) нашим методом еще раз, но без использования матрицы $X2$. Сохраним те же обозначения, но возьмем правую часть уравнения (4) в виде вектора $r = (E + X \cdot X)^{-1} \cdot e$. Тогда

$$|r_k - \tilde{r}_k| \approx \{6 \times 10^{-14}, 4 \times 10^{-13}, 7 \times 10^{-13}, 1 \times 10^{-11}, 4 \times 10^{-11}, 2 \times 10^{-10}, 2 \times 10^{-9}, 3 \times 10^{-9}, 3 \times 10^{-8}, 2 \times 10^{-7}\},$$



Фиг. 1. Аппроксимация $\arctan x$ полиномами Лежандра ($N = 10$).

т.е. точность аппроксимации правой части (4) несколько хуже, чем ранее. Это же справедливо для нового вектора y . Однако теперь

$$r_1 = \frac{140675049238504}{179113035648015}, \quad \frac{\pi}{4} - r_1 \approx -6.326259326040 \times 10^{-14}, \quad (8)$$

что соответствует 10-точечной квадратуре Гаусса (см. выше).

Данное соответствие нашего рационального приближения интеграла его квадратуре Гаусса на самом деле является полным совпадением.

Возьмем произвольную рациональную функцию $r(x)$ с рациональными коэффициентами (можно и комплексно-рациональными или даже символьными). Рассмотрим задачу Коши

$$z'(x) = r(x) = \frac{p(x)}{q(x)}, \quad x \in [0, 1], \quad z(0) = 0. \quad (9)$$

Ее решением является функция $z(x) = \int_0^x r(t)dt$, если она существует. Но нас интересуют формальные символьные вычисления в конечномерной аппроксимации.

Для задачи (9) в случае общего положения (см. ниже) вычислим следующие массивы для размерностей аппроксимации $N = 1, 2, \dots$

$S_1[N] = r_1$, где вектор $r = \langle r_1, \dots, r_N \rangle^t$ вычисляется по формуле

$$r = p(X).q(X)^{-1}e = q(X)^{-1}p(X).e,$$

где коммутирующие матрицы $p(X)$ и $q(X)$ получаются подстановкой матрицы X вместо независимой переменной x в полиномы $p(x)$ и $q(x)$. Напомним, что X^2 мы теперь не используем.

$S_2[N] = \sum_{k=1}^N z_k$, где вектор z вычисляется так же, как вектор y для (4). То есть берем матрицу $A = D$, заменяем последнюю (нулевую) строку на вектор i_0 . Далее заменяем последний (ненулевой) элемент вектора r на ноль, т.е. берем $r[N] = 0$, согласно начальному значению решения. Затем вычисляем вектор $z = A^{-1}.r$, который дает приближенное разложение функции $z(x)$, если она существует. Иначе это просто формальный вектор. Отметим, что $S_2[1] = 0$ в силу способа вычисления.

Этот массив можно также вычислять альтернативным способом с помощью оператора формального интегрирования, как это описано в разд. 2.

$S_3[N] = G_N(r)$, т.е. $S_3[N]$ — это значение N -точечной квадратуры Гаусса функции $r(x)$ на интервале $[0, 1]$.

Напомним, что корни x_n полиномов Лежандра $P(N, x)$ могут быть вычислены (в принципе) в виде радикалов только до $N = 9$ включительно. Это же относится к весам w_n квадратурной формулы Гаусса,

$$w_n = \frac{4x_n(1-x_n)}{N^2 P^2(N-1, x_n)}, \quad n = 1, \dots, N. \quad (10)$$

Разумеется, практически делать это бессмысленно, поэтому массив $S_3[N]$ (в отличие от предыдущих) дается в плавающей арифметике, но с любым (в принципе) числом десятичных разрядов.

Наложим очевидное ограничение на полюса функции $r(x)$, т.е. если $q(x)$ делится на полином Лежандра $P(M, x)$, то в массивах S_1, S_2, S_3 следует пропустить размерность $N = M$ (см. предложение 1). А в случае если $q(x)$ имеет корень $x = 1/2$, то эти массивы вычисляются только для четных N .

Заметим, что функция $r(x)$ не может иметь полюс ($\neq 1/2$) в одном из корней полинома $P(M, x)$ и не иметь полюсов во всех остальных корнях.

С учетом этих условий справедлива

Теорема. Массивы $S_k, k = 1, 2, 3$, совпадают.

Доказательство. Напомним, что $x = Xe$ по определению, т.е. вектор $x \in \mathcal{P}_N$, соответствующий функции x , получается применением матрицы X к вектору e . Аналогично получается вектор $p = p(X)e$, соответствующий любому полиному $p(x)$. Но эти же рассуждения применимы для общей функции от матрицы, $f(X)$, т.е. вектор $f = f(X)e$ — это коэффициенты приближенного разложения функции $f(x)$ по смещенным полиномам Лежандра при условии, что матрица $f(X)$ определена.

Функция от матрицы $f(X)$ однозначно определяется значениями функции $f(x)$ на спектре матрицы X (см. [10]), т.е. значениями этой функции в корнях полинома $P(N, x)$ (так как матрица X диагонализируема, т.е. производные $f(x)$ не нужны).

Для рациональной функции $r(x)$ не требуется вычислять ее значения на спектре матрицы X , а можно формально подставить матрицу X вместо переменной x (см. [10]). Поэтому совпадение массивов S_1 и S_2 следует из совпадения двух интегралов.

Величина $S_1[N]$ — это первая компонента вектора правой части (9), т.е. нулевой коэффициент ее аппроксимации полиномами $P(n, x)$. Иными словами, это аппроксимация интеграла $r(x)$ на отрезке $[0, 1]$. Величина $S_2[N]$ получается из аппроксимации интеграла $\int_0^x r(t)dt$ как функции на отрезке $[0, 1]$ подстановкой $x = 1$ в эту функцию, т.е. $S_1 = S_2$.

Вторая часть утверждения теоремы, т.е. $S_1 = S_3$, следует из представления матричной функции $r(X)$ интерполяционным полиномом Лагранжа—Сильвестра (см. [10]). Имеем

$$r(X) = \sum_{n=1}^N r(x_n) l_{N,n}(X), \quad l_{N,n}(x) = \prod_{k \neq n} \frac{x - x_k}{x_n - x_k},$$

где $\{x_n, n = 1, \dots, N\}$ — это спектр матрицы X (корни полинома $P(N, x)$), а $l_{N,n}(x)$ — это фундаментальные полиномы лагранжевой интерполяции.

Поскольку $l_{N,n}(x) \in \mathcal{P}_N$, то

$$l_{N,n}(X).e[1] = \int_0^1 l_{N,n}(x) dx = w_n,$$

где w_n — это веса (10) квадратурной формулы Лежандра—Гаусса (см. [9]). Таким образом, вектор правой части (9) представим в виде

$$r(X).e = \sum_{n=1}^N r(x_n) l_{N,n}(X).e,$$

поэтому $S_1 = r(X).e[1]$ — это квадратура Лежандра—Гаусса функции $r(x)$. Что и требовалось доказать.

Следствие. Любая квадратура Лежандра–Гаусса на интервале $[0, 1]$ от любой рациональной функции $r(x)$ (с учетом сделанных оговорок) является рациональным числом, т.е.

$$\sum_{n=1}^N w_n r(x_n) \in \mathbb{Q}. \quad (11)$$

Этот результат может быть (относительно легко) проверен в CAS до размерности $N = 5$ и, в принципе, до размерности $N = 9$, хотя выражение для корня полинома $\tilde{P}(8, x)$ занимает страницу. Для $N > 9$ этот результат может быть проверен только численно, но зато с любой, в принципе, точностью.

Возьмем, например, $r(x) = 1/x$, для которой решение задачи Коши (9) $z(x)$ не существует. Но это не мешает нам вычислить квадратуры Гаусса, так как все корни x_n лежат внутри интервала. Получим в точности числа $2H(N)$, $N = 1, 2, \dots$, где $H(N)$ – это гармоническое число.

Возьмем теперь $r(x) = q/(1 + qx)$. Тогда $z(x) = \ln(1 + qx)$ является формальным решением задачи Коши (9). По нашему алгоритму получим для $N = 1, 2, \dots$,

$$S_1[N] = \left\{ \frac{2q}{2+q}, \frac{3q(2+q)}{(6+6q+q^2)}, \frac{q(60+60q+11q^2)}{3(2+q)(10+10q+q^2)}, \frac{5q(2+q)(42+42q+5q^2)}{6(70+140q+90q^2+20q^3+q^4)}, \dots \right\}. \quad (12)$$

Квадратуры Гаусса (11) получаются весьма громоздкими и в плавающей арифметике, но при подстановке численных значений параметра q совпадают с рациональными функциями (12).

Формула (12) также показывает, почему полюс $x = 1/2$ при $q = -2$ неприемлем. Он соответствует корню $x = 0$ полиномов Лежандра нечетной степени. Других рациональных корней у полиномов Лежандра не существует.

Предложение 3. Последовательность рациональных приближений (12), т.е. $S_1[N]$, $N = 1, 2, \dots$, совпадает с диагональными паде-аппроксимациями $P[N, N](z(x))|_{x=1} = P[N, N](\ln(1 + q))$.

Доказательство. Докажем другое утверждение, которое равносильно этому согласно доказанной теореме. Последовательность квадратур Гаусса функции $r(x) = q/(1 + qx)$, т.е. $S_3[N]$, $N = 1, 2, \dots$, совпадает с этими паде-аппроксимациями. В самом деле,

$$G_N\left(\frac{q}{1+qx}\right) = \sum_{k=0}^{\infty} G_N((-1)^k q^{k+1} x^k)$$

в силу линейности квадратуры. Но квадратуры Гаусса точны на полиномах степени $M \leq 2N - 1$, поэтому

$$G_N\left(\frac{q}{1+qx}\right) = \sum_{k=0}^{2N-1} \frac{(-1)^k q^{k+1}}{k+1} + O(q^{2N+1}) = \ln(1 + q) + O(q^{2N+1}).$$

Слева здесь стоит рациональная функция от q нужного вида, т.е. отношение полиномов степени N . Что и требовалось доказать.

Например, подставив $q = 1$ в (12), получаем последовательность рациональных приближений для $\ln 2$:

$$\left\{ \frac{2}{3}, \frac{9}{13}, \frac{131}{189}, \frac{445}{642}, \frac{34997}{50490}, \frac{62307}{62307}, \frac{2359979}{3404730}, \frac{25786503}{37202060}, \dots \right\}, \quad (13)$$

которая совпадает с последовательностью, полученной в [11] с помощью ϵ -алгоритма Винна (Wynn). Пользуясь случаем, исправляем там опечатку: $(8/13 \rightarrow 9/13)$.

Эта последовательность соответствовала в [11] поддиагональной паде-аппроксимации, так как там рассматривалась функция $\ln(1 + x)/x$. Отметим также, что в [11] было выбрано неудачное преобразование эквивалентности, в результате чего полученная цепная дробь выглядела сложнее, чем она есть (см. разд. 4).

Совпадение паде-аппроксимаций функции, представленной рядом Стильеса, т.е. где коэффициенты разложения – это моменты интеграла Стильеса, с квадратурами Гаусса не является новым результатом (см. [12]) (авторы там перепутали $[n - 1, n]$ - и $[n, n - 1]$ -аппроксимации).

Таким образом, существенно нелинейная операция — вычисление диагональной паде-аппроксимации $\ln(1+x)$ с помощью ε -алгоритма Винна сводится нашим методом к линейным операциям с матрицами.

Также теперь ясно, почему не удалось получить точную паде-аппроксимацию для арктангенса. Дело в том, что наш алгоритм линеен, поэтому мы получили сумму двух различных паде-аппроксимаций. А именно,

$$\frac{1}{1+x^2} = \frac{1}{2(1+ix)} + \frac{1}{2(1-ix)}, \quad \int_0^x \frac{dt}{2(1+it)} = \frac{1}{2} \arctan x - \frac{i}{4} \ln(1+x^2) = v(x),$$

т.е. последовательность $\{S_i[N]\}$ соответствует сумме двух паде-аппроксимаций,

$$S_i[N] = P[N, N](v(x))\big|_{x=1} + \overline{P[N, N](v(x))\big|_{x=1}}, \quad N = 1, 2, \dots$$

Хотя любая рациональная функция представима в виде суммы простых дробей, следствие, которое мы вывели из теоремы, не следует из совпадения квадратур Гаусса с паде-аппроксимациями простых дробей по очевидным причинам.

Как показано в [11], из последовательности (13) следует иррациональность числа $\ln 2$ по критерию Дирихле, т.е.

$$|p - q \ln 2| \rightarrow 0,$$

где p/q — это рациональные приближения в (13). Этим свойством обладают некоторые (но далеко не все) рациональные последовательности, полученные с помощью паде-аппроксимаций, и мы займемся этим в следующем разделе.

Таким образом, наш метод дает весьма точные приближения коэффициентов разложения интегралов некоторых рациональных функций по полиномам Лежандра (см. фиг. 1).

Оказалось, что если эти коэффициенты исходно рациональны, то наш метод в ряде случаев вычисляет их точно в рациональной арифметике. Рассмотрим уравнение

$$xy'(x) = 1,$$

решением которого является функция $y(x) = \ln x$ при выборе краевого условия $y(1) = 0$. Это также первый пример, где мы рассматриваем более сложный оператор, чем просто дифференцирование. Однако способ решения этого уравнения точно такой же.

Берем матрицу $A = X \cdot D$ и заменяем последнюю строку на вектор b_0 . Последний элемент вектора $r = e$ здесь ноль. Вычисляем вектор $y = A^{-1} \cdot r$, который дает приближенное разложение $\ln x$. Однако в данном случае все коэффициенты разложения, кроме последнего, $y[N]$, оказываются точными. Таким образом, получаем

$$\ln x = -1 + \sum_{n=1}^{\infty} (-1)^{n-1} \frac{2n+1}{n(n+1)} P(n, x), \quad x \in (0, 1]. \quad (14)$$

Если бы мы решали уравнение $y'(x) = 1/x$ этим способом, что возможно, так как матрица X обратима, то мы получили бы почти такой же результат, т.е. все коэффициенты находятся правильно, кроме (по неизвестным пока причинам) предпоследнего.

Приведем еще один пример. Рассмотрим уравнение

$$xy'(x) - y(x) = x,$$

общим решением которого является функция $y(x) = x(C + \ln x)$. Поэтому возьмем краевое условие $y(1) = 0$ и решим эту задачу как предыдущую. Тогда при каждом N мы получим все коэффициенты разложения $y(x) = x \ln x$, кроме последнего, $y[N]$. Таким образом, получаем

$$x \ln x = -\frac{1}{4} + \frac{1}{12} P(1, x) + \sum_{n=2}^{\infty} (-1)^n \frac{2n+1}{n(n^2-1)(n+2)} P(n, x), \quad x \in [0, 1].$$

Полученные разложения могут быть использованы для решения неалгебраических ОДУ нашим методом (см. разд. 5). Хотя равномерная аппроксимация решений окажется при этом соответствующей ожидаемой, т.е. (вообще говоря) не может быть лучше, чем аппроксимация $\ln x$ по-

линомами Лежандра, но аппроксимация решений в концах интервала может быть очень хорошей, что мы уже наблюдали (см. фиг. 1).

4. ЦЕПНЫЕ ДРОБИ И ИРРАЦИОНАЛЬНОСТЬ

Как известно, быстро сходящиеся последовательности рациональных чисел играют ключевую роль в доказательствах иррациональности их предела (см. [11] и ссылки там).

При этом, как оказалось во всех без исключения случаях (а их можно пересчитать по пальцам одной руки), если последовательность $\{p_n/q_n\}$ дает иррациональность своего предела x (например, константы Апери) по критерию Дирихле, то она сходится к x быстрее (в обычном, численном смысле), чем обыкновенная цепная дробь числа x . А значит (см. [11]), эта последняя может быть однозначно восстановлена по последовательности $\{p_n/q_n\}$ с помощью функций $K(x, n)$, которые мы определили в [11].

Иными словами, в этом случае последовательность $K(p_n/q_n, n)$ совпадает с последовательностью подходящих дробей числа x , т.е. $K(p_n/q_n, n) = K(x, n)$, $n > \text{const}$.

Но это последнее свойство может также выполняться, если критерий Дирихле для последовательности $\{p_n/q_n\}$ не выполняется (см. [11]). Например, для последовательности $\{S_1[N] = r_1\}$ рациональных приближений числа $\pi/4$, которую мы получили в задаче (4), $K(S_1[N], N) = K(\pi/4, N)$, $N = 1, 2, \dots$, т.е. мы получаем все подходящие дроби числа $\pi/4$. При этом, например,

$$\frac{\pi}{4} - S_1[30] \approx 1.36 \times 10^{-40}, \quad \frac{\pi}{4} - K(\pi/4, 30) \approx 9.51 \times 10^{-31}.$$

Однако наибольший интерес все же представляют последовательности, в структуре которых удается выявить некоторые закономерности. Этим мы и займемся в этом разделе.

Последовательности диагональных, а также над- или поддиагональных паде-аппроксимаций и их комбинации (qd -алгоритм) некоторых функций могут давать очень хорошие рациональные приближения иррациональных чисел. Например, последовательность (13) дает иррациональность числа $\ln 2$ по критерию Дирихле (см. [11]).

Но у нас также имеется более общая последовательность (12). Как оказалось, эта последняя значительно более удобна для выявления закономерностей в ее структуре, чем ее частный вариант при $q = 1$. Это связано с тем, что параметр q не дает CAS приводить дроби к наименьшему знаменателю, что в данном случае контрпродуктивно.

Закономерность в последовательности (12) становится видна, если преобразовать ее в цепную дробь Эйлера (по формулам Д. Бернулли). Обозначим (12) (заменяв там $q \rightarrow t$) как последовательность $s(n) = p_n/q_n$ подходящих дробей, где p_n и q_n — это, соответственно, числитель и знаменатель дробей (12). Вычислим

$$\begin{aligned} a_1 &= p_1, & a_n &= \frac{p_{n-1}q_n - p_nq_{n-1}}{p_{n-1}q_{n-2} - p_{n-2}q_{n-1}}, & n > 1, \\ b_1 &= q_1, & b_n &= \frac{p_nq_{n-2} - p_{n-2}q_n}{p_{n-1}q_{n-2} - p_{n-2}q_{n-1}}, & n > 1, \end{aligned} \quad (15)$$

где $p_0 = 0$ и $q_0 = 1$. Тогда

$$s(n) = \mathbf{K}_{k=1}^n \frac{a_k}{b_k},$$

или

$$\ln(1+t) = \frac{2t}{2+t - \frac{t^2}{6+3t - \frac{4t^2}{10+5t - \frac{9t^2}{14+7t - \frac{16t^2}{18+9t - \dots}}}}}, \quad (16)$$

где закономерность в частных дробях очевидна.

Хотя нам не удалось получить Паде-аппроксимации арктангенса с помощью нашего алгоритма напрямую, но применение формулы

$$\arctan x = \frac{i}{2} \ln \left(\frac{1-ix}{1+ix} \right)$$

в (16) дает такое разложение. После некоторых преобразований получаем известное разложение

$$\arctan x = \frac{x}{1 + \frac{x^2}{3 + \frac{4x^2}{5 + \frac{9x^2}{7 + \dots}}}}.$$

Применим наш подход к некоторым ОДУ, зависящим от параметра. Рассмотрим задачу Коши

$$y'(x) - ty(x) = 0, \quad y(0) = 1, \quad y(x) = \exp(tx),$$

где t — это независимый параметр. Это также пример, как формально однородное уравнение решается как неоднородное за счет начального условия, т.е. вектор в правой части $r = 0$, но затем $r[N] = 1$.

Применив наш алгоритм и преобразование (15), получим

$$e^t = \frac{t}{1 - \frac{2t}{2 + t + \frac{t^2}{6 + \frac{t^2}{10 + \frac{t^2}{14 + \frac{t^2}{18 + \frac{t^2}{22 + \dots}}}}}}}, \quad (17)$$

т.е. после начального куска цепная дробь становится “квазипериодической”.

Однако в данном случае мы получили феномен, который ранее (насколько нам известно) не наблюдался. То есть мы получили разложение, которое (при $t = 1$) содержит обыкновенную цепную дробь некоторой константы. Более того, дробь (17) (при $t = 1$) сходится на несколько порядков быстрее, чем обыкновенная цепная дробь константы e .

Данный феномен объясняется довольно просто. Избавившись от начального куска дроби (17) и построив ее сверху согласно очевидным закономерностям, получим известное разложение

$$\frac{e^t - 1}{e^t + 1} = \frac{t}{2 + \frac{t^2}{6 + \frac{t^2}{10 + \frac{t^2}{14 + \frac{t^2}{18 + \frac{t^2}{22 + \dots}}}}}. \quad (18)$$

Таким образом, преобразование Мёбиуса с рациональными коэффициентами иррационального числа может радикальным образом изменить скорость сходимости обыкновенной цепной дроби новой константы по сравнению с исходной. Например,

$$e - K(e, 10) \approx 1.102 \times 10^{-7}, \quad \frac{e-1}{e+1} - K\left(\frac{e-1}{e+1}, 10\right) \approx 4.216 \times 10^{-26}.$$

Это свойство могло бы найти применение в доказательствах иррациональности, но (насколько нам известно) никогда не применялось.

Применим теперь формулу

$$i \tan x = \frac{\exp(2ix) - 1}{\exp(2ix) + 1}$$

в (18). Тогда (после преобразования эквивалентности) получим классическое разложение для тангенса:

$$\tan x = \frac{x}{1 - \frac{x^2}{3 - \frac{x^2}{5 - \frac{x^2}{7 - \frac{x^2}{9 - \dots}}}}}. \quad (19)$$

Это разложение Ламберт и Лежандр использовали в доказательстве иррациональности числа π (см. [13], р. 289–296). А именно, согласно (очень простому) критерию иррациональности Ламберта–Лежандра формула (19) может давать только иррациональные числа для рациональных x . Но $\tan(\pi/4) = 1$, поэтому π не может быть рациональным.

По тем же причинам из дроби (18) следует иррациональность $\exp(r)$ для $0 \neq r \in \mathbb{Q}$. Поэтому и $\ln(r)$ иррационально для $1 \neq r \in \mathbb{Q}$.

Заметим, что основную часть доказательства Лежандра в [13] занимал как раз вывод цепной дроби (19). Лежандр использовал для этого функции Бесселя, причем, по-видимому, до самого Бесселя. А мог бы, как теперь ясно, применить свои полиномы.

Приведем еще один пример. Рассмотрим задачу Коши

$$(1 + tx)y'(x) - sty(x) = st, \quad y(0) = 0, \quad y(x) = (1 + tx)^s - 1,$$

где s и t — это независимые параметры. Применив наш алгоритм и преобразование (15), получим

$$(1 + t)^s = 1 + \frac{st}{1 + \frac{(1-s)t}{2 - \frac{2(s+1)t}{st - 2t - 6 - \frac{(s^2-4)t^2}{10 + 5t + \frac{(s^2-9)t^2}{14 + 7t + \frac{(s^2-16)t^2}{18 + 9t \dots}}}}}}, \quad (20)$$

т.е., как и раньше, после начального куска цепная дробь становится “квазипериодической”. Избавившись от начального куска дроби (20) и построив ее сверху согласно очевидным закономерностям, получим разложение

$$\frac{(1+t)^s - 1}{(1+t)^s + 1} = \frac{st}{2 + t + \frac{(s^2-1)t^2}{6 + 3t + \frac{(s^2-4)t^2}{10 + 5t + \frac{(s^2-9)t^2}{14 + 7t + \frac{(s^2-16)t^2}{18 + 9t + \dots}}}}}, \quad (21)$$

которое весьма напоминает (18).

Применим теперь тождество $i^{-2i} = e^\pi$, т.е. сделаем подстановку $s = -2i$, $t = i - 1$ в (21). Тогда получим (после преобразования эквивалентности)

$$\frac{e^\pi - 1}{e^\pi + 1} = \frac{1}{2} \prod_{n=1}^{\infty} \frac{n^2 - 2n + 5}{2n - 1} = \frac{2}{1 + \frac{5}{3 + \frac{8}{5 + \frac{13}{7 + \frac{20}{9 + \dots}}}}}. \quad (22)$$

Цепная дробь для e^π была получена недавно в [14]. Причем, как полагает автор в [14], это было сделано впервые. Однако наша дробь устроена проще.

Заметим также, что цепная дробь (22) сходится к своему пределу медленнее, чем его подходящие дроби, хотя число $e^\pi = (-1)^{-i}$ трансцендентно (как алгебраическое число в степени мнимая квадратичная иррациональность (см. [15] и ссылки там)).

5. КОНСТАНТЫ ЭЙЛЕРА И ЭЙЛЕРА–ГОМПЕРТЦА

Как было отмечено в [16], константа Эйлера γ и константа Эйлера–Гомпертца δ имеют весьма похожие интегральные представления. В этом разделе мы продолжим эту аналогию.

Рассмотрим функцию

$$y(x) = \int_0^{\infty} \ln(x+t) \exp(-t) dt = \ln(x) + \exp(x) \operatorname{Ei}(1, x),$$

где $\operatorname{Ei}()$ — это интегральная экспонента. Функция $y(x)$ удовлетворяет ОДУ:

$$y'(x) = y(x) - \ln(x), \quad (23)$$

общим решением которого является функция

$$y(x) = \ln(x) + \exp(x) \operatorname{Ei}(1, x) + C \exp(x).$$

Легко проверить, что $y(0) = -\gamma + C$ и $y(1) = y'(1) = \delta + C \exp(1)$. Поэтому, к сожалению, из уравнения (23) нельзя получить сразу обе константы, γ и δ , а только их линейную комбинацию с константой e . Иными словами, задача Коши для уравнения (23), $y(0) = 0$, дает рациональные приближения для константы

$$\eta = \delta + e\gamma = \sum_{n=1}^{\infty} \frac{H(n)}{n!} \approx 2.1653822153269. \quad (24)$$

Эта формула следует из представления (см. [17])

$$\delta = e(F([1,1],[2,2],-1) - \gamma),$$

где $F()$ — это гипергеометрическая функция, т.е., как легко проверить,

$$\exp(x) F([1,1],[2,2],-x) = \sum_{n=1}^{\infty} \frac{H(n)x^{n-1}}{n!}.$$

Задача Коши для уравнения (23) решается точно так же, как мы уже делали неоднократно. Отличие состоит в том, что эта задача сингулярна ($y'(0) = +\infty$), а также что ОДУ не является алгебраическим.

Таким образом, решается линейная алгебраическая задача $Ay = r$, где $A = D - E$ с последующей корректировкой последней строки, а вектор r получается с помощью формулы (14), и $r[N] = 0$, согласно начальному условию.

Мы проделали эти вычисления в рациональной арифметике до размерности аппроксимации $N = 100$. Полученная в результате последовательность рациональных приближений $\{s(n), n = 1, \dots, 100\}$ константы η не дает иррациональности этой константы по критерию Дирихле, но приближает ее лучше в численном смысле, чем последовательность подходящих дробей

константы η , т.е., согласно [11], $K(\eta, n) = K(s(n), n)$, $n > 11$. Иными словами, получена последовательность подходящих дробей константы η начиная с 12-й. Например,

$$\eta - s(50) \approx 4.3461 \times 10^{-95}, \quad \eta - K(\eta, 50) \approx 1.8460 \times 10^{-59}. \quad (25)$$

Здесь мы воспользовались модификацией алгоритма, т.е. замедлили скорость сходимости полученной последовательности рациональных приближений с помощью функций $K(x, n)$, введенных в [11].

Данный модифицированный алгоритм (т.е. вычисление n -й подходящей дроби от n -го рационального приближения) обладает сертификатом качества (см. [11]). Он гарантированно дает подходящие дроби нужной константы либо не дает, о чем сигнализирует. При этом не надо знать численного значения данной константы, а все вычисления проводятся в рациональной арифметике.

В случае если алгоритм “дает сбой”, т.е. очередное рациональное приближение оказывается хуже, чем приближение подходящей дробью этой константы (что может быть, например, если в цепной дроби встретился нетипично большой знаменатель), то это никак не влияет на последующие подходящие дроби, если данный член просто опустить. Такой случай встречался в разложении $\zeta(6)$ (см. [11]). Полученная по формулам (15) цепная дробь просто окажется некоторым сжатием исходной.

Таким образом, свойство последовательности рациональных чисел $\{s(n)\}$ сходиться к своему пределу x быстрее (в численном смысле), чем последовательность подходящих дробей числа x , $\{K(x, n)\}$, является внутренним свойством данной последовательности и проверяется в рациональной арифметике независимо от знания ее предела x .

Воспользуемся этими соображениями для проверки скорости сходимости ряда (24). Этот ряд весьма напоминает ряд Энгеля (см. [18]), а такие ряды обычно сходятся к своему пределу быстрее, чем подходящие дроби этого предела (см. [11]).

Гармоническое число в числителе (24) не должно сильно повлиять на скорость сходимости, так как оно растет как $\ln n$, что пренебрежимо мало по сравнению с факториалом.

Как оказалось, последовательность частичных сумм ряда (24), $\{h(m)\}$, и в самом деле стремится к η быстрее, чем подходящие дроби этой константы, т.е. $K(\eta, m) = K(h(m), m)$, $m > 40$. Например,

$$\eta - h(100) \approx 5.5684 \times 10^{-160}, \quad \eta - K(\eta, 100) \approx 9.0078 \times 10^{-112}.$$

В завершение этого примера решим задачу Коши $y(0) = -\gamma$ для уравнения (23), но на сей раз в символьном виде. То есть γ и δ теперь просто символы. Найдем последовательность приближений для $y(1) = \delta$, $N = 1, 2, \dots$. Эта последовательность имеет вид $\tilde{\eta}(N) - \tilde{\epsilon}(N)\gamma$, где тильда означает, что это рациональные приближения соответствующих констант. Они оказались в точности равны полученным ранее, т.е. в формулах (25) и (17).

Таким образом, если иметь последовательность хороших рациональных приближений для γ (или δ), то по данному алгоритму можно получить (почти) столь же хорошую последовательность рациональных приближений для δ (или γ).

6. СУММИРОВАНИЕ ГОЛОНОМНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Как мы видели, наш метод дает эффективные рациональные приближения коэффициентов разложения Фурье–Лежандра для функций, которые являются решениями линейных однородных ОДУ с полиномиальными коэффициентами. Такие функции называются *голономными* (см., например, [19]). Те неоднородные ОДУ, которые мы рассматривали, очевидно приводятся к однородным.

Оригинальный τ -метод Ланцоша также предназначался для этого класса уравнений (см. [4]).

Последовательность чисел $\{a_n, n = 0, 1, \dots\}$ называется *голономной*, если она удовлетворяет линейному разностному уравнению с полиномиальными (от n) коэффициентами.

Если в качестве функции понимать формальный степенной ряд, то каждой голономной функции соответствует голономная последовательность чисел — коэффициентов ее формального разложения, а также наоборот: каждой голономной последовательности соответствует голо-

номная функция — обыкновенная производящая функция этой последовательности (см. [19, р. 140]).

Хотя многие математики, продолжая традиции 19-го столетия, полагают, что “не каждый формальный ряд соответствует аналитической функции” (см. [19, р. 29]), однако это не так. Каждый формальный ряд соответствует аналитической функции, правда, возможно, не одной (теорема Бореля—Ритта, см. [20, р. 43]). Все зависит от интерпретации слова “соответствует”.

Возьмем ряд из [19, р. 29] в качестве примера. Мы использовали этот ряд в [21] как пример “голономного” суммирования последовательности. Рассмотрим формальный ряд

$$y(x) = \sum_{n=0}^{\infty} n! x^n \quad (26)$$

и найдем для него аналитическую функцию $y(x)$, такую, что $y(1) = 1$.

Формальный голономный ряд (26) удовлетворяет дифференциальному уравнению

$$x^2 y' - (1 - x)y + 1 = 0, \quad (27)$$

которое неголономно, но приводится к нужному виду. Функция

$$y(x) = \frac{1}{x} \exp\left(-\frac{1}{x}\right) \left(C - \operatorname{Ei}\left(1, -\frac{1}{x}\right)\right),$$

где C — это константа интегрирования, является общим аналитическим решением этого уравнения.

Для любой константы C эта аналитическая функция “соответствует” своему формальному разложению (26), которое является асимптотическим в полуплоскости $\operatorname{Re}(x) > 0$, с трансфинитной экспоненциально малой добавкой.

Это означает, что $y(1)$, например, может принимать любое значение. Если взять

$$C = \operatorname{Ei}(1, -1) + \exp(1) \approx 0.8231640121032 - i\pi,$$

то получим $y(1) = 1$.

Чтобы получить разложение этой функции по полиномам Лежандра нашим методом, нужно взять матрицу $X2.D - E + X$ и заменить ее последнюю строку на вектор b_0 (см. разд. 2). Вектор правой части — это $r = -e$ с подстановкой $r[N] = 1$, согласно выбранному краевому значению. Заметим, что $X2.D = X.X.D$. Для $N = 10$ получим

$$y(0) = 2693025253/2682336916, \quad y(0) - 1 \approx 3.98471 \times 10^{-3},$$

что весьма близко к верхней равномерной оценке погрешности для этой функции на интервале $[0, 1]$.

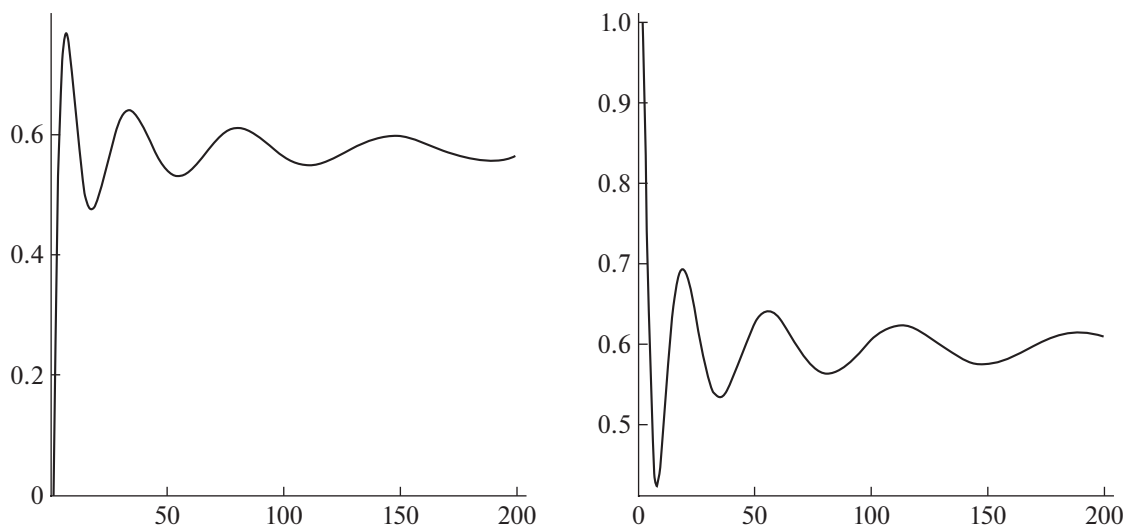
Отметим, что неединственность функции, которая соответствует своему формальному степенному разложению (а такие разложения всегда асимптотические), не имеет никакого отношения к сходимости или расходимости этого разложения. Это же относится к наличию либо отсутствию экспоненциально малой (трансфинитной) добавки. Предыдущий пример переносится без существенных изменений на уравнение

$$x^2 y' - y = -\frac{1+x+x^2}{(1+x)^2}, \quad y(x) = \frac{1}{1+x} + C \exp\left(-\frac{1}{x}\right).$$

Таким образом, мы имеем способ суммирования голономных последовательностей и рядов, причем необязательно сходящихся в обычном смысле, который сводится к решению соответствующих голономных ОДУ и подстановке $x = 1$ в полученное решение.

Здесь наш метод приближенного решения голономных ОДУ может быть весьма полезен, так как эти уравнения, вообще говоря, неинтегрируемы. Но даже в случае их интегрируемости в квадратурах (с применением CAS) эти квадратуры могут быть очень сложно устроены и (почти) совершенно бесполезны.

Вычислим для примера константу Эйлера—Гомпертца путем суммирования расходящегося ряда (26) при $x = -1$. Это классическая задача, которая решалась многими способами. Ланцош ее также использовал для демонстрации эффективности τ -метода (см. [4]).



Фиг. 2. Сходимость частичных сумм $u(n)$ (слева) и $v(n)$ (справа).

$u(n)$ и $v(n)$ соответственно. Пользуясь известным рекуррентным соотношением для полиномов Лагерра, получаем $u(1) = 0$, $u(2) = 1/4$, $u(3) = 17/36$ и

$$n^2 u(n) = (3n^2 - 4n + 2)u(n-1) - (n-1)(3n-4)u(n-2) + (n-1)(n-2)u(n-3),$$

а также $v(1) = 1$, $v(2) = 1$, $v(3) = 5/6$ и

$$(n-1)nv(n) = (n-1)(3n-4)v(n-1) - (3n-4)(n-2)v(n-2) + (n-2)^2 v(n-3).$$

На фиг. 2 показаны графики частичных сумм $u(n)$ и $v(n)$ до $n = 200$. Из фиг. 2 видно, что эти суммы осциллируют с медленно уменьшающейся амплитудой и увеличивающимся периодом. Такие ряды не могут суммироваться обычными средствами.

В данном случае мы уже просуммировали ряды (29), пользуясь явным видом полученных решений. Однако, как было отмечено, интегрируемость ОДУ для голономных функций является, вообще говоря, редким явлением. Кроме того, явные формулы не дают рациональных приближений. Поэтому применим наш метод аппроксимации решений.

Мы уже неоднократно описывали алгоритм решения таких уравнений. Уравнения (31) решаются точно так же. В результате для размерностей аппроксимации $N = 10$ и $N = 20$ получим для γ :

$$\gamma - u(1) \approx 1.1167 \times 10^{-5}, \quad \gamma - u(1) \approx -1.2006 \times 10^{-7},$$

а также для δ :

$$\delta - v(1) \approx -2.8034 \times 10^{-8}, \quad \delta - v(1) \approx 2.9407 \times 10^{-13}.$$

Таким образом, хотя голономные последовательности для γ и δ выглядят почти идентично (см. (29)), сам вид уравнений (31), а также результаты аппроксимации их решений (32) свидетельствуют, что константа γ имеет более сложную природу в некотором “голономном” смысле. На это же указывает тот факт, что для рациональных приближений константы δ существуют линейные формы второго порядка, а для γ пока что только третьего (см. [16]).

В заключение заметим, что наши методы обобщаются также на другие системы ортогональных полиномов, например, на полиномы Чебышёва. Последние гораздо удобнее для численного решения голономных ОДУ, для итерационного решения нелинейных уравнений, а также для решения разностных уравнений, которые возникают при обобщенном суммировании рядов (см. [21]). Однако это является темой отдельной работы.

СПИСОК ЛИТЕРАТУРЫ

1. *Pashkovskii S.* Computational Application of Chebyshev Polynomials and Series. Moscow: Nauka, 1983.
2. *Бабенко К.И.* Основы численного анализа. Ижевск: РИХД, 2002.
3. *Gottlieb D., Orszag S.A.* Numerical analysis of spectral methods: theory and applications. CBMS Regional Conference Series in Applied Mathematics. 6th printing, 1996.
4. *Lanczos C.* Applied Analysis. New-York: Dover Publications, 1956.
5. *Варин В.П., Петров А.Г.* Гидродинамическая модель ушной улитки // Ж. вычисл. матем. и матем. физ. 2009. Т.49. № 9. С. 1708–1723.
6. *Wilf H.S.* Mathematics for the physical sciences. NewYork: Wiley, 1962.
7. *Johnson D.* Chebyshev Polynomials in the Spectral Tau Method and Applications to Eigenvalue Problems // NASA Contractor Report 198451. 1996.
8. *Ortiz E.L., Samara H.* An Operational Approach to the Tau Method for the Numerical Solution of Non-Linear Differential Equations // Computing. 1981. V. 27. P. 15–25.
9. *Krylov V.I.* Approximate calculation of integrals. New-York, London: Macmillan, 1962.
10. *Gantmacher F.R.* Application of the Theory of Matrices. New-York: Chelsea Press, 1960.
11. *Варин В.П.* Преобразование последовательностей в доказательствах иррациональности некоторых фундаментальных констант // Ж. вычисл. матем. и матем. физ. 2022. Т. 62. № 10. С. 1587–1614.
12. *Allen G.D. et al.* Padé approximation and Gaussian quadrature // Bull. Austral. Math. Soc. 1974. V. 11. P. 63–69.
13. *Legendre A.M.* Éléments de Géométrie. (2em. ed). Chez Fermin Didot Père et Fils. Paris, 1823.
14. *Rivoal T.* Polynomial continued fractions for $\exp(\pi)$ // hal-03269677v3. 2022. <https://hal.archives-ouvertes.fr/hal-03269677v3>.
15. *Кузьмин Р.О.* Об одном новом классе трансцендентных чисел // Известия Акад. наук СССР. VII сер. Отд. физ.-мат. наук. 1930. Вып. 6. С. 585–597.
16. *Aptekarev A.I.* On linear forms containing the Euler constant // [arXiv:0902.1768v2]. 2009. <http://arxiv.org/abs/0902.1768v2>.
17. *Варин В.П.* Факториальное преобразование некоторых классических комбинаторных последовательностей // Ж. вычисл. матем. и матем. физ. 2018. Т. 59. № 6. С. 1747–1770.
18. *Perron O.* Irrationalzahlen. Berlin, Leipzig: Göschen's Lehrbücherei, 1921.
19. *Kauers M., Paule P.* The Concrete Tetrahedron. Symbolic Sums, Recurrence Equations, Generating Functions, Asymptotic Estimates. Wien: Springer, 2011.
20. *Wasow W.* Asymptotic expansions for ordinary differential equations. New-York: Dover Publications, 1987.
21. *Варин В.П.* Функциональное суммирование рядов // Ж. вычисл. матем. и матем. физ. 2023. Т. 63. № 1. С. 3–17.
22. *Wimp J.* Sequence transformations and their applications. New-York, etc.: Academic Press, 1981.

ОБЩИЕ ЧИСЛЕННЫЕ МЕТОДЫ

УДК 519.63

УЛУЧШЕННАЯ РАЗНОСТНАЯ СХЕМА ДЛЯ ЗАДАЧИ КОШИ В СЛУЧАЕ УРАВНЕНИЯ ПЕРЕНОСА

© 2023 г. Г. И. Шишкин^{1,*}, Л. П. Шишкина¹¹ 620108 Екатеринбург, ул. С. Ковалевской, 16, ИММ УрО РАН, Россия

*e-mail: shishkin@imm.uran.ru

Поступила в редакцию 04.04.2023 г.
Переработанный вариант 04.04.2023 г.
Принята к публикации 28.04.2023 г.

Рассматривается задача Коши для регулярного уравнения переноса. Для этой задачи с использованием техники Ричардсона строится улучшенная разностная схема, сходящаяся в равномерной норме со вторым порядком скорости сходимости. Библ. 6.

Ключевые слова: уравнение переноса, задача Коши, стандартная разностная схема, равномерная сетка, невязка, разложение невязки, монотонность дифференциальной и сеточной задач, техника Ричардсона, улучшенная разностная схема, сходимость в равномерной норме.

DOI: 10.31857/S0044466923080136, EDN: WTFQVG

1. ВВЕДЕНИЕ

В настоящее время большое внимание уделяется разработке численных методов для уравнений математической физики повышенного порядка точности. Здесь сложились два направления. Первое направление связано с повышением порядка аппроксимации разностных уравнений. Подходы, связанные с этим направлением, рассматривались в исследованиях Г.И. Марчука [1], Г.И. Марчука и В.В. Шайдурова [2], А.А. Самарского [3], Г.И. Шишкина и Л.П. Шишкиной [4]. Второе направление связано с построением решений на основе разностных уравнений сравнительно невысокого порядка на последовательности равномерных вложенных сеток. Эти методы получили название экстраполяции по Ричардсону, см., например, [5]. В настоящей работе для задачи Коши для уравнения переноса строятся разложения обратных разностных производных и соответствующих невязок по степеням шагов основной и разреженной сеток соответственно. Это позволяет применить технику Ричардсона на двух вложенных сетках и построить разностную схему, сходящуюся в равномерной норме со вторым порядком скорости сходимости.

О содержании работы. Постановка задачи Коши для уравнения переноса и цель исследования приводятся в разд. 2. Априорные оценки решения и производных устанавливаются в разд. 3. Стандартная разностная схема для задачи Коши для уравнения переноса рассматривается в разд. 4 и устанавливается ее сходимость в равномерной норме с первым порядком скорости сходимости. В разд. 5 и 6 построены разложения обратных разностных производных по степеням шагов основной и разреженной сеток соответственно. С использованием этих разложений получены разложения для соответствующих невязок. В разд. 7 на основе разложений невязок сеточных решений с применением техники Ричардсона строится схема, сходящаяся в равномерной норме со вторым порядком скорости сходимости. Выводы приводятся в разд. 8.

2. ПОСТАНОВКА ЗАДАЧИ КОШИ ДЛЯ УРАВНЕНИЯ ПЕРЕНОСА. ЦЕЛЬ ИССЛЕДОВАНИЯ

На полосе $\bar{G}$:

$$\bar{G} = G \cup S, \quad (2.1a)$$

где

$$G = D \times (0, T], \quad D = (-\infty < x < \infty); \quad S = \{(x, t) : -\infty < x < \infty, t = 0\}, \quad (2.1b)$$

рассмотрим задачу Коши для регулярного уравнения переноса

$$Lu(x, t) = f(x, t), \quad (x, t) \in G; \quad u(x, t) = \varphi(x), \quad (x, t) \in S. \quad (2.2)$$

Здесь

$$L = b(x, t) \frac{\partial}{\partial x} + c(x, t) + p(x, t) \frac{\partial}{\partial t}, \quad (x, t) \in G, \quad (2.3a)$$

функции $b(x, t)$, $c(x, t)$, $p(x, t)$, $f(x, t)$ предполагаются достаточно гладкими на $\bar{G}$, функция $\varphi(x)$ предполагается достаточно гладкой на множестве S , причем,

$$b_0 \leq b(x, t) \leq b^0, \quad 0 \leq c(x, t) \leq c^0, \quad p_0 \leq p(x, t) \leq p^0, \quad |f(x, t)| \leq M, \quad (x, t) \in \bar{G}; \quad (2.3б)$$

$$|\varphi(x)| \leq M, \quad x \in D; \quad b_0, p_0 > 0. \quad (2.3в)$$

Через M (через m) обозначаем достаточно большие (малые) положительные постоянные. В случае сеточных задач эти постоянные не зависят от шаблонов разностных схем.

Наша цель — для задачи Коши для уравнения переноса (2.2), (2.1) построить разностную схему, сходящуюся в равномерной норме с порядком скорости сходимости выше первого.

3. АПРИОРНЫЕ ОЦЕНКИ РЕШЕНИЯ И ПРОИЗВОДНЫХ

Рассмотрим ряд априорных оценок решения задачи (2.2), (2.1), необходимых при построении разностных схем и обосновании их сходимости. В разд. 5 и 6 нам потребуется ограниченность производных $\partial^k / \partial x^k u(x, t)$, $\partial^{k_0} / \partial t^{k_0} u(x, t)$, $(x, t) \in \bar{G}$, при $k + k_0 \leq 3$. Вывод оценок подобен выводу оценок решения и производных регулярных и сингулярных компонент решения сингулярно возмущенного уравнения переноса в [6].

3.1. Для задачи Коши для уравнения переноса (2.2), (2.1) справедлив *принцип максимума*, подобный принципу максимума для сингулярно возмущенного уравнения переноса в [6].

Теорема 3.1. Пусть для данных задачи Коши для уравнения переноса (2.2), (2.1) выполняется условие

$$Lu(x, t) \geq 0, \quad (x, t) \in G; \quad u(x, t) \geq 0, \quad (x, t) \in S.$$

Тогда для функции $u(x, t)$ справедлива оценка $u(x, t) \geq 0$, $(x, t) \in \bar{G}$.

3.2. Приведем априорные оценки для решения задачи Коши (2.2), (2.1).

Применяя технику мажорантных функций (подобную приведенной в [6] для начально-краевой задачи для сингулярно возмущенного уравнения переноса), находим оценку решения задачи (2.2), (2.1)

$$|u(x, t)| \leq M, \quad (x, t) \in \bar{G}. \quad (3.1)$$

При исследовании производных решения задачи Коши считаем, что коэффициенты и правая часть уравнения являются достаточно гладкими на $\bar{G}$: $b, c, p, f \in C^{k, k_0}(\bar{G})$, $k + k_0 \leq K$, $K \leq 3$, начальная функция — достаточно гладкая на множестве S , производные $\frac{\partial^k}{\partial x^k} \varphi(x)$, $k \leq K$, ограничены при $K = 3$. При этих условиях выполняется включение

$$u \in C^{k, k_0}(\bar{G}), \quad k + k_0 \leq K. \quad (3.2)$$

Тогда для функции $u(x, t)$ справедлива оценка

$$\left| \frac{\partial^{k+k_0}}{\partial x^k \partial t^{k_0}} u(x, t) \right| \leq M, \quad (x, t) \in \bar{G}, \quad k + k_0 \leq K. \quad (3.3)$$

Справедлива следующая

Теорема 3.2. Пусть для данных задачи Коши для уравнения переноса (2.2), (2.1) выполняется условие $b, c, p, f \in C^{k, k_0}(\bar{G})$, $k + k_0 \leq K$, $K = 3$. Тогда для решения задачи Коши $u(x, t)$ и его производных справедливы оценки (3.1), (3.3).

4. СТАНДАРТНАЯ РАЗНОСТНАЯ СХЕМА

Построим *стандартную разностную схему* на основе монотонной сеточной аппроксимации задачи Коши для уравнения переноса (2.2), (2.1) и исследуем ее сходимость.

4.1. На множестве $\bar{G}$ введем сетку

$$\bar{G}_{h\tau} = \bar{\omega} \times \bar{\omega}_0. \quad (4.1)$$

Здесь $\bar{\omega}$ и $\bar{\omega}_0$ – равномерные сетки на множествах $(-\infty, \infty)$ и $[0, T]$ соответственно. Пусть h и τ – шаги сеток $\bar{\omega}$ и $\bar{\omega}_0$; узел $(0, 0)$ принадлежит сетке $\bar{G}_{h\tau}$.

Предполагаем выполненным условие $h \leq M N^{-1}$, $\tau \leq M N_0^{-1}$, где $N + 1$ и $N_0 + 1$ – число узлов на отрезке единичной длины на множестве $(-\infty, \infty)$ и число узлов сетки $\bar{\omega}_0$ соответственно.

Задачу (2.2), (2.1) аппроксимируем *стандартной разностной схемой* [3]

$$\Lambda z(x, t) = f(x, t), \quad (x, t) \in G_{h\tau}; \quad z(x, t) = \varphi(x), \quad (x, t) \in S_{h\tau}. \quad (4.2a)$$

Здесь $G_{h\tau} = G \cap \bar{G}_{h\tau}$, $S_{h\tau} = S \cap \bar{G}_{h\tau}$,

$$\Lambda \equiv b(x, t)\delta_{\bar{x}} + c(x, t) + p(x, t)\delta_{\bar{t}}, \quad (4.2b)$$

$\delta_{\bar{x}} z(x, t)$ и $\delta_{\bar{t}} z(x, t)$ – первые обратные разностные производные (производные назад) по x и t соответственно.

Разностная схема (4.2), (4.1) монотонна (определение монотонности разностных схем см., например, в [3]). Для схемы (4.2), (4.1) справедлив *сеточный принцип максимума*.

Теорема 4.1. Пусть для разностной схемы (4.2), (4.1) выполняются условия $\Lambda z(x, t) \geq 0$, $(x, t) \in G_{h\tau}$; $z(x, t) \geq 0$, $(x, t) \in S_{h\tau}$. Тогда для функции $z(x, t)$ справедлива оценка $z(x, t) \geq 0$, $(x, t) \in \bar{G}_{h\tau}$.

4.2. С учетом априорных оценок устанавливаем сходимость в равномерной норме разностной схемы (4.2), (4.1).

С использованием оценок (3.3), где $K = 2$, подобно [6], получаем, что решение разностной схемы (4.2), (4.1) сходится в равномерной норме к решению задачи Коши (2.2), (2.1) с оценкой

$$|u(x, t) - z(x, t)| \leq M[N^{-1} + N_0^{-1}], \quad (x, t) \in \bar{G}_{h\tau}. \quad (4.3)$$

Теорема 4.2. Пусть для решения задачи Коши (2.2), (2.1) выполняется оценка (3.3) при $K = 2$. Тогда решение разностной схемы (4.2), (4.1) сходится в равномерной норме и для него справедлива оценка (4.3).

5. РАЗЛОЖЕНИЯ РАЗНОСТНЫХ ПРОИЗВОДНЫХ И НЕВЯЗОК ПО СТЕПЕНЯМ ШАГОВ ПРОСТРАНСТВЕННОЙ И ВРЕМЕННОЙ СЕТОК

В случае разностной схемы (4.2), (4.1) построим разложения первых обратных разностных производных по x и t по степеням шагов h и τ равномерных сеток $\bar{\omega}$ и $\bar{\omega}_0$. С использованием этих разложений получим разложения по степеням шагов h и τ для соответствующих невязок, которые будем использовать в разд. 7 при построении схемы, сходящейся со вторым порядком скорости сходимости.

5.1. Рассмотрим разложение обратной разностной производной $\delta_{\bar{x}} u(x, t)$ по шагу h в узле (x, t) , где

$$\delta_{\bar{x}} u(x, t) = \frac{u(x, t) - u(x - h, t)}{h}, \quad (x, t) \in \bar{G}_{h\tau}.$$

Заметим, что узел $(0, 0)$ принадлежит сетке $\bar{G}_{h\tau}$.

Заметим, что для $u(x - h, t)$ выполняется следующее разложение по степеням шага h :

$$u(x - h, t) = u(x, t) - h \frac{\partial}{\partial x} u(x, t) + 2^{-1} h^2 \frac{\partial^2}{\partial x^2} u(x, t) - 6^{-1} h^3 \frac{\partial^3}{\partial x^3} u(x, t), \quad (5.1)$$

$$(x, t) \in G_{h\tau}, \quad \vartheta \in [x - h, x].$$

С учетом разложения (5.1) для $u(x-h, t)$ получаем разложение обратной разностной производной $\delta_{\bar{x}}u(x, t)$ по степеням шага h в узле (x, t)

$$\delta_{\bar{x}}u(x, t) = \frac{\partial}{\partial x}u(x, t) - 2^{-1}h \frac{\partial^2}{\partial x^2}u(x, t) + 6^{-1}h^2 \frac{\partial^3}{\partial x^3}u(\vartheta, t), \quad (5.2)$$

$$(x, t) \in G_{h\tau}, \quad \vartheta \in [x-h, x].$$

Теперь, с учетом разложения (5.2) получаем разложение по степеням шага h в узле (x, t) для разности $\delta_{\bar{x}}u(x, t) - \partial/\partial x u(x, t)$, которую назовем *невязкой* по аналогии с определением у Н.Н. Калиткина [5]

$$\delta_{\bar{x}}u(x, t) - \frac{\partial}{\partial x}u(x, t) = -2^{-1}h \frac{\partial^2}{\partial x^2}u(x, t) + 6^{-1}h^2 \frac{\partial^3}{\partial x^3}u(\vartheta, t), \quad (x, t) \in G_{h\tau}. \quad (5.3)$$

5.2. Рассмотрим разложение обратной разностной производной $\delta_{\bar{t}}u(x, t)$ по степеням шага τ в узле (x, t) , где

$$\delta_{\bar{t}}u(x, t) = \frac{u(x, t) - u(x, t-\tau)}{\tau}, \quad (x, t) \in G_{h\tau}.$$

Узел $(0, 0)$ принадлежит сетке $\bar{G}_{h\tau}$.

Подобно разложению (5.1), для $u(x, t-\tau)$ выполняется следующее разложение по степеням шага τ :

$$u(x, t-\tau) = u(x, t) - \tau \frac{\partial}{\partial t}u(x, t) + 2^{-1}\tau^2 \frac{\partial^2}{\partial t^2}u(x, t) - 6^{-1}\tau^3 \frac{\partial^3}{\partial t^3}u(x, \eta), \quad (5.4)$$

$$(x, t) \in G_{h\tau}, \quad \eta \in [t-\tau, t].$$

С учетом разложения (5.4) для $u(x, t-\tau)$ получаем разложение обратной разностной производной $\delta_{\bar{t}}u(x, t)$ по степеням шага τ

$$\delta_{\bar{t}}u(x, t) = \frac{\partial}{\partial t}u(x, t) - 2^{-1}\tau \frac{\partial^2}{\partial t^2}u(x, t) + 6^{-1}\tau^2 \frac{\partial^3}{\partial t^3}u(x, \eta), \quad (x, t) \in G_{h\tau}, \quad \eta \in [t-\tau, t]. \quad (5.5)$$

Аналогично (5.3), с учетом разложения (5.5), получаем разложение по степеням шага τ в узле (x, t) для невязки $\delta_{\bar{t}}u(x, t) - \partial/\partial t u(x, t)$:

$$\delta_{\bar{t}}u(x, t) - \frac{\partial}{\partial t}u(x, t) = -2^{-1}\tau \frac{\partial^2}{\partial t^2}u(x, t) + 6^{-1}\tau^2 \frac{\partial^3}{\partial t^3}u(x, \eta), \quad (x, t) \in G_{h\tau}. \quad (5.6)$$

5.3. Основные результаты о разложении обратных разностных производных по x и t и невязок по степеням сеточных шагов h и τ равномерных сеток $\bar{\omega}$ и $\bar{\omega}_0$.

Справедлива следующая

Теорема 5.1. Пусть для решения задачи Коши для уравнения переноса (2.2), (2.1) выполняется оценка (3.3) при $K = 3$. Тогда в случае разностной схемы (4.2), (4.1) для разностных производных $\delta_{\bar{x}}u(x, t)$ и $\delta_{\bar{t}}u(x, t)$, $(x, t) \in G_{h\tau}$ справедливы разложения (5.2) и (5.5) соответственно.

Также справедлива следующая

Теорема 5.2. Пусть выполняется условие теоремы 5.1. Тогда для невязок $\delta_{\bar{x}}u(x, t) - \frac{\partial}{\partial x}u(x, t)$ и $\delta_{\bar{t}}u(x, t) - \frac{\partial}{\partial t}u(x, t)$ справедливы разложения (5.3) и (5.6).

6. РАЗЛОЖЕНИЯ РАЗНОСТНЫХ ПРОИЗВОДНЫХ И НЕВЯЗОК ПО СТЕПЕНЯМ ШАГОВ РАЗРЕЖЕННЫХ СЕТОК

На основе сетки $\bar{G}_{h\tau(4.1)}$, введенной в (4.1), строим *разреженную сетку*, на которой для разностной схемы (4.2), аналогично построениям в разд. 5, строим разложения разностных производных по x и t по степеням шагов разреженных сеток и разложения для соответствующих невязок, ко-

торые будут использованы в разд. 7 при построении схемы улучшенного порядка точности, сходящейся со вторым порядком скорости сходимости.

6.1. На основе сетки $\bar{G}_{h\tau(4.1)}$ строим следующую *разреженную сетку*:

$$\bar{G}_{h^*\tau^*}^* = \bar{\omega}^* \times \bar{\omega}_0^*. \quad (6.1)$$

Здесь $\bar{\omega}^*$ и $\bar{\omega}_0^*$ – равномерные сетки на множествах $(-\infty, \infty)$ и $[0, T]$ соответственно. Величины шагов h^* и τ^* сеток $\bar{\omega}^*$ и $\bar{\omega}_0^*$ в два раза больше величины шагов h и τ сеток $\bar{\omega}$ и $\bar{\omega}_0$. Узел $(0, 0)$ принадлежит исходной сетке $\bar{G}_{h\tau}$.

На сетке $\bar{G}_{h^*\tau^*}^*$ рассмотрим *разностную схему*

$$\begin{aligned} \Lambda^* z^*(x, t) &\equiv \{b(x, t)\delta_{\bar{x}^*} + p(x, t)\delta_{\bar{t}^*} + c(x, t)\} z^*(x, t) = f(x, t), \quad (x, t) \in G_{h^*\tau^*}^*, \\ z^*(x, t) &= \varphi(x), \quad (x, t) \in S_{h^*\tau^*}^*. \end{aligned} \quad (6.2)$$

Здесь $G_{h^*\tau^*}^* = G \cap \bar{G}_{h^*\tau^*}^*$, $S_{h^*\tau^*}^* = S \cap \bar{G}_{h^*\tau^*}^*$; $\delta_{\bar{x}^*} z^*(x, t)$ и $\delta_{\bar{t}^*} z^*(x, t)$ – первые обратные разностные производные на разреженной сетке

$$\delta_{\bar{x}^*} z^*(x, t) = \frac{z^*(x, t) - z^*(x - h^*, t)}{h^*}, \quad (x, t) \in \bar{G}_{h^*\tau^*}^*.$$

$$\delta_{\bar{t}^*} z^*(x, t) = \frac{z^*(x, t) - z^*(x, t - \tau^*)}{\tau^*}, \quad (x, t) \in G_{h^*\tau^*}^*.$$

Для функции

$$w^*(x, t) = z^*(x, t) - u(x, t), \quad (x, t) \in \bar{G}_{h^*\tau^*}^*$$

выполняется следующее соотношение:

$$\begin{aligned} \Lambda^* w^*(x, t) &= \Lambda^* z^*(x, t) - \Lambda^* u(x, t) = (L - \Lambda^*)u(x, t) = b(x, t) \left[\frac{\partial}{\partial x} u(x, t) - \delta_{\bar{x}^*} u(x, t) \right] + \\ &+ p(x, t) \left[\frac{\partial}{\partial t} u(x, t) - \delta_{\bar{t}^*} u(x, t) \right], \quad (x, t) \in G_{h^*\tau^*}^*, \end{aligned} \quad (6.3)$$

причем

$$w^*(x, t) = 0, \quad (x, t) \in S_{h^*\tau^*}^*.$$

6.2. Рассмотрим разложение обратной разностной производной $\delta_{\bar{x}^*} u(x, t)$ по шагу h^* в узле (x, t) .

Для производной $\delta_{\bar{x}^*} u(x, t)$ имеем следующее представление:

$$\delta_{\bar{x}^*} u(x, t) = \frac{u(x, t) - u(x - h^*, t)}{h^*}, \quad (x, t) \in \bar{G}_{h^*\tau^*}^*.$$

Узел $(0, 0)$ принадлежит сетке $\bar{G}_{h^*\tau^*}^*$.

Заметим, что для $u(x - h^*, t)$ выполняется следующее разложение по степеням шага h^* :

$$\begin{aligned} u(x - h^*, t) &= u(x, t) - h^* \frac{\partial}{\partial x} u(x, t) + 2^{-1} h^{*2} \frac{\partial^2}{\partial x^2} u(x, t) - 6^{-1} h^{*3} \frac{\partial^3}{\partial x^3} u(x, t), \\ (x, t) &\in G_{h^*\tau^*}^*, \quad \vartheta \in [x - h^*, x]. \end{aligned} \quad (6.4)$$

С учетом разложения (6.4) для $u(x - h^*, t)$ получаем разложение обратной разностной производной $\delta_{\bar{x}^*}u(x, t)$ по степеням шага h^*

$$\delta_{\bar{x}^*}u(x, t) = \frac{\partial}{\partial x}u(x, t) - 2^{-1}h^* \frac{\partial^2}{\partial x^2}u(x, t) + 6^{-1}h^{*2} \frac{\partial^3}{\partial x^3}u(\vartheta, t), \quad (6.5)$$

$$(x, t) \in G_{h^*\tau^*}^*, \quad \vartheta \in [x - h^*, x].$$

6.3. Подобно (6.4), для $u(x, t - \tau^*)$ выполняется следующее разложение по степеням шага τ^* :

$$u(x, t - \tau^*) = u(x, t) - \tau^* \frac{\partial}{\partial t}u(x, t) + 2^{-1}\tau^{*2} \frac{\partial^2}{\partial t^2}u(x, t) - 6^{-1}\tau^{*3} \frac{\partial^3}{\partial t^3}u(x, \eta), \quad (6.6)$$

$$(x, t) \in G_{h^*\tau^*}^*, \quad \eta \in [t - \tau^*, t].$$

С учетом разложения (6.6) для $u(x, t - \tau^*)$, получаем разложение обратной разностной производной $\delta_{\bar{t}^*}u(x, t)$ по степеням шага τ^*

$$\delta_{\bar{t}^*}u(x, t) = \frac{\partial}{\partial t}u(x, t) - 2^{-1}\tau^* \frac{\partial^2}{\partial t^2}u(x, t) + 6^{-1}\tau^{*2} \frac{\partial^3}{\partial t^3}u(x, \eta), \quad (6.7)$$

$$(x, t) \in G_{h^*\tau^*}^*, \quad \eta \in [t - \tau^*, t].$$

6.4. Основным результатом о разложении разностных производных по x и t по степеням сеточных шагов разреженной сетки $G_{h^*\tau^*}^*$.

В силу разложений (6.5) и (6.7), для невязок $\delta_{\bar{x}^*}u(x, t) - \frac{\partial}{\partial x}u(x, t)$ и $\delta_{\bar{t}^*}u(x, t) - \frac{\partial}{\partial t}u(x, t)$ справедливы следующие разложения по степеням сеточных шагов h^* и τ^* :

$$\delta_{\bar{x}^*}u(x, t) - \frac{\partial}{\partial x}u(x, t) = -2^{-1}h^* \frac{\partial^2}{\partial x^2}u(x, t) + 6^{-1}h^{*2} \frac{\partial^3}{\partial x^3}u(\vartheta, t), \quad (x, t) \in G_{h^*\tau^*}^*, \quad \vartheta \in [x - h^*, x]; \quad (6.8)$$

$$\delta_{\bar{t}^*}u(x, t) - \frac{\partial}{\partial t}u(x, t) = -2^{-1}\tau^* \frac{\partial^2}{\partial t^2}u(x, t) + 6^{-1}\tau^{*2} \frac{\partial^3}{\partial t^3}u(x, \eta), \quad (x, t) \in G_{h^*\tau^*}^*, \quad \eta \in [t - \tau^*, t].$$

Таким образом, справедлива следующая

Теорема 6.1. Пусть для данных задачи Коши для уравнения переноса (2.2), (2.1) выполняется условие $b, c, p, f \in C^{k, k_0}(\bar{G})$, $k + k_0 \leq K$, $K = 3$. Тогда для невязок $\delta_{\bar{x}^*}u(x, t) - \frac{\partial}{\partial x}u(x, t)$ и $\delta_{\bar{t}^*}u(x, t) - \frac{\partial}{\partial t}u(x, t)$ справедливы разложения (6.8) на разреженной сетке $G_{h^*\tau^*}^*$.

7. РАЗНОСТНАЯ СХЕМА РИЧАРДСОНА

Отметим, что структура разложений невязок сеточных решений относительно шагов сеток на основной и разреженной равномерных сетках подобна. Эти разложения позволяют применить технику Ричардсона и построить сеточное решение, сходящееся со вторым порядком скорости сходимости.

7.1. Рассмотрим линейную комбинацию решения $z(x, t)$ разностной схемы (4.2), (4.1) на основной сетке и решения $z^*(x, t)$ разностной схемы (6.2), (6.1) на разреженной сетке $\bar{G}_{h^*\tau^*}^*$, введенной в (6.1):

$$\hat{z}(x, t) = \hat{z}(x, t; \alpha) = \alpha z(x, t) + (1 - \alpha)z^*(x, t), \quad (x, t) \in \bar{G}_{h^*\tau^*}^*, \quad (7.1)$$

где параметр α выбираем из условия, чтобы разложения по h и τ для невязки $\hat{z}(x, t)$ не содержали линейных членов разложения.

С учетом разложений невязок (5.6) и (6.8), получаем

$$\alpha = 2. \quad (7.2)$$

В силу соотношений (7.1) и (7.2), функция $\hat{z}(x, t)$, $(x, t) \in \bar{G}_{h^*\tau^*}^*$, при $\alpha = 2$ переходит в функцию $\tilde{z}(x, t)$, определяемую следующим соотношением:

$$\tilde{z}(x, t) = \hat{z}(x, t; \alpha = 2) = 2z(x, t) - z^*(x, t), \quad (x, t) \in \bar{G}_{h^*\tau^*}^*. \quad (7.3)$$

Заметим, что для функции $\tilde{z}(x, t)$ разложения невязок по h и τ не содержат линейных членов, что влечет следующую оценку:

$$|\tilde{z}(x, t) - u(x, t)| \leq M(N^{-2} + N_0^{-2}), \quad (x, t) \in \bar{G}_{h^*\tau^*}^*. \quad (7.4)$$

Здесь мы учли, что при $t = 0$ и $\alpha = 2$ соотношение (7.3) также выполняется.

Таким образом, построенная разностная схема — будем говорить — схема Ричардсона (7.3), (6.1), при $N, N_0 \rightarrow \infty$ сходится в равномерной норме со вторым порядком скорости сходимости; для решения разностной схемы справедлива оценка (7.4).

Справедлива следующая

Теорема 7.1. Пусть для данных задачи Коши для уравнения переноса (2.2), (2.1) выполняется условие $b, c, p, f \in C^{k, k_0}(\bar{G})$, $k + k_0 \leq K$, $K = 3$. Тогда решение $\tilde{z}(x, t)$, $(x, t) \in \bar{G}_{h^*\tau^*}^*$, разностной схемы Ричардсона (7.3), (6.1) сходится в равномерной норме со вторым порядком скорости сходимости; для решения $\tilde{z}(x, t)$, $(x, t) \in \bar{G}_{h^*\tau^*}^*$, справедлива оценка (7.4).

8. ВЫВОДЫ

1. Рассмотрена постановка задачи Коши для уравнения переноса и определена цель исследования.

2. Для задачи Коши для уравнения переноса построена монотонная стандартная разностная схема и установлена ее сходимость в равномерной норме с первым порядком скорости сходимости.

3. Построены разложения первых обратных разностных производных по степеням шагов основных равномерных пространственной и временной сеток. С использованием этих разложений получены разложения для соответствующих невязок сеточных решений.

4. Построены разложения первых обратных разностных производных по степеням шагов разреженных равномерных пространственной и временной сеток. С использованием этих разложений получены разложения для соответствующих невязок сеточных решений.

5. На основе линейной комбинации решений разностных схем на основной и разреженной сетках, с учетом разложений соответствующих невязок на основной и разреженной сетках, построена разностная схема Ричардсона, сходящаяся в равномерной норме со вторым порядком скорости сходимости.

СПИСОК ЛИТЕРАТУРЫ

1. Марчук Г.И. Методы вычислительной математики. М.: Наука, 1989. 608 с.
2. Марчук Г.И., Шайдуров В.В. Повышение точности решений разностных схем. М.: Наука, 1979. 320 с.
3. Самарский А.А. Теория разностных схем. М.: Наука, 1989. 616 с.
4. Shishkin G.I., Shishkina L.P. Difference Methods for Singular Perturbation Problems. V. 140 of Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics. Boca Raton: CRC Press, 2009. 408 p.
5. Калиткин Н.Н. Численные методы. М.: Наука, 1978. 512 с.
6. Шишкин Г.И. Разностная схема для начально-краевой задачи для сингулярно возмущенного уравнения переноса // Ж. вычисл. матем. и матем. физ. 2017. Т. 57. № 11. С. 1824—1830.

ОБЩИЕ ЧИСЛЕННЫЕ МЕТОДЫ

УДК 519.626

ОБ ОДНОВРЕМЕННОМ ОПРЕДЕЛЕНИИ КОЭФФИЦИЕНТА ТЕПЛОПРОВОДНОСТИ И ОБЪЕМНОЙ ТЕПЛОЕМКОСТИ ВЕЩЕСТВА¹⁾

© 2023 г. А. Ю. Горчаков¹, В. И. Зубов^{1,*}

¹ 119333 Москва, ул. Вавилова, 40, ФИЦ ИУ РАН, Россия

*e-mail: vladimir.zubov@mail.ru

Поступила в редакцию 15.02.2023 г.

Переработанный вариант 15.02.2023 г.

Принята к публикации 28.04.2023 г.

Изучение нелинейных проблем, связанных с процессом теплопередачи в веществе, очень важно для практики. Ранее авторами был предложен эффективный алгоритм определения коэффициента теплопроводности вещества на основе результатов экспериментального наблюдения за динамикой температурного поля в объекте. В данной работе исследуется возможность расширения применения предложенного алгоритма для получения численного решения задачи одновременной идентификации зависящих от температуры объемной теплоемкости и коэффициента теплопроводности исследуемого вещества. Рассмотрение осуществляется на основе первой краевой задачи для одномерного нестационарного уравнения теплопроводности. Рассматриваемая обратная коэффициентная задача сводится к вариационной задаче, которая решается градиентными методами, основанными на применении методологии быстрого автоматического дифференцирования. Исследуется вопрос единственности решения обратной задачи. Библ. 26. Фиг. 9.

Ключевые слова: теплопроводность, обратные коэффициентные задачи, градиент, уравнение теплопроводности.

DOI: 10.31857/S0044466923080070, **EDN:** WSEUWS

ВВЕДЕНИЕ

Классическое уравнение теплопроводности часто используют для описания и математического моделирования разнообразных тепловых процессов. При этом входящие в уравнение теплопроводности плотность вещества, его удельная теплоемкость и коэффициент теплопроводности считаются известными функциями координат и температуры. Задание дополнительных краевых условий позволяет определить динамику температурного поля в изучаемом объекте.

Однако не всегда свойства вещества бывают известны. Так, при создании новых материалов приходится встречаться с ситуацией, когда объемная теплоемкость и коэффициент теплопроводности вещества зависят только от температуры, и эта зависимость неизвестна. В связи с этим возникает задача определения зависимости объемной теплоемкости и коэффициента теплопроводности вещества от температуры по результатам экспериментального наблюдения за динамикой температурного поля. К такой же задаче приходят и в том случае, когда хотят описать сложный тепловой процесс некоторой упрощенной математической моделью. Примером может служить математическое моделирование процессов распространения тепла в сложных пористых композиционных материалах, где заметную роль играет радиационная теплопередача (см. [1, 2]).

Задача определения только коэффициента теплопроводности вещества рассматривается довольно давно. Подтверждением этому может служить большое количество публикаций, посвященных указанному вопросу (см., например, [3–11]). Большое внимание в этих работах уделяется не только теоретическому исследованию обратных коэффициентных задач, но и разработке численных методов их решения. Что касается одновременного определения объемной теплоем-

¹⁾ Работа выполнялась с использованием инфраструктуры Центра коллективного пользования “Высокопроизводительные вычисления и большие данные” (ЦКП “Информатика”) ФИЦ ИУ РАН (г. Москва).

кости и коэффициента теплопроводности вещества, то работ, посвященных этому вопросу, немало (см., например, [12–14]).

Все указанные выше работы посвящены решению одной и той же задачи — определению теплофизических параметров исследуемого вещества. Однако используемые математические постановки этой общей задачи в разных работах разные. Разные и методы, применяемые для решения этой задачи.

В настоящей работе рассматривается одна из возможных постановок такой обратной коэффициентной задачи. Ее рассмотрение проводится на основе первой краевой задачи для одномерного нестационарного уравнения теплопроводности. Обратная коэффициентная задача сводится к вариационной задаче. В качестве целевого функционала выбрано среднеквадратичное отклонение рассчитанного поля температуры в образце от его экспериментального значения. Предложен алгоритм численного решения обратной коэффициентной задачи. В основе этого алгоритма лежит современная методология быстрого автоматического дифференцирования (БАД-методология, см. [15, 16]), позволившая успешно решить ряд сложных задач оптимального управления динамическими системами (см., например, [17–19]). Приведенные в работе примеры решения обратной коэффициентной задачи демонстрируют работоспособность предложенного алгоритма. В статье, кроме того, исследуется вопрос единственности получаемого решения сформулированной обратной задачи.

1. ПОСТАНОВКА ЗАДАЧИ

Рассматривается слой материала ширины L . Температура этого слоя в начальный момент времени известна. Также известно, как изменяется со временем температура на краях этого слоя. Распределение температурного поля в слое в каждый момент времени описывается решением следующей начально-краевой (смешанной) задачи:

$$C(T) \frac{\partial T(x, t)}{\partial t} - \frac{\partial}{\partial x} \left(K(T) \frac{\partial T(x, t)}{\partial x} \right) = 0, \quad (x, t) \in Q, \quad (1.1)$$

$$T(x, 0) = w_0(x) \quad (0 \leq x \leq L), \quad (1.2)$$

$$T(0, t) = w_1(t), \quad T(L, t) = w_2(t) \quad (0 \leq t \leq \Theta). \quad (1.3)$$

Здесь x — декартова координата точки в слое; t — время; $Q = \{(0 < x < L) \times (0 < t < \Theta)\}$; $T(x, t)$ — температура вещества в точке с координатой x в момент времени t ; $C(T)$ — объемная теплоемкость вещества, $K(T)$ — коэффициент теплопроводности; $w_0(x)$ — заданная температура слоя в начальный момент времени; $w_1(t)$ — заданная температура на левом краю слоя; $w_2(t)$ — заданная температура на правом краю слоя.

Если зависимость объемной теплоемкости вещества $C(T)$ и его коэффициента теплопроводности $K(T)$ от температуры T известна, то, решая смешанную задачу (1.1)–(1.3), найдем распределение температуры $T(x, t)$ в $\bar{Q}$. Задачу (1.1)–(1.3) ниже будем называть *прямой* задачей.

Если же зависимость объемной теплоемкости вещества и его коэффициента теплопроводности от температуры неизвестна, то представляет большой интерес задача определения этих зависимостей. Одной из возможных постановок такой задачи (эта задача относится к классу задач идентификации параметров модели) является следующая: найти такую зависимость объемной теплоемкости вещества $C(T)$ и коэффициента теплопроводности $K(T)$ от температуры T , при которой температурное поле $T(x, t)$, полученное в результате решения сформулированной задачи (1.1)–(1.3), мало отличается от поля $Y(x, t)$, полученного экспериментально. Мерой отклонения этих функций может служить величина

$$\Phi(C(T), K(T)) = \int_0^\Theta \int_0^L [T(x, t) - Y(x, t)]^2 \mu(x, t) dx dt, \quad (1.4)$$

где $\mu(x, t) \geq 0$ — заданная весовая функция. Таким образом, задача оптимального управления состоит в определении оптимального управления $\{C(T), K(T)\}$ и соответствующего оптимального решения $T(x, t)$ задачи (1.1)–(1.3), при котором функционал (1.4) достигает минимального значения.

Следует отметить, что в представленной постановке задача оптимального управления всегда имеет неединственное решение. Действительно, если $\{C^*(T), K^*(T)\}$ – решение сформулированной обратной задачи, то при любом действительном числе λ управление $\{\lambda C^*(T), \lambda K^*(T)\}$ также является решением этой задачи. Для выделения единственного решения сформулированной задачи идентификации необходимо использовать дополнительное условие. В качестве такого условия может выступать, например, условие $K(T_*) = K_*$ с заданными числами T_* и K_* . Можно также при решении задачи определять отношение $\chi(T) = C(T)/K(T)$.

Задачи оптимального управления, подобные этой, обычно решаются численно с помощью некоторого метода спуска, который требует знания градиента функционала (1.4).

2. ЗАДАЧА ОПТИМАЛЬНОГО УПРАВЛЕНИЯ В ДИСКРЕТНОМ СЛУЧАЕ

При численном решении сформулированной задачи оптимального управления искомые функции $C(T)$ и $K(T)$, $T \in [a, b]$, аппроксимировались непрерывными кусочно-линейными функциями следующим образом. Отрезок $[a, b]$, на котором идентифицируются функции $C(T)$ и $K(T)$, разбивался точками $\tilde{T}_0 = a, \tilde{T}_1, \tilde{T}_2, \dots, \tilde{T}_N = b$ на N частей (они могут быть как равными, так и неравными). Каждой из точек $\tilde{T}_n$, ($n = 0, \dots, N$) ставились в соответствие числа $c_n = C(\tilde{T}_n)$, $k_n = K(\tilde{T}_n)$. Искомые функции $C(T)$ и $K(T)$ аппроксимировались непрерывными кусочно-линейными функциями с узлами в точках $\{(\tilde{T}_n, c_n)\}_{n=0}^N$ и $\{(\tilde{T}_n, k_n)\}_{n=0}^N$, так что

$$C(T) = c_n + \frac{c_{n+1} - c_n}{\tilde{T}_{n+1} - \tilde{T}_n} (T - \tilde{T}_n) \quad \text{при} \quad \tilde{T}_n \leq T \leq \tilde{T}_{n+1} \quad (n = 0, \dots, N-1),$$

$$K(T) = k_n + \frac{k_{n+1} - k_n}{\tilde{T}_{n+1} - \tilde{T}_n} (T - \tilde{T}_n) \quad \text{при} \quad \tilde{T}_n \leq T \leq \tilde{T}_{n+1} \quad (n = 0, \dots, N-1).$$

Область $Q = \{(x, t) : (0 < x < L) \times (0 < t < \Theta)\}$ разбивалась сеточными линиями $\{\tilde{x}_i\}_{i=0}^I$ и $\{\tilde{t}^j\}_{j=0}^J$ на ряд прямоугольников с шагами $h_i = \tilde{x}_{i+1} - \tilde{x}_i$, ($i = \overline{0, I-1}$), $\tau^j = \tilde{t}^j - \tilde{t}^{j-1}$, ($j = \overline{1, J}$). В каждом узле $(\tilde{x}_i, \tilde{t}^j)$ расчетной области Q , характеризуемом парой индексов (i, j) , все функции задавались своими значениями в точке $(\tilde{x}_i, \tilde{t}^j)$ (например, $T(\tilde{x}_i, \tilde{t}^j) = T_i^j$). В каждом из полученных прямоугольников требовалось выполнение закона теплового баланса. В результате получалась конечно-разностная схема, аппроксимирующая смешанную задачу (1.1)–(1.3). Полученная система нелинейных алгебраических уравнений решалась итерационным методом с привлечением метода прогонки [20]. С помощью описанного подхода решалась смешанная задача (1.1)–(1.3) и находилась функция $T(x, t)$ (точнее, ее аппроксимация T_i^j).

Целевой функционал (1.4) аппроксимировался функцией $F(c_0, c_1, \dots, c_N, k_0, k_1, \dots, k_N)$ конечного числа переменных следующим образом:

$$\Phi(C(T), K(T)) \approx F = \sum_{j=1}^J \sum_{i=1}^{I-1} (T_i^j - Y_i^j)^2 \mu_i^j h_i \tau^j.$$

Для минимизации целевого функционала в работе использовался градиентный метод. Хорошо известно, что для эффективной работы градиентных методов следует использовать точное значение градиента целевой функции. Вычислить значение градиента эффективно и с машинной точностью позволяет БАД-методология (см. [15, 16, 21]). Именно применение БАД-методологии позволило успешно решить ряд сложных задач оптимального управления. В работах [17–19, 21] использовался “ручной” вариант БАД-методологии, при котором все необходимые формулы выводились вручную и использовались в оптимизационном коде. Однако появились стандартные пакеты программ, в автоматическом режиме выполняющие примерно ту же работу (см. [22–24]). В статье [25] представлены результаты сравнения градиентов, вычисленных с помощью ручной версии БАД-методологии, с помощью стандартных пакетов программ, и сделан вывод, что стандартные пакеты программ позволяют практически с той же точностью вычислять градиенты сложных функций, что и ручная версия БАД-методологии. Поэтому в настоящей работе градиент целевой функции вычислялся с помощью пакета прикладного программного обеспече-

ния Adept (см. [23]), при этом использовался обратный (Reverse, Adjoint) метод автоматического дифференцирования. Для поиска минимума целевой функции применялся метод L-BFGS-B (см. [26]).

3. ЧИСЛЕННЫЕ РЕЗУЛЬТАТЫ

Минимальное значение целевого функционала (1.4), получаемое при решении сформулированной обратной задачи, зависит от заданного экспериментального поля $Y(x, t)$. Если экспериментальное поле $Y(x, t)$ принадлежит области достижимости функций $T(x, t)$, построенных в результате решения прямой задачи (1.1)–(1.3) при допустимых $C(T)$ и $K(T)$, то значение функционала равно нулю. В противном случае значение функционала больше нуля. При проведении численных экспериментов экспериментальное поле температур $Y(x, t)$ выбиралось из области достижимости решений прямой задачи.

Для проверки работоспособности предложенного алгоритма было выполнено большое число тестовых расчетов. Эти расчеты условно можно разделить на две группы.

Все примеры первой группы базировались на том, что функция

$$T_*(x, t) = \left(\frac{n+1}{\beta} \right)^{\frac{1}{n+1}} [\alpha(t + \gamma - x)]^{\frac{1}{m-n}},$$

$$\alpha = \frac{(m-n)}{(n+1)}, \quad \beta = (n+1)^{\frac{m-n}{m+1}},$$

является аналитическим решением уравнения (1.1) при $\gamma = \text{const}$, $C(T) = \beta T^n$, $K(T) = T^m$.

В соответствии с этим использовались следующие входные данные задачи:

$$L = 1, \quad \Theta = 1, \quad \gamma = 1.5, \quad Q = (0, 1) \times (0, 1),$$

$$w_0(x) = T_*(x, 0) = \left(\frac{n+1}{\beta} \right)^{\frac{1}{n+1}} [\alpha(\gamma - x)]^{\frac{1}{m-n}} \quad (0 \leq x \leq 1),$$

$$w_1(t) = T_*(0, t) = \left(\frac{n+1}{\beta} \right)^{\frac{1}{n+1}} [\alpha(t + \gamma)]^{\frac{1}{m-n}} \quad (0 \leq t \leq 1),$$

$$w_2(t) = T_*(1, t) = \left(\frac{n+1}{\beta} \right)^{\frac{1}{n+1}} [\alpha(t + \gamma - 1)]^{\frac{1}{m-n}} \quad (0 \leq t \leq 1),$$

$$Y(x, t) = T_*(x), \quad (x, t) \in Q,$$

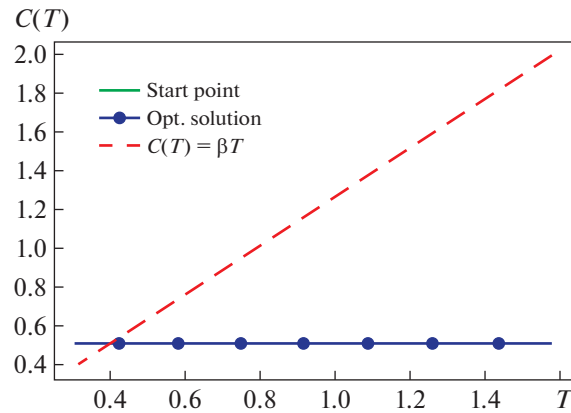
$$a = \min_{(x,t) \in Q} Y(x, t), \quad b = \max_{(x,t) \in Q} Y(x, t).$$
(3.1)

При решении прямой задачи использовалась равномерная сетка с параметрами $I = 40$ (количество интервалов вдоль оси x), $J = 700$ (количество интервалов вдоль оси t), которая обеспечивает достаточную точность вычисления поля температур и поля импульсов (сопряженных переменных). Отрезок $[a, b]$ разбивался на 80 интервалов ($N = 80$).

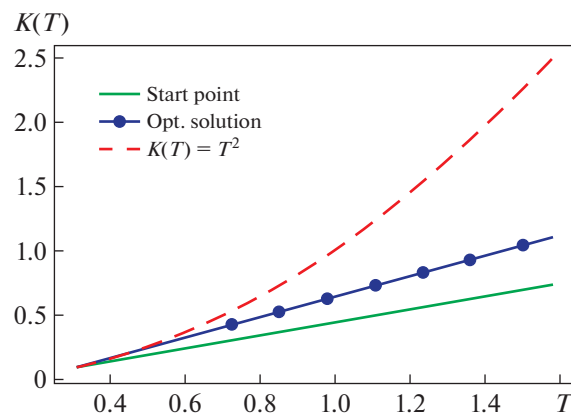
В первом примере первой группы расчетов рассматривалась задача нахождения объемной теплоемкости и коэффициента теплопроводности с входными данными (3.1) при $m = 2$ и $n = 1$. При этом полагалось, что в целевом функционале (1.4) весовая функция $\mu(x) \equiv 1$. Как отмечалось в разд. 1, решение сформулированной задачи идентификации неединственно. Поэтому при решении задачи идентификации объемной теплоемкости и коэффициента теплопроводности использовалось дополнительное условие $K(a) = a^m$.

Вначале при решении задачи оптимизации функционала (1.4) в качестве начального приближения выбирались функции $C(T) = 0.5$ и $K(T) = 0.5 \cdot T$. Результаты полученного решения обратной задачи представлены на фиг. 1 (функция $C(T)$) и фиг. 2 (функция $K(T)$).

В результате минимизации целевого функционала (1.4) оптимальная функция $C(T)$ (Opt. solution) не изменилась по сравнению с начальным приближением (Start point). На фиг. 1 они нераз-



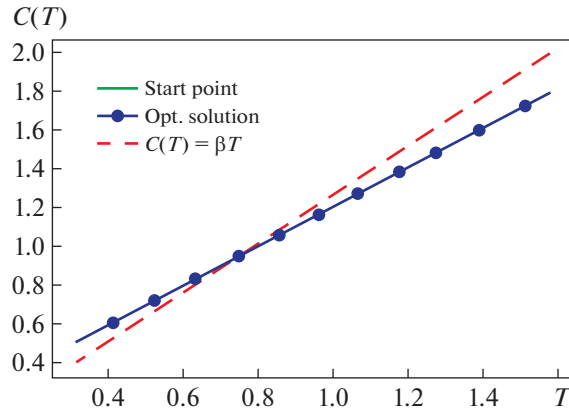
Фиг. 1. Распределение объемной теплоемкости.



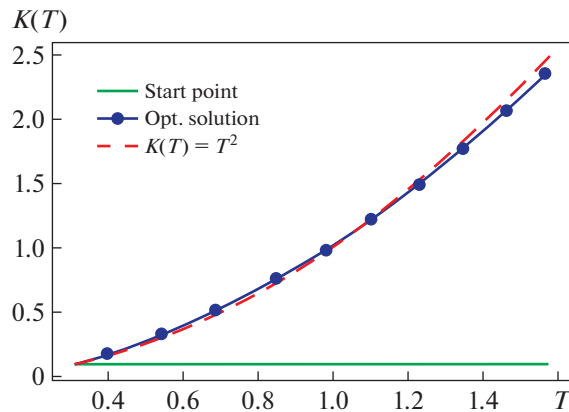
Фиг. 2. Распределение коэффициента теплопроводности.

личимы. Однако полученная функция $C(T)$ отличается от ожидавшегося вида $C(T) = \beta T$ (штриховая линия). Что касается функции $K(T)$, то здесь оптимальное решение (Opt. solution) уже отличается от начального приближения, оставаясь при этом линейной функцией. И опять оптимальное решение отличается от ожидавшегося вида $K(T) = T^2$ (штриховая линия). Отметим при этом, что полученное распределение по температуре отношения $\frac{K'(T)}{C(T)}$ хорошо совпадает с теоретическим значением $\frac{2}{\beta} = 2^{2/3}$, получаемым при $C(T) = \beta T$ и $K(T) = T^2$.

На следующем шаге при решении той же задачи оптимизации функционала (1.4) в качестве начального приближения для функции $C(T)$ была выбрана линейная функция (на фиг. 3 она обозначена как “Start point” и совпадает с “Opt. solution”), а для функции $K(T)$ — постоянная функция, проходящая через точку (a, a^2) (на фиг. 4 она обозначена как “Start point”). В результате решения задачи оптимизации получены оптимальные управления, обозначенные на фиг. 3 и фиг. 4 как “Opt. solution”. Эти фигуры показывают следующее. Оптимальная функция $C(T)$ не отличается от начального приближения, но не совпадает с ожидаемой функцией $C(T) = \beta T$ (штриховая линия на фиг. 3). Оптимальная функция $K(T)$ представляет собой параболу, проходящую через точку (a, a^2) , но также не совпадает с ожидаемой функцией $K(T) = T^2$ (штриховая линия на фиг. 4). Интересным также представляется тот факт, что отношение $\frac{K'(T)}{C(T)}$, как и в



Фиг. 3. Распределение объемной теплоемкости.



Фиг. 4. Распределение коэффициента теплопроводности.

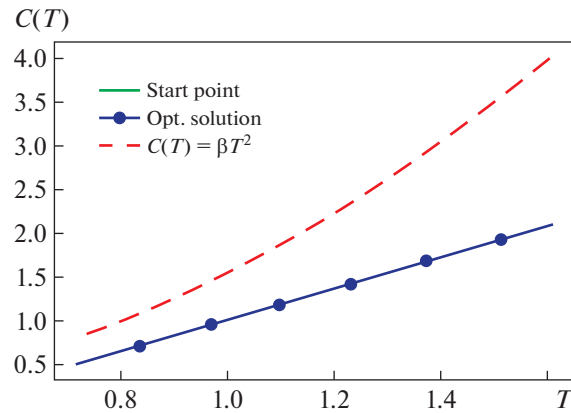
предыдущем случае, остается практически постоянным на отрезке $[a, b]$ изменения температуры и хорошо совпадает с теоретическим значением $\frac{2}{\beta} = 2^{2/3}$, получаемым при $C(T) = \beta T$ и $K(T) = T^2$.

Наконец, при решении той же задачи оптимизации в качестве начальных приближений выбирались “ожидаемые” функции, а именно, $C(T) = \beta T$, $K(T) = T^2$, $\beta = \sqrt[3]{2}$ (эти функции на предыдущих фигурах выделялись штриховыми линиями). В результате оптимизации эти функции практически не изменились. Небольшие изменения, на графиках совершенно не видные, связаны с аппроксимацией как самих функций, так и дифференциального уравнения. Отношение $\frac{K'(T)}{C(T)}$, построенное на основе полученного оптимального решения, отличается от точного не более, чем на 10^{-10} .

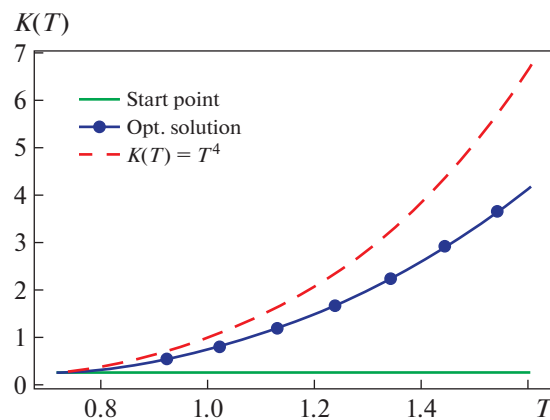
Отметим также, что во всех рассмотренных случаях отношение $\chi(T) = C(T)/K(T)$ представляло собой разные функции для разных вариантов.

Во втором примере первой группы расчетов рассматривалась задача нахождения объемной теплоемкости и коэффициента теплопроводности с входными данными (3.1) при $m = 4$ и $n = 2$. При этом, как и раньше, полагалось, что в целевом функционале (1.4) весовая функция $\mu(x) \equiv 1$. При решении задачи идентификации объемной теплоемкости и коэффициента теплопроводности использовалось дополнительное условие $K(a) = a^m$.

Вначале при решении сформулированной задачи оптимизации функционала (1.4) в качестве начального приближения для функции $C(T)$ была выбрана линейная функция (на фиг. 5 она



Фиг. 5. Распределение объемной теплоемкости.

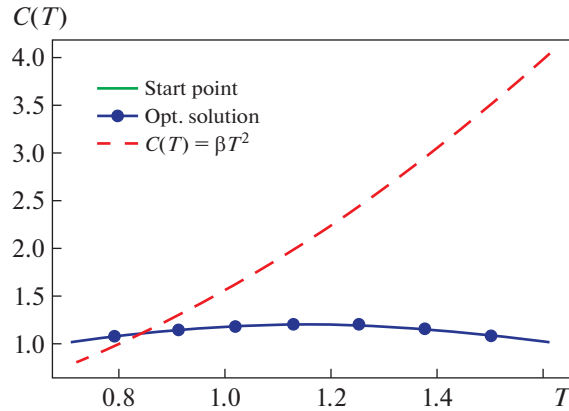


Фиг. 6. Распределение коэффициента теплопроводности.

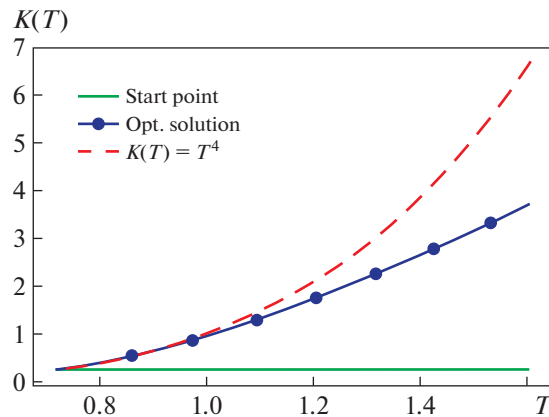
обозначена как “Start point” и совпадает с “Opt. solution”), а для функции $K(T)$ — постоянная функция, проходящая через точку (a, a^4) (на фиг. 6 она обозначена как “Start point”). В результате решения задачи оптимизации получены оптимальные управления, обозначенные на фиг. 5 и фиг. 6 как “Opt. solution”. Эти фигуры показывают следующее. Оптимальная функция $C(T)$ не отличается от начального приближения, но не совпадает с ожидаемой функцией $C(T) = \beta T^2$ (штриховая линия на фиг. 5). Оптимальная функция $K(T)$ представляет собой кубическую параболу, проходящую через точку (a, a^4) . Она также не совпадает с ожидаемой функцией $K(T) = T^4$ (штриховая линия на фиг. 6).

Далее, задача из второго примера первой группы расчетов решалась в случае, когда в качестве начальных приближений выбирались следующие функции, а именно, $C(T) = (T - a)(b - T) + 1$ (см. фиг. 7), $K(T) = a^4$ (см. фиг. 8) (эти функции на фиг. 7 и фиг. 8 выделялись сплошными линиями). В результате оптимизации функция $C(T)$ практически не изменилась (на фиг. 7 начальное приближение и оптимальное решение неразличимы). Что касается функции $K(T)$, то оптимальное решение приближенно имеет вид кубической параболы (см. фиг. 8).

Наконец, при решении задачи оптимизации в качестве начальных приближений выбирались “ожидаемые” функции, а именно, $C(T) = \beta T^2$, $K(T) = T^4$, $\beta = \sqrt[5]{9}$ (эти функции на предыдущих фигурах выделялись штриховыми линиями). В результате оптимизации эти функции практически не изменились. Небольшие изменения, на графиках совершенно не видимые, связаны с ап-



Фиг. 7. Распределение объемной теплоемкости.



Фиг. 8. Распределение коэффициента теплопроводности.

проксимацией как самих функций, так и дифференциального уравнения. Отношение $\frac{K'(T)}{C(T)}$, построенное на основе оптимального решения, отличается от точного не более, чем на 10^{-10} .

Как и в первом примере первой группы, здесь во всех рассмотренных случаях отношение $\chi(T) = C(T)/K(T)$ представляло собой разные функции для разных вариантов.

Задача нахождения объемной теплоемкости и коэффициента теплопроводности с входными данными (3.1) решалась и при других значениях параметров m и n . На основе результатов, полученных при решении всех обратных задач первой группы, можно сделать вывод о том, что решение задачи определения объемной теплоемкости и коэффициента теплопроводности является неединственным.

В качестве первого примера второй группы рассмотрим обратную задачу (1.1)–(1.4) со следующими входными данными:

$$\begin{aligned}
 L &= 1, \quad \Theta = 1, \quad Q = (0, 1) \times (0, 1), \\
 w_0(x) &= x^2 \quad (0 \leq x \leq 1), \\
 w_1(t) &= T_*(0, t) = t, \quad w_2(t) = T_*(1, t) = t + 1 \quad (0 \leq t \leq 1), \\
 Y(x, t) &= T_*(x, t) = x^2 + t, \quad (x, t) \in Q. \\
 a &= 0, \quad b = 2.
 \end{aligned} \tag{3.2}$$

Нетрудно проверить, что функция $T_*(x, t) = x^2 + t$ является аналитическим решением уравнения (1.1) при $C(T) = 2$ и $K(T) = 1$, так что пара $\{C(T), K(T)\} = \{2, 1\}$ и функция $T_*(x, t) = x^2 + t$ являются оптимальным управлением и соответствующим оптимальным решением обратной задачи (1.1)–(1.4) с данными (3.2).

Сформулированная задача решалась численно при разных начальных приближениях, например, $C(T) = 2T$, $K(T) = \frac{1}{4}T^2$, $K(2) = 1$. Независимо от начального приближения получалось одно и то же решение $C(T) = 2$, $K(T) = 1$ с точностью 10^{-12} в норме C . Как оказалось, это связано со спецификой используемого экспериментального поля температур. Объяснить единственность получаемого решения можно следующим образом.

В предположении, что функция $T(x, t) \in C^2(Q) \cap C(\bar{Q})$, $C(T) \in C([a, b])$, $K(T) \in C^1([a, b])$, уравнение (1.1) можно записать в виде

$$C(T) \frac{\partial T(x, t)}{\partial t} - K'(T) \left(\frac{\partial T(x, t)}{\partial x} \right)^2 - K(T) \frac{\partial^2 T(x, t)}{\partial x^2} = 0. \quad (3.3)$$

Подставив функцию $T_*(x, t) = x^2 + t$ в уравнение (3.3), получим равенство, справедливое для всех $(x, t) \in Q$

$$C(T_*) \times 1 - K'(T_*) \times 4x^2 - K(T_*) \times 2 = 0.$$

Отсюда следует, что вдоль произвольной изотермы $x^2 + t = \tau$, ($\tau \in (0, 2)$) поля $T_*(x, t)$ должно быть выполнено соотношение

$$[C(\tau) - 2K(\tau)] = K'(\tau) \times 4x^2, \quad \tau = \text{const}, \quad x \in \begin{cases} (0, \sqrt{\tau}), & 0 < \tau \leq 1, \\ (\sqrt{\tau-1}, 1), & 1 < \tau < 2, \end{cases}$$

что возможно лишь при $K'(\tau) = 0$, $C(\tau) = 2K(\tau)$. Принимая во внимание условие $K(2) = 1$, получаем решение обратной задачи $C(T) = 2$, $K(T) = 1$.

Во втором примере второй группы рассматривалась обратная задача (1.1)–(1.4) с такими входными данными:

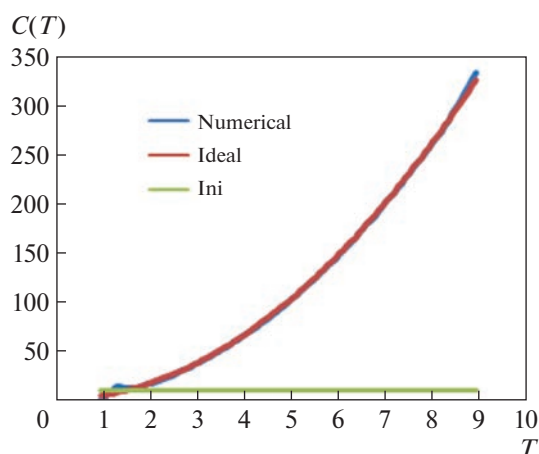
$$\begin{aligned} L &= 1, \quad \Theta = 1, \quad Q = (0, 1) \times (0, 1), \\ w_0(x) &= 9/(x+1) \quad (0 \leq x \leq 1), \\ w_1(t) &= 9/(5t+1) \quad (0 \leq t \leq 1), \\ w_2(t) &= 9/(7t+2) \quad (0 \leq t \leq 1), \\ Y(x, t) &= T_{**}(x, t) \quad (x, t) \in Q, \\ a &= 1, \quad b = 9, \quad K(1) = 1, \end{aligned}$$

где экспериментальное поле $Y(x, t) = T_{**}(x, t)$ строилось численно как решение следующей смешанной задачи (1.1)–(1.3):

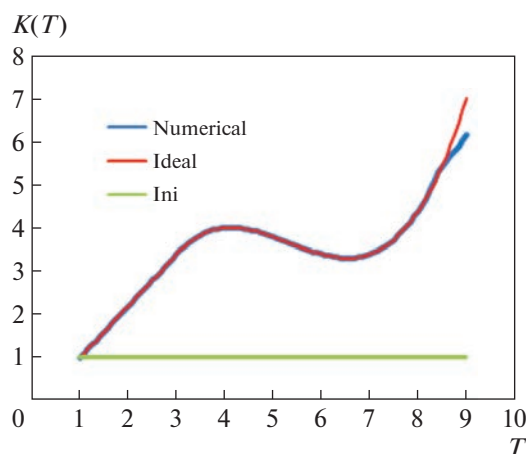
$$\begin{aligned} C(T) \frac{\partial T(x, t)}{\partial t} - \frac{\partial}{\partial x} \left(K(T) \frac{\partial T(x, t)}{\partial x} \right) &= 0 \quad (0 < x < 1), \quad (0 < t < 1), \\ T(x, 0) &= 9/(x+1) \quad (0 \leq x \leq 1), \\ T(0, t) &= 9/(5t+1) \quad (0 \leq t \leq 1), \\ T(L, t) &= 9/(7t+2) \quad (0 \leq t \leq 1), \end{aligned}$$

при

$$K(T) = \begin{cases} 0.1(T-3)(T-6)(T-7) + 3.4, & T \geq 3, \\ 1.2(T-3) + 3.4, & T < 3. \end{cases}$$



Фиг. 9. Распределение объемной теплоемкости.



Фиг. 10. Распределение коэффициента теплопроводности.

$$C(T) = 4T^2, \quad T \in [a, b] \quad a = 1, \quad b = 9.$$

Здесь, как и в первом примере второй группы, обратная задача решалась численно при разных начальных приближениях. Независимо от начального приближения получалось одно и то же решение с высокой точностью. На фиг. 9 и фиг. 10 представлены результаты решения задачи при начальном приближении $C(T) = 10$, $K(T) = 1$, которые подтверждают, что единственное решение определяется предложенным алгоритмом с высокой точностью.

Вторая группа рассмотренных примеров позволяет заключить, что существуют такие поля из области достижимости, при которых решение обратной задачи определения объемной теплоемкости и коэффициента теплопроводности является единственным.

4. АНАЛИЗ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

Проведенные вычислительные эксперименты показали, что единственность решения задачи одновременной идентификации объемной теплоемкости и коэффициента теплопроводности в предложенной постановке существенно зависит от используемого экспериментального поля. Эти эксперименты позволили сделать некоторые выводы.

Обоснование этих выводов проведем при следующих предположениях.

Искомые объемная теплоемкость $C(T)$ и коэффициент теплопроводности $K(T)$ должны принадлежать области определения, т.е. удовлетворять условиям: $C(T) \in C([a, b])$, $K(T) \in C^1([a, b])$, $C(T) > 0$, $K(T) > 0$, для всех $T \in [a, b]$. Под областью достижимости будем понимать множество функций $T(x, t) \in C_{x,t}^{2,1}(\bar{Q}) \cap C(\bar{Q})$ таких, что они удовлетворяют заданным условиям (1.2), (1.3) и для каждой из них в области определения найдутся функции $C(T)$ и $K(T)$, при которых $T(x, t)$ является решением смешанной задачи (1.1)–(1.3).

При таких предположениях уравнение (1.1) можно представить в виде (3.3).

Далее, под τ -изолинией функции $T(x, t)$ будем понимать, как это принято, множество точек (x, t) области Q , в которых значение функции $T(x, t)$ имеет одно и то же значение τ . При этом отметим, что изолиния (множество) может представлять собой объединение нескольких, не связанных между собой линий.

Справедливы следующие утверждения.

Лемма 1. Пусть функция $T(x, t)$ принадлежит области достижимости. Если на некоторой τ -изолинии, $T(x, t) = \tau \in [a, b]$, функции $\frac{\partial T(x, t)}{\partial t}$ и $\left(\frac{\partial T(x, t)}{\partial x}\right)^2$, рассматриваемые как функции некоторого параметра s вдоль этой изолинии, являются линейно независимыми, то отношение $\chi(T) = \frac{C(T)}{K(T)}$ имеет единственное значение при $T = \tau$.

Действительно, пусть у обратной задачи имеются два решения: $\{C_1(T), K_1(T)\}$ и $\{C_2(T), K_2(T)\}$ с двумя разными отношениями: $\chi_1(T) = \frac{C_1(T)}{K_1(T)}$ и $\chi_2(T) = \frac{C_2(T)}{K_2(T)}$. Это означает, что для всех $(x, t) \in Q$ справедливы два равенства типа (3.3):

$$C_1(T) \frac{\partial T(x, t)}{\partial t} - K_1'(T) \left(\frac{\partial T(x, t)}{\partial x}\right)^2 - K_1(T) \frac{\partial^2 T(x, t)}{\partial x^2} = 0, \quad (4.1)$$

$$C_2(T) \frac{\partial T(x, t)}{\partial t} - K_2'(T) \left(\frac{\partial T(x, t)}{\partial x}\right)^2 - K_2(T) \frac{\partial^2 T(x, t)}{\partial x^2} = 0. \quad (4.2)$$

Разделим обе части соотношения (4.1) на $K_1(T) > 0$, а обе части соотношения (4.2) на $K_2(T) > 0$. Разность получившихся при этом соотношений имеет вид, справедливый для всех $(x, t) \in Q$:

$$\left(\frac{C_1(T)}{K_1(T)} - \frac{C_2(T)}{K_2(T)}\right) \frac{\partial T(x, t)}{\partial t} - \left(\frac{K_1'(T)}{K_1(T)} - \frac{K_2'(T)}{K_2(T)}\right) \left(\frac{\partial T(x, t)}{\partial x}\right)^2 = 0. \quad (4.3)$$

В точках τ -изолинии $T(x, t) = \tau$ линейная независимость функций $\frac{\partial T(x, t)}{\partial t}$ и $\left(\frac{\partial T(x, t)}{\partial x}\right)^2$ вместе с условием (4.3) позволяют заключить, что

$$\frac{C_1(\tau)}{K_1(\tau)} = \frac{C_2(\tau)}{K_2(\tau)}, \quad \frac{K_1'(\tau)}{K_1(\tau)} = \frac{K_2'(\tau)}{K_2(\tau)}, \quad (4.4)$$

т.е. значения отношений $\chi(\tau) = \frac{C(\tau)}{K(\tau)}$ и $v(\tau) = \frac{K'(\tau)}{K(\tau)}$ одинаковые для всех возможных решений обратной задачи.

Теорема 1. Пусть функция $T(x, t)$ принадлежит области достижимости. Если на каждой τ -изолинии отрезка $\tau \in [\alpha, \beta] \subset [a, b]$ функции $\frac{\partial T(x, t)}{\partial t}$ и $\left(\frac{\partial T(x, t)}{\partial x}\right)^2$, рассматриваемые как функции некоторого параметра s вдоль этой изолинии, являются линейно независимыми, то для любого решения $\{C(T), K(T)\}$ обратной задачи отношение $\chi(T) = \frac{C(T)}{K(T)}$, $T \in [\alpha, \beta]$, представляет собой одну и ту же функцию.

Для доказательства теоремы 1 допустим существование двух решений $\{C_1(T), K_1(T)\}$ и $\{C_2(T), K_2(T)\}$ обратной задачи (1.1)–(1.4) с разными отношениями $\chi_1(T)$ и $\chi_2(T)$. На основании леммы 1 можно утверждать, что для всех $\tau \in [\alpha, \beta]$ справедливы равенства (4.4), т.е.

$$\frac{C_1(\tau)}{K_1(\tau)} = \frac{C_2(\tau)}{K_2(\tau)}, \quad \frac{K_1'(\tau)}{K_1(\tau)} = \frac{K_2'(\tau)}{K_2(\tau)}, \quad \tau \in [\alpha, \beta].$$

Первое из равенств показывает, что вдоль всего отрезка $T \in [\alpha, \beta]$ отношение $\chi(T) = \frac{C(T)}{K(T)}$ является функцией, не зависящей от решения обратной задачи. Интегрирование второго равенства позволяет заключить, что на отрезке $T \in [\alpha, \beta]$ отношение разных коэффициентов теплопроводности остается постоянным, т.е. $\frac{K_1(T)}{K_2(T)} = \text{const}$. Учитывая замечание, сделанное в конце разд. 1,

можно утверждать, что при сформулированных в теореме 1 условиях все решения рассматриваемой обратной задачи определяются с точностью до множителя. Если же при формулировке задачи идентификации (1.1)–(1.4) дополнительно задано условие $K(T_*) = K_*$ с заданными числами $T_* \in [\alpha, \beta]$ и K_* , то решение обратной задачи на отрезке $T \in [\alpha, \beta]$ будет единственным, хотя вне этого отрезка $T \in [a, b] \setminus [\alpha, \beta]$ решения $\{C_1(T), K_1(T)\}$ и $\{C_2(T), K_2(T)\}$ могут различаться.

Замечание 1. Примером, иллюстрирующим теорему 1, может служить первый пример второй группы (см. (3.2)). В нем функция $T(x, t) = x^2 + t$, производные $\frac{\partial T(x, t)}{\partial t} = 1$, $\frac{\partial T(x, t)}{\partial x} = 2x$, и функции 1 и $4x^2$ линейно независимы вдоль любой изолинии $x^2 + t = \tau$, ($\tau \in (0, 2)$) поля $T(x, t)$. Как показано в разд. 3, при дополнительном условии решение обратной задачи единственно.

Теорема 2. Пусть функция $T(x, t)$ принадлежит области достижимости. Если на каждой τ -изолинии отрезка $\tau \in [\alpha, \beta] \subset [a, b]$ функции $\frac{\partial T(x, t)}{\partial t}$ и $\left(\frac{\partial T(x, t)}{\partial x}\right)^2$, рассматриваемые как функции некоторого параметра s вдоль этой изолинии, являются линейно зависимыми и функция $\frac{\partial T(x, t)}{\partial x}$ не равна тождественно нулю, то решение $\{C(T), K(T)\}$ обратной задачи неединственно (имеются решения с разными отношениями $\chi(T) = \frac{C(T)}{K(T)}$).

Для доказательства теоремы 2 вначале обратим внимание на линейную зависимость функций $\frac{\partial T(x, t)}{\partial t}$ и $\left(\frac{\partial T(x, t)}{\partial x}\right)^2$ вдоль каждой τ -изолинии отрезка $\tau \in [\alpha, \beta] \subset [a, b]$. Это означает, что существуют функции $A(T)$ и $B(T)$, $T \in [\alpha, \beta]$, такие что

$$A(T) \frac{\partial T(x, t)}{\partial t} + B(T) \left(\frac{\partial T(x, t)}{\partial x}\right)^2 = 0,$$

причем $A^2(T) + B^2(T) > 0$. Более того, $A(T) \neq 0$ для всех $T \in [\alpha, \beta]$. Действительно, предположив, что $A(\tau) = 0$ для какого-то значения $\tau \in [\alpha, \beta]$, придем к противоречию с условием отличия от тождественного нуля производной $\frac{\partial T(x, t)}{\partial x}$. Таким образом, справедливо следующее равенство:

$$\frac{\partial T(x, t)}{\partial t} = \tilde{B}(T) \left(\frac{\partial T(x, t)}{\partial x}\right)^2, \quad (4.5)$$

где $\tilde{B}(T) = -\frac{B(T)}{A(T)} \in C([\alpha, \beta])$.

Так как функция $T(x, t)$ принадлежит области достижимости, то существуют функции $C_*(T)$ и $K_*(T)$ из области определения, при которых функция $T(x, t)$ является решением смешанной задачи (1.1)–(1.3). Покажем, что в области определения кроме них найдутся другие функции $C(T)$ и $K(T)$, при которых эта функция $T(x, t)$ является решением смешанной задачи (1.1)–(1.3).

Пусть $\psi(T) \in \overset{0}{C}([\alpha, \beta])$ – произвольная положительная функция. Определим функции $f(T) \in C([\alpha, \beta])$ и $\sigma(T) \in C^1([\alpha, \beta])$ соотношениями

$$f(T) = \frac{K_*'(T)}{K_*(T)} + \tilde{B}(T)\psi(T), \quad \sigma(T) = \exp \left[\int_{\alpha}^T f(\xi) d\xi \right].$$

Построим новые объемную теплоемкость $C(T)$ и коэффициент теплопроводности $K(T)$:

$$K(T) = \begin{cases} K_*(T), & a \leq T < \alpha, \\ K_*(\alpha)\sigma(T), & \alpha \leq T \leq \beta, \\ K_*(T)K_*(\alpha)\sigma(\beta)/K_*(\beta), & \beta < T \leq b, \end{cases} \quad (4.6)$$

$$C(T) = \begin{cases} C_*(T), & a \leq T < \alpha, \\ K(T)(C_*(T)/K_*(T) + \psi(T)), & \alpha \leq T \leq \beta, \\ C_*(T)K(T)/K_*(T), & \beta < T \leq b, \end{cases}$$

и покажем, что они также являются решением обратной задачи для рассматриваемой функции-поля $T(x, t)$ из области достижимости.

Во-первых, непосредственной проверкой нетрудно убедиться в том, что построенные функции $K(T) \in C^1([a, b])$, $C(T) \in C([a, b])$ и что при всех $T \in [\alpha, \beta]$ справедливо соотношение

$$\frac{K'(T)}{K(T)} - \tilde{B}(T) \frac{C(T)}{K(T)} = \frac{K_*'(T)}{K_*(T)} - \tilde{B}(T) \frac{C_*(T)}{K_*(T)}, \quad (4.7)$$

а при всех $T \in [a, \alpha]$ и $T \in [\beta, b]$ – соотношения

$$\frac{K'(T)}{K(T)} = \frac{K_*'(T)}{K_*(T)}, \quad \frac{C(T)}{K(T)} = \frac{C_*(T)}{K_*(T)}. \quad (4.8)$$

Во-вторых, используя эти соотношения, покажем, что при заданной в условии теоремы 2 функции $T(x, t)$ уравнение (3.3) выполняется не только при заданных $(C_*(T), K_*(T))$, но и при построенных $(C(T), K(T))$. Действительно, в тех точках области Q , в которых $T \in [a, \alpha]$ или $T \in [\beta, b]$, имеем

$$\begin{aligned} \Psi(x, t) &\equiv C(T) \frac{\partial T(x, t)}{\partial t} - K'(T) \left(\frac{\partial T(x, t)}{\partial x} \right)^2 - K(T) \frac{\partial^2 T(x, t)}{\partial x^2} = K(T) \left[\frac{C(T)}{K(T)} \frac{\partial T(x, t)}{\partial t} - \right. \\ &\quad \left. - \frac{K'(T)}{K(T)} \left(\frac{\partial T(x, t)}{\partial x} \right)^2 - \frac{\partial^2 T(x, t)}{\partial x^2} \right] = K(T) \left[\frac{C_*(T)}{K_*(T)} \frac{\partial T(x, t)}{\partial t} - \frac{K_*'(T)}{K_*(T)} \left(\frac{\partial T(x, t)}{\partial x} \right)^2 - \frac{\partial^2 T(x, t)}{\partial x^2} \right] = \\ &= \frac{K(T)}{K_*(T)} \left[C_*(T) \frac{\partial T(x, t)}{\partial t} - K_*'(T) \left(\frac{\partial T(x, t)}{\partial x} \right)^2 - K_*(T) \frac{\partial^2 T(x, t)}{\partial x^2} \right] = 0. \end{aligned}$$

Здесь использовались соотношения (4.8).

В тех же точках области Q , в которых $T \in [\alpha, \beta]$, имеем

$$\begin{aligned} \Psi(x, t) &\equiv C(T) \frac{\partial T(x, t)}{\partial t} - K'(T) \left(\frac{\partial T(x, t)}{\partial x} \right)^2 - K(T) \frac{\partial^2 T(x, t)}{\partial x^2} = \\ &= K(T) \left[\frac{C(T)}{K(T)} \frac{\partial T(x, t)}{\partial t} - \frac{K'(T)}{K(T)} \left(\frac{\partial T(x, t)}{\partial x} \right)^2 - \frac{\partial^2 T(x, t)}{\partial x^2} \right] = \\ &= (\text{используем условие (4.7)}) = \end{aligned}$$

$$\begin{aligned}
&= K(T) \left\{ \frac{C(T)}{K(T)} \frac{\partial T(x,t)}{\partial t} - \left[\frac{K'_*(T)}{K_*(T)} - \tilde{B}(T) \left(\frac{C_*(T)}{K_*(T)} - \frac{C(T)}{K(T)} \right) \right] \left(\frac{\partial T(x,t)}{\partial x} \right)^2 - \frac{\partial^2 T(x,t)}{\partial x^2} \right\} = \\
&= K(T) \left\{ \frac{C(T)}{K(T)} \left[\frac{\partial T(x,t)}{\partial t} - \tilde{B}(T) \left(\frac{\partial T(x,t)}{\partial x} \right)^2 \right] - \left[\frac{K'_*(T)}{K_*(T)} - \tilde{B}(T) \frac{C_*(T)}{K_*(T)} \right] \left(\frac{\partial T(x,t)}{\partial x} \right)^2 - \frac{\partial^2 T(x,t)}{\partial x^2} \right\} = \\
&= (\text{используем условие (4.5) в первой квадратной скобке}) = \\
&= K(T) \left\{ \frac{C_*(T)}{K_*(T)} \tilde{B}(T) \left(\frac{\partial T(x,t)}{\partial x} \right)^2 - \frac{K'_*(T)}{K_*(T)} \left(\frac{\partial T(x,t)}{\partial x} \right)^2 - \frac{\partial^2 T(x,t)}{\partial x^2} \right\} = \\
&= (\text{используем условие (4.5) в первом слагаемом}) = \\
&= K(T) \left\{ \frac{C_*(T)}{K_*(T)} \frac{\partial T(x,t)}{\partial t} - \frac{K'_*(T)}{K_*(T)} \left(\frac{\partial T(x,t)}{\partial x} \right)^2 - \frac{\partial^2 T(x,t)}{\partial x^2} \right\} = \\
&= \frac{K(T)}{K_*(T)} \left[C_*(T) \frac{\partial T(x,t)}{\partial t} - K'_*(T) \left(\frac{\partial T(x,t)}{\partial x} \right)^2 - K_*(T) \frac{\partial^2 T(x,t)}{\partial x^2} \right] = 0.
\end{aligned}$$

Таким образом, обратная задача, кроме функций $C_*(T)$ и $K_*(T)$, имеет другое, построенное решение $C(T)$ и $K(T)$, причем $\frac{C(T)}{K(T)} \neq \frac{C_*(T)}{K_*(T)}$.

Замечание 2. Пусть заданная функция $T(x, t)$ принадлежит области достижимости. Пусть на каждой τ -изолинии отрезка $\tau \in [a, b]$ функции $\frac{\partial T(x, t)}{\partial t}$ и $\left(\frac{\partial T(x, t)}{\partial x} \right)^2$, рассматриваемые как функции некоторого параметра s вдоль этой изолинии, являются линейно зависимыми и связаны соотношением (4.5). Тогда для всех $T \in [a, b]$ функция $\frac{\partial^2 T(x, t)}{\partial x^2}$ обязательно имеет вид

$$\frac{\partial^2 T(x, t)}{\partial x^2} = R(T) \left(\frac{\partial T(x, t)}{\partial x} \right)^2, \quad (4.9)$$

где $R(T) \in C([a, b])$ однозначно определяется полем $T(x, t)$. Действительно, используя в уравнении (3.3) выражение (4.5), можно получить

$$\frac{\partial^2 T(x, t)}{\partial x^2} = \left(\frac{C(T)}{K(T)} \tilde{B}(T) - \frac{K'(T)}{K(T)} \right) \left(\frac{\partial T(x, t)}{\partial x} \right)^2 = R(T) \left(\frac{\partial T(x, t)}{\partial x} \right)^2.$$

Так как функции $\tilde{B}(T)$ и $R(T)$ однозначно определяются заданным полем $T(x, t)$ из области достижимости, то $C(T)$ и $K(T)$ должны быть связаны следующим соотношением:

$$K'(T) + R(T)K(T) - C(T)\tilde{B}(T) = 0. \quad (4.10)$$

Следовательно, функции $C(T)$ и $K(T)$ могут быть любыми, удовлетворяющими уравнению (4.10). Задав произвольно $K(T) \in C^1([a, b])$, из (4.10) можно однозначно определить $C(T) \in C([a, b])$. И наоборот, задав $C(T) \in C([a, b])$ и значение $K(T)$ в некоторой точке отрезка $[a, b]$, можно однозначно определить $K(T) \in C^1([a, b])$ как решение задачи Коши. Именно с такой неоднозначностью и приходилось сталкиваться при идентификации $C(T)$ и $K(T)$ в первой группе расчетов.

Зададимся теперь обратным вопросом: какие поля температуры $T(x, t)$ из области достижимости удовлетворяют условиям теоремы 2? Как отмечено в замечании 2, если функции $\frac{\partial T(x, t)}{\partial t}$ и $\left(\frac{\partial T(x, t)}{\partial x} \right)^2$ являются линейно зависимыми вдоль каждой τ -изолинии отрезка $\tau \in [a, b]$, то спра-

ведливо равенство (4.9). С помощью функции $V(T) = \int_a^T R(\xi) d\xi$ равенство (4.9) можно преобразовать следующим образом:

$$\frac{\partial^2 T(x, t)}{\partial x^2} = R(T) \frac{\partial T(x, t)}{\partial x} \frac{\partial T(x, t)}{\partial x} = \frac{\partial T(x, t)}{\partial x} \frac{\partial V(T(x, t))}{\partial x}, \quad (x, t) \in Q,$$

$$\frac{\partial}{\partial x} \left(\ln \left| \frac{\partial T(x, t)}{\partial x} \right| - V(T(x, t)) \right) = 0, \quad \text{откуда} \quad \frac{\partial T(x, t)}{\partial x} = \rho_1(t) \exp(V(T(x, t))).$$

Если ввести новую функцию $G(T) = \int_a^T \exp(-V(\xi)) d\xi$, то последнее равенство может быть записано в виде

$$\exp(-V(T(x, t))) \frac{\partial T(x, t)}{\partial x} = \frac{\partial G(T(x, t))}{\partial x} = \rho_1(t),$$

интегрируя которое, получим $G(T(x, t)) = q(x, t) = \rho_1(t)x + \rho_2(t)$, где $\rho_1(t)$ и $\rho_2(t)$ — произвольные гладкие функции. Так как $G(T) \in C^2([a, b])$ и является монотонной, то существует функция $S(T) = G^{-1}(T) \in C^2([a, b])$, обратная к $G(T)$. Поэтому

$$T(x, t) = S(q) = S(\rho_1(t)x + \rho_2(t)). \quad (4.11)$$

По условию теоремы 2 функции $\frac{\partial T(x, t)}{\partial t}$ и $\left(\frac{\partial T(x, t)}{\partial x}\right)^2$ являются линейно зависимыми на каждой изотерме. Следовательно, они связаны соотношением (4.5). Вычислив производные функции $T(x, t)$ с помощью равенства (4.11) и подставив в соотношение (4.5), получим

$$S'(q)(\rho_1'(t)x + \rho_2'(t)) = \tilde{B}(T)(S'(q))^2 \rho_1^2(t) \quad \text{или} \quad \rho_1'(t)x + \rho_2'(t) = \tilde{B}(S(q))S'(q)\rho_1^2(t). \quad (4.12)$$

Так как левая часть равенства (4.12) представляет собой линейную функцию переменной x , то функция $\tilde{B}(S(q))S'(q)$ должна быть линейной по переменной q , т.е. иметь вид $\tilde{B}(S(q))S'(q) = \omega_1 q + \omega_2$, где ω_1 и ω_2 — произвольные постоянные. Используя полученное представление функции $\tilde{B}(S(q))S'(q)$ в равенстве (4.12), найдем

$$\rho_1'(t)x + \rho_2'(t) = [\omega_1(\rho_1(t)x + \rho_2(t)) + \omega_2]\rho_1^2(t). \quad (4.13)$$

Равенство (4.13) справедливо для всех $(x, t) \in Q$, если неизвестные функции удовлетворяют системе уравнений

$$\rho_1'(t) = \omega_1 \rho_1^3(t), \quad \rho_2'(t) = \rho_1^2(t)(\omega_1 \rho_2(t) + \omega_2). \quad (4.14)$$

При $\omega_1 = 0$ система уравнений (4.14) имеет следующее решение:

$$\rho_1(t) = \lambda_1, \quad \rho_2(t) = \omega_2 \lambda_1^2 t + \lambda_2, \quad (4.15)$$

а при $\omega_1 \neq 0$

$$\rho_1(t) = \frac{1}{\sqrt{\sigma_1 - 2\omega_1 t}},$$

$$\rho_2(t) = \frac{\sigma_2}{\sqrt{\sigma_1 - 2\omega_1 t}} - \frac{\omega_2}{\omega_1}, \quad \sigma_1 - 2\omega_1 t > 0 \quad (4.16)$$

где $\lambda_1, \lambda_2, \sigma_1, \sigma_2$ — произвольные константы.

Таким образом, определенная выше функция $q(x, t) = \rho_1(t)x + \rho_2(t)$ может иметь только такой вид $q(x, t) = \varepsilon_1 x + \varepsilon_2 t + \varepsilon_3$, либо такой вид $q(x, t) = \frac{x + \varepsilon_3}{\sqrt{\varepsilon_1 - \varepsilon_2 t}} + \varepsilon_4$, $\varepsilon_1 - \varepsilon_2 t > 0$, где $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$ — произвольные постоянные, а поле $T(x, t)$ из области достижимости, при котором решение обратной задачи неединственно — вид $T(x, t) = S(q(x, t))$. Непосредственной проверкой нетрудно убедиться

ся в том, что это утверждение верно для любых констант $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$, лишь бы в области $(x, t) \in Q$ выполнялось условие $(\varepsilon_1 - \varepsilon_2 t) > 0$.

Замечание 3. Все примеры первой группы расчетов иллюстрируют полученное утверждение при $q(x, t) = \varepsilon_1 x + \varepsilon_2 t + \varepsilon_3$. Действительно, использовавшиеся там экспериментальные поля имели вид $T(x, t) = S(q(x, t)) = S(\varepsilon_1 x + \varepsilon_2 t + \varepsilon_3)$ и получалось неединственное решение. Что касается случая $q(x, t) = \frac{x + \varepsilon_3}{\sqrt{\varepsilon_1 - \varepsilon_2 t}} + \varepsilon_4$, то, как подтвердили проведенные расчеты, при экспериментальном поле $T(x, t) = \frac{x^2}{t} = S(q(x, t)) = q^2(x, t) = \left(\frac{x}{\sqrt{t}}\right)^2$ обратная задача имеет неединственное решение.

ЗАКЛЮЧЕНИЕ

В работе рассмотрена одна из возможных постановок обратной коэффициентной задачи, позволяющая по результатам наблюдения за динамикой температурного поля одновременно идентифицировать зависящие от температуры объемную теплоемкость вещества и его коэффициент теплопроводности. Отмечено, что в такой постановке искомые коэффициенты могут быть найдены с точностью до одного и того же множителя. Предложен численный алгоритм решения этой задачи. Результатом решения обратной задачи являются как искомые параметры вещества, так и температурное поле, в общем случае, не совпадающее с экспериментально заданным полем. Найденное поле принадлежит области достижимости и является ближайшим к экспериментальному полю в норме пространства L_2 . Свойства именно этого поля из области достижимости определяют, будет ли решение обратной задачи единственно. Найден класс температурных полей из области достижимости, для которых обратная задача имеет неединственное решение.

СПИСОК ЛИТЕРАТУРЫ

1. *Зверев В.Г., Гольдин В.Д., Назаренко В.А.* Радиационно-кондуктивный теплоперенос в волокнистой термостойкой изоляции при тепловом воздействии // Теплофиз. высоких температур. 2008. Т. 46. № 1. С. 119–125.
2. *Алифанов О.М., Черепанов В.В.* Математическое моделирование высокопористых волокнистых материалов и определение их физических свойств // Теплофиз. высоких температур. 2009. Т. 47. № 3. С. 463–472.
3. *Алифанов О.М.* Обратные задачи теплообмена. М.: Машиностр., 1988.
4. *Yeung W.K., Lam T.T.* Second-order finite difference approximation for inverse determination of thermal conductivity // Internat. Journal of Heat Mass Transfer. 1996. V. 39. P. 3685–3693.
5. *Chen H.T., Lin J.Y., Wu C.H., Huang C.H.* Numerical algorithm for estimating temperature-dependent thermal conductivity // Numerical Heat Transfer. 1996. V. B29. P. 509–522.
6. *Самарский А.А., Вабищевич П.Н.* Вычислительная теплопередача. М.: Едиториал УРСС, 2003.
7. *Kim S., Kim M.C., Kim K.Y.* Non-iterative estimation of temperature dependent thermal conductivity without internal measurements // Internat. Journal of Heat Mass Transfer. 2003. V. 46. P. 1891–1810.
8. *Majchrzak E., Freus K., Freus S.* Identification of temperature dependent thermal conductivity using the gradient method // J. of Appl. Math. and Comput. Mech. 2006. V. 5. Issue 1. P. 114–123.
9. *Balazs Czel, Gyula Grof.* Inverse identification of temperature-dependent thermal conductivity via genetic algorithm with cost function-based rearrangement of genes // Internat. Journal of Heat and Mass Transfer. 2012. V. 55. № 15. P. 4254–4263.
10. *Matsevityi Yu.M., Alekhina S.V., Borukhov V.T., Zayats G.M., Kostikov A.O.* Identification of the Thermal Conductivity Coefficient for Quasi-Stationary Two-Dimensional Heat Conduction Equations // J. of Engng Phys. and Thermophys. 2017. V. 90. № 6. P. 1295–1301.
11. *Evtushenko Y., Zubov V., Albu A.* Inverse coefficient problems and fast automatic differentiation // J. of Inverse and Ill-Posed Problems. 2022. V. 30 (3). P. 447–460.
12. *Huang C.H., Yan J.Y.* An inverse problem in simultaneously measuring temperature-dependent thermal conductivity and heat capacity // Internat. Journal of Heat Mass Transfer. 1995. V. 38. P. 3433–3441.
13. *Imani A., Ranjbar A.A., Esmkhani M.* Simultaneous estimation of temperature-dependent thermal conductivity and heat capacity based on modified genetic algorithm // Inverse Problems in Sci. and Engng. 2006. V. 14. № 7. P. 767–783.

14. *Miao Cui, Kai Yang, Xiao-liang Xu, Sheng-dong Wang, Xiao-wei Gao.* A modified Levenberg-Marquardt algorithm for simultaneous estimation of multi-parameters of boundary heat flux by solving transient nonlinear inverse heat conduction problems // *Internat. Journal of Heat and Mass Transfer.* 2016. V. 97. P. 908–916.
15. *Евтушенко Ю.Г.* Оптимизация и быстрое автоматическое дифференцирование. М.: ВЦ им. А.А. Дородницына РАН, 2013. 144 с.
16. *Евтушенко Ю.Г., Zubov B.И.* Об обобщенной методологии быстрого автоматического дифференцирования // *Ж. вычисл. матем. и матем. физ.* 2016. Т. 56. № 11. С. 1847–1862.
17. *Евтушенко Ю.Г., Засухина Е.С., Zubov B.И.* О численном подходе к оптимизации решения задачи Бюргера с помощью граничных условий // *Ж. вычисл. матем. и матем. физ.* 1997. Т. 37. № 12. С. 1449–1458.
18. *Албу А.Ф., Zubov B.И.* Исследование задачи оптимального управления процессом кристаллизации вещества в новой постановке для объекта сложной геометрической формы // *Ж. вычисл. матем. и матем. физ.* 2014. Т. 54. № 12. С. 1879–1893.
19. *Евтушенко Ю.Г., Zubov B.И., Албу А.Ф.* Оптимальное управление тепловыми процессами с фазовыми переходами. Коллективная монография. М.: МАКС Пресс, 2021. С. 248.
<https://doi.org/10.29003/m2449.978-5-317-06677-2>
20. *Самарский А.А.* Теория разностных схем. М.: Наука, 1989.
21. *Албу А.Ф., Zubov B.И.* Об эффективности решения задач оптимального управления с помощью методологии быстрого автоматического дифференцирования // *Труды Института матем. и механ. УрО РАН.* 2015. Т. 21. № 4. С. 20–29.
22. *Hascoet L., Pascual V.* The Tapenade automatic differentiation tool: principles, model, and specification // *ACM Transact. on Math. Software (TOMS).* 2013. V. 39. № 3. P. 1–43.
23. *Hogan R.J.* Fast reverse-mode automatic differentiation using expression templates in C++ // *ACM Transact. on Math. Software (TOMS).* 2014. V. 40. № 4. P. 26–42.
24. *Горчаков А.Ю.* О программных пакетах быстрого автоматического дифференцирования // *Информационные технологии и вычисл. системы.* 2018. № 1. P. 30–36.
25. *Albu A., Gorchakov A., Zubov V.* On the effectiveness of the fast automatic differentiation methodology // *Commun. in Comput. and Informat. Sci.* 2019. V. 974. P. 264–276.
26. *Yixuan Qiu,* L-BFGS++, 2021. <https://github.com/yixuan/LBFGSpp/>

ОБЫКНОВЕННЫЕ ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ

УДК 517.911.5+517.927

О РЕШЕНИЯХ КРАЕВОЙ ЗАДАЧИ ДЛЯ ОДНОГО ДИФФЕРЕНЦИАЛЬНОГО УРАВНЕНИЯ ВТОРОГО ПОРЯДКА С ПАРАМЕТРОМ И РАЗРЫВНОЙ ПРАВОЙ ЧАСТЬЮ¹⁾

© 2023 г. О. В. Басков¹, Д. К. Потапов^{1,*}¹ 199034 Санкт-Петербург, Университетская наб., 7/9,
Санкт-Петербургский государственный университет, Россия

*e-mail: d.potapov@spbu.ru

Поступила в редакцию 25.02.2023 г.
Переработанный вариант 15.03.2023 г.
Принята к публикации 30.03.2023 г.

Рассматривается краевая задача для обыкновенного дифференциального уравнения второго порядка с параметром и разрывной правой частью. Устанавливаются теоремы о числе решений для исследуемой задачи. Полученные решения иллюстрируются графиками. Описывается процесс численного решения изучаемой краевой задачи. Библ. 13. Фиг. 9.

Ключевые слова: краевая задача, обыкновенное дифференциальное уравнение, разрывная правая часть, аналитическое решение, численное решение.

DOI: 10.31857/S0044466923080021, EDN: IAYWK

1. ВВЕДЕНИЕ. ПОСТАНОВКА ЗАДАЧИ

На протяжении ряда лет дифференциальные уравнения с разрывными правыми частями привлекают внимание многих исследователей. Такие уравнения представляют интерес как с теоретической точки зрения, так и с точки зрения приложений. Последние десять лет проблема существования решений обыкновенных дифференциальных уравнений второго порядка с разрывными правыми частями изучалась в [1–12]. Настоящая работа является продолжением этих исследований.

Рассмотрим модельную краевую задачу для обыкновенного дифференциального уравнения второго порядка с параметром и разрывной правой частью, в которой решения выписываются в явном виде. А именно,

$$-u'' = \lambda g(x, u(x)), \quad x \in (0, 1), \quad (1)$$

$$u(0) = u(1) = 0. \quad (2)$$

Здесь λ — положительный параметр, правая часть

$$g(x, u) = \begin{cases} a, & \text{если } u \leq 1, \\ b, & \text{если } u > 1, \end{cases}$$

имеет разрыв первого рода вдоль линии $u = 1$, a и b — вещественные числа.

Отметим, что если $u(x)$ — решение задачи (1), (2), то $v(x) = u(1 - x)$ также является решением этой задачи. Действительно, имеем

$$-v''(x) = -u''(1 - x) = \lambda g(1 - x, u(1 - x)) = \lambda g(x, v(x))$$

(последнее равенство справедливо в силу вида правой части), а также

$$v(0) = u(1) = 0, \quad v(1) = u(0) = 0.$$

¹⁾ Работа выполнена при финансовой поддержке РНФ (код проекта № 23-21-00069, <https://rscf.ru/project/23-21-00069/>).

Таким образом, решения краевой задачи (1), (2) либо симметричны относительно прямой $x = 1/2$ (т.е. $u(x) = u(1-x)$), либо имеются парные решения, получаемые отражением относительно данной прямой.

2. АНАЛИТИЧЕСКОЕ РЕШЕНИЕ ЗАДАЧИ

В зависимости от параметра λ решим искомую задачу аналитически при различных значениях a и b . Имеем нижеследующие теоремы 1–4 о числе решений краевой задачи (1), (2).

Теорема 1. Пусть $0 < a < b$. Тогда задача (1), (2) имеет одно решение при $\lambda \in (0, 8(2b-a)/b^2) \cup (8/a, +\infty)$, три решения при $\lambda \in (8(2b-a)/b^2, 8/a)$ и два решения при $\lambda = 8(2b-a)/b^2$ или $\lambda = 8/a$.

Доказательство. Сначала установим вид решения. Будем искать решение задачи (1), (2) через задачу Коши для уравнения (1) с начальными условиями $u(0) = 0$, $u'(0) = t$.

Пока $u(x) \leq 1$, уравнение (1) дает $-u'' = \lambda a$, и

$$u(x) = tx - \frac{1}{2}\lambda ax^2. \quad (3)$$

Если $t \leq 0$, то $u(1) = t - (1/2)\lambda a < 0$, и решений, удовлетворяющих условию $u(1) = 0$, нет. Поэтому $t > 0$.

Функция (3) достигает максимума при $x = t/(\lambda a)$, равного $u(t/(\lambda a)) = t^2/(2\lambda a)$. Если $t \leq \sqrt{2\lambda a}$, то этот максимум не превосходит единицы, а значит, решение $u(x) \leq 1$ при всех $x \in [0, 1]$. Таким образом, возможно существование решения в виде параболы (3).

Рассмотрим теперь случай $t > \sqrt{2\lambda a}$. Тогда существует x_1 такое, что $u(x_1) = 1$. Подставляя (3), получим квадратное уравнение $(1/2)\lambda ax_1^2 - tx_1 + 1 = 0$ с решениями $x_1 = (t \pm \sqrt{t^2 - 2\lambda a})/(\lambda a)$. Однако следует учесть, что после перехода решения через линию $u = 1$ правая часть $g(x, u)$ меняет значение, и решение перестает описываться формулой (3). Так что рассматривать следует лишь $x_1 = (t - \sqrt{t^2 - 2\lambda a})/(\lambda a)$. Из (3) найдем $u'(x_1) = t - \lambda ax_1 = \sqrt{t^2 - 2\lambda a} > 0$.

Пока $u(x) > 1$, уравнение (1) имеет вид $-u'' = \lambda b$, и решение описывается формулой

$$u(x) = u(x_1) + u'(x_1)(x - x_1) - \frac{1}{2}\lambda b(x - x_1)^2. \quad (4)$$

Поскольку мы ищем решения, удовлетворяющие условию $u(1) = 0$, то должна найтись точка $x_2 \in (x_1, 1)$, при которой решение вновь пересечет линию, на которой правая часть терпит разрыв, т.е. $u(x_2) = 1$. Полагая $x = x_2$ в (4) и учитывая $u(x_1) = u(x_2) = 1$, $x_2 > x_1$, находим $x_2 - x_1 = 2u'(x_1)/(\lambda b)$. А тогда $u'(x_2) = u'(x_1) - \lambda b(x_2 - x_1) = -u'(x_1) = -\sqrt{t^2 - 2\lambda a} < 0$.

Наконец, при $x \geq x_2$ вновь $u < 1$, и решение дается формулой

$$u(x) = u(x_2) + u'(x_2)(x - x_2) - \frac{1}{2}\lambda a(x - x_2)^2. \quad (5)$$

В силу $u(x_2) = 1$, $u'(x_2) < 0$ получим, что $u(x) \leq 1$ при $x \geq x_2$, так что больше решение не пересечет линию $u = 1$. Таким образом, возможно существование решения в виде объединения трех парабол (3)–(5).

Перейдем теперь к нахождению самих решений. Рассмотрим решения вида (3). Ограничение $u(1) = 0$ приводит к значению $t = (1/2)\lambda a$, так что решение описывается уравнением $u(x) = (1/2)\lambda ax(1-x)$. Оно достигает максимума при $x = 1/2$, равного $u(1/2) = (1/8)\lambda a$. Условие $t \leq \sqrt{2\lambda a}$ равносильно условию $u(1/2) \leq 1$, откуда можно получить значения параметра, при которых это решение существует: $\lambda \leq 8/a$.

Рассмотрим теперь решения, состоящие из трех парабол. Положим

$$u'(x) = \begin{cases} -\lambda ax + t, & 0 \leq x \leq x_1, \\ -\lambda bx + r, & x_1 \leq x \leq x_2, \\ -\lambda ax + s, & x_2 \leq x \leq 1. \end{cases}$$

Такой выбор $u'(x)$ удовлетворяет уравнению (1), если $u(x) > 1 \Leftrightarrow x_1 < x < x_2$. Обозначим $u'(1/2) = c$. Тогда $r = (1/2)\lambda b + c$. В точке $x = x_1$ должно быть выполнено равенство $-\lambda ax_1 + t = -\lambda bx_1 + r$, откуда $t = r - \lambda(b-a)x_1 = (1/2)\lambda b - \lambda(b-a)x_1 + c$. По аналогии, в точке $x = x_2$ должно быть $-\lambda bx_2 + r = -\lambda ax_2 + s$, откуда $s = (1/2)\lambda b - \lambda(b-a)x_2 + c$. Итак, учет непрерывности $u'(x)$ привел к выражению

$$u'(x) = \begin{cases} -\lambda ax - \lambda(b-a)x_1 + \frac{1}{2}\lambda b + c, & 0 \leq x \leq x_1, \\ -\lambda b\left(x - \frac{1}{2}\right) + c, & x_1 \leq x \leq x_2, \\ -\lambda ax - \lambda(b-a)x_2 + \frac{1}{2}\lambda b + c, & x_2 \leq x \leq 1. \end{cases}$$

Проинтегрируем его с учетом ограничений (2). Имеем

$$u(x) = \begin{cases} -\frac{1}{2}\lambda ax^2 - \lambda(b-a)x_1x + \frac{1}{2}\lambda bx + cx, & 0 \leq x \leq x_1, \\ -\frac{1}{2}\lambda b\left(x - \frac{1}{2}\right)^2 + c\left(x - \frac{1}{2}\right) + d, & x_1 \leq x \leq x_2, \\ \frac{1}{2}\lambda a(1-x^2) + \lambda(b-a)x_2(1-x) - \frac{1}{2}\lambda b(1-x) - c(1-x), & x_2 \leq x \leq 1. \end{cases}$$

Учтем, что $u(x_1) = u(x_2)$:

$$-\frac{1}{2}\lambda b\left(x_1 - \frac{1}{2}\right)^2 + c\left(x_1 - \frac{1}{2}\right) + d = -\frac{1}{2}\lambda b\left(x_2 - \frac{1}{2}\right)^2 + c\left(x_2 - \frac{1}{2}\right) + d,$$

откуда

$$-\frac{1}{2}\lambda b(x_2 - x_1)(x_1 + x_2 - 1) = c(x_2 - x_1),$$

и в силу $x_2 > x_1$ получаем $c = -(1/2)\lambda b(x_1 + x_2 - 1)$.

Теперь учтем непрерывность $u(x)$. При $x = x_1$ потребуем

$$-\frac{1}{2}\lambda ax_1^2 - \lambda(b-a)x_1^2 + \frac{1}{2}\lambda bx_1 + cx_1 = -\frac{1}{2}\lambda b\left(x_1 - \frac{1}{2}\right)^2 + c\left(x_1 - \frac{1}{2}\right) + d,$$

откуда

$$d = \frac{1}{8}\lambda b + \frac{1}{2}c - \frac{1}{2}\lambda(b-a)x_1^2.$$

При $x = x_2$ должно выполняться

$$-\frac{1}{2}\lambda b\left(x_2 - \frac{1}{2}\right)^2 + c\left(x_2 - \frac{1}{2}\right) + d = \frac{1}{2}\lambda a + \frac{1}{2}\lambda ax_2^2 - \lambda bx_2^2 - \lambda ax_2 - \frac{1}{2}\lambda b + \frac{3}{2}\lambda bx_2 - c + cx_2,$$

откуда

$$d = \frac{1}{2}\lambda a - \frac{3}{8}\lambda b - \frac{1}{2}c + \lambda(b-a)x_2 - \frac{1}{2}\lambda(b-a)x_2^2.$$

Приравнявая выражения для d , находим

$$\frac{1}{8}\lambda b + \frac{1}{2}c - \frac{1}{2}\lambda(b-a)x_1^2 = \frac{1}{2}\lambda a - \frac{3}{8}\lambda b - \frac{1}{2}c + \lambda(b-a)x_2 - \frac{1}{2}\lambda(b-a)x_2^2,$$

откуда

$$c = \frac{1}{2}\lambda(b-a)x_1^2 - \frac{1}{2}\lambda(b-a)(1-x_2)^2.$$

Приравнивание выражений для c дает

$$-\frac{1}{2}\lambda b(x_1 + x_2 - 1) = \frac{1}{2}\lambda(b-a)(x_1 - x_2 + 1)(x_1 + x_2 - 1),$$

и либо $x_1 + x_2 = 1$, либо $-(1/2)\lambda b = (1/2)\lambda(b-a)(x_1 - x_2 + 1)$, откуда $x_2 - x_1 = (2b-a)/(b-a) > 1$, что невозможно в силу $0 < x_1 < x_2 < 1$. Таким образом, $x_2 = 1 - x_1$. Отсюда следует $c = 0$.

Наконец, учтем требование $u(x_1) = 1$. Тогда $(1/2)\lambda ax_1^2 - \lambda bx_1^2 + (1/2)\lambda bx_1 = 1$, откуда

$$\begin{aligned} (2b-a)x_1^2 - bx_1 + \frac{2}{\lambda} &= 0, \\ x_1 &= \frac{b \pm \sqrt{b^2 - (8/\lambda)(2b-a)}}{2(2b-a)}. \end{aligned} \quad (6)$$

Эти корни имеются в случае $b^2 - (8/\lambda)(2b-a) \geq 0$ или $\lambda \geq 8(2b-a)/b^2$. Однако следует еще убедиться в выполнении условия $0 < x_1 < x_2 < 1$. Положительность x_1 очевидна. В силу $x_2 = 1 - x_1$ остается проверить, что $x_1 < 1/2$. Это условие приводит к неравенству

$$\pm \sqrt{b^2 - \frac{8}{\lambda}(2b-a)} < b-a,$$

справедливому для нижнего знака в силу условия $a < b$. Для верхнего знака получаем $b^2 - (8/\lambda)(2b-a) < b^2 - a(2b-a)$, откуда $\lambda < 8/a$. Отметим, что в силу положительности a ограничения на λ совместны: $8(2b-a)/b^2 < 8/a$.

Таким образом, при $0 < \lambda < 8(2b-a)/b^2$ задача (1), (2) имеет единственное решение

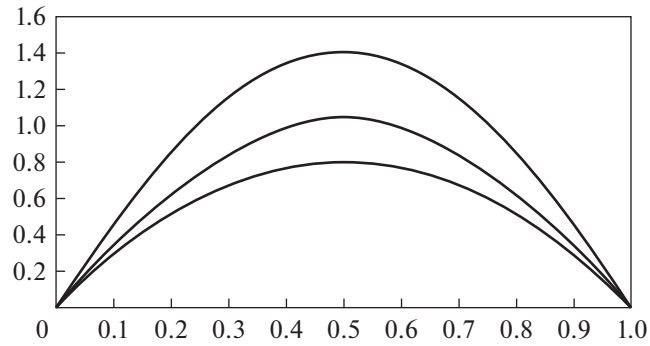
$$u(x) = \frac{1}{2}\lambda ax(1-x). \quad (7)$$

При $\lambda = 8(2b-a)/b^2$ помимо решения (7) появляется решение

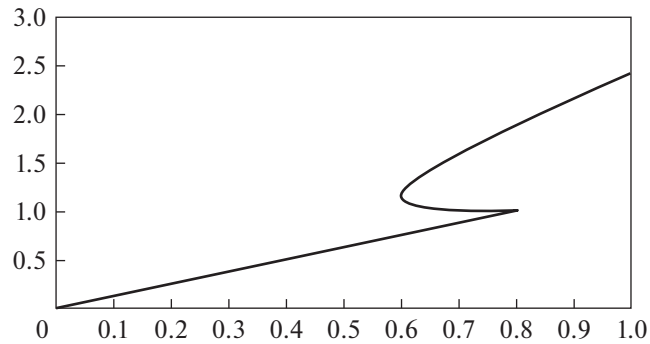
$$u(x) = \begin{cases} \frac{4}{b^2}x(b^2 - a(2b-a)x), & 0 \leq x \leq \frac{b}{2(2b-a)}, \\ \frac{4(2b-a)}{b}x(1-x) - \frac{b-a}{2b-a}, & \frac{b}{2(2b-a)} \leq x \leq \frac{3b-2a}{2(2b-a)}, \\ \frac{4}{b^2}(1-x)((b-a)^2 + a(2b-a)x), & 1 - \frac{b}{2(2b-a)} \leq x \leq 1. \end{cases}$$

При $8(2b-a)/b^2 < \lambda < 8/a$ существуют три решения, одно из которых дается формулой (7), а два других имеют вид

$$u(x) = \begin{cases} \frac{1}{2}\lambda x(b - 2(b-a)x_1 - ax), & 0 \leq x \leq x_1, \\ \frac{1}{2}\lambda bx(1-x) - \frac{1}{2}\lambda(b-a)x_1^2, & x_1 \leq x \leq x_2, \\ \frac{1}{2}\lambda(1-x)(b-a - 2(b-a)x_1 + ax), & x_2 \leq x \leq 1, \end{cases} \quad (8)$$



Фиг. 1. Графики решений $u(x)$ при $a = 1$, $b = 2$, $\lambda = 6.4$.



Фиг. 2. Зависимость $u(0.5)$ от λ при $a = 1$, $b = 2$.

где $x_2 = 1 - x_1$, а x_1 дается выражением (6). При $\lambda = 8/a$ остаются два решения $u(x) = 4x(1 - x)$ и

$$u(x) = \begin{cases} 4x \left(\frac{2b^2}{a(2b-a)} - 1 - x \right), & 0 \leq x \leq \frac{a}{2(2b-a)}, \\ \frac{4b}{a} x(1-x) - \frac{a(b-a)}{2b-a}, & \frac{a}{2(2b-a)} \leq x \leq \frac{4b-3a}{2(2b-a)}, \\ 4(1-x) \left(\frac{2(b-a)^2}{a(2b-a)} + x \right), & 1 - \frac{a}{2(2b-a)} \leq x \leq 1. \end{cases}$$

Наконец, при $\lambda > 8/a$ остается единственное решение (8), в котором $x_2 = 1 - x_1$ и

$$x_1 = \frac{b - \sqrt{b^2 - (8/\lambda)(2b-a)}}{2(2b-a)}.$$

Теорема 1 доказана.

Проиллюстрируем полученный результат. На фиг. 1 изображены графики трех решений краевой задачи (1), (2) при $a = 1$, $b = 2$, $\lambda = 6.4$. Нижний график соответствует решению, не проходящему через точки разрыва правой части, а два верхних графика проходят через область $u > 1$.

Каждое решение задачи (1), (2) при $a > 0$ достигает своего максимума при $x = 1/2$. На фиг. 2 изображена зависимость максимального значения $u(0.5)$ от параметра λ . По этому графику легко проследить изменение количества решений.

Теорема 2. Пусть $a = 0 < b$. Тогда задача (1), (2) имеет одно решение при $\lambda \in (0, 16/b)$, два решения при $\lambda = 16/b$ и три решения при $\lambda \in (16/b, +\infty)$.

Доказательство. Сначала заметим, что при $u \leq 1$ уравнение (1) принимает вид $u'' = 0$, и его решениями являются линейные функции. Существует единственная линейная функция, удовлетворяющая условиям (2): $u = 0$. Это тривиальное решение, существующее при всех значениях параметра $\lambda > 0$.

Теперь будем искать решения, пересекающие линию $u = 1$. Рассмотрим задачу Коши для уравнения (1) с начальными условиями $u(0) = 0$, $u'(0) = t$. В начале $u < 1$, уравнение (1) имеет вид $u'' = 0$, и решение $u(x) = tx$. Очевидно, что при $t \leq 0$ решение не достигнет значения $u = 1$. Так что $t > 0$. Потребуем существование $x_1 \in (0, 1)$ такого, что $u(x_1) = 1$. Тогда $t = 1/x_1$.

При $u > 1$ уравнение (1) приобретает вид $-u'' = \lambda b$, так что $u'(x) = u'(x_1) - \lambda b(x - x_1) = 1/x_1 - \lambda b(x - x_1)$, а

$$u(x) = 1 + \frac{1}{x_1}(x - x_1) - \frac{1}{2}\lambda b(x - x_1)^2. \quad (9)$$

Чтобы удовлетворить ограничению $u(1) = 0$, решение снова должно пересечь прямую $u = 1$, поэтому существует $x_2 \in (x_1, 1)$ такое, что $u(x_2) = 1$. Подстановка $x = x_2$ в (9) с учетом $x_2 > x_1$ дает

$$x_2 - x_1 = \frac{2}{\lambda b x_1}. \quad (10)$$

Отсюда $u'(x_2) = -u'(x_1) = -1/x_1$.

Далее $u < 1$ и $u'(x)$ сохраняет свое значение, так что больше решение не примет значения $u = 1$. Таким образом, решение должно иметь вид

$$u(x) = \begin{cases} \frac{x}{x_1}, & 0 \leq x \leq x_1, \\ -\frac{1}{2}\lambda b x^2 + cx + d, & x_1 \leq x \leq x_2, \\ \frac{1-x}{x_1}, & x_2 \leq x \leq 1. \end{cases}$$

Из условия непрерывности производной $u'(x_1) = 1/x_1 = -\lambda b x_1 + c$ находим $c = 1/x_1 + \lambda b x_1$. Из условия $u(x_1) = 1$ находим $d = 1 + (1/2)\lambda b x_1^2 - c x_1 = -(1/2)\lambda b x_1^2$. Далее, условие $u(x_2) = 1$ дает $(1 - x_2)/x_1 = 1$ или $x_1 = 1 - x_2$. С учетом (10) получаем $1 - 2x_1 = 2/(\lambda b x_1)$ или $2x_1^2 - x_1 + 2/(\lambda b) = 0$, откуда

$$x_1 = \frac{1 \pm \sqrt{1 - 16/(\lambda b)}}{4}. \quad (11)$$

Данное решение возможно при значениях параметра $\lambda \geq 16/b$. Отметим, что по этой формуле справедливы неравенства $0 < x_1 < 1/2$ и $0 < x_1 < x_2 < 1$.

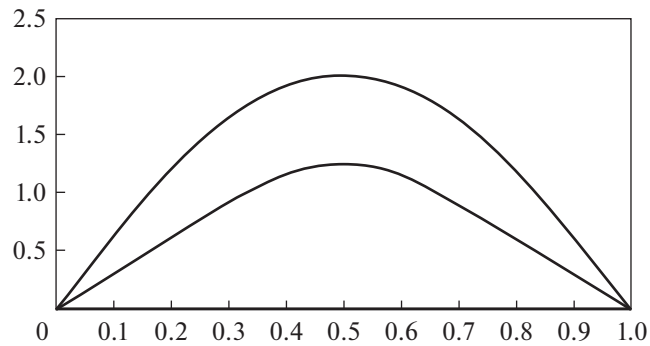
Таким образом, при $0 < \lambda < 16/b$ существует лишь тривиальное решение $u = 0$. При $\lambda = 16/b$ помимо тривиального появляется еще решение

$$u(x) = \begin{cases} 4x, & 0 \leq x \leq \frac{1}{4}, \\ 8x(1-x) - \frac{1}{2}, & \frac{1}{4} \leq x \leq \frac{3}{4}, \\ 4(1-x), & \frac{3}{4} \leq x \leq 1. \end{cases}$$

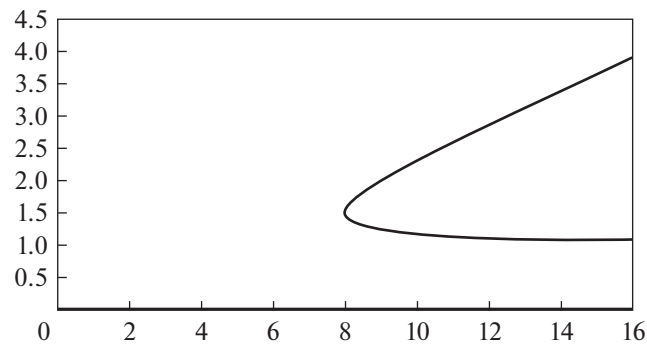
Наконец, при $\lambda > 16/b$ имеем три решения, одно из которых тривиальное $u = 0$, а два других даются формулой

$$u(x) = \begin{cases} \frac{x}{x_1}, & 0 \leq x \leq x_1, \\ \frac{1}{2}\lambda b x(1-x) - \frac{1}{2}\lambda b x_1^2, & x_1 \leq x \leq x_2, \\ \frac{1-x}{x_1}, & x_2 \leq x \leq 1, \end{cases}$$

в которой $x_2 = 1 - x_1$, а два значения x_1 находятся из (11). Теорема 2 доказана.



Фиг. 3. Графики решений $u(x)$ при $a = 0$, $b = 2$, $\lambda = 9$.



Фиг. 4. Зависимость $u(0.5)$ от λ при $a = 0$, $b = 2$.

Также проиллюстрируем полученные решения. На фиг. 3 изображены графики трех решений задачи (1), (2) при $a = 0$, $b = 2$, $\lambda = 9$. Нетривиальные решения достигают максимума при $x = 1/2$. На фиг. 4 изображена зависимость значений $u(0.5)$ от параметра λ , по которой можно проследить изменение количества решений.

Теорема 3. Пусть $a < 0 < b$. Тогда существует счетное множество бифуркационных значений параметра $\lambda_1, \lambda_2, \dots$ таких, что при $\lambda \in (0, \lambda_1)$ задача (1), (2) имеет единственное решение, при $\lambda \in (\lambda_k, \lambda_{k+1})$ существует $2k + 1$ решение, при $\lambda = \lambda_k$ имеется $2k$ решений в случае нечетного k и $2k - 1$ решение в случае четного k . Здесь $\lambda_{2k-1} = -8(k^2(b-a)^2 - b^2)/(ab^2)$, $\lambda_{2k} = -8k^2(b-a)^2/(ab^2)$, $k \geq 1$.

Доказательство. Сначала рассмотрим случай, когда $u \leq 1$. Уравнение (1) имеет вид $-u'' = \lambda a$, его решение $u(x) = (1/2)\lambda ax(1-x)$ удовлетворяет условиям (2) и притом $u \leq 0$. Таким образом, это решение существует при всех $\lambda > 0$.

Теперь рассмотрим случай, когда решение пересекает линию $u = 1$. Пусть $x_1, x_2, \dots, x_{2k}$ — все точки, при которых $u(x) = 1$, $u'(x) \neq 0$. Этих точек четное количество, поскольку для того, чтобы выполнить условия (2), решение должно перейти между областями $u \leq 1$ и $u > 1$ четное число раз.

Рассмотрим задачу Коши для уравнения (1) с начальными условиями $u(0) = 0$, $u'(0) = t$. Тогда при $0 < x < x_1$ будем иметь $u'(x) = t - \lambda ax$, так что $u'(x_1) = t - \lambda ax_1$. При этом $\int_0^{x_1} u'(x) dx = u(x_1) - u(0) = 1$, так что $tx_1 - (1/2)\lambda ax_1 = 1$, откуда $t = 1/x_1 + (1/2)\lambda ax_1$, а $u'(x_1) = 1/x_1 - (1/2)\lambda ax_1$.

При $x_{2i+1} < x < x_{2i+2}$ будем иметь $u'(x) = u'(x_{2i+1}) - \lambda b(x - x_{2i+1})$, причем в силу условий $u(x_{2i+2}) = u(x_{2i+1}) = 1$ должно выполняться равенство $\int_{x_{2i+1}}^{x_{2i+2}} u'(x) dx = 0$, откуда $u'(x_{2i+1})(x_{2i+2} - x_{2i+1}) -$

$-(1/2)\lambda b(x_{2i+2} - x_{2i+1})^2 = 0$, и в силу неравенства $x_{2i+1} < x_{2i+2}$ имеем $x_{2i+2} - x_{2i+1} = 2u'(x_{2i+1})/(\lambda b)$. При этом $u'(x_{2i+2}) = u'(x_{2i+1}) - \lambda b(x_{2i+2} - x_{2i+1}) = -u'(x_{2i+1})$.

При $x_{2i} < x < x_{2i+1}$ будем иметь $u'(x) = u'(x_{2i}) - \lambda a(x - x_{2i})$. Аналогично, $\int_{x_{2i}}^{x_{2i+1}} u'(x)dx = 0$, откуда $x_{2i+1} - x_{2i} = 2u'(x_{2i})/(\lambda a)$, и $u'(x_{2i+1}) = -u'(x_{2i})$.

Таким образом, в точках $x_1, x_2, \dots, x_{2k}$ значения производной имеют чередующиеся знаки и одинаковы по абсолютной величине: $u'(x_{2i+1}) = u'(x_1)$, $u'(x_{2i}) = -u'(x_1)$.

Наконец, рассмотрим промежуток $x_{2k} < x < 1$. Здесь $u'(x) = u'(x_{2k}) - \lambda a(x - x_{2k})$, и в силу условий $u(x_{2k}) = 1$, $u(1) = 0$ должно выполняться равенство $\int_{x_{2k}}^1 u'(x)dx = -1$ или $u'(x_{2k})(1 - x_{2k}) - (1/2)\lambda a(1 - x_{2k})^2 = -1$. Отсюда $u'(x_{2k}) = -1/(1 - x_{2k}) + (1/2)\lambda a(1 - x_{2k})$.

Из равенства $u'(x_{2k}) = -u'(x_1)$ получаем уравнение

$$\frac{1}{x_1} - \frac{1}{2}\lambda a x_1 = \frac{1}{1 - x_{2k}} - \frac{1}{2}\lambda a(1 - x_{2k})$$

или

$$\frac{1 - x_{2k} - x_1}{x_1(1 - x_{2k})} = -\frac{1}{2}\lambda a(1 - x_1 - x_{2k}).$$

Отсюда либо $x_{2k} = 1 - x_1$, либо $x_1(1 - x_{2k}) = 2/(\lambda a)$.

Теперь вспомним, что

$$x_{2i+1} - x_{2i} = \frac{2u'(x_{2i})}{\lambda a} = -\frac{2u'(x_1)}{\lambda a}, \quad x_{2i+2} - x_{2i+1} = \frac{2u'(x_{2i+1})}{\lambda b} = \frac{2u'(x_1)}{\lambda b}.$$

Тогда $1 = x_1 + (x_2 - x_1) + \dots + (x_{2k} - x_{2k-1}) + (1 - x_{2k}) = x_1 + 2cu'(x_1)/\lambda + 1 - x_{2k}$, где $c = k/b - (k-1)/a$. Отсюда $x_{2k} - x_1 = 2cu'(x_1)/\lambda = 2c/(\lambda x_1) - acx_1$.

Если $x_{2k} = 1 - x_1$, то $1 - 2x_1 = 2c/(\lambda x_1) - acx_1$, $(2 - ac)x_1^2 - x_1 + 2c/\lambda = 0$ и

$$x_1 = \frac{1 \pm \sqrt{1 - (8c/\lambda)(2 - ac)}}{2(2 - ac)}. \quad (12)$$

Данное решение возможно при условии

$$1 - \frac{8c}{\lambda}(2 - ac) \geq 0 \quad \text{или} \quad \lambda \geq 8c(2 - ac) = -\frac{8(k^2(b-a)^2 - b^2)}{ab^2}, \quad k \geq 1.$$

Если $x_1(1 - x_{2k}) = -2/(\lambda a)$, то

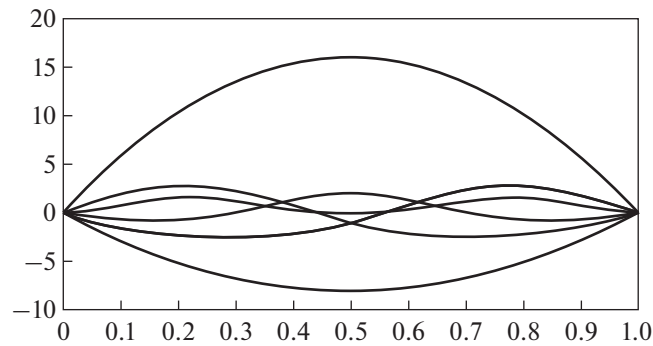
$$1 + \frac{2}{\lambda a x_1} - x_1 = \frac{2c}{\lambda x_1} - acx_1, \quad (1 - ac)x_1^2 - x_1 - \frac{2(1 - ac)}{\lambda a} = 0, \\ x_1 = \frac{1 \pm \sqrt{1 + 8/(\lambda a)(1 - ac)^2}}{2(1 - ac)}. \quad (13)$$

Данное решение существует при условии

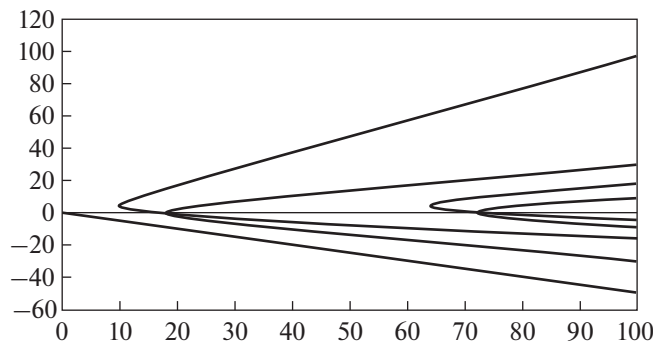
$$1 + \frac{8}{\lambda a}(1 - ac)^2 \geq 0, \quad \lambda \geq -\frac{8}{a}(1 - ac)^2 = -\frac{8k^2(b-a)^2}{ab^2}, \quad k \geq 1.$$

Следует выяснить, могут ли формулы (12) и (13) давать одинаковые значения x_1 . Это важно для количества решений, поскольку каждому x_1 однозначно соответствует свое значение $u'(0)$ и, таким образом, свое решение задачи $u(x)$. Равенство

$$\frac{1 \pm \sqrt{1 - (8c/\lambda)(2 - ac)}}{2(2 - ac)} = \frac{1 \pm \sqrt{1 + (8/(\lambda a))(1 - ac)^2}}{2(1 - ac)}$$



Фиг. 5. Графики решений $u(x)$ при $a = -1$, $b = 2$, $\lambda = 64$.



Фиг. 6. Зависимость $u'(0)$ от λ при $a = -1$, $b = 2$.

равносильно

$$\pm(1-ac)\sqrt{1-\frac{8c}{\lambda}(2-ac)} = 1 \pm (2-ac)\sqrt{1+\frac{8}{\lambda a}(1-ac)^2},$$

откуда следует

$$-\frac{8c}{\lambda}(1-ac)^2(2-ac) = \pm 2(2-ac)\sqrt{1+\frac{8}{\lambda a}(1-ac)^2} + 2(2-ac) + \frac{8}{\lambda a}(1-ac)^2(2-ac)^2,$$

$$-\frac{4c}{\lambda}(1-ac)^2 = \pm \sqrt{1+\frac{8}{\lambda a}(1-ac)^2} + 1 + \frac{4}{\lambda a}(1-ac)^2(2-ac),$$

$$-1 - \frac{8}{\lambda a}(1-ac)^2 = \pm \sqrt{1+\frac{8}{\lambda a}(1-ac)^2},$$

$$1 + \frac{16}{\lambda a}(1-ac)^2 + \frac{64}{\lambda^2 a^2}(1-ac)^4 = 1 + \frac{8}{\lambda a}(1-ac)^2,$$

$$\lambda = -\frac{8}{a}(1-ac)^2.$$

При таком значении $1 - (8c/\lambda)(2-ac) = 1/(1-ac)^2$, и при выборе верхнего знака в формуле (12) получим $x_1 = (1 + 1/(1-ac))/(2(2-ac)) = 1/(2(1-ac))$, что совпадает с формулой (13). Поскольку решение однозначно определяется значением $u'(0)$, которое выражено через x_1 , при $\lambda = -(8/a)(1-ac)^2$ решение, задаваемое формулой (13), совпадает с решением, вытекающим из формулы (12). Интересно также отметить, что при этом значении параметра $u'(0) = 1/x_1 + (1/2)\lambda a x_1 = 0$. Теорема 3 доказана.

Проиллюстрируем полученный результат. На фиг. 5 изображены шесть решений задачи (1), (2) при $a = -1$, $b = 2$, $\lambda = 64$. В данном случае значение $u(0.5)$ не обязательно является точкой максимума решений, и отследить количество решений по этому значению не удастся. Для этой цели на фиг. 6 приведен график зависимости $u'(0)$ от λ .

Теорема 4. Пусть $0 < b \leq a$. Тогда задача (1), (2) имеет единственное решение при всех значениях параметра $\lambda \in (0, +\infty)$.

Доказательство. Случай $a = b$ очевиден, поскольку правая часть непрерывна, и единственное решение $u(x) = (1/2)\lambda ax(1-x)$.

Рассмотрим случай $a > b$. Сначала будем искать решения, лежащие в области $u \leq 1$. Они имеют вид (7) и удовлетворяют условию $u \leq 1$ при $\lambda \leq 8/a$.

Теперь предположим, что решение заходит в область $u > 1$, другими словами, существует точка $x_1 \in (0, 1)$ такая, что $u(x_1) = 1$, $u'(x_1) > 0$. Тогда из-за краевого условия $u(1) = 0$ должна найтись точка $x_2 \in (x_1, 1)$ такая, что $u(x_2) = 1$, $u'(x_2) < 0$. Затем, в силу отрицательности второй производной, $u'(x)$ убывает и остается отрицательной, так что решение больше не примет значение $u = 1$. Поэтому оно состоит из трех парабол.

Положим

$$u'(x) = \begin{cases} -\lambda a(x - x_1) + s, & 0 \leq x \leq x_1, \\ -\lambda b(x - x_1) + s, & x_1 \leq x \leq x_2, \\ -\lambda a(x - x_2) - \lambda b(x_2 - x_1) + s, & x_2 \leq x \leq 1. \end{cases}$$

Здесь уже учтено требование непрерывности $u'(x)$. Интегрирование с учетом начального условия $u(0) = 0$ и непрерывности решения приводит к выражению

$$u(x) = \begin{cases} -\frac{\lambda}{2}ax^2 + \lambda ax_1x + sx, & 0 \leq x \leq x_1, \\ -\frac{\lambda}{2}b(x - x_1)^2 + sx + \frac{\lambda}{2}ax_1^2, & x_1 \leq x \leq x_2, \\ -\frac{\lambda}{2}a(x - x_2)^2 - \frac{\lambda}{2}b(x_2 - x_1)(2x - x_1 - x_2) + sx + \frac{\lambda}{2}ax_1^2, & x_2 \leq x \leq 1. \end{cases}$$

Учет краевого условия $u(1) = 0$ позволяет выразить

$$s = \frac{1}{2}\lambda a(1 - x_1 - x_2)(1 + x_1 - x_2) + \frac{1}{2}\lambda b(x_2 - x_1)(2 - x_1 - x_2).$$

В силу равенства $u(x_1) = u(x_2)$ имеем

$$sx_1 + \frac{1}{2}\lambda ax_1^2 = -\frac{1}{2}\lambda b(x_2 - x_1)^2 + sx_2 + \frac{1}{2}\lambda ax_1^2,$$

$$s(x_2 - x_1) = \frac{1}{2}\lambda b(x_2 - x_1)^2,$$

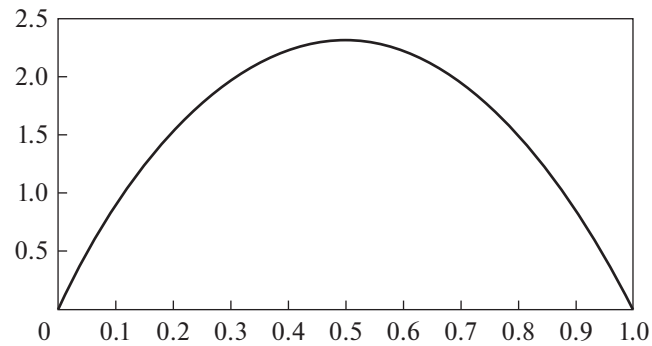
откуда с учетом $x_2 > x_1$ и выражения для s получаем

$$\frac{1}{2}\lambda a(1 - x_1 - x_2)(1 + x_1 - x_2) + \frac{1}{2}\lambda b(x_2 - x_1)(1 - x_1 - x_2) = 0.$$

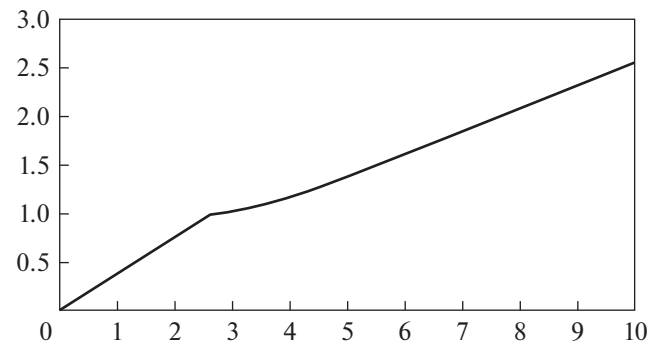
Тогда либо $x_1 + x_2 = 1$, либо

$$\frac{1}{2}\lambda a(1 + x_1 - x_2) + \frac{1}{2}\lambda b(x_2 - x_1) = 0,$$

откуда $x_2 - x_1 = a/(a - b) > 1$, что невозможно.



Фиг. 7. График решения $u(x)$ при $a = 3$, $b = 2$, $\lambda = 9$.



Фиг. 8. Зависимость $u(0.5)$ от λ при $a = 3$, $b = 2$.

Таким образом, $x_2 = 1 - x_1$ и $s = (1/2)\lambda b(1 - 2x_1)$. Остается рассмотреть условие $u(x_1) = 1$:

$$\frac{1}{2}\lambda b x_1(1 - 2x_1) + \frac{1}{2}\lambda a x_1^2 = 1,$$

$$(a - 2b)x_1^2 + b x_1 - \frac{2}{\lambda} = 0.$$

Если $a > 2b$, то

$$x_1 = \frac{-b + \sqrt{b^2 + (8/\lambda)(a - 2b)}}{2(a - 2b)},$$

второй корень не удовлетворяет условию $x_1 > 0$. В силу $0 < x_1 < x_2 < 1$ мы должны иметь $x_1 < 1/2$, откуда $\lambda > 8/a$.

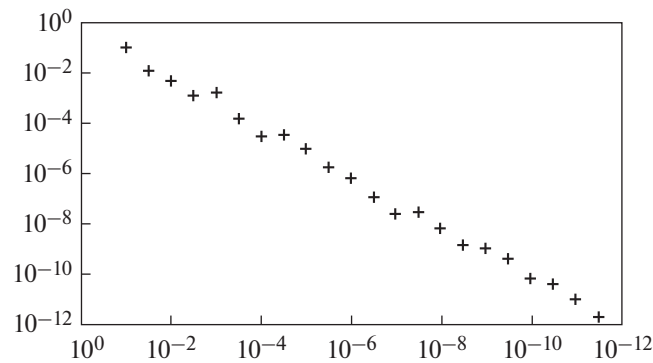
Если $a = 2b$, то $x_1 = 2/(\lambda b)$, и условие $x_1 < 1/2$ справедливо при $\lambda > 8/a$.

Если $b < a < 2b$, то

$$x_1 = \frac{b \pm \sqrt{b^2 - (8/\lambda)(2b - a)}}{2(2b - a)}.$$

Условие $x_1 < 1/2$ приводит к неравенству $\pm \sqrt{b^2 - (8/\lambda)(2b - a)} < b - a$, в котором правая часть отрицательна, поэтому верхний знак не подходит. Для нижнего знака условие выполняется при $\lambda > 8/a$.

Таким образом, при $a > b$ и $\lambda > 8/a$, имеется единственное значение x_1 , которое позволяет получить решение вида (8).

Фиг. 9. Зависимость ошибки от ϵ .

Объединяя рассмотренные случаи, получаем существование единственного решения при всех $\lambda > 0$. Теорема 4 доказана.

На фиг. 7 изображен график решения задачи (1), (2) при $a = 3$, $b = 2$, $\lambda = 9$. Решение достигает максимума при $x = 1/2$. Зависимость значений $u(0.5)$ от параметра λ приведена на фиг. 8.

Отметим, что при $a \in \mathbb{R}$ и $b \in (-\infty, 0]$ остаются лишь решения $u(x)$ вида (7), не выходящие из области $u \leq 1$ и существующие при $\lambda a \leq 8$.

Таким образом, рассмотрены все возможные соотношения между числами a и b .

3. ЧИСЛЕННОЕ РЕШЕНИЕ ЗАДАЧИ

Приближенное решение краевой задачи (1), (2) будем искать методом пристрелки (см. [13]). Для этого сначала выбираем наугад значения t_0 и t_1 — начальные приближения значения $u'(0)$. Решаем задачу Коши для уравнения $-u_i'' = \lambda g(x, u_i)$ с начальными условиями $u_i(0) = 0$, $u_i'(0) = t_i$, находим значения $u_i(1)$, $i = 0, 1$. Затем получаем следующее приближение:

$$t_{i+1} = t_i - \frac{t_i - t_{i-1}}{u_i(1) - u_{i-1}(1)} u_i(1), \quad i = 1, 2, \dots$$

Процесс продолжается, пока не будет выполнено неравенство $|u_i(1)| < \epsilon$, где $\epsilon > 0$ — выбранная желаемая точность.

Задачу Коши для уравнения (1) с начальными условиями $u(0) = 0$, $u'(0) = t$ решаем методом Эйлера. Выбираем количество промежутков N , шаг $h = 1/N$. Обозначим $x_i = ih$, u_i будут приближать значения $u(x_i)$, а u_i' соответственно $u'(x_i)$. Полагаем $u_0 = 0$, $u_0' = t$. На каждом шаге считываем $u_{i+1}' = u_i' - \lambda g(x_i, u_i)h$, $u_{i+1} = u_i + (1/2)(u_i' + u_{i+1}')h$.

За счет простой левой части уравнения (1) данный метод дает точное решение задачи Коши при условии, что точки, в которых правая часть испытывает скачок на решении, попадают на узлы сетки. Однако численный метод, как правило, используется тогда, когда нет точной информации об этих точках. Поэтому предлагается следующая корректировка.

Инициализация: $u_0 = 0$, $u_0' = t$.

Рассчитываем $u_{i+1}' = u_i' - \lambda g(x_i, u_i)h$, $u_{i+1} = u_i + (1/2)(u_i' + u_{i+1}')h$. Если выполнено условие $(u_i - 1)(u_{i+1} - 1) \geq 0$, то продолжаем. В противном случае проводим коррекцию. Будем искать на отрезке $[0, h]$ такой подшаг h_* , что при вычислении

$$u_*' = u_i' - \lambda g(x_i, u_i)h_*, \quad u_* = u_i + \frac{1}{2}(u_i' + u_*')h \quad (14)$$

получим значение $u_* = 1$. Для этого предлагается использовать бинарный поиск. Начинаем с $h_l = 0$, $h_r = h$, $u_l = u_i$, $u_r = u_{i+1}$. На каждой итерации выбираем среднее значение $h_c = (1/2)(h_l + h_r)$ и вычисляем соответствующие $u_c' = u_i' - \lambda g(x_i, u_i)h_c$, $u_c = u_i + (1/2)(u_i' + u_c')h$. Если $(u_i - 1)(u_c - 1) < 0$, то присваиваем $h_l = h_c$, $u_l = u_c$, иначе полагаем $h_r = h_c$, $u_r = u_c$. Процесс про-

должается до тех пор, пока значение подшага не будет найдено с заданной точностью: $|h_r - h_l| < \delta$, где $\delta > 0$. Пусть $h_* = (1/2)(h_l + h_r)$. Тогда вычисляем первый подшаг по формулам (14), а затем вычисляем второй подшаг $u'_{i+1} = u'_* - \lambda g(x_i, u_{i+1})(h - h_*)$, $u_{i+1} = u_* + (1/2)(u'_* + u'_{i+1})(h - h_*)$. Здесь в правую часть необходимо подставить вычисленное перед запуском коррекции значение u_{i+1} , чтобы получить ее значение после скачка.

Если h_* удается находить точно, то такой метод позволяет получать точное решение рассматриваемой задачи Коши.

На фиг. 9 в логарифмических осях приведены результаты численных расчетов зависимости ошибки $\max_i |u_i - u(x_i)|$ приближения точного решения $u(x)$ краевой задачи (1), (2), изображенного на верхнем графике на фиг. 1, значениями u_i , вычисленными по описанной схеме, от выбора ε при значениях $a = 1$, $b = 2$, $\lambda = 6.4$, $N = 1000$, $\delta = \varepsilon$. В данном случае ошибка имеет порядок $O(\varepsilon)$.

СПИСОК ЛИТЕРАТУРЫ

1. *Потанов Д.К.* Задача Штурма–Лиувилля с разрывной нелинейностью // Дифференц. ур-ния. 2014. Т. 50. № 9. С. 1284–1286.
2. *Kamachkin A.M., Potapov D.K., Yevstafyeva V.V.* Solution to second-order differential equations with discontinuous right-hand side // Electron. J. Differ. Equat. 2014. № 221. P. 1–6.
3. *Llibre J., Teixeira M.A.* Periodic solutions of discontinuous second order differential systems // J. Singularities. 2014. V. 10. P. 183–190.
4. *Потанов Д.К.* Существование решений, оценки дифференциального оператора и “разделяющее” множество в краевой задаче для дифференциального уравнения второго порядка с разрывной нелинейностью // Дифференц. ур-ния. 2015. Т. 51. № 7. С. 970–974.
5. *Самойленко А.М., Нижник И.Л.* Дифференциальные уравнения с биустойчивой нелинейностью // Укр. матем. ж. 2015. Т. 67. № 4. С. 517–554.
6. *Bonanno G., D’Agui G., Winkert P.* Sturm–Liouville equations involving discontinuous nonlinearities // Minimax Theory Appl. 2016. V. 1. № 1. P. 125–143.
7. *Kamachkin A.M., Potapov D.K., Yevstafyeva V.V.* Non-existence of periodic solutions to non-autonomous second-order differential equation with discontinuous nonlinearity // Electron. J. Differ. Equat. 2016. № 04. P. 1–8.
8. *Kamachkin A.M., Potapov D.K., Yevstafyeva V.V.* Existence of solutions for second-order differential equations with discontinuous right-hand side // Electron. J. Differ. Equat. 2016. № 124. P. 1–9.
9. *Bensid S., Diaz J.I.* Stability results for discontinuous nonlinear elliptic and parabolic problems with a S-shaped bifurcation branch of stationary solutions // Disc. Contin. Dyn. Syst. Ser. B. 2017. V. 22. № 5. P. 1757–1778.
10. *da Silva C.E.L., da Silva P.R., Jacquemard A.* Sliding solutions of second-order differential equations with discontinuous right-hand side // Math. Meth. Appl. Sci. 2017. V. 40. № 14. P. 5295–5306.
11. *Павленко В.Н., Постникова Е.Ю.* Задача Штурма–Лиувилля для уравнения с разрывной нелинейностью // Челяб. физ.-матем. ж. 2019. Т. 4. Вып. 2. С. 142–154.
12. *da Silva C.E.L., Jacquemard A., Teixeira M.A.* Periodic solutions of a class of non-autonomous discontinuous second-order differential equations // J. Dyn. Contr. Syst. 2020. V. 26. № 1. P. 17–44.
13. *Калиткин Н.Н.* Численные методы. М.: Наука, 1978.

УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

УДК 517.95

О НЕКОТОРЫХ ЭЛЛИПТИЧЕСКИХ КРАЕВЫХ ЗАДАЧАХ В КОНИЧЕСКИХ ОБЛАСТЯХ

© 2023 г. В. Б. Васильев^{1,*}

¹ 308015 Белгород, ул. Победы, 85, Белгородский государственный
национальный исследовательский университет, Россия

*e-mail: vbv57@inbox.ru

Поступила в редакцию 09.02.2023 г.
Переработанный вариант 25.03.2023 г.
Принята к публикации 28.03.2023 г.

Рассматривается модельное эллиптическое псевдодифференциальное уравнение в многогранном конусе и исследуется ситуация, когда некоторые параметры конуса стремятся к своим предельным значениям. В пространствах Соболева–Слободецкого решение уравнения в конусе строится при наличии специальной волновой факторизации эллиптического символа. Показано, что предельное решение краевой задачи с дополнительным интегральным условием может существовать только при дополнительных ограничениях на граничное условие. Библ. 15.

Ключевые слова: псевдодифференциальное уравнение, эллиптический символ, конус, волновая факторизация, область с разрезом.

DOI: 10.31857/S0044466923080161, EDN: WTRDTP

1. ВВЕДЕНИЕ

В теории псевдодифференциальных уравнений в негладких областях (областях с негладкой границей) модельные уравнения в канонических негладких областях (конусах) играют особую роль: однозначная разрешимость такого уравнения гарантирует фредгольмовость общего псевдодифференциального уравнения в области с соответствующей конической точкой на границе. Этот факт называется локальным принципом и в той или иной степени общности фигурирует во всех вариантах теорий псевдодифференциальных уравнений и связанных с ними краевых задач в негладких областях.

Теория псевдодифференциальных уравнений широко используется во многих разделах математики и физики, такие уравнения появляются, в частности, в задачах дифракции электромагнитных волн (см., например, [1–3]), где широко применяется метод факторизации. Мы также используем многомерный вариант этого метода для получения интегральных представлений решений рассматриваемых краевых задач.

В настоящей работе рассматриваются некоторые краевые задачи для модельных псевдодифференциальных уравнений, когда конус вырождается в конус меньшей размерности, другими словами, когда какие-то параметры исходного конуса стремятся к нулю. Основой этих исследований служат теория краевых задач для эллиптических псевдодифференциальных уравнений (см. [4]), теория одномерных сингулярных интегральных уравнений (см. [5–7]), многомерный комплексный анализ (см. [8]) и разработанный автором метод волновой факторизации, который привел к многочисленным результатам о разрешимости краевых задач для эллиптических псевдодифференциальных уравнений в канонических негладких областях евклидова пространства $\mathbb{R}^m$ (см. [9–13]).

2. ПСЕВДОДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ В КОНУСЕ

2.1. Модельные операторы в канонической области

Пусть $D \subset \mathbb{R}^m$ — область в m -мерном пространстве, функция определена на $\mathbb{R}^m$. Модельным псевдодифференциальным оператором A в области D мы называем оператор вида

$$(Au)(x) = \int_D \int_{\mathbb{R}^m} A(\xi) u(y) e^{i(y-x)\xi} d\xi dy, \quad x \in D,$$

функция $A(\xi)$ называется его символом. Здесь рассматривается класс символов, удовлетворяющих условию

$$c_1 \leq |A(\xi)(1 + |\xi|)^{-\alpha}| \leq c_2, \quad \xi \in \mathbb{R}^m.$$

Число $\alpha \in \mathbb{R}$ — это порядок псевдодифференциального оператора A .

Область $D \subset \mathbb{R}^m$ мы называем канонической областью, если это конус $C \subset \mathbb{R}^m$, не содержащий целой прямой в пространстве $\mathbb{R}^m$.

С конусами связаны важные области многомерного комплексного пространства и понятие специальной факторизации эллиптического символа, с помощью которого удастся описать картину разрешимости псевдодифференциального уравнения в конусе.

Определение 1. *Радиальной трубчатой областью над конусом C называется область в многомерном комплексном пространстве $\mathbb{C}^m M$ следующего вида:*

$$T(C) \equiv \{z \in \mathbb{C}^m : z = x + iy, x \in \mathbb{R}^m, y \in C\}.$$

Сопряженным конусом C^ называется такой конус, для всех точек которого выполняется условие*

$$x \cdot y > 0 \quad \forall y \in C,$$

$x \cdot y$ обозначает скалярное произведение x и y .

Всюду ниже мы будем предполагать, что символ $A(\xi)$ допускает волновую факторизацию относительно конуса C (см. [8], [14]) с индексом α :

$$A(\xi) = A_+(\xi) \cdot A_-(\xi),$$

где $A_+(\xi)$ допускает аналитическое продолжение в $T(-C^*)$ и $A_-(\xi)$ — в $T(C^*)$. Мы приводим ниже точное определение такой факторизации, поскольку величина индекса существенно влияет на картину разрешимости псевдодифференциального уравнения.

Определение 2. *Волновой факторизацией символа $A(\xi)$ относительно конуса C называется его представление в виде*

$$A(\xi) = A_+(\xi) A_-(\xi),$$

где сомножители $A_+(\xi)$, $A_-(\xi)$ должны удовлетворять следующим условиям:

- 1) $A_+(\xi)$, $A_-(\xi)$ определены для всех значений $\xi \in \mathbb{R}^m$, кроме, возможно, точек $\partial(C^* \cup (-C^*))$;
- 2) $A_+(\xi)$, $A_-(\xi)$ допускают аналитическое продолжение в радиальные трубчатые области $T(C^*)$, $T(-C^*)$ соответственно, для которых справедливы оценки

$$c'_1(1 + |\xi| + |\tau|)^{\alpha} \leq |A_+(\xi + i\tau)| \leq c_1(1 + |\xi| + |\tau|)^{\alpha},$$

$$c'_2(1 + |\xi| + |\tau|)^{\alpha-\alpha} \leq |A_-(\xi - i\tau)| \leq c_2(1 + |\xi| + |\tau|)^{\alpha-\alpha} \quad \forall \tau \in C^*.$$

Число $\alpha \in \mathbb{R}$ называется *индексом волновой факторизации*.

Класс символов, допускающих волновую факторизацию, достаточно богат. Отдельная глава книги [14] целиком посвящена этому вопросу, там приведено достаточно много примеров.

2.2. Конструкция решения

Мы рассмотрим здесь многогранный конус и уравнение

$$(Au)(x) = 0, \quad x \in \mathbb{R}^3 \setminus \overline{C_+^{ab}}, \quad (1)$$

в пространстве Соболева—Слободецкого $H^s(\mathbb{R}^3 \setminus \overline{C_+^{ab}})$, где

$$C_+^{ab} = \{x \in \mathbb{R}^3 : x = (x_1, x_2, x_3), x_3 < a|x_1| + b|x_2|, a, b > 0\}.$$

Мы рассмотрим специальный случай $s - \varkappa = 1 + \delta$, $|\delta| < 1/2$, для которого автором были получены ранее результаты о структуре решения уравнения (1) (см. [11–13]). Введем следующие одномерные сингулярные интегральные операторы (см. [5–7]):

$$(S_1 u)(\xi_1, \xi_2, \xi_3) = \text{v.p.} \frac{i}{2\pi} \int_{-\infty}^{+\infty} \frac{u(\tau, \xi_2, \xi_3) d\tau}{\xi_1 - \tau}, \quad (S_2 u)(\xi_1, \xi_2, \xi_3) = \text{v.p.} \frac{i}{2\pi} \int_{-\infty}^{+\infty} \frac{u(\xi_1, \eta, \xi_3) d\eta}{\xi_2 - \eta}.$$

В терминах этих операторов была получена формула

$$A_{\varkappa}(\xi) \tilde{u}(\xi) = \tilde{C}_1(\xi_1 - a\xi_3, \xi_2 - b\xi_3) + \tilde{C}_2(\xi_1 - a\xi_3, \xi_2 + b\xi_3) + \tilde{C}_3(\xi_1 + a\xi_3, \xi_2 - b\xi_3) + \tilde{C}_4(\xi_1 + a\xi_3, \xi_2 + b\xi_3), \quad (2)$$

где

$$\begin{aligned} \tilde{C}_1(\xi_1 - a\xi_3, \xi_2 - b\xi_3) &= \frac{1}{4} \tilde{c}_0(\xi_1 - a\xi_3, \xi_2 - b\xi_3) - \frac{1}{2} (S_1 \tilde{c}_0)(\xi_1 - a\xi_3, \xi_2 - b\xi_3) - \\ &\quad - \frac{1}{2} (S_2 \tilde{c}_0)(\xi_1 - a\xi_3, \xi_2 - b\xi_3) + (S_1 S_2 \tilde{c}_0)(\xi_1 - a\xi_3, \xi_2 - b\xi_3); \\ \tilde{C}_2(\xi_1 - a\xi_3, \xi_2 + b\xi_3) &= \frac{1}{4} \tilde{c}_0(\xi_1 - a\xi_3, \xi_2 + b\xi_3) - \frac{1}{2} (S_1 \tilde{c}_0)(\xi_1 - a\xi_3, \xi_2 + b\xi_3) + \\ &\quad + \frac{1}{2} (S_2 \tilde{c}_0)(\xi_1 - a\xi_3, \xi_2 + b\xi_3) - (S_1 S_2 \tilde{c}_0)(\xi_1 - a\xi_3, \xi_2 + b\xi_3); \\ \tilde{C}_3(\xi_1 + a\xi_3, \xi_2 - b\xi_3) &= \frac{1}{4} \tilde{c}_0(\xi_1 + a\xi_3, \xi_2 - b\xi_3) + \frac{1}{2} (S_1 \tilde{c}_0)(\xi_1 + a\xi_3, \xi_2 - b\xi_3) - \\ &\quad - \frac{1}{2} (S_2 \tilde{c}_0)(\xi_1 + a\xi_3, \xi_2 - b\xi_3) - (S_1 S_2 \tilde{c}_0)(\xi_1 + a\xi_3, \xi_2 - b\xi_3); \\ \tilde{C}_4(\xi_1 + a\xi_3, \xi_2 + b\xi_3) &= \frac{1}{4} \tilde{c}_0(\xi_1 + a\xi_3, \xi_2 + b\xi_3) + \frac{1}{2} (S_1 \tilde{c}_0)(\xi_1 + a\xi_3, \xi_2 + b\xi_3) + \\ &\quad + \frac{1}{2} (S_2 \tilde{c}_0)(\xi_1 + a\xi_3, \xi_2 + b\xi_3) + (S_1 S_2 \tilde{c}_0)(\xi_1 + a\xi_3, \xi_2 + b\xi_3), \end{aligned}$$

и $c_0(x_1, x_2)$ — произвольная функция из пространства $H^{s-\varkappa+1/2}(\mathbb{R}^2)$. Таким образом, ядро оператора A является одномерным подпространством пространства $H^s(\mathbb{R}^3 \setminus \overline{C_+^{ab}})$.

Для выделения единственного решения (однозначного определения функции $c_0(\xi_1, \xi_2)$) требуется какое-то дополнительное условие. Исходя из вида общего решения, наиболее удобным представляется задание сужения $\tilde{u}(\xi_1, \xi_2, 0)$ или, другими словами, задание интегрального условия

$$\int_{-\infty}^{+\infty} u(x_1, x_2, x_3) dx_3 \equiv g(x_1, x_2), \quad (3)$$

которое в образах Фурье выглядит следующим образом:

$$\tilde{u}(\xi_1, \xi_2, 0) = \tilde{g}(\xi_1, \xi_2), \quad (4)$$

$g(x_1, x_2)$ — известная функция.

Полагая $\xi_3 = 0$ в формуле (2), получим

$$\begin{aligned} \sum_{k=1}^4 \tilde{C}_k(\xi_1, \xi_2) &= \frac{1}{4} \tilde{c}_0(\xi_1, \xi_2) - \frac{1}{2} (S_1 \tilde{c}_0)(\xi_1, \xi_2) - \frac{1}{2} (S_2 \tilde{c}_0)(\xi_1, \xi_2) + (S_1 S_2 \tilde{c}_0)(\xi_1, \xi_2) + \\ &+ \frac{1}{4} \tilde{c}_0(\xi_1, \xi_2) - \frac{1}{2} (S_1 \tilde{c}_0)(\xi_1, \xi_2) + \frac{1}{2} (S_2 \tilde{c}_0)(\xi_1, \xi_2) - (S_1 S_2 \tilde{c}_0)(\xi_1, \xi_2) + \\ &+ \frac{1}{4} \tilde{c}_0(\xi_1, \xi_2) + \frac{1}{2} (S_1 \tilde{c}_0)(\xi_1, \xi_2) - \frac{1}{2} (S_2 \tilde{c}_0)(\xi_1, \xi_2) - (S_1 S_2 \tilde{c}_0)(\xi_1, \xi_2) + \\ &+ \frac{1}{4} \tilde{c}_0(\xi_1, \xi_2) + \frac{1}{2} (S_1 \tilde{c}_0)(\xi_1, \xi_2) + \frac{1}{2} (S_2 \tilde{c}_0)(\xi_1, \xi_2) + (S_1 S_2 \tilde{c}_0)(\xi_1, \xi_2) = \tilde{c}_0(\xi_1, \xi_2). \end{aligned}$$

С учетом условия (4) находим

$$\tilde{c}_0(\xi') = \tilde{A}_\#(\xi', 0) \tilde{g}(\xi'). \quad (5)$$

Итог наших вычислений содержится в следующем результате, который станет отправной точкой для дальнейших исследований (см. [11], [13]).

Теорема 1. Пусть $\varkappa - s = 1 + \delta$, $|\delta| < 1/2$, $g \in H^{s+1/2}(\mathbb{R}^2)$. Тогда единственное решение задачи (1), (3) дается формулой (2), где $c_0(x_1, x_2)$ определено формулой (5).

В конусе присутствуют два параметра a, b , и интерес представляют случаи, когда эти параметры устремляются к своим предельным значениям 0 или ∞ . Получающаяся при этом область будет областью с разрезом, и этот разрез представляет собой конус размерности меньшей, чем размерность пространства. Нас будет интересовать поведение решения задачи (1), (3) в этих предельных случаях. Как будет показано ниже, предельные решения могут существовать только при некоторых дополнительных требованиях к функции g .

Стоит отметить, что в двумерном случае имеется только один конус, и предельные решения рассмотрены в [15]. В многомерных ситуациях конусов гораздо больше, и, в частности, предельная ситуация $a \rightarrow 0$, $b = \text{const}$ и $a = \text{const}$, $b \rightarrow 0$ описана в [13]. Отметим также, что случай $a, b \rightarrow 0$ соответствует полупространству и полностью разобран в книге [4].

Мы разберем здесь некоторые предельные ситуации, опираясь на формулу (2), и покажем, какие условия следует наложить на функцию g для существования предельных решений.

Рассмотрим равенство (2). Используя замену переменных $\xi_1 - a\xi_3 = t_1$, $\xi_1 + a\xi_3 = t_3$ и находя $\xi_1 = (t_3 + t_1)/2$, $\xi_3 = (t_3 - t_1)/(2a)$, мы можем определить функцию $\tilde{c}_0$ в новых переменных t_1, ξ_2, t_3 с учетом условия (4). Переписывая формулу (2) в новых переменных t_1, ξ_2, t_3 , мы получим

$$\begin{aligned} A_\# \left(\frac{t_2 + t_1}{2}, \xi_2, \frac{t_3 - t_1}{2a} \right) \tilde{u} \left(\frac{t_2 + t_1}{2}, \xi_2, \frac{t_3 - t_1}{2a} \right) &= \tilde{C}_1 \left(t_1, \xi_2 - b \frac{t_3 - t_1}{2a} \right) + \\ &+ \tilde{C}_2 \left(t_1, \xi_2 + b \frac{t_3 - t_1}{2a} \right) + \tilde{C}_3 \left(t_3, \xi_2 - b \frac{t_3 - t_1}{2a} \right) + \tilde{C}_4 \left(t_3, \xi_2 + b \frac{t_3 - t_1}{2a} \right). \end{aligned} \quad (6)$$

Устремляя $a \rightarrow +\infty$, получаем

$$A_\# \left(\frac{t_2 + t_1}{2}, \xi_2, 0 \right) \tilde{u} \left(\frac{t_2 + t_1}{2}, \xi_2, 0 \right) = \tilde{C}_1(t_1, \xi_2) + \tilde{C}_2(t_1, \xi_2) + \tilde{C}_3(t_3, \xi_2) + \tilde{C}_4(t_3, \xi_2).$$

Дальше идут вычисления

$$\tilde{C}_1(t_1, \xi_2) + \tilde{C}_2(t_1, \xi_2) + \tilde{C}_3(t_3, \xi_2) + \tilde{C}_4(t_3, \xi_2) = \frac{\tilde{c}_0(t_1, \xi_2) + \tilde{c}_0(t_3, \xi_2)}{2} - (S_1 \tilde{c}_0)(t_1, \xi_2) + (S_1 \tilde{c}_0)(t_3, \xi_2).$$

Принимая во внимание условие (4), формулу (5) в равенстве (6) и введя новое обозначение

$$\tilde{A}_\#(\xi_1, \xi_2, 0) \tilde{g}(\xi_1, \xi_2) \equiv h(\xi_1, \xi_2),$$

приходим к следующему уравнению с параметром ξ_2 для функции g :

$$h \left(\frac{t_2 + t_1}{2}, \xi_2 \right) = \frac{h(t_1, \xi_2) + h(t_3, \xi_2)}{2} - (S_1 h)(t_1, \xi_2) + (S_1 h)(t_3, \xi_2). \quad (7)$$

Собирая проведенные выкладки, приходим к следующей формулировке результата.

Теорема 2. Если символ $A(\xi)$ допускает волновую факторизацию относительно конуса C_+^{ab} с индексом $\varkappa$ таким, что $\varkappa - s = 1 + \delta$, $|\delta| < 1/2$ для всех достаточно больших a , то единственное решение краевой задачи (1), (3) имеет предел при $a \rightarrow +\infty$ тогда и только тогда, когда функция в граничном условии $g \in H^{s+1/2}(R^2)$ является решением уравнения (7).

2.3. Двумерная ситуация

Мы опишем здесь, что происходит на плоскости. Уравнение (1) записывается как

$$(Au)(x) = 0, \quad x \in \mathbb{R}^2 \setminus \overline{C_+^a}, \quad (8)$$

и рассматривается в пространстве $H^s(\mathbb{R}^2 \setminus \overline{C_+^a})$. Число $\varkappa$ обозначает индекс волновой факторизации (как и выше, предполагается, что такая факторизация для символа $A(\xi)$ существует) относительно угла C_+^a (см. [14]), и предполагается выполненным условие $1/2 < \varkappa - s < 3/2$, угол C_+^a имеет вид

$$C_+^a = \{x \in \mathbb{R}^2 : x_2 > a|x_1|, a > 0\}.$$

Общее решение уравнения (8) в пространстве Соболева–Слободецкого $H^s(\mathbb{R}^2 \setminus \overline{C_+^a})$ имеет следующий вид (см. [11]):

$$\tilde{u}(\xi) = \frac{\tilde{c}_0(\xi_1 + a\xi_2) + \tilde{c}_0(\xi_1 - a\xi_2)}{2A_\#(\xi_1, \xi_2)} + A_\#^{-1}(\xi_1, \xi_2) \left(\text{v.p.} \frac{i}{2\pi} \int_{-\infty}^{+\infty} \frac{\tilde{c}_0(\eta) d\eta}{\xi_1 + a\xi_2 - \eta} - \text{v.p.} \frac{i}{2\pi} \int_{-\infty}^{+\infty} \frac{\tilde{c}_0(\eta) d\eta}{\xi_1 - a\xi_2 - \eta} \right),$$

где c_0 — произвольная функция из $H^{s-\varkappa+1/2}(\mathbb{R})$.

Вводим обозначения

$$\text{v.p.} \frac{i}{\pi} \int_{-\infty}^{+\infty} \frac{\tilde{c}_0(\eta) d\eta}{\xi_1 + a\xi_2 - \eta} \equiv \tilde{d}_0(\xi_1 + a\xi_2), \quad \text{v.p.} \frac{i}{\pi} \int_{-\infty}^{+\infty} \frac{\tilde{c}_0(\eta) d\eta}{\xi_1 - a\xi_2 - \eta} \equiv \tilde{d}_0(\xi_1 - a\xi_2) \quad (9)$$

и получаем

$$\tilde{u}(\xi_1, \xi_2) = \frac{\tilde{c}_0(\xi_1 + a\xi_2) + \tilde{c}_0(\xi_1 - a\xi_2) + \tilde{d}_0(\xi_1 + a\xi_2) - \tilde{d}_0(\xi_1 - a\xi_2)}{2A_\#(\xi_1, \xi_2)} \equiv \frac{\tilde{c}(\xi_1 + a\xi_2) + \tilde{d}(\xi_1 - a\xi_2)}{2A_\#(\xi_1, \xi_2)}, \quad (10)$$

где $\tilde{c}(\xi_1 + a\xi_2) \equiv \tilde{c}_0(\xi_1 + a\xi_2) + \tilde{d}_0(\xi_1 + a\xi_2)$, $\tilde{d}(\xi_1 - a\xi_2) \equiv \tilde{c}_0(\xi_1 - a\xi_2) - \tilde{d}_0(\xi_1 - a\xi_2)$.

Далее мы добавляем интегральное условие

$$\int_{-\infty}^{+\infty} u(x_1, x_2) dx_2 \equiv g(x_1), \quad (11)$$

которое в образах Фурье выглядит как $\tilde{u}(\xi_1, 0) = \tilde{g}(\xi_1)$. С учетом формулы (10) можно заключить, что

$$\frac{\tilde{c}_0(\xi_1)}{A_\#(\xi_1, 0)} = \tilde{g}(\xi_1).$$

Следовательно, мы можем определить функцию $\tilde{c}_0(\xi_1) = A_\#(\xi_1, 0)\tilde{g}(\xi_1)$. Тогда по формуле (9) мы можем определить $\tilde{d}_0(\xi_1)$ так, что формула (10) дает решение уравнения (8). Подытожив, заключаем, что решение уравнения (8) при условии (11) дается формулой

$$\begin{aligned} \tilde{u}(\xi_1, \xi_2) &= \frac{A_\#(\xi_1 + a\xi_2, 0)\tilde{g}(\xi_1 + a\xi_2) + A_\#(\xi_1 - a\xi_2, 0)\tilde{g}(\xi_1 - a\xi_2)}{2A_\#(\xi_1, \xi_2)} + \\ &+ \frac{1}{2A_\#(\xi_1, \xi_2)} \text{v.p.} \frac{i}{\pi} \int_{-\infty}^{+\infty} \frac{A_\#(\eta, 0)\tilde{g}(\eta) d\eta}{\xi_1 + a\xi_2 - \eta} - \frac{1}{2A_\#(\xi_1, \xi_2)} \text{v.p.} \frac{i}{\pi} \int_{-\infty}^{+\infty} \frac{A_\#(\eta, 0)\tilde{g}(\eta) d\eta}{\xi_1 - a\xi_2 - \eta} \end{aligned}$$

(подробности см. в [12]).

Теперь попытаемся выяснить, что произойдет с решением, когда $a \rightarrow \infty$, эта ситуация соответствует плоскости с разрезом по лучу.

Вводя новые переменные, обозначим

$$a_{\neq}(t_1, t_2) \equiv A_{\neq}\left(\frac{t_1 + t_2}{2}, \frac{t_1 - t_2}{2a}\right), \quad \tilde{U}(t_1, t_2) \equiv \tilde{u}\left(\frac{t_1 + t_2}{2}, \frac{t_1 - t_2}{2a}\right)$$

и запишем

$$\begin{aligned} \tilde{U}(t_1, t_2) = & \frac{A_{\neq}(t_1, 0)\tilde{g}(t_1) + A_{\neq}(t_2, 0)\tilde{g}(t_2)}{2a_{\neq}(t_1, t_2)} + \\ & + \frac{1}{2a_{\neq}(t_1, t_2)} \text{v.p.} \frac{i}{\pi} \int_{-\infty}^{+\infty} \frac{A_{\neq}(\eta, 0)\tilde{g}(\eta)d\eta}{t_1 - \eta} - \frac{1}{2a_{\neq}(t_1, t_2)} \text{v.p.} \frac{i}{\pi} \int_{-\infty}^{+\infty} \frac{A_{\neq}(\eta, 0)\tilde{g}(\eta)d\eta}{t_2 - \eta}. \end{aligned}$$

Устремляя $a \rightarrow +\infty$, вводя новые обозначения $A_{\neq}(t, 0)\tilde{g}(t) \equiv G(t)$, $\lim_{a \rightarrow +\infty} a_{\neq}(t_1, t_2) \equiv h(t_1, t_2)$, мы можем записать

$$\begin{aligned} \tilde{u}\left(\frac{t_1 + t_2}{2}, 0\right) = \tilde{U}(t_1, t_2) = & \frac{A_{\neq}(t_1, 0)\tilde{g}(t_1) + A_{\neq}(t_2, 0)\tilde{g}(t_2)}{2h(t_1, t_2)} + \\ & + \frac{1}{2h(t_1, t_2)} \text{v.p.} \frac{i}{\pi} \int_{-\infty}^{+\infty} \frac{A_{\neq}(\eta, 0)\tilde{g}(\eta)d\eta}{t_1 - \eta} - \frac{1}{2h(t_1, t_2)} \text{v.p.} \frac{i}{\pi} \int_{-\infty}^{+\infty} \frac{A_{\neq}(\eta, 0)\tilde{g}(\eta)d\eta}{t_2 - \eta}. \end{aligned} \quad (12)$$

С учетом условия (11) имеем

$$G\left(\frac{t_1 + t_2}{2}\right) = \frac{G(t_1) + G(t_2)}{2} + (SG)(t_1) - (SG)(t_2), \quad (13)$$

где

$$(SG)(t) = \text{v.p.} \frac{i}{2\pi} \int_{-\infty}^{+\infty} \frac{G(\eta)d\eta}{t - \eta}.$$

Итог наших рассуждений содержится в следующем результате.

Теорема 3. Если символ $A(\xi_1, \xi_2)$ допускает волновую факторизацию относительно конуса C_+^a для всех достаточно больших a , то предел решения (10) краевой задачи (8), (11) при $a \rightarrow +\infty$ существует тогда и только тогда, когда выполнено условие (13).

3. РАЗРЕЗЫ В МНОГОМЕРНОМ ПРОСТРАНСТВЕ

Опишем здесь несколько многомерных ситуаций, основываясь на материале предыдущих разделов. Точнее, мы покажем, какие многомерные области с разрезами могут быть получены аналогичным предельным переходом и приведем формулировки соответствующих постановок краевых задач.

Основная идея заключается в следующем. Пусть C_1 и C_2 — конусы в $\mathbb{R}^m$ и $\mathbb{R}^n$ соответственно, не содержащие целой прямой. Очевидно, что $C_1 \times C_2$ — это конус в пространстве $\mathbb{R}^{m+n}$, не содержащий целой прямой в $\mathbb{R}^{m+n}$. Тогда мы можем поставить краевую задачу, аналогичную (1), (5) в области $\mathbb{R}^{m+n} \setminus (C_1 \times C_2)$. Записывая формулу для решения этой задачи (при наличии волновой факторизации относительно “большого” конуса), мы можем рассматривать краевые задачи в областях с многомерными разрезами разной геометрии, устремляя к предельным значениям параметры конусов C_1 и C_2 .

Вот первый вариант такой краевой задачи:

$$(Au)(x) = 0, \quad x \in \mathbb{R}^5 \setminus \overline{(C_+^a \times C_+^{bd})}, \quad \int_{\mathbb{R}^2} u(x_1, x_2, x_3, x_4, x_5) dx_2 dx_5 = g(x_1, x_3, x_4), \quad (14)$$

где $C_+^a \subset \mathbb{R}^2$, $C_+^{bd} \subset \mathbb{R}^3$.

Второй вариант

$$(Au)(x) = 0, \quad x \in \mathbb{R}^4 \setminus \overline{(C_+^a \times C_+^b)}, \quad \int_{\mathbb{R}^2} u(x_1, x_2, x_3, x_4) dx_2 dx_4 = g(x_1, x_3), \quad (15)$$

где $C_+^a \subset \mathbb{R}^2$, $C_+^b \subset \mathbb{R}^2$.

Можно рассмотреть и такой вариант с двумя многогранными углами в $\mathbb{R}^6$, точнее, следующую краевую задачу:

$$(Au)(x) = 0, \quad x \in \mathbb{R}^6 \setminus \overline{(C_+^{ab} \times C_+^{dl})}, \quad \int_{\mathbb{R}^2} u(x_1, x_2, x_3, x_4, x_5, x_6) dx_3 dx_6 = g(x_1, x_2, x_4, x_5), \quad (16)$$

где $C_+^{ab} \subset \mathbb{R}^3$, $C_+^{dl} \subset \mathbb{R}^3$.

Наконец, можно вернуться к краевой задаче (14) и рассмотреть такую же краевую задачу. Разумеется, решение будет таким же, но предельный вариант рассмотрим другой: $a \rightarrow \infty$, $b \rightarrow \infty$, $d = \text{const}$, получая разрез другой геометрии. Более подробно мы рассмотрим эти ситуации в последующих публикациях.

СПИСОК ЛИТЕРАТУРЫ

1. Ильинский А.С., Смирнов Ю.Г. Дифракция электромагнитных волн на проводящих тонких экранах (Псевдодифференциальные операторы в задачах дифракции). М.: ИПРЖР, 1996.
2. Speck F.-O. From Sommerfeld diffraction problems to operator factorisation // Constr. Math. Anal. 2019. V. 2. № 4. P. 183–216.
3. Castro L., Duduchava R., Speck F.-O. Mixed impedance boundary value problems for the Laplace-Beltrami equation // J. Integral Equations Appl. 2020. V. 32. № 3. P. 275–292.
4. Эскин Г.И. Краевые задачи для эллиптических псевдодифференциальных уравнений. М.: Наука, 1973.
5. Гахов Ф.Д. Краевые задачи. М.: Наука, 1977.
6. Мухелишвили Н.И. Сингулярные интегральные уравнения. М.: Наука, 1968.
7. Milkhin S.G., Pröbldorf S. Singular Integral Operators. Akademie-Verlag, Berlin, 1986.
8. Владимиров В.С. Методы теории функций многих комплексных переменных. М.: Наука, 1964.
9. Vasilyev V.B. Pseudo-differential equations on manifolds with non-smooth boundaries. In: Pinelas S. editor. Differential and Difference Equations (and Applications. Springer Proc. Math. & Stat. **47**). Berlin: Springer, 2013. P. 625–637.
10. Vasilyev V.B. Elliptic equations, manifolds with non-smooth boundaries, and boundary value problems. In: Dang P., Ku M., Qian T., Rodino L. eds. New Trends in Analysis and Interdisciplinary Applications. Basel: Birkhäuser, 2017. P. 337–344.
11. Vasilyev V.B. Pseudo-differential equations, wave factorization, and related problems // Math. Meth. Appl. Sci. 2018. V. 41. № 18. P. 9252–9263.
12. Vasilyev V.B. Pseudo-differential equations and conical potentials: 2-dimensional case // Opusc. Math. 2019. V. 39. № 1. P. 109–124.
13. Vasilyev V.B. On certain 3-dimensional limit boundary value problems // Lobachevskii J. Math. 2020. V. 41. № 5. P. 917–925.
14. Vasil'ev V.B. Wave Factorization of Elliptic Symbols: Theory and Applications. Introduction to the Theory of Boundary Value Problems in Non-Smooth Domains. Dordrecht–Boston–London: Kluwer Acad. Publ., 2000.
15. Kutaiba Sh., Vasilyev V. On limit behavior of a solution to boundary value problem in a plane sector // Math. Meth. Appl. Sci. 2021. V. 44. № 15. P. 11904–11912.

**УРАВНЕНИЯ
В ЧАСТНЫХ ПРОИЗВОДНЫХ**

УДК 519.63

**APPLICATION OF IBSEF METHOD TO CHAFFEE-INFANTE EQUATION
IN $(1 + 1)$ AND $(2 + 1)$ DIMENSIONS¹⁾**

© 2023 г. U. Demirbilek^{1,*}, Kh. R. Mamedov^{2,**}

¹ Department of Mathematics, Mersin University, 33110, Mersin, Turkiye

² Department of Mathematics, Iğdir University, 76000, Iğdir, Turkiye

*e-mail: udemirbilek@mersin.edu.tr

**e-mail: hanlar.residoglu@igdir.edu.tr

Поступила в редакцию 04.04.2023 г.

Переработанный вариант 04.04.2023 г.

Принята к публикации 28.04.2023 г.

Применение улучшенного метода Бернулли к решению уравнения Чаффи–Инфанта для размерности $(1 + 1)$ и $(2 + 1)$. Рассмотрен модифицированный метод решения нестационарного уравнения Чаффи–Инфанта с помощью сведения к уравнению Бернулли в случае размерности $(1 + 1)$ и $(2 + 1)$. Используя представление решения в виде бегущей волны для некоторых частных случаев, удается получить явный вид решения.

Ключевые слова: модифицированный метод Бернулли, уравнение Чаффи–Инфанта, бегущая волна.

DOI: 10.31857/S0044466923080057, **EDN:** WSDAWP

¹⁾ Полный текст статьи печатается в английской версии журнала.

УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

УДК 519.635

О ЕДИНСТВЕННОСТИ РЕШЕНИЯ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ, К КОТОРЫМ РЕДУЦИРУЮТСЯ ОБРАТНЫЕ ЛИНЕЙНЫЕ ЗАДАЧИ ГРАВИМЕТРИИ И МАГНИТОМЕТРИИ: ЛОКАЛЬНЫЙ СЛУЧАЙ¹⁾

© 2023 г. И. И. Колотов¹, Д. В. Лукьяненко¹, И. Э. Степанова^{1,2,*}, А. Г. Ягола¹

¹ 119992 Москва, Ленинские горы, 1, МГУ им. М.В. Ломоносова, физический факультет, Россия

² 123995 Москва, ул. Б. Грузинская, 10, с. 1, Институт физики Земли им. О.Ю. Шмидта РАН, Россия

*e-mail: tet@ifz.ru

Поступила в редакцию 06.02.2023 г.
Переработанный вариант 19.03.2023 г.
Принята к публикации 28.04.2023 г.

Рассматриваются вопросы однозначной разрешимости систем линейных алгебраических уравнений, к которым редуцируются многие обратные задачи геофизики в результате дискретизации. Приводятся примеры вырожденных и невырожденных систем разных размерностей, возникающих при интерпретации гравиметрических и магнитометрических данных. Библ. 13. Фиг. 4.

Ключевые слова: вырожденные системы линейных алгебраических уравнений, интегральные представления, однозначная разрешимость.

DOI: 10.31857/S0044466923080094, EDN: WSNXOD

ВВЕДЕНИЕ

При интерпретации геофизических данных часто используются интегральные представления магнитного и гравитационного полей (см. [1–6]). Такого рода представления позволяют не только эффективно решать задачи по определению параметров геологической среды (магнитной восприимчивости, плотности породы, коэффициентов акустической жесткости и т.п.), но и строить аналитические продолжения полей, данные о которых получены с помощью дистанционного зондирования (см. [1]). Практически все постановки для описанного класса проблем редуцируются к системам линейных алгебраических уравнений в результате дискретизации.

1. ТЕОРЕМЫ ЕДИНСТВЕННОСТИ В СЛУЧАЕ ВОССТАНОВЛЕНИЯ МАГНИТНОЙ ИНДУКЦИИ

Приведем сначала выражения для магнитного поля (см. [2], [3]) в предположении, что среда является парамагнитной и измерения выполняются в некоторой декартовой системе координат. Если рассматриваются небольшие по площади полигоны (территории, на которых проводятся геофизические исследования), то можно считать поверхность Земли плоской, а саму планету – нижним полупространством. Постановки обратных задач при таких допущениях будем в дальнейшем называть локальными. Сферичность Земли обычно не учитывается, если линейный размер полигона не превышает 111 км, что соответствует приблизительно одному градусу по широте:

$$\vec{B}(\vec{r}_s) = \frac{\mu_0}{4\pi} \iiint_V K(\vec{r}_s, \vec{r}) M(\vec{r}) dV, \quad \vec{r} = (x, y, z); \quad \vec{r}_s = (x_s, y_s, z_s),$$

¹⁾ Работа выполнена при финансовой поддержке РНФ (проект № 23-41-00002).

$$K(\bar{r}_s, \bar{r}) = \frac{1}{\rho^5} \begin{bmatrix} 3(x - x_s)^2 - \rho^2 & 3(x - x_s)(y - y_s) & 3(x - x_s)(z - z_s) \\ 3(x - x_s)(y - y_s) & 3(y - y_s)^2 - \rho^2 & 3(y - y_s)(z - z_s) \\ 3(x - x_s)(z - z_s) & 3(y - y_s)(z - z_s) & 3(z - z_s)^2 - \rho^2 \end{bmatrix}, \quad (1.1)$$

$$\rho \equiv |\bar{r} - \bar{r}_s| = \sqrt{(x - x_s)^2 + (y - y_s)^2 + (z - z_s)^2}.$$

Здесь $\bar{r}_s$ — радиус-вектор точки наблюдения P_s в декартовой системе координат, $\bar{r}$ — радиус-вектор точки внутри носителя V , $\bar{B}(\bar{r}_s)$ — магнитная индукция в системе СИ, измеренная в точке, характеризующейся радиус-вектором $\bar{r}_s$, μ_0 — магнитная постоянная в системе СИ. Через $K(\bar{r}_s, \bar{r})$ обозначена матричная функция координат точки наблюдения и точки, принадлежащей носителю, $M(\bar{r})$ — вектор-функция, характеризующая намагниченность среды.

Поле, создаваемое одним магнитным диполем, выражается формулой

$$B(\bar{r}_s) = \frac{\mu_0}{4\pi} \frac{3(\bar{m}, \bar{r} - \bar{r}_s)(\bar{r} - \bar{r}_s) - \rho^2 \bar{m}}{\rho^5}. \quad (1.2)$$

Смысл обозначений в (1.2) такой же, как и в (1.1). Через $\bar{m} = (m_1, m_2, m_3)^T$ обозначен магнитный момент точечного диполя. Будем считать, что в точках наблюдения установлены условные “сенсоры”. Возникает вопрос об однозначной разрешимости следующей системы линейных алгебраических уравнений, возникающей, как уже отмечалось выше, в процессе дискретизации задачи (1.1):

$$\begin{aligned} \sum_{i=1}^N \left(\frac{3\{m_{ix}x_{ki} + m_{iy}y_{ki} + m_{iz}z_{ki}\}x_{ki}}{\rho_{ki}^5} - \frac{m_{ix}}{\rho_{ki}^3} \right) &= B_{kx}, \\ \sum_{i=1}^N \left(\frac{3\{m_{ix}x_{ki} + m_{iy}y_{ki} + m_{iz}z_{ki}\}y_{ki}}{\rho_{ki}^5} - \frac{m_{iy}}{\rho_{ki}^3} \right) &= B_{ky}, \\ \sum_{i=1}^N \left(\frac{3\{m_{ix}x_{ki} + m_{iy}y_{ki} + m_{iz}z_{ki}\}z_{ki}}{\rho_{ki}^5} - \frac{m_{iz}}{\rho_{ki}^3} \right) &= B_{kz}, \quad k = 1, \dots, N. \end{aligned} \quad (1.3)$$

В (1.3) через ρ_{ki} обозначено расстояние между k -м сенсором и i -м диполем:

$$\rho_{kl} = \sqrt{(x_k - x_l)^2 + (y_k - y_l)^2 + (z_k - z_l)^2},$$

$\bar{m}_i = (m_{ix}, m_{iy}, m_{iz})^T$ — магнитный момент i -го диполя в декартовой системе координат. Можно утверждать, что система (1.3) имеет блочный вид

$$\begin{bmatrix} K_{11} & K_{12} & \dots & K_{1N} \\ K_{21} & K_{22} & \dots & K_{2N} \\ \dots & \dots & \dots & \dots \\ K_{N1} & K_{N2} & \dots & K_{NN} \end{bmatrix} \begin{pmatrix} \bar{m}_1 \\ \bar{m}_2 \\ \dots \\ \bar{m}_N \end{pmatrix} = \begin{pmatrix} \bar{B}_1 \\ \bar{B}_2 \\ \dots \\ \bar{B}_N \end{pmatrix}, \quad (1.4)$$

где блок K_{il} соответствует i -му сенсору и l -му диполю, $\bar{m}_i$ — вектор компонент i -го диполя, $\bar{B}_l$ — вектор-функция, измеренная в l -м сенсоре. Блок K_{il} имеет следующий вид:

$$K_{il} = \begin{bmatrix} \frac{2x_{il}^2 - y_{il}^2 - z_{il}^2}{r_{il}^5} & \frac{3x_{il}y_{il}}{r_{il}^5} & \frac{3x_{il}z_{il}}{r_{il}^5} \\ \frac{3x_{il}y_{il}}{r_{il}^5} & \frac{2y_{il}^2 - x_{il}^2 - z_{il}^2}{r_{il}^5} & \frac{3y_{il}z_{il}}{r_{il}^5} \\ \frac{3x_{il}z_{il}}{r_{il}^5} & \frac{3y_{il}z_{il}}{r_{il}^5} & \frac{2z_{il}^2 - x_{il}^2 - y_{il}^2}{r_{il}^5} \end{bmatrix}, \quad x_{il} = x_i - x_l, \quad y_{il} = y_i - y_l, \quad z_{il} = z_i - z_l.$$

Рассмотрим сначала случай двух сенсоров и двух магнитных диполей.



Фиг. 1. Случай, в котором диполи и сенсоры не разделяют друг друга.

1.1. Два диполя и два сенсора на одной прямой

Пусть и сенсоры, и диполи располагаются на одной прямой (примем ее за ось X некоторой декартовой системы координат). Координаты диполей обозначим через x_1, x_2 , а координаты сенсоров — через x_1^s, x_2^s . Предположим сначала, что диполи и сенсоры не разделяют друг друга (см. фиг. 1). В начале координат расположен первый диполь.

Тогда система (1.4) принимает вид

$$A_{22}\bar{x} = \bar{B}_2, \quad (1.5)$$

$$A_{22} = \begin{bmatrix} \frac{2}{r_{11}^3} & 0 & 0 & \frac{2}{r_{12}^3} & 0 & 0 \\ 0 & -\frac{1}{r_{11}^3} & 0 & 0 & -\frac{1}{r_{12}^3} & 0 \\ 0 & 0 & -\frac{1}{r_{11}^3} & 0 & 0 & -\frac{1}{r_{12}^3} \\ \frac{2}{r_{21}^3} & 0 & 0 & \frac{2}{r_{22}^3} & 0 & 0 \\ 0 & -\frac{1}{r_{21}^3} & 0 & 0 & -\frac{1}{r_{22}^3} & 0 \\ 0 & 0 & -\frac{1}{r_{12}^3} & 0 & 0 & -\frac{1}{r_{22}^3} \end{bmatrix},$$

$$r_{11} = |x_1 - x_{1s}|, \quad r_{12} = |x_1 - x_{2s}|, \quad r_{21} = |x_2 - x_{1s}|, \quad r_{22} = |x_2 - x_{2s}|, \quad (1.6)$$

$$\bar{B}_2 = \begin{pmatrix} B_{1x} \\ B_{1y} \\ B_{1z} \\ B_{2x} \\ B_{2y} \\ B_{2z} \end{pmatrix}, \quad \bar{x} = \begin{pmatrix} m_{1x} \\ m_{1y} \\ m_{1z} \\ m_{2x} \\ m_{2y} \\ m_{2z} \end{pmatrix}.$$

Переобозначим элементы в матрице системы (6):

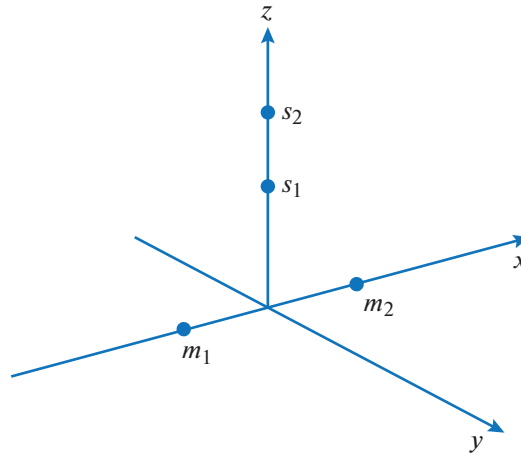
$$\tilde{A}_{22} = \begin{bmatrix} a_{11} & 0 & 0 & a_{12} & 0 & 0 \\ 0 & b_{11} & 0 & 0 & b_{12} & 0 \\ 0 & 0 & b_{11} & 0 & 0 & b_{12} \\ a_{21} & 0 & 0 & a_{22} & 0 & 0 \\ 0 & b_{21} & 0 & 0 & b_{22} & 0 \\ 0 & 0 & b_{21} & 0 & 0 & b_{22} \end{bmatrix}. \quad (1.7)$$

Для того чтобы детерминант матрицы (1.7) был равен нулю, необходимо и достаточно, чтобы координаты сенсоров совпадали (при различных координатах диполей):

$$\begin{aligned} \det A = 0 &\Rightarrow a_{11}a_{22} - a_{12}a_{21} = (b_{11}b_{22} - b_{12}b_{21})^2 = 0 \Rightarrow a_{ij} = -2b_{ij} \Rightarrow \\ &\Rightarrow \det A = 0 \Leftrightarrow b_{11}b_{22} - b_{12}b_{21} = 0 \Leftrightarrow (x_{1s} - x_1)(x_{2s} - x_2) = (x_{1s} - x_2)(x_{2s} - x_1) \Rightarrow \\ &\Rightarrow x_{1s}(x_1 - x_2) = x_{2s}(x_1 - x_2) \Rightarrow (x_{1s} - x_{2s})(x_1 - x_2) = 0. \end{aligned} \quad (1.8)$$



Фиг. 2. Случай, в котором диполи и сенсоры разделяют друг друга.



Фиг. 3. Случай взаимного расположения сенсоров и диполей в трехмерном пространстве.

Таким образом, нами доказана

Теорема 1. *Решение системы (1.5) единственно, если два сенсора и два диполя расположены в различных точках на одной и той же прямой. Диполи и сенсоры при этом не “разделяют друг друга”.*

Рассмотрим теперь случай, когда в начале координат расположен первый диполь, затем идет второй сенсор, потом – второй диполь, и последний – первый сенсор (см. фиг. 2). В этом случае для равенства нулю детерминанта матрицы получаем равенство $x_{1s}x_{2s} = 0.5x_2(x_{1s} + x_{2s})$.

Если координата второго диполя связана с координатами сенсоров данным соотношением, то в таком случае система уравнений вырождена, и компоненты двух диполей определить однозначно нельзя.

Замечание. Можно, безусловно, рассмотреть сразу общий случай расположения диполей. Детерминант обращается в нуль при выполнении либо условия $(x_{1s} - x_1)(x_{2s} - x_2) = (x_{1s} - x_2)(x_{2s} - x_1)$, что соответствует случаю “неразделенных” диполей, либо условия $(x_{1s} - x_1)(x_{2s} - x_2) = (x_{1s} - x_2)(x_1 - x_{2s})$, и в последнем случае система уравнений может быть вырождена.

1.2. Два сенсора и два диполя, расположенные в одной плоскости

Если провести через два диполя прямую и расположить два сенсора на какой-нибудь прямой в плоскости, перпендикулярной прямой, содержащей диполи и проходящей через середину соединяющего диполи отрезка (см. фиг. 3), то получаем систему из шести уравнений с шестью неизвестными ($N = 2$ в системе (1.3)):

$$\begin{aligned} \sum_{i=1}^2 \left(\frac{3\{m_{ix}x_{ki} + m_{iy}y_{ki} + m_{iz}z_{ki}\}x_{ki}}{\rho_{ki}^5} - \frac{m_{ix}}{\rho_{ki}^3} \right) &= B_{kx}, \\ \sum_{i=1}^2 \left(\frac{3\{m_{ix}x_{ki} + m_{iy}y_{ki} + m_{iz}z_{ki}\}y_{ki}}{\rho_{ki}^5} - \frac{m_{iy}}{\rho_{ki}^3} \right) &= B_{ky}, \\ \sum_{i=1}^2 \left(\frac{3\{m_{ix}x_{ki} + m_{iy}y_{ki} + m_{iz}z_{ki}\}z_{ki}}{\rho_{ki}^5} - \frac{m_{iz}}{\rho_{ki}^3} \right) &= B_{kz}, \quad k = 1, 2. \end{aligned} \quad (1.9)$$

Систему (1.9) можно переписать в блочном виде

$$\hat{K}\hat{m} = \hat{B}, \quad (1.10)$$

где $\hat{K} = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}$, $\hat{m} = \begin{pmatrix} \bar{m}_1 \\ \bar{m}_2 \end{pmatrix}$, $\hat{B} = \begin{pmatrix} \bar{B}_1 \\ \bar{B}_2 \end{pmatrix}$. Выражения для блоков матрицы (1.10) выглядят следующим образом:

$$\begin{aligned} K_{11} &= \begin{bmatrix} \frac{2x_{11}^2 - z_{11}^2}{r_{11}^5} & 0 & \frac{-3x_{11}z_{11}}{r_{11}^5} \\ 0 & -\frac{1}{r_{11}^3} & 0 \\ \frac{-3x_{11}z_{11}}{r_{11}^5} & 0 & \frac{2z_{11}^2 - x_{11}^2}{r_{11}^5} \end{bmatrix}, & K_{12} &= \begin{bmatrix} \frac{2x_{12}^2 - z_{12}^2}{r_{12}^5} & 0 & \frac{-3x_{12}z_{12}}{r_{12}^5} \\ 0 & -\frac{1}{r_{12}^3} & 0 \\ \frac{-3x_{12}z_{12}}{r_{12}^5} & 0 & \frac{2z_{12}^2 - x_{12}^2}{r_{12}^5} \end{bmatrix}, \\ K_{21} &= \begin{bmatrix} \frac{2x_{21}^2 - z_{21}^2}{r_{21}^5} & 0 & \frac{-3x_{21}z_{21}}{r_{21}^5} \\ 0 & -\frac{1}{r_{21}^3} & 0 \\ \frac{-3x_{21}z_{21}}{r_{21}^5} & 0 & \frac{2z_{21}^2 - x_{21}^2}{r_{21}^5} \end{bmatrix}, & K_{22} &= \begin{bmatrix} \frac{2x_{22}^2 - z_{22}^2}{r_{22}^5} & 0 & \frac{-3x_{22}z_{22}}{r_{22}^5} \\ 0 & -\frac{1}{r_{22}^3} & 0 \\ \frac{-3x_{22}z_{22}}{r_{22}^5} & 0 & \frac{2z_{22}^2 - x_{22}^2}{r_{22}^5} \end{bmatrix}. \end{aligned} \quad (1.11)$$

Расстояния от расположенного таким образом сенсора до каждого из двух диполей будут одинаковыми.

Если в (1.11) координаты сенсоров связаны соотношениями

$$\begin{aligned} y_{1i} &= \alpha z_{1i}; & y_{2i} &= \alpha z_{2i}, & i &= 1, 2, \\ \rho_{1i} &= \rho_{i2}, & i &= 1, 2, \end{aligned}$$

или

$$\begin{aligned} y_{1i} &= y_{2i} = 0, & i &= 1, 2, \\ \rho_{1i} &= \rho_{i2}, & i &= 1, 2, \end{aligned}$$

то получаем, что в матрице (1.11) две строки пропорциональны:

$$\begin{bmatrix} 0 & -\frac{1}{\rho_{11}^3} & 0; 0 & -\frac{1}{\rho_{12}^3} & 0 \\ 0 & -\frac{1}{\rho_{21}^3} & 0; 0 & -\frac{1}{\rho_{22}^3} & 0 \end{bmatrix}. \quad (1.12)$$

В (1.12) записи, соответствующие различным диполям, отделяются точкой с запятой. Аналогичный результат мы получим, если расположим два сенсора на одном расстоянии от прямой, соединяющей два диполя (при этом не требуется, чтобы расстояния от сенсора до каждого из двух диполей были одинаковыми). Сенсоры s_1 , s_2 и диполи m_1 , m_2 на фиг. 3 при этом меняются местами. Квазирешение в этом случае также определяется неоднозначно.

1.3. Два диполя и N сенсоров, расположенных на прямой, перпендикулярной отрезку, соединяющему диполи и проходящей через его середину

Предположим теперь, что измерения компонент магнитной индукции выполняются в точках прямой, проходящей через середину отрезка, соединяющего два диполя, и перпендикулярной этому отрезку (этот случай аналогичен описанному в предыдущем пункте). Система (1.3) принимает вид

$$\sum_{i=1}^2 \left(\frac{3\{m_{ix}x_{ki} + m_{iy}y_{ki} + m_{iz}z_{ki}\}x_{ki}}{\rho_{ki}^5} - \frac{m_{ix}}{\rho_{ki}^3} \right) = B_{kx},$$

$$\sum_{i=1}^2 \left(\frac{3\{m_{ix}x_{ki} + m_{iy}y_{ki} + m_{iz}z_{ki}\}y_{ki}}{\rho_{ki}^5} - \frac{m_{iy}}{\rho_{ki}^3} \right) = B_{ky}, \quad (1.13)$$

$$\sum_{i=1}^2 \left(\frac{3\{m_{ix}x_{ki} + m_{iy}y_{ki} + m_{iz}z_{ki}\}z_{ki}}{\rho_{ki}^5} - \frac{m_{iz}}{\rho_{ki}^3} \right) = B_{kz}, \quad k = 1, \dots, N.$$

Очевидно, (1.13) является переопределенной системой линейных алгебраических уравнений и, в общем случае, несовместна. Как известно (см. [7]), в конечномерном случае всегда существует квазирешение системы, полученной из исходной с помощью первой трансформации Гаусса:

$$Ax = f \rightarrow A^T Ax = A^T f. \quad (1.14)$$

Поскольку ранг $A^T A$ равен рангу исходной матрицы A , то при описанном выше расположении диполей и сенсоров мы получим систему линейных алгебраических уравнений с двумя пропорциональными друг другу строками:

$$\begin{bmatrix} 0 & -\frac{1}{\rho_{11}^3} & 0; 0 & -\frac{1}{\rho_{12}^3} & 0 \\ 0 & -\frac{1}{\rho_{21}^3} & 0; 0 & -\frac{1}{\rho_{22}^3} & 0 \end{bmatrix}, \quad \rho_{11} = \rho_{12}, \quad \rho_{21} = \rho_{22}. \quad (1.15)$$

Таким образом, нами доказана

Теорема 2. Если все сенсоры расположены на прямой, лежащей в плоскости симметрии двух неизвестных диполей $\vec{m}_1, \vec{m}_2$, то квазирешение системы (1.14) определяется неоднозначно. Однако нормальное квазирешение будет единственным.

1.4. Два диполя и N сенсоров, расположенных на одной прямой

Для данного случая верен аналог теоремы 1.

Теорема 3. Если два диполя и N сенсоров расположены на одной прямой, координаты двух каких-либо сенсоров не совпадают, и эти два сенсора не разделяют выделенные два диполя, то квазирешение системы (1.8) единственно, независимо от того, какие значения принимают координаты остальных $N - 2$ сенсоров.

Доказательство. Система уравнений, из которой в описанном случае определяются компоненты двух диполей, лежащих на одной прямой с сенсорами, имеет вид (1.13), первые два блока которой выражаются с помощью формулы (1.6). Ранг такой системы не может быть больше 6.

Рассмотрим два диполя и два сенсора, которые оба расположены либо правее, либо левее обоих диполей. Согласно теореме 1, компоненты диполей определяются однозначно, и, следовательно, остальные сенсоры не добавят информации о магнитном поле.

1.5. N диполей и N сенсоров в трехмерном пространстве

Если требуется определить компоненты N магнитных диполей $\vec{m}_1, \vec{m}_2, \dots, \vec{m}_N$ по измеренным в N произвольных точках трехмерного пространства R^3 трем компонентам вектора магнитной индукции (т.е. известным векторам $\vec{B}_1, \vec{B}_2, \dots, \vec{B}_N$), то исследование, аналогичное проведенному в п. 1.2, позволяет сделать вывод о том, что решение системы (1.3) будет неединственным, если для каких-либо двух индексов k и l справедливы следующие соотношения:

$$y_{ki} = y_{li} = 0, \quad i = 1, \dots, N. \quad (1.16)$$

$$\rho_{ki} = \rho_{li}, \quad i = 1, \dots, N.$$

Тогда две строки в (1.3) будут совпадать (в k -м и l -м блоках для соответствующих сенсоров):

$$\begin{bmatrix} 0 - \frac{1}{3}; & 0 - \frac{1}{3}; & \dots; & 0 - \frac{1}{3} & 0 \\ \rho_{k1} & \rho_{k2} & & \rho_{kN} & \\ 0 - \frac{1}{3}; & 0 - \frac{1}{3}; & \dots; & 0 - \frac{1}{3} & 0 \\ \rho_{l1} & \rho_{l2} & & \rho_{lN} & \end{bmatrix}. \quad (1.17)$$

Если в более общем случае выполняются соотношения

$$y_{ki} = \alpha z_{ki}; \quad y_{li} = \alpha z_{li}, \quad \text{и} \quad \rho_{ki} = \rho_{li}, \quad i = 1, \dots, N, \quad (1.18)$$

то умножая третьи строки блоков матрицы (1.3) на α и вычитая их из вторых строк соответственно для двух блоков с индексами k и l , мы получим аналог (1.17).

Таким образом, нами доказана

Теорема 4. При выполнении условий (1.16) или (1.18) решение системы (1.3) определяется неоднозначно.

Замечание. Для практических целей более важными являются необходимое и достаточное условия однозначной разрешимости системы (1.3), однако отсутствие пропорциональности каких-либо двух строк матрицы не означает, что система не вырождена.

Требуется дальнейшие исследования в этом направлении.

2. ТЕОРЕМЫ ЕДИНСТВЕННОСТИ В МЕТОДЕ S-АППРОКСИМАЦИЙ. ЛОКАЛЬНЫЙ ВАРИАНТ

Если известны компоненты магнитного или гравитационного полей (например, первая производная потенциала по z на некотором рельефе), то можно представить потенциал поля в виде суммы простого и двойного слоев, создаваемых совокупностью горизонтальных плоскостей, расположенных ниже заданного рельефа. Если систему координат выбрать так, чтобы дневная поверхность (поверхность Земли в локальном варианте метода S-аппроксимаций, см. [4], [5], [9–13]) задавалась уравнением $z = 0$, то потенциал представляется в виде (см. [8])

$$V(M) = \sum_{l=1}^L \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{\rho_{1,l}(\xi_1, \xi_2) d\xi_1 d\xi_2}{\sqrt{(x - \xi_1)^2 + (y - \xi_2)^2 + (z - H_l)^2}} + \sum_{l=1}^L \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{\rho_{2,l}(\xi_1, \xi_2)(z - H_l) d\xi_1 d\xi_2}{[\sqrt{(x - \xi_1)^2 + (y - \xi_2)^2 + (z - H_l)^2}]^3}, \quad (2.1)$$

$$M = (x, y, z), \quad \hat{x} = (x, y), \quad \hat{\xi} = (\xi_1, \xi_2, \xi_3), \quad \hat{\xi} = (\xi_1, \xi_2).$$

Тогда производная по z потенциала V , взятая с обратным знаком, будет иметь вид

$$\begin{aligned} -\frac{\partial V}{\partial z}(M) &= \sum_{l=1}^L \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{\rho_{1,l}(\xi) (z - H_l) z d\xi_1 d\xi_2}{[\sqrt{(x - \xi_1)^2 + (y - \xi_2)^2 + (z - H_l)^2}]^3} + \\ &+ \sum_{l=1}^L \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{\rho_{2,l}(\xi) (2(z - H_l)^2 - (x - \xi_1)^2 - (y - \xi_2)^2) d\xi_1 d\xi_2}{[\sqrt{(x - \xi_1)^2 + (y - \xi_2)^2 + (z - H_l)^2}]^5}, \quad M = (x, y, z). \end{aligned} \quad (2.2)$$

Функции $\rho_{1,l}$, $\rho_{2,l}$ неизвестны. Пусть компоненты поля заданы в конечном множестве точек M_i , $M_i = (x_i, y_i, z_i)$, $i = 1, \dots, N$. Обозначим подынтегральную функцию в первом слагаемом для l -го слоя в (2.2) в точке M_i через $Q_{1,l}^{(i)}$, а во втором слагаемом — через $Q_{2,l}^{(i)}$. Тогда получим

$$-\frac{\partial V(M_i)}{\partial z} \equiv f_i = \sum_{l=1}^L \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (\rho_{1,l}(\hat{\xi}) Q_{1,l}^{(i)}(\hat{\xi}) + \rho_{2,l}(\hat{\xi}) Q_{2,l}^{(i)}(\hat{\xi})) d\hat{\xi}, \quad i = 1, \dots, N. \quad (2.3)$$

На практике компоненты поля бывают заданы с некоторой погрешностью δ , поэтому входной информацией являются значения $f_{i,\delta}$. С помощью решения вариационной задачи

$$\Omega(\rho) = \sum_{l=1}^L \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (\rho_{1,l}^2(\hat{\xi}) + \rho_{2,l}^2(\hat{\xi})) d\hat{\xi} = \min_{\rho}, \quad (2.4)$$

$$f_{i,\delta} - \sum_{l=1}^L \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (\rho_{1,l}(\hat{\xi}) Q_{1,l}^{(i)}(\hat{\xi}) + \rho_{2,l}(\hat{\xi}) Q_{2,l}^{(i)}(\hat{\xi})) d\hat{\xi} = 0, \quad i = 1, \dots, N, \quad (2.5)$$

получим, что искомые функции должны иметь вид (см. [4], [5], [9])

$$\begin{aligned} \rho_{1,l}^{(a)}(\hat{\xi}) &= \tilde{\rho}_{1,l}(\hat{\xi}, \lambda), \quad \rho_{2,l}^{(a)}(\hat{\xi}) = \tilde{\rho}_{2,l}(\hat{\xi}, \lambda), \\ \tilde{\rho}_{1,l}(\hat{\xi}, \lambda) &= \sum_{i=1}^N \lambda_i Q_{1,l}^{(i)}(\hat{\xi}), \quad \tilde{\rho}_{2,l}(\hat{\xi}, \lambda) = \sum_{i=1}^N \lambda_i Q_{2,l}^{(i)}(\hat{\xi}), \quad l = 1, \dots, L. \end{aligned} \quad (2.6)$$

Таким образом, приходим к следующей системе линейных уравнений:

$$A\lambda = f_\delta, \quad (2.7)$$

элементы матрицы которой в нашем случае имеют вид

$$a_{ij} = \sum_{l=1}^L \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (Q_{1,l}^{(i)}(\hat{\xi}) Q_{1,l}^{(j)}(\hat{\xi}) + Q_{2,l}^{(i)}(\hat{\xi}) Q_{2,l}^{(j)}(\hat{\xi})) d\hat{\xi}, \quad 1 \leq i \leq N, \quad 1 \leq j \leq N. \quad (2.8)$$

Элементы a_{ij} матрицы A могут быть вычислены явно с помощью интеграла Пуассона:

$$\begin{aligned} a_{ij} &= 2\pi \sum_{l=1}^L \left\{ \frac{z_i + z_j - 2H_l}{\left(\sqrt{(z_i + z_j - 2H_l)^2 + (x_i - x_j)^2 + (y_i - y_j)^2} \right)^3} - \right. \\ &\quad \left. - \frac{(z_i + z_j - 2H_l)(9[(x_i - x_j)^2 + (y_i - y_j)^2] - 6(z_i + z_j - 2H_l)^2)}{\left(\sqrt{(z_i + z_j - 2H_l)^2 + (x_i - x_j)^2 + (y_i - y_j)^2} \right)^7} \right\}, \\ &\quad 1 \leq i \leq N, \quad 1 \leq j \leq N. \end{aligned} \quad (2.9)$$

Поясним, как от формулы (2.8) перейти к выражению (2.9). На каждой из плоскостей H_l распределены источники гравитационного поля. Первое слагаемое в сумме (2.2) при фиксированном значении индекса l представляет собой двойной слой с плотностью $\rho_{1,l}(\hat{\xi})$, а второе слагаемое – производную этого двойного слоя по z , плотность распределения этой производной равна $\rho_{2,l}(\hat{\xi})$. Далее, в формуле (2.8) при каждом значении индекса l выражение

$$Q_{1,l}^{(i)}(\hat{\xi}) Q_{1,l}^{(j)}(\hat{\xi}) = \frac{(z_i - H_l)(z_j - H_l)}{\left(\sqrt{(z_i - H_l)^2 + (\xi_1 - x_i)^2 + (\xi_2 - y_i)^2} \right)^3 \left(\sqrt{(z_j - H_l)^2 + (\xi_1 - x_j)^2 + (\xi_2 - y_j)^2} \right)^3}$$

есть произведение ядра интеграла Пуассона, вычисленного в точке M_i с координатами (x_i, y_i, z_i) и точке плоскости $\hat{\xi} = (\xi_1, \xi_2)$, на значение плотности $\rho_{1,l}(\hat{\xi})$, которая зависит от координат точки $M_j = (x_j, y_j, z_j)$, следующее:

$$\rho_{1,l}(\hat{\xi}) = \frac{(z_j - H_l)}{\left(\sqrt{(z_j - H_l)^2 + (\xi_1 - x_j)^2 + (\xi_2 - y_j)^2} \right)^3}.$$

Таким образом, на плоскости H_l значение искомой гармонической в верхнем полупространстве функции равно $\rho_{1,l}(\hat{\xi})$ в точке с координатами $\hat{\xi} = (\xi_1, \xi_2)$. Если теперь “подняться выше”, в точку $M_i = (x_i, y_i, z_i)$, то к числителю дроби в выражении для $\rho_{1,l}(\hat{\xi})$ нужно прибавить $z_j - H_l$, и на такое же значение изменится z -координата в знаменателе. Аналогичным образом трактуется выражение $Q_{2,l}^{(i)}(\hat{\xi}) Q_{2,l}^{(j)}(\hat{\xi})$ в (2.8). Разница заключается только в том, что одна производная “перебрасывается” с ядра Пуассона на функцию, описывающую плотность масс в интеграле. Поэтому в (2.9) при каждом значении индекса, по которому выполняется суммирование, складываются первые и третьи производные потенциалов точечных источников.

Иначе говоря, в (2.9) при фиксированном значении индекса r стоит сумма первых и третьих производных по z от потенциала точечного источника, расположенного в точке $\tilde{M}_j = (x_j, y_j, 2H_l - z_j)$, вычисленных в точке $M_i = (x_i, y_i, z_i)$.

По найденным из решения системы (2.7)–(2.9) множителям λ_i , $i = 1, \dots, N$, можно далее определить величины некоторых функционалов p_s , $s = 1, \dots, S$ (см. [4], [9–13]). Эти функционалы могут представлять собой высшие производные потенциала, аналитическое продолжение гравитационного поля и т.д. Для приложений очень важно иметь возможность пересчета значений поля на некоторую регулярную сеть наблюдений.

Рассмотрим случай, когда носитель потенциалов простого и двойного слоев только один. Пусть D – плоскость, описываемая уравнением

$$z = \text{const} = -H, \quad H > 0.$$

По найденным из решения системы (2.7)–(2.9) плотностям $\rho_1(\xi_1, \xi_2) \in L_2(D)$ и $\rho_2(\xi_1, \xi_2) \in L_2(D)$ простого и двойного слоев, распределенных на этой плоскости, мы можем найти совокупность ограниченных линейных функционалов вида

$$p_s = \int_D \rho_1(\xi_1, \xi_2) P_1^{(s)}(\xi_1, \xi_2) d\xi_1 d\xi_2 + \int_D \rho_2(\xi_1, \xi_2) P_2^{(s)}(\xi_1, \xi_2) d\xi_1 d\xi_2, \quad s = 1, \dots, S, \quad (2.10)$$

где для всех s

$$\|P_1^{(s)}\|^2 = \int_D (P_1^{(s)})^2(\xi_1, \xi_2) d\xi_1 d\xi_2 < +\infty, \quad \|P_2^{(s)}\|^2 = \int_D (P_2^{(s)})^2(\xi_1, \xi_2) d\xi_1 d\xi_2 < +\infty. \quad (2.11)$$

Действительно, значения аномалий силы тяжести $g(M_i; \rho)$ выражаются соотношением (2.10) при $s = i$:

$$P_1^{(i)}(\xi_1, \xi_2, \xi_3) = \gamma \frac{(\xi_3 - z_i)}{R^3(\xi - \mathbf{r}_i)}, \quad P_2^{(i)}(\xi_1, \xi_2, \xi_3) = \gamma \frac{((\xi_1 - x_i)^2 + (\xi_2 - y_i)^2 - 2(\xi_3 - z_i)^2)}{R^5(\xi - \mathbf{r}_i)}, \quad (2.12)$$

а значения элементов аномального поля $u(M_i; \rho)$ – соотношением вида

$$u(M_i; \rho) = D_{\mathbf{r}}^{\alpha} g(\mathbf{r}; \rho) \Big|_{\mathbf{r}=\mathbf{r}_i}, \quad (2.13)$$

где $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ – мультииндекс, $|\alpha| = \alpha_1 + \alpha_2 + \alpha_3 \geq 1$,

$$D_x^{\alpha} = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \partial x_3^{\alpha_3}}, \quad \rho = (\rho_1, \rho_2), \quad \xi = (\xi_1, \xi_2, \xi_3), \quad \hat{\xi} = (\xi_1, \xi_2), \quad (2.14)$$

и соотношением (2.10) при $s = i$:

$$P_1^{(i)}(\xi_1, \xi_2, \xi_3) = \gamma (-1)^{|\alpha|} D_{\hat{\xi}}^{\alpha} \left(\frac{\xi_3 - z}{R^3(\hat{\xi} - \mathbf{r})} \right) \Big|_{\mathbf{r}=\mathbf{r}_i}, \quad (2.15)$$

$$P_2^{(i)}(\xi_1, \xi_2, \xi_3) = \gamma (-1)^{|\alpha|} D_{\hat{\xi}}^{\alpha} \left(\frac{((\xi_1 - x)^2 + (\xi_2 - y)^2 - 2(\xi_3 - z)^2)}{R^5(\hat{\xi} - \mathbf{r})} \right) \Big|_{\mathbf{r}=\mathbf{r}_i}.$$

Везде γ – гравитационная постоянная ($\gamma = 6.67 \times 10^{-11}$ Н · м²/кг²). Здесь точка M_i имеет координаты $M_i = (x_i, y_i, z_i)$, $\mathbf{r}_i = (x_i, y_i, z_i)$. Причем $z \geq -H$, т.е. мы вычисляем значения производных потенциала выше носителя, которым в нашем случае является плоскость D . Здесь $R(\hat{\xi} - \mathbf{r}_i)$ – длина вектора $(\hat{\xi} - \mathbf{r}_i)$.

Исследование матрицы системы (2.7) на однозначную разрешимость является весьма трудоемким, поэтому мы ограничимся в настоящей работе рассмотрением двух частных случаев: а) представлений физических полей (гравитационного и магнитного) в виде потенциала простого слоя и б) представления компонент вектор-функций в виде потенциала двойного слоя. В выражении (2.10) это будут первое и второе слагаемые соответственно.

Если какая-либо высшая производная гравитационного потенциала (см. (2.10)–(2.13)) аппроксимируется потенциалом простого слоя, то матрица системы (2.7)–(2.9) принимает вид

$$A = \begin{bmatrix} \frac{1}{4(z_1 + H)^2} & \frac{z_1 + z_2 + 2H}{((z_1 + z_2 + 2H)^2 + r_{12}^2)^{3/2}} \cdots & \frac{z_1 + z_N + 2H}{((z_1 + z_N + 2H)^2 + r_{1N}^2)^{3/2}} \\ \cdots & \cdots & \cdots \\ \frac{z_1 + z_N + 2H}{((z_1 + z_N + 2H)^2 + r_{1N}^2)^{3/2}} & \frac{z_2 + z_N + 2H}{((z_2 + z_N + 2H)^2 + r_{2N}^2)^{3/2}} \cdots & \frac{1}{4(z_N + H)^2} \end{bmatrix}. \quad (2.16)$$

Попытаемся выяснить, по аналогии с разд. 1 настоящей статьи, в каких случаях матрица (2.16) является вырожденной. Для того чтобы определитель матрицы был равен нулю, достаточно, чтобы какие-либо две строки этой матрицы были пропорциональны друг другу, например, i -я и j -я. Рассмотрим более подробно элементы указанных строк и запишем условие пропорциональности:

$$A_{ij} = \begin{bmatrix} \cdots \frac{1}{4(z_i + H)^2} \frac{z_i + z_j + 2H}{((z_i + z_j + 2H)^2 + r_{ij}^2)^{3/2}} \cdots \frac{z_i + z_k + 2H}{((z_i + z_k + 2H)^2 + r_{ij}^2)^{3/2}} \\ \cdots \frac{z_i + z_j + 2H}{((z_i + z_j + 2H)^2 + r_{ij}^2)^{3/2}} \frac{1}{4(z_j + H)^2} \cdots \frac{z_k + z_j + 2H}{((z_k + z_j + 2H)^2 + r_{ij}^2)^{3/2}} \end{bmatrix} \times$$

$$\times \frac{(z_i + z_j + 2H)^2}{((x_i - x_j)^2 + (y_i - y_j)^2 + (z_i + z_j + 2H)^2)^3} = \frac{1}{16(z_i + H)^2(z_j + H)^2},$$

$$r_{ij}^2 \equiv (x_i - x_j)^2 + (y_i - y_j)^2, \quad (2.17)$$

$$r_{ij}^2 = \left(16(z_i + z_j + 2H)^2(z_i + H)^2(z_j + H)^2\right)^{1/3} - (z_i + z_j + 2H)^2 \times$$

$$\times \frac{z_i + z_k + 2H}{((z_i + z_k + 2H)^2 + r_{ik}^2)^{3/2}} = C \frac{z_k + z_j + 2H}{((z_k + z_j + 2H)^2 + r_{kj}^2)^{3/2}}, \quad (2.18)$$

$$C = \left\{ \frac{z_i + z_j + 2H}{((x_i - x_j)^2 + (y_i - y_j)^2 + (z_i + z_j + 2H)^2)^{3/2}} \right\} : \left\{ \frac{1}{4(z_i + H)^2} \right\}.$$

Из (2.18) следует, что r_{ij}^2 всегда меньше или равно нулю:

$$r_{ij}^2 = \left(16(z_i + z_j + 2H)^2(z_i + H)^2(z_j + H)^2\right)^{1/3} - (z_i + z_j + 2H)^2,$$

$$16(z_i + H)^2(z_j + H)^2 \geq (z_i + z_j + 2H)^4 \Rightarrow$$

$$\Rightarrow 4(z_i + H)(z_j + H) \geq (z_i + z_j + 2H)^2 \Rightarrow z_i = z_j, \quad r_{ij}^2 = 0.$$

Следовательно, диагональные элементы указанных двух строк не могут быть пропорциональны внедиагональным ни при каких значениях z -координат точек наблюдения.

Таким образом, мы можем сделать вывод о том, что верна

Теорема 5. Матрица системы (2.7)–(2.9) при представлении поля с помощью потенциала простого слоя имеет ранг не меньше двух; таким образом, для двух различных точек наблюдения система является невырожденной.

Перейдем теперь к случаю б) представления элементов гравитационного и магнитного полей в виде потенциала двойного слоя.

Матрица системы (2.9) приобретает следующий вид:

$$A = \begin{bmatrix} \frac{6}{16(z_1 + H)^4} & \dots & \frac{(z_1 + z_N + 2H)(6(z_1 + z_N + 2H)^2 - 9r_{1N}^2)}{((z_1 + z_N + 2H)^2 + r_{1N}^2)^{7/2}} \\ \dots & \dots & \dots \\ \frac{(z_1 + z_N + 2H)(6(z_1 + z_N + 2H)^2 - 9r_{1N}^2)}{((z_1 + z_N + 2H)^2 + r_{1N}^2)^{7/2}} & \dots & \frac{6}{16(z_N + H)^4} \end{bmatrix}. \quad (2.19)$$

Запишем, как и для случая простого слоя, условие пропорциональности диагональных и вне-диагональных элементов каких-нибудь двух строк:

$$\frac{(z_i + z_j + 2H)(6(z_i + z_j + 2H)^2 - 9r_{ij}^2)}{((z_i + z_j + 2H)^2 + r_{ij}^2)^{7/2}} \equiv C(z_j) = \frac{6}{16(z_i + H)^2(z_j + H)^2} = \text{const} > 0, \quad (2.20)$$

$$r_{ij}^2 = f(R, C(z_j)), \quad 6R^2 \geq C(z_j)R^6 \Rightarrow R \leq \left(\frac{6}{C(z_j)} \right)^{1/4},$$

$$R = \sqrt{(z_i + z_j + 2H)^2 + r_{ij}^2}$$

(мы рассматриваем константу C как функцию координаты: $C = C(z_j)$).

Явную зависимость $r_{ij}^2 = f(R, C(z_j))$ найти трудно, поэтому попробуем определить, выполняется ли условие (2.20) на какой-либо сфере радиуса $R = \sqrt{(z_i + z_j + 2H)^2 + r_{ij}^2}$. Если некоторая точка трехмерного пространства лежит на такой сфере, то мы можем выразить ее координату z :

$$z_i + z_j + 2H = \sqrt{R^2 - r_{ij}^2}.$$

Будем считать, что z_i, z_j известны. Тогда $C_1 \equiv z_i + z_j + 2H = C_1(C)$, $C = \frac{6}{16(z_i + H)^2(z_j + H)^2}$.

Исследуем, всегда ли существует такое $r_{ij}^2 > 0$ (при всех ли $C_1(C)$ и C), что справедливо равенство

$$\frac{C_1(6C_1^2 - 9r_{ij}^2)}{(\sqrt{C_1^2 + r_{ij}^2})^7} = C.$$

Попробуем изобразить поиск решения графически (см. фиг. 4).

Если $C_1^2 + r_{ij}^2 = R^2$, то, согласно (2.20), на сфере радиуса R мы будем иметь:

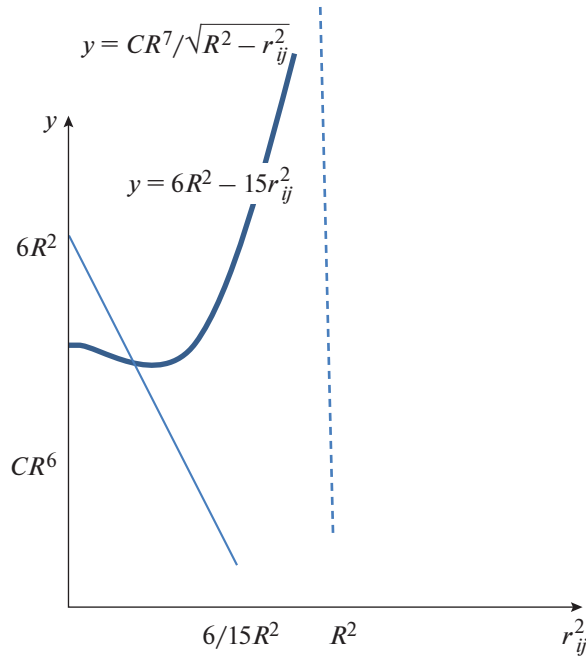
$$\sqrt{R^2 - r_{ij}^2}(6R^2 - 15r_{ij}^2) = CR^7 \Rightarrow r_{ij}^2 > 0 \quad \text{существует, если} \quad CR^6 < 6R^2 \Rightarrow R^4 < \frac{6}{C}.$$

Но $R \geq C_1$. Поэтому $C_1^4 \leq \frac{6}{C}$. Получаем противоречие. Следовательно, верна

Теорема 6. Матрица системы (2.7)–(2.9) в случае представления элементов полей потенциалом двойного слоя имеет ранг не меньше двух, если две точки наблюдения различны.

Замечание. Теоремы 5 и 6 нельзя обобщить на случай произвольного числа точек N , поскольку условие, что никакие две строки матрицы не являются линейно зависимыми, не является достаточным для невырожденности системы линейных алгебраических уравнений.

Однако можно привести два примера однозначной разрешимости системы (2.7)–(2.9) при $N = 3$ и $N = 4$.



Фиг. 4. Графический способ решения уравнения (2.20).

Пример 1: $N = 3$.

Если число точек наблюдения равно трем, то матрица системы (2.7)–(2.9) для потенциала простого слоя приобретает вид

$$A_{33} = \begin{bmatrix} a & b & b \\ b & a & b \\ b & b & a \end{bmatrix}, \quad a = \frac{1}{4(z_1 + H)^2} = \frac{1}{4(z_2 + H)^2} = \frac{1}{4(z_3 + H)^2}, \quad (2.21)$$

$$b = \frac{z_1 + z_2 + 2H}{((z_1 + z_2 + 2H)^2 + r_{12}^2)^{3/2}} = \frac{z_1 + z_3 + 2H}{((z_1 + z_3 + 2H)^2 + r_{13}^2)^{3/2}} = \frac{z_2 + z_3 + 2H}{((z_2 + z_3 + 2H)^2 + r_{23}^2)^{3/2}}.$$

Такие соотношения между элементами матрицы возможны, если координаты z всех трех точек наблюдения одинаковы, а сами точки расположены в вершинах равностороннего треугольника на соответствующей плоскости. При выполнении указанных условий все три точки лежат на поверхности в трехмерном пространстве, описываемой уравнением

$$\frac{z}{(z^2 + r^2)^{3/2}} = C = \text{const} > 0.$$

Запишем выражение для детерминанта матрицы (2.21):

$$\det A_{33} = a^3 + 2b^3 - 3ab^2.$$

Выясним, может ли он быть равен нулю. Имеем

$$a^3 + 2b^3 - 3ab^2 = 0 \Rightarrow a^3 - ab^2 = 2ab^2(a - b) \Rightarrow a(a - b)(a + b) = 2b^2(a - b).$$

Числа a и b всегда больше нуля при построении S-аппроксимаций. Поэтому последнее равенство может выполняться только в том случае, когда эти числа равны. Но они не могут быть равны, согласно теореме 5. Следовательно, в данном частном случае трехмерной системы уравнений решение определяется однозначно.

Пример 2: $N = 4$.

1. Рассмотрим систему (2.7)–(2.9) для четырех точек наблюдения при аппроксимации элементов поля с помощью потенциала простого слоя. Матрица системы имеет в этом случае вид

$$A_{44} = \begin{bmatrix} a & c & d_1 & d_2 \\ c & b & d_3 & d_4 \\ d_1 & d_3 & a_1 & c_1 \\ d_2 & d_4 & c_1 & b_1 \end{bmatrix}, \quad a = \frac{1}{4(z_1 + H)^2}, \quad c = \frac{z_1 + z_2 + 2H}{((z_1 + z_2 + 2H)^2 + r_{12}^2)^{3/2}}, \quad b = \frac{1}{4(z_2 + H)^2},$$

$$a_1 = \frac{1}{4(z_3 + H)^2}, \quad b_1 = \frac{1}{4(z_4 + H)^2}, \quad c_1 = \frac{z_3 + z_4 + 2H}{((z_3 + z_4 + 2H)^2 + r_{34}^2)^{3/2}},$$

$$d_1 = \frac{z_1 + z_3 + 2H}{((z_1 + z_3 + 2H)^2 + r_{13}^2)^{3/2}}, \quad d_2 = \frac{z_1 + z_4 + 2H}{((z_1 + z_4 + 2H)^2 + r_{14}^2)^{3/2}},$$

$$d_3 = \frac{z_2 + z_3 + 2H}{((z_2 + z_3 + 2H)^2 + r_{23}^2)^{3/2}}, \quad d_4 = \frac{z_2 + z_4 + 2H}{((z_2 + z_4 + 2H)^2 + r_{24}^2)^{3/2}}. \quad (2.22)$$

Выберем координаты точек наблюдения так, чтобы элементы матрицы A_{44} , обозначенные через $d_i, i = 1, \dots, 4$, были одинаковыми. Почему этого можно добиться?

Рассмотрим поверхность $C = \frac{z}{(z^2 + r^2)^{3/2}}$ в пространстве координат x, y, z . Выразим r^2 через z явно:

$$C = \frac{z}{(z^2 + r^2)^{3/2}} \Rightarrow r^2 = \frac{z^{2/3}}{C^{2/3}} - z^2 \equiv f(z).$$

Производная функции $f(z)$ положительна при $z < \frac{1}{(3C')^{3/4}}, C' = C^{2/3}$.

Функция $f(z)$ является непрерывной функцией своего аргумента и $f(z) \geq 0$, если

$$z \leq \frac{1}{(C')^{3/4}}.$$

Отсюда следует, что при выполнении условий

$$z_k < \frac{1}{\sqrt{C}} - \max(z_1, z_2) - 2H$$

существуют такие r_{1k}, r_{2k} , что точки $(z_1 + z_k + 2H, r_{1k}), (z_2 + z_k + 2H, r_{2k}), k = 3, 4$, лежат на указанной выше поверхности, и матрица системы имеет вид (2.22). Можно представить выпуклый четырехугольник, в вершинах которого расположены четыре точки наблюдения, z -координаты этих точек различны, но суммы $z_i + z_k$ могут быть одинаковыми при каких-то значениях индексов (но не при всех: четыре точки на плоскости не могут располагаться попарно на одинаковых расстояниях). Вычислим детерминант этой матрицы с помощью формулы Лапласа разложения по двумерным минорам первых двух строк:

$$\det A_{44} = \sum_{i=1}^6 (-1)^{\varepsilon_i} M_i \bar{M}_i, \quad M_1 = \begin{vmatrix} a & c \\ c & b \end{vmatrix}, \quad M_2 = \begin{vmatrix} a & d \\ c & d \end{vmatrix},$$

$$M_3 = \begin{vmatrix} a & d \\ c & d \end{vmatrix}, \quad M_4 = M_5 = \begin{vmatrix} c & d \\ b & d \end{vmatrix}, \quad M_6 = \begin{vmatrix} d & d \\ d & d \end{vmatrix} = 0,$$

ε_i — знак перестановки, соответствующий i -му минору, $\bar{M}_i$ — дополнительный минор.

В итоге получаем

$$\det A_{44} = (ab - c^2)(a_1 b_1 - c_1^2) + d^2(a + b - 2c)(a_1 + b_1 - 2c_1). \quad (2.23)$$

Согласно теореме 5, первое слагаемое всегда положительно, что касается второго слагаемого, то оно тоже положительно в силу соотношения

$$(a^2 + b^2)(a + b)^2 \geq 8a^2b^2 \quad \text{при любых } a, b > 0, \quad \text{так как}$$

$$(a^2 + b^2)(a + b)^2 \geq 2ab(a + b)^2 \geq 8a^2b^2.$$

Таким образом, в данном частном случае четырех точек наблюдения решение системы (2.4) определяется однозначно.

2. Предположим теперь, что четыре точки в трехмерном пространстве расположены таким образом, что матрица системы (2.7)–(2.9) в случае потенциала простого слоя является тёплицевой:

$$A_{44} = \begin{bmatrix} a & b & c & d \\ b & a & b & c \\ c & b & a & b \\ d & c & b & a \end{bmatrix},$$

$$a = \frac{1}{4(z_1 + H)^2}, \quad b = \frac{z_1 + z_2 + 2H}{((z_1 + z_2 + 2H)^2 + r_{12}^2)^{3/2}}, \quad c = \frac{z_1 + z_3 + 2H}{((z_1 + z_3 + 2H)^2 + r_{13}^2)^{3/2}}, \quad (2.24)$$

$$d = \frac{z_1 + z_4 + 2H}{((z_1 + z_4 + 2H)^2 + r_{14}^2)^{3/2}} = c, \quad z_1 = z_2 = z_3 = z_4, \quad r_{12} = r_{23} = r_{34}, \quad r_{13} = r_{24}.$$

Можно считать, что точки наблюдения находятся в вершинах ромба с углом в 60 градусов. Для определителя (2.21) получаем следующее выражение:

$$\det A_{44} = (b^2 - ac)^2 - b^2(a - c)^2.$$

Считаем, что $a > b > c$ (выполнения этого условия можно добиться с помощью указанного выше расположения точек наблюдения). Имеем

$$b^2(b^2 - a^2) > c^2(b^2 - a^2).$$

Таким образом, мы показали, что в рассмотренном случае матрица системы (2.4) не вырождена, и, следовательно, система имеет единственное решение.

ЗАКЛЮЧЕНИЕ

При обработке данных детальной съемки можно применять различные варианты метода линейных интегральных представлений. Тот же принцип справедлив и при решении возникающих при этом систем линейных алгебраических уравнений. Метод S-аппроксимаций позволяет построить за сравнительно короткое время аналитические аппроксимации элементов аномальных потенциальных полей в “первом приближении” по точности. Такая аппроксимация позволяет составить общее представление о природе источников, характере аномалии и т.п. Затем можно строить более точные модели среды. Приведенные в настоящей статье теоремы позволяют выбирать наиболее эффективные в каждом конкретном случае алгоритмы регуляризации при решении обратных задач геофизики, в частности, путем разбиения исходной матрицы системы на блоки, для которых векторы неизвестных параметров можно определить однозначно.

Авторы выражают глубокую благодарность А.С. Леонову за полезные рекомендации и внимание к работе.

СПИСОК ЛИТЕРАТУРЫ

1. *Salnikov A.M., Stepanova I.E., Gudkova T.V., Batov A.V.* Analytical modeling of the magnetic field of Mars from satellite data using modified S-approximations // *Doklady Earth Sciences*. 2021. V. 499. P. 575–579.
2. *Lukyanenko D.V., Yagola A.G.* Some methods for solving of 3D inverse problem of magnetometry // *Eurasian Journal of Mathematical and Computer Applications*. 2016. V. 4. № 3. P. 4–14.
3. *Kolotov I.I., Lukyanenko D.V., Stepanova I.E. et al.* Recovering the magnetic image of Mars from satellite observations // *Journal of Imaging*. 2021. V. 7. № 11. P. 234.
4. *Страхов В.Н.* Геофизика и математика. М.: ОИФЗ РАН, 1999. 64 с.

5. *Страхов В.Н., Степанова И.Э.* Метод S-аппроксимаций и его использование при решении задач гравиметрии (локальный вариант) // *Физика Земли*. 2002. № 2. С. 3–19.
6. *Страхов В.Н., Степанова И.Э.* Метод S-аппроксимаций и его использование при решении задач гравиметрии (региональный вариант) // *Физика Земли*. 2002. № 7. С. 3–12.
7. *Ягола А.Г., Степанова И.Э., Ван Янфей, Титаренко В.Н.* Обратные задачи и методы их решения. Приложения к геофизике. М.: Бином, 2014. 214 с.
8. *Кошляков Н.С., Глинер Э.Б., Смирнов М.М.* Основные дифференциальные уравнения математической физики. М.: Физматгиз, 1962. 767 с.
9. *Stepanova I.E.* On the S-approximation of the Earth's gravity field // *Inverse Problems in Science and Engineering*. 2008. V. 16. № 5. P. 535–544.
10. *Stepanova I.E.* On the S-approximation of the Earth's gravity field. Regional version // *Inverse Problems in Science and Engineering*. 2009. V. 16. № 5. P. 1095–1111.
11. *Stepanova I.E., Shchepetilov A.V., Mikhailov P. S.* Analytical Models of the Physical Fields of the Earth in Regional Version with Ellipticity // *Izvestiya, Physics of the Solid Earth*. 2022. V. 58. № 3. P. 406–419.
12. *Gudkova T.V., Stepanova I.E., Batov A.V.* Density anomalies in subsurface layers of mars: model estimates for the site of the InSight mission seismometer // *Solar System Research volume*. 2020. V. 54. P. 15–19.
13. *Степанова И.Э., Щепетилов А.В., Погорелов В.В., Михайлов П.С.* Структурно-параметрический подход при построении цифровых моделей рельефа и гравитационного поля Земли с использованием аналитических S-аппроксимаций // *Геофизические процессы и биосфера*. 2020. Т. 19. № 2. С. 107–116.

УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

УДК 517.95

АНАЛОГ ТЕОРЕМЫ КЕЛЛОГА ДЛЯ КУСОЧНО-ЛЯПУНОВСКИХ ОБЛАСТЕЙ

© 2023 г. А. П. Солдатов^{1,2,3,*}¹ 119333 Москва, ул. Вавилова, 40, ФИЦ ИУ РАН, Россия² Московский центр фундаментальной и прикладной математики при МГУ, Москва, Россия³ 111250 Москва, Красноказарменная ул., 14, НИУ МЭИ, Россия

*e-mail: soldatov48@gmail.com

Поступила в редакцию 16.01.2023 г.
Переработанный вариант 16.01.2023 г.
Принята к публикации 28.04.2023 г.

В рамках весовых гёльдеровых пространств введены классы гладких дуг и кусочно-гладких контуров, инвариантные относительно отображений степенными функциями. В терминах этих классов по аналогии с классической теоремой Келлога описаны граничные свойства конформных отображений. Библ. 3.

Ключевые слова: конформное отображение, кусочно-ляпуновский контур, радиальная дуга, весовое пространство Гёльдера.

DOI: 10.31857/S0044466923080148, EDN: WTNCGU

По определению гладкая дуга на комплексной плоскости принадлежит классу $C^{n,v}$, $0 < v < 1$, $n = 1, 2, \dots$, если она допускает параметризацию из аналогичного класса. Пусть область D ограничена контуром Γ , составленным из дуг $\Gamma_1, \dots, \Gamma_m \in C^{n,v}$, и конечное множество F образовано концами этих дуг. В этом случае говорим, что пара (Γ, F) принадлежит классу $C^{n,v}$. Очевидно, F состоит из m точек, к каждой из которых сходятся две дуги. Запись $\Gamma \in C^{n,v}$ означает, что контур Γ гладкий, и любая дуга $\Gamma_0 \subseteq \Gamma$ принадлежит классу $C^{n,v}$.

Рассмотрим какую-либо каноническую область $\mathbb{D}$, ограниченную гладким контуром класса $C^{n,v}$, причем число связных компонент обоих контуров Γ и $\partial\mathbb{D}$ одно и то же. В частности, существует конформное отображение ω области D на $\mathbb{D}$. Если контур $\Gamma \in C^{n,v}$, то согласно классической теореме Келлога (см. [1], [2]) функция $\omega \in C^{n,v}(\bar{D})$, и аналогичным свойством обладает обратное отображение ω^{-1} . В общем случае, когда $(\Gamma, F) \in C^{n,v}$, функция ω принадлежит классу $C^{n,v}(\bar{D}_0)$ в любой подобласти $D_0 \subseteq D$, лежащей вне окрестности множества F .

Рассмотрим это конформное отображение в окрестности точек $\tau \in F$. Положим $S_\tau = D \cap \{|z - \tau| < \varrho\}$, где ϱ достаточно мало. Тогда S_τ представляет собой криволинейный сектор с вершиной τ и боковыми сторонами $G_\tau^\pm$. Раствор θ_τ этого сектора предполагаем положительным, так что $0 < \theta_\tau \leq 2\pi$. Таким образом, все внутренние углы области D в точках $\tau \in F$ положительны.

Следующий результат хорошо известен (см. [2]) и описывает поведение функции ω в окрестности τ .

Теорема 1. Пусть (Γ, F) принадлежит классу $C^{1,v}$ и $\delta_\tau = \pi/\theta_\tau$, $\tau \in F$. Тогда

$$\omega(z) - \omega(\tau) = A_\tau(z)(z - \tau)^{\delta_\tau}, \quad \omega'(z) = B_\tau(z)(z - \tau)^{\delta_\tau - 1}, \quad (1)$$

где $A_\tau, B_\tau \in C^\varepsilon(\bar{S}_\tau)$ с достаточно малым $\varepsilon > 0$, причем $A_\tau(\tau) \neq 0$, $B_\tau(\tau) \neq 0$.

Аналогичный результат справедлив и для обратного отображения ω^{-1} по отношению к $\delta_\tau = \theta_\tau/\pi$.

Доказательство почти очевидно и основывается на теореме Келлога. Не ограничивая общности, можно считать, что $\tau = \omega(\tau) = 0$, и символ τ в обозначениях опущен. Рассмотрим область $\tilde{S} = \{z^\delta, z \in S\}$, которая является “криволинейным полукругом”. При этом гладкая дуга

$$\partial\tilde{S} \cap \{|\zeta| \leq \varrho^\delta\}$$

принадлежит классу $C^{1,\varepsilon}$ с некоторым малым ε . Более точно, можно положить $\varepsilon = \min(v, v/\delta)$. По теореме Келлога функция $\tilde{\omega}(\zeta) = \omega(\zeta^{1/\delta}) \in C^{1,\varepsilon}(\tilde{S}_0)$, где $\tilde{S}_0 = \tilde{S} \cap \{|\zeta| < \varrho_0\}$, $\varrho_0 < \varrho^\delta$. Поэтому остается заметить, что $\omega(z) = \tilde{\omega}(z^\delta)$, $z \in S$.

Основная цель статьи состоит в построении классов гладких дуг, которые были бы инвариантны относительно степенных функций, выпрямляющих контур в окрестности его угловых точек. Использование подобных классов позволило бы в рамках соответствующих весовых пространств описать единообразно коэффициенты A и B в представлениях (2). В основе всех построений лежат весовые гильбертовы пространства $C_\lambda^{n,v}$, подробно описанные в [3].

Предварительно введем следующее понятие. Гладкую дугу G с концом τ назовем *радиальной* (по отношению к этому концу), если она допускает параметризацию вида

$$\gamma(r) = \tau + re^{ih(r)}, \quad 0 \leq r \leq \varrho, \quad (2)$$

где функция $h(r) \in C[0, \varrho] \cap C^1(0, \varrho]$ и $rh'(r) \rightarrow 0$ при $r \rightarrow 0$. Эквивалентное определение состоит в том, что каждая окружность радиуса $r \leq \varrho$ с центром в τ пересекает эту дугу в единственной точке и при том некасательно.

Нетрудно видеть, что любая разомкнутая гладкая дуга G в некоторой окрестности своего конца τ радиальна.

В самом деле, пусть $\delta(t)$, $0 \leq t \leq 1$, — некоторая ее гладкая параметризация, причем $\delta(0) = \tau$. Тогда функция $\alpha(t) = |\delta(t) - \delta(0)|$ непрерывно дифференцируема, и $\alpha'(0) = |\delta'(0)|$. Поэтому на отрезке $[0, \varepsilon]$, где $\varepsilon < 1$ достаточно мало, функция $r = \alpha(t)$ имеет обратную $t = \beta(r)$, $0 \leq r \leq \varrho = \alpha(\varepsilon)$, и функция $\gamma(r) = \delta[\beta(r)]$ есть радиальная параметризация соответствующей дуги.

Класс $C^{n,v}$ радиальных дуг можно описать в терминах функции $\theta(r)$, фигурирующей в представлении (2).

Лемма 1. Радиальная дуга (G, τ) принадлежит классу $C^{n,v}$ тогда и только тогда, когда

$$h(r) \in C^{n-1,v}[0, \varrho], \quad rh^{(n)}(r) \in C^v[0, \varrho]. \quad (3)$$

Доказательство. Если выполнены условия (3), то функция γ в (2) принадлежит классу $C^{n,v}[0, \varrho]$ и, следовательно, дуга $G \in C^{n,v}$. Обратно, пусть эта дуга допускает параметризацию $\delta : [0, \varrho] \rightarrow G$ класса $C^{n,v}[0, \varrho]$ и $\delta(0) = \tau$. Не ограничивая общности, можно считать, что $\tau = 0$. Функция

$$\delta_0(t) = t^{-1}\delta(t) = \int_0^1 \delta'(tu)du \in C^{n-1,v}$$

и всюду отлична от нуля, так что вместе с ней классу $C^{n-1,v}$ принадлежат и функции $|\delta_0|^{\pm 1}$. Функция $\alpha(t) = |\delta(t)|$ осуществляет гомеоморфизм отрезка $[0, \varrho]$ на себя с положительной производной

$$\alpha' = \frac{\operatorname{Re}(\delta' \bar{\delta})}{|\delta|} = \frac{\operatorname{Re}(\delta' \bar{\delta}_0)}{|\delta_0|},$$

принадлежащей классу $C^{n-1,v}$. Поэтому $\alpha \in C^{n,v}$, и аналогичным свойством обладает обратное отображение $\beta = \alpha^{-1}$. Но тогда и функция $\delta \circ \beta \in C^{n,v}$. Поскольку она также является параметризацией дуги и $|\delta[\beta(r)]| = \alpha[\beta(r)] = r = |\gamma(r)|$, отсюда следует, что $\gamma = \delta \circ \beta$ и, значит, $\gamma \in C^{n,v}[0, \varrho]$. Но тогда $r^{-1}\gamma(r) = e^{ih(r)} \in C^{n-1,v}[0, \varrho]$, так что и $h(r) \in C^{n-1,v}[0, \varrho]$. После n -кратного дифференцирования равенства (2) приходим к справедливости и второго условия в (3).

В дальнейшем без ограничения общности можно считать, что в теореме 1 боковые стороны $G_\tau^\pm$ криволинейных секторов S_τ радиальны и определяются формулой (2) с соответствующими функциями $h_\tau^\pm(r)$. Зафиксируем τ , опуская символ τ в обозначениях. Рассмотрим функцию $\zeta = \ln(z - \tau)$, которая отображает S на криволинейную полуполосу Q в полуплоскости $\operatorname{Re} \zeta \leq \ln \varrho$. В явном виде

$$Q = \{\zeta = \xi + i\eta, |g^-(\xi) < \eta < g^+(\xi), \xi < \ln \varrho\} \quad (4)$$

с функциями $g^\pm(\xi) = h^\pm(e^\xi)$. Очевидно, эти функции имеют конечные пределы $\theta^\pm$ на $-\infty$, причем $(g^\pm)'(\xi) \rightarrow 0$ при $\xi \rightarrow -\infty$.

Обозначим $C^{n,v}(\bar{Q})$, $0 \leq v \leq 1$, $n = 0, 1, \dots$, класс аналитических в Q функций, которые вместе со своими производными до n -го порядка включительно непрерывны и ограничены в замкнутой области $\bar{Q}$ и при $0 < v < 1$ ($v = 1$) удовлетворяют условию Гёльдера с показателем v (условию Липшица). Относительно соответствующей нормы это пространство является банаховой алгеброй по умножению.

По определению пространство $C_0^{n,\mu}(\bar{S}, \tau)$ состоит из всех функций

$$\phi(z) = \psi[\ln(z - \tau)], \quad \psi \in C^{n,v}(\bar{Q}). \quad (5)$$

Оно снабжается перенесенной нормой, относительно которой является банаховой алгеброй. Весовое пространство $C_\lambda^{n,\mu}(\bar{S}, \tau)$, $\lambda \in \mathbb{R}$, состоит из функций

$$\phi(z) = (z - \tau)^\lambda \phi_0(z), \quad \phi_0 \in C_0^{n,v}(\bar{S}, \tau). \quad (6)$$

Относительно перенесенной нормы это пространство банахово. Полученное семейство банаховых пространств монотонно убывает (в смысле вложений) по каждому из параметров:

$$C_\lambda^{n,v_2} \subseteq C_\lambda^{n,v_1}, \quad v_1 \leq v_2 \leq 1; \quad C_{\lambda_2}^{n,v} \subseteq C_{\lambda_1}^{n,v}, \quad \lambda_1 \leq \lambda_2; \quad C_\lambda^{n+1,0} \subseteq C_\lambda^{n,1}. \quad (7)$$

В самом деле, первое вложение является следствием аналогичного свойства пространств $C^{n,v}$. Второе вложение вытекает из того, что при $\varepsilon > 0$ функция $(z - \tau)^\varepsilon$ принадлежит $C_0^{n,v}$. Наконец, третье вложение выводится из аналогичного вложения $C^{n+1,0}(\bar{Q}) \subseteq C^{n,1}(\bar{Q})$. Нетрудно убедиться, что оно справедливо для любой области Q , которая равномерно связна в следующем смысле. Существует такая постоянная $M \geq 1$, что любые две точки $\zeta_1, \zeta_2 \in Q$ можно соединить в этой области гладкой дугой, длина которой не превосходит $M|\zeta_1 - \zeta_2|$. Из описания (4) рассматриваемой криволинейной полуполосы Q легко видеть, что она равномерно линейно связна.

Исходя из пространств $C_\lambda^{0,v}$, аналогичные пространства для $n \geq 1$ можно ввести индуктивно по n условиям

$$\phi \in C_\lambda^{n-1,v}, \quad \phi' \in C_{\lambda-1}^{n-1,v}. \quad (8)$$

Это следует из аналогичного описания пространств $C^{n,v}(\bar{Q})$ и того факта, что равенство $\phi(z) = \psi[\ln(z - \tau)]$ влечет $(z - \tau)\phi'(z) = \psi'[\ln(z - \tau)]$.

В свою очередь, пространство $C_0^{0,\nu}$ при $0 < \nu \leq 1$ может быть определено независимо как класс всех ограниченных аналитических в S функций $\phi(z)$, для которых $(z - \tau)\phi(z) \in C^{0,\nu}(\bar{S})$. Последнее условие равносильно тому, что функция ϕ удовлетворяет оценке

$$|z_1 - \tau|^\nu |\phi(z_1) - \phi(z_2)| \leq C_0 |z_1 - z_2|^\nu, \quad (9)$$

где постоянная $C_0 > 0$ не зависит от $z_1, z_2 \in S$.

Именно подобным образом семейство пространств $C_\lambda^{n,\nu}$ вводилось в [3]. Эквивалентность этих двух подходов обеспечивается теоремой 2.7.2 этой работы.

Отметим связь весовых пространств с преобразованиями области Q на себя.

Лемма 2. Пусть аналитическая в S функция α осуществляет гомеоморфизм $\bar{S}$ на себя, оставляя точку τ неподвижной, и удовлетворяет двустороннему условию Липшица

$$C^{-1} |z_1 - z_2| \leq |\alpha(z_1) - \alpha(z_2)| \leq C |z_1 - z_2|, \quad (10)$$

где постоянная $C \geq 1$ не зависит от z_1, z_2 .

Тогда оператор $\phi \rightarrow \phi \circ \alpha$ ограничен в пространстве $C_\lambda^{0,\nu}(\bar{S}, \tau)$.

Эта лемма установлена в [3], однако чтобы проиллюстрировать приведенные выше свойства весовых пространств, приведем ее доказательство. Достаточно ограничиться случаем $0 < \nu \leq 1$. Функция $\alpha(z) - \tau = \alpha(z) - \alpha(\tau)$ удовлетворяет условию Липшица и обращается в нуль в точке τ , так что она принадлежит $C_1^{0,1}(\bar{S}, \tau)$ и, следовательно,

$$a(z) = (z - \tau)^{-1} [\alpha(z) - \tau] \in C_0^{0,1}(\bar{S}, \tau).$$

В частности, функция a удовлетворяет оценке (9) с $\nu = 1$. Она показывает, что если область D содержит множество значений этой функции и $f \in C^{0,1}(\bar{D})$, то суперпозиция $f \circ a$ также удовлетворяет этой оценке и, значит, принадлежит $C_0^{0,1}$. В силу (10) область D можно выбрать односвязной вне некоторой окрестности нуля, так что функция $f(\zeta) = \zeta^\lambda$ принадлежит $C^{0,1}(\bar{D})$. Следовательно, функция

$$b(z) = (z - \tau)^{-\lambda} [\alpha(z) - \tau]^\lambda \in C_0^{0,1}(\bar{S}, \tau)$$

для любого $\lambda \in \mathbb{R}$. Таким образом, в соответствии с (6) лемму достаточно установить хотя бы для одного весового порядка λ . Пространство $C_\nu^{0,\nu}(\bar{S}, \tau)$ совпадает с подпространством $C^{0,\nu}(\bar{S})$ функций, обращающихся в точке τ в нуль. Но в силу (10) для этого пространства утверждение леммы очевидно.

С каждым целым числом $l \leq n$ свяжем модифицированный весовой класс

$$C_{(l)}^{n,\nu}(\bar{S}, \tau) = P_l + \bigcup_{0 < \varepsilon < 1} C_{l+\varepsilon}^{n,\nu}(\bar{S}, \tau), \quad (11)$$

где P_l означает класс всех многочленов степени не выше l (так что $P_l = 0$ при $l < 0$). Из определения видно, что

$$C^{l,\nu}(\bar{S}) \subseteq C_{(l)}^{n,\nu}(\bar{S}, \tau) \subseteq \bigcup_{0 < \varepsilon < 1} C^{l,\varepsilon}(\bar{S}), \quad 0 \leq l \leq n. \quad (12)$$

В частности, класс $C_{(l)}^{n,\nu}(\bar{S}, \tau)$ состоит из всех функций $\phi \in C^l(\bar{S})$, для которых

$$\phi(z) - \sum_{k=0}^l \frac{\phi^{(k)}(\tau)}{k!} (z - \tau)^k \in C_{l+\varepsilon}^{n,\nu}(\bar{S}, \tau)$$

с некоторым малым положительным ε .

Заметим, что класс $C_{(l)}^{n,v}$ выдерживает умножение на функции $a \in C_{(0)}^{n,v}$, в частности, при $l = 0$ он является алгеброй. В дальнейшем главный интерес представляют случаи $l = 0$ и $l = 1$.

Пространства $C_{\lambda}^{n,v}$ и классы $C_{(l)}^{n,v}$ совершенно аналогично можно ввести и в случае, когда роль S играет отрезок прямой, а точка τ — один из его концов. При этом условие аналитичности в S заменяется дифференцируемостью на этом интервале. Ясно, что все перечисленные свойства весовых пространств сохраняются без изменений.

По определению гладкая дуга G с концом τ принадлежит классу $C_{(1)}^{n,v}$, если она допускает параметризацию $\delta: [0,1] \rightarrow G$, $\delta(0) = \tau$, принадлежащую соответствующему пространству $C_{(1)}^{n,v}([0,1], 0)$. В силу (12) имеем вложение $C_{(1)}^{n,v} \subseteq C^{1,\varepsilon}$ с некоторым малым $\varepsilon > 0$, так что дуги этого класса являются ляпуновскими. При этом вне любой окрестности своего конца τ они принадлежат классу $C^{n,v}$.

Обратимся к криволинейным секторам S_{τ} с боковыми сторонами $G_{\tau}^{\pm}$, фигурирующим в теореме 1. По определению пара (Γ, F) принадлежит классу $C_{(1)}^{n,v}$, если все дуги $G_{\tau}^{\pm} \in C_{(1)}^{n,v}$ и любая дуга $\Gamma_0 \subseteq \Gamma \setminus F$ принадлежит классу $C^{n,v}$.

В принятых обозначениях основной результат статьи, уточняющий теорему 1, можем сформулировать следующим образом.

Теорема 2. Пусть (Γ, F) принадлежит классу $C_{(1)}^{n,v}$, $0 < v < 1$.

Тогда в представлении (1) функции

$$A_{\tau}(z) \in C_{(0)}^{n,v}(\bar{S}_{\tau}, \tau), \quad B_{\tau}(z) \in C_{(0)}^{n-1,v}(\bar{S}_{\tau}, \tau). \quad (13)$$

Аналогичный результат справедлив и для обратного отображения ω^{-1} по отношению к $\delta_{\tau} = \theta_{\tau}/\pi$.

Заметим, что в действительности достаточно установить первое соотношение в (13). В самом деле, дифференцирование первого представления в (1) дает равенство

$$B_{\tau}(z) = \delta_{\tau} A_{\tau}(z) + A_{\tau}(z),$$

согласно которому функция $B_{\tau}(z) \in C_{(0)}^{n-1,v}(\bar{S}_{\tau}, \tau)$.

Доказательству этой теоремы предпошлем несколько важных свойств класса $C_{(l)}^{n,v}$.

Теорема 3. (а) Условия $\phi \in C_{(l)}^{n,v}(\bar{S}, \tau)$ и $\phi' \in C_{(l-1)}^{n-1,v}(\bar{S}, \tau)$ равносильны.

(б) Пусть $\phi \in C_{(0)}^{n,v}(\bar{S}, \tau)$, область D содержит значения $\phi(z)$, $z \in S$, и функция $f(\zeta) \in C^{n,1}(\bar{D})$.

Тогда суперпозиция $f \circ \phi \in C_{(l)}^{n,v}(\bar{S}, \tau)$.

(в) Пусть θ есть разворот сектора S с вершиной $\tau = 0$, так что функция $z \rightarrow z^{\delta}$, где $\delta = \pi/\theta$, переводит S на криволинейный сектор S_0 раствора π .

Тогда условия $\phi(z) \in C_{(0)}^{n,v}(S_0, 0)$ и $\phi(z^{\delta}) \in C_{(0)}^{n,v}(S, 0)$ равносильны.

Доказательство. (а) В одну сторону утверждение следует непосредственно из определения (11). Обратно, пусть $\phi' \in C_{(l-1)}^{n-1,v}(\bar{S}, \tau)$, так что для некоторого многочлена $p_1(z) \in P_{l-1}$ функция $\phi'(z) - p_1(z) \in C_{(l-1+\varepsilon)}^{n-1,v}$ с некоторым малым $\varepsilon > 0$. Заменяя ϕ на $\phi - p$, где $p' = p_1$, можем считать, что

$$\phi'(z) \in C_{(l-1+\varepsilon)}^{n-1,v}. \quad (14)$$

Остается показать, что тогда с точностью до постоянного слагаемого функция $\phi \in C_{(l+\varepsilon)}^{n,v}$. Не ограничивая общности, можно считать $\tau = 0$.

Из (14), в частности, следует оценка

$$|\phi'(z)| \leq C |z|^{\lambda-1}, \quad \lambda = l + \varepsilon. \quad (15)$$

Поскольку $0 < \varepsilon < 1$, число λ здесь отлично от нуля. Зафиксируем точку $z_0 \in S$ и рассмотрим функцию $\phi_0(\zeta) = \phi(|z_0|\zeta)$ в области D_0 , представляющей пересечения сектора $\{\zeta, |z_0|\zeta \in S\}$ с кольцом $\{1/2 < |\zeta| < 1\}$. В полярной системе координат $\zeta = re^{i\theta}$ она описывается неравенством

$$h_0^-(r) < \theta < h_0^+(r), \quad 1/2 < r < 1,$$

с непрерывно дифференцируемыми функциями $h_0^\pm(r) = h^\pm(|z_0|r)$. Эти функции равномерно ограничены по $|z_0|$ в пространстве $C^1[1/2, 1]$ и стремятся к постоянным функциям $\theta^\pm$ по норме этого пространства. Поэтому область D_0 равномерно линейно связна, причем соответствующая постоянная $M \geq 1$ не зависит от z_0 .

В силу (15) производная функции $\phi_0(\zeta)$ в области D_0 допускает оценку

$$|\phi_0'(\zeta)| = |z_0| |\phi'(|z_0|\zeta)| \leq C |z_0|^\lambda \max(1, 2^{1-\lambda}).$$

В силу равномерной линейной связности области D_0 отсюда

$$|\phi_0(\zeta_1) - \phi_0(\zeta_2)| \leq C_0 |z_0|^\lambda |\zeta_1 - \zeta_2|, \quad \zeta_j \in D_0,$$

где здесь и ниже $C_0, C_1, \dots$ означают положительные постоянные, зависящие только от C, λ и M . Пусть точка $z_1 \in S$ выбрана по условию $|z_1| = |z_0|/2$. Тогда $z_j = |z_0|\zeta_j$ с $\zeta_j \in D_0$ и предыдущая оценка переходит в

$$|\phi(z_1) - \phi(z_0)| \leq C_1 |z_0|^\lambda. \quad (16)$$

Далее рассмотрим случаи $\lambda > 0$ и $\lambda < 0$ отдельно. В первом случае рассмотрим последовательность точек $z_0, z_1, \dots$ области S , выбранную по условию $|z_{k+1}| = |z_k|/2$, которая, очевидно, сходится к нулю. В силу (16) отсюда

$$|\phi(z_{k+s}) - \phi(z_k)| \leq C_1 (|z_k|^\lambda + \dots + |z_{k+s-1}|^\lambda) \leq \frac{C_1 M}{1 - 2^{-\lambda}} |z_k|^\lambda,$$

так что существует предел $c = \lim \phi(z_k)$ при $k \rightarrow \infty$. Полагая $k = 0$ и переходя в этом неравенстве к пределу при $s \rightarrow \infty$, приходим к оценке

$$|\phi(z_0) - c| \leq C_2 |z_0|^\lambda.$$

Вместе с (14) она означает, в частности, что функция $\phi(z) - c$ принадлежит классу $C_{\lambda}^{1,0}(\bar{S}, \tau)$, так что с учетом (7) она принадлежит и пространству $C_{\lambda}^{0,v}(\bar{S}, \tau)$. Совместно с (14) отсюда следует, что $\phi(z) - c \in C_{\lambda}^{n,v}(\bar{S}, \tau)$. Тем самым рассматриваемое утверждение теоремы для случая $\lambda > 0$ установлено.

Пусть $\lambda < 0$. Выберем последовательность точек $z_0, z_{-1}, \dots, z_{-m} \in S$, удовлетворяющих условию $|z_{-k-1}| = 2|z_{-k}|$, где натуральное m определяется из условия $|z_{-m}| \leq \varrho$, но $|z_{-m-1}| > \varrho$. Тогда рассуждения, аналогичные предыдущим, приводят к оценке

$$|\phi(z_0)| \leq C_2 |z_0|^\lambda + \sup_{\varrho/2 < |z| < \varrho} |\phi(z)|.$$

Как и выше, из нее совместно с (14) следует, что $\phi \in C_{\lambda}^{n,v}$ и в этом случае.

(б) Рассмотрим сначала случай $n = 0$, предполагая для простоты $\tau = 0$. Пусть $\phi \in C_{(0)}^{0,v}$, так что для некоторой постоянной $c \in \mathbb{C}$ имеем представление

$$\phi(z) = c + z^\varepsilon \phi_0(z), \quad \phi_0 \in C_0^{0,v}, \quad (17)$$

с некоторым малым $\varepsilon > 0$.

Нетрудно видеть, что тогда ϕ удовлетворяет аналогичной (9) оценке

$$|z_1|^{v-\varepsilon} |\phi(z_1) - \phi(z_2)| \leq C |z_1 - z_2|^v, \quad (18)$$

справедливой для любых точек $z_1, z_2 \in S$, подчиненных условию $|z_1| \leq |z_2|$.

В самом деле, величина $|z_1|^{v-\varepsilon} |\phi(z_1) - \phi(z_2)|$ не превосходит

$$|z_1|^v |\phi_0(z_1) - \phi_0(z_2)| + |z_1|^{v-\varepsilon} |z_1^\varepsilon - z_2^\varepsilon| |\phi_0(z_2)| \leq C |z_1 - z_2|^v$$

с постоянной

$$C = C_0 + \sup_{|\zeta| \geq 1} \frac{|1 - \zeta^\varepsilon|}{|1 - \zeta|^v} \sup_{z \in S} |\phi_0(z)|,$$

где C_0 фигурирует в (9) для ϕ_0 .

Убедимся, что верно и обратное: из оценки (18) следует представление (17). Как и при доказательстве (а), зафиксируем точку $z_0 \in S$ и рассмотрим последовательность $z_0, z_1, \dots$ точек S , выбранных по условию $|z_{k+1}| = |z_k|/2$. Тогда на основании (18) имеем неравенства

$$|\phi(z_k) - \phi(z_{k+1})| \leq C |z_{k+1}|^\varepsilon |1 - z_{k+1}^{-1} z_k|^\varepsilon \leq C_0 |z_k|^\varepsilon.$$

Как и при доказательстве (а), отсюда заключаем, что существует предел $c = \lim \phi(z_k)$ при $k \rightarrow \infty$, и справедлива оценка

$$|\phi(z_0) - c| \leq C_1 |z_0|^\varepsilon. \quad (19)$$

В частности, с помощью равенства (17) можем ввести функцию $\phi_0(z)$, по модулю не превосходящую C_1 . Остается убедиться, что для нее справедлива оценка (9). Считая для определенности $|z_1| \leq |z_2|$, рассмотрим величину $|z_1|^v |\phi_0(z_1) - \phi_0(z_2)|$. Очевидно, она не превосходит выражения

$$|z_1|^{v-\varepsilon} |\phi(z_1) - \phi(z_2)| + C_1 |z_1|^v |z_2|^\varepsilon |z_1^{-\varepsilon} - z_2^{-\varepsilon}| \leq C_0 |z_1 - z_2|^v$$

с постоянной

$$C_0 = C + C_1 \sup_{|\zeta| \leq 1} \frac{|\zeta|^v |1 - \zeta^{-\varepsilon}|}{|1 - \zeta|^v},$$

где C_1 фигурирует в (19).

Рассмотрим теперь функцию f , удовлетворяющую условию Липшица $|f(\zeta_1) - f(\zeta_2)| \leq L |\zeta_1 - \zeta_2|$ на множестве значений функции ϕ . В силу (18) функция $f[\phi(z)]$ также удовлетворяет аналогичной оценке с постоянной LC , так что она принадлежит классу $C_{(0)}^{0,v}$, тем самым для $n = 0$ утверждение теоремы установлено.

В общем случае воспользуемся индукцией по n , т.е. пусть утверждение справедливо для класса $C_{(0)}^{n-1,v}$ и $\phi \in C_{(0)}^{n,v}$. По предположению индукции в равенстве $(f \circ \phi)' = (f' \circ \phi)\phi'$ множитель $f' \circ \phi$ принадлежит $C_{(0)}^{n-1,v}$, а функция $\phi' \in C_{(-1)}^{n-1,v}$. Поэтому и $(f \circ \phi)' \in C_{(-1)}^{n-1,v}$, так что на основании (а) функция $f \circ \phi \in C_{(0)}^{n,v}$.

(в) Подстановка $\zeta = \ln z$ преобразование $z \rightarrow z^\delta$ переводит в $\zeta \rightarrow \delta\zeta$. Поэтому по определению (5) оператор $\phi(z) \rightarrow \phi(z^\delta)$ осуществляет изоморфизм банаховых пространств $C_0^{n,v}(D_0, 0) \rightarrow C_0^{n,v}(D, 0)$ и, следовательно, $C_\lambda^{n,v}(D_0, 0) \rightarrow C_{\delta\lambda}^{n,v}(D, 0)$. В соответствии с определением (11) класса $C_{(0)}^{n,v}$ отсюда утверждение теоремы получается непосредственно.

Теорема 3 позволяет получить аналог леммы 1 для дуг класса $C_{(1)}^{n,v}$.

Лемма 3. Радиальная дуга (G, τ) принадлежит классу $C_{(1)}^{n,v}$ тогда и только тогда, когда

$$h(r) \in C_{(0)}^{n,v}([0, \varrho], 0). \quad (20)$$

В частности, преобразование $z \rightarrow z^\delta$ не выводит из класса $C_{(1)}^{n,v}$ радиальных дуг с концом $\tau = 0$.

Доказательство осуществляется по той же схеме, что и лемма 1. Пусть выполнено условие (20), тогда на основании теоремы 3(б) функция $e^{ih(r)} \in C_{(0)}^{n,v}$, так что и производная

$$\gamma'(r) = [1 + irh'(r)]e^{ih(r)} \in C_{(0)}^{n-1,v}.$$

Поэтому в силу теоремы 3(а) параметризация $\gamma \in C_{(1)}^{n,v}$.

Обратно, пусть эта дуга допускает параметризацию $\delta : [0, \varrho] \rightarrow G$ класса $C_{(1)}^{n,v}([0, \varrho], 0)$ и $\delta(0) = \tau$. Не ограничивая общности можно считать, что $\tau = 0$. Функция $\delta_0(t) = t^{-1}\delta(t)$ принадлежит $C_{(0)}^{n,v}$ и всюду отлична от нуля, так что на основании теоремы 3(б) и функции $|\delta_0|^{\pm 1} \in C_{(0)}^{n,v}$. Функция $\alpha(t) = |\delta(t)|$ осуществляет гомеоморфизм отрезка $[0, \varrho]$ на себя с положительной производной

$$\alpha' = \frac{\operatorname{Re}(\delta' \bar{\delta})}{|\delta|} = \frac{\operatorname{Re}(\delta' \bar{\delta}_0)}{|\delta_0|},$$

которая на основании теоремы 3(б) принадлежит классу $C_{(0)}^{n-1,v}$. Поэтому $\alpha \in C_{(1)}^{n,v}$.

Утверждается, что аналогичным свойством обладает и обратное отображение $\beta = \alpha^{-1}$. В самом деле, функция β удовлетворяет двустороннему условию Липшица, фигурирующему в лемме 2, поэтому на основании этой леммы функция $\alpha' \circ \beta \in C_{(0)}^{0,v}$ и, значит, $\beta' = (\alpha' \circ \beta)^{-1}$ принадлежит этому же классу. Тогда $\beta \in C_{(1)}^{1,v}$, поскольку при $n \geq 2$ функция $\alpha' \in C_{(0)}^{1,v}$, отсюда и суперпозиция $\alpha' \circ \beta$, а вместе с ней и функция β' принадлежат классу $C_{(0)}^{1,v}$. При $n \geq 3$ аналогично убеждаемся, что $\beta' \in C_{(0)}^{1,v}$. В результате после конечного числа шагов придем ко включениям $\beta' \in C_{(0)}^{n-1,v}$ и $\beta \in C_{(1)}^{n,v}$.

Но тогда и функция $\delta \circ \beta \in C_{(1)}^{n,v}$. Поскольку она также является параметризацией дуги и $|\delta[\beta(r)]| = \alpha[\beta(r)] = r = |\gamma(r)|$, отсюда следует, что $\gamma = \delta \circ \beta$ и, значит, $\gamma \in C_{(1)}^{n,v}([0, \varrho], 0)$. Но тогда $r^{-1}\gamma(r) = e^{ih(r)} \in C_{(0)}^{n,v}([0, \varrho], 0)$, так что на основании теоремы 3(б) и $h(r) \in C_{(0)}^{n,v}([0, \varrho], 0)$.

Что касается последнего утверждения леммы, то заметим, что образ G_0 радиальной дуги G с концом $\tau = 0$ при преобразовании $z \rightarrow z^\delta$ является радиальной дугой, которая описывается аналогично (2) по отношению к функции $h_0(r) = \delta h(r^{1/\delta})$. Поэтому, как отмечено при доказательстве теоремы 3(в), она принадлежит классу $C_{(0)}^{n,v}$ вместе с h .

При $n = 0$ весовые пространства $C_\lambda^{0,v}$ можно вводить и на дугах G с концом τ , согласно лемме 3 гладкая параметризация $\delta : [0, 1] \rightarrow G$, $\delta(0) = \tau$, осуществляет изоморфизм $C_\lambda^{0,v}(G, \tau) \rightarrow C_\lambda^{0,v}([0, 1], 0)$. При $n \geq 1$ на дугах G класса $C_1^{n,v}$ пространство $C_\lambda^{n,v}(G, \tau)$ естественно вводить с помощью параметризации δ того же класса условием $\phi \circ \delta \in C_\lambda^{n,v}([0, 1], 0)$. Из доказательства леммы 3 видно, что так определенный класс не зависит от выбора параметризации. Аналогично понимаются и классы $C_{(l)}^{n,v}(G, \tau)$.

Если кусочно-гладкий контур Γ , ограничивающий область D , имеет единственную угловую точку, т.е. $F = \{\tau\}$, то совершенно аналогично криволинейным секторам определяются пространства $C_\lambda^{n,v}(\bar{D}, \tau)$ и классы $C_{(1)}^{n,v}(\bar{D}, \tau)$ с сохранением всех описанных выше их свойств. В случае $(\Gamma, \tau) \in C_{(1)}^{n,v}$ эти пространства и классы можно рассматривать и на контуре Γ .

Рассмотрим связь классов $C_{(l)}^{n,v}$ с интегралами типа Коши на единичной окружности.

Лемма 4. Пусть $\mathbb{D}$ есть единичный круг с границей $\mathbb{T}$, ориентированной против часовой стрелки, и точка $\tau \in \mathbb{T}$. Пусть функция $\varphi \in C_{(l)}^{n,v}(\mathbb{T}, \tau)$, $-1 \leq l \leq n$.

Тогда функция

$$\phi(z) = \frac{1}{2\pi i} \int_{\mathbb{T}} \frac{\varphi(t) dt}{t - z}, \quad z \in \mathbb{D}, \quad (21)$$

принадлежит $C_{(l)}^{n,v}(\bar{\mathbb{D}}, \tau)$.

Доказательство проведем сначала для случая $n = 0$, когда l принимает одно из значений $0, -1$. Пусть сначала $l = -1$, тогда $\varphi \in C_{\varepsilon-1}^{0,v}(\mathbb{T}, \tau)$. Как установлено в [3], в этом случае интеграл (21) определяет ограниченный оператор $C_{\varepsilon-1}^{0,v}(\mathbb{T}, \tau) \rightarrow C_{\varepsilon-1}^{0,v}(\bar{\mathbb{D}}, \tau)$.

Пусть далее $l = n = 0$. Тогда $\varphi_0 = \varphi - c \in C_{\varepsilon}^{0,v}(\mathbb{T}, \tau)$. Из очевидного тождества

$$\frac{1}{t - z} - \frac{1}{t - \tau} = \frac{z - \tau}{(t - \tau)(t - z)}$$

следует, что $\phi(z) = c + c_0 + \phi_0$, где положено

$$c_0 = \frac{1}{2\pi i} \int_{\mathbb{T}} \frac{\varphi_0(t) dt}{t - \tau}, \quad \phi_0(z) = \frac{z - \tau}{2\pi i} \int_{\mathbb{T}} \frac{\varphi_0(t) dt}{(t - \tau)(t - z)}.$$

Поскольку $(t - \tau)^{-1} \varphi_0(t) \in C_{\varepsilon-1}^{0,v}(\mathbb{T}, \tau)$, из тех же соображений утверждение леммы справедливо и в этом случае.

Пусть далее $n \geq 1$. Условимся производную φ' на $\mathbb{T}$ понимать по отношению к комплексному параметру, т.е.

$$\varphi'(t_0) = \lim_{t \rightarrow t_0} \frac{\varphi(t) - \varphi(t_0)}{t - t_0}, \quad t_0 \in \mathbb{T}.$$

Тогда, дифференцируя равенство (21) и интегрируя по частям, получим аналогичное выражение

$$\phi'(z) = \frac{1}{2\pi i} \int_{\mathbb{T}} \frac{\varphi'(t) dt}{t - z}, \quad z \in \mathbb{D},$$

для производной функции ϕ . Поэтому на основании теоремы 2(а) дело сводится к случаю, когда n и l уменьшены на единицу. Повторяя этот процесс, приходим к двум случаям, когда либо $l = n = 0$, либо $l = -1, n \geq 0$.

Первый случай, а также случай $l = -1, n = 0$ разобраны выше. Пусть $l = -1$ и $n \geq 1$, так что $\varphi \in C_{\varepsilon-1}^{n,v}$. Вводя обозначение $\mathcal{D}\varphi = (z - \tau)\varphi'(z)$ для весовой производной, индуктивное определение (8) весового пространства $C_{\varepsilon-1}^{n,v}$ можем записать в форме $\phi, \mathcal{D}\phi \in C_{\varepsilon-1}^{n-1,v}$ или, что равносильно, в форме

$$\mathcal{D}^k \phi \in C_{\varepsilon-1}^{0,v}, \quad 0 \leq k \leq n. \quad (22)$$

Воспользуемся тождеством

$$(z - \tau) \frac{\partial}{\partial z} \left(\frac{1}{t - z} \right) = - \frac{\partial}{\partial t} \left(\frac{t - \tau}{t - z} \right).$$

Как и выше, из него с помощью интегрирования по частям приходим к равенству

$$(\mathcal{D}\phi)(z) = \frac{1}{2\pi i} \int_{\mathbb{T}} \frac{(\mathcal{D}\varphi)(t) dt}{t - z}.$$

Повторяя эту процедуру, в соответствии с (22) отсюда приходим к справедливости леммы и в рассматриваемом случае.

Прежде чем переходить к доказательству теоремы 2 в общем случае, рассмотрим отдельно ситуацию, когда область D ограничена простым гладким контуром Γ , множество F состоит из одной точки τ , и $\mathbb{D}$ является единичным кругом.

Теорема 4. Пусть контур Γ гладкий и $(\Gamma, \tau) \in C_{(1)}^{n,v}$, $0 < v < 1$.

Тогда конформное отображение $\omega: D \rightarrow \mathbb{D}$, $\omega(\tau) = 1$, принадлежит $C_{(1)}^{n,v}(\bar{D}, \tau)$ и, соответственно, обратное отображение $\sigma = \omega^{-1} \in C_{(1)}^{n,v}(\bar{\mathbb{D}}, 1)$.

Доказательство достаточно провести для σ , тогда принадлежность ω классу $C_{(1)}^{n,v}(\bar{D}, \tau)$ устанавливается совершенно аналогично лемме 3.

Действительно, пусть $\sigma \in C_{(1)}^{n,v}(\bar{\mathbb{D}}, 1)$, так что по теореме 3(б) функция

$$(\sigma')^{-1} \in C_{(0)}^{n-1,v}(\bar{\mathbb{D}}, 1). \quad (23)$$

Воспользуемся равенством

$$\omega' = (\sigma')^{-1} \circ \omega. \quad (24)$$

По теореме Келлога функция ω удовлетворяет двустороннему условию Гёльдера, и можно воспользоваться леммой 2, согласно которой $\omega' \in C_{(0)}^{0,v}(\bar{D}, \tau)$ и, значит, $\omega \in C_{(1)}^{1,v}(\bar{D}, \tau)$. При $n \geq 2$ в силу (23) обе функции в правой части (24) принадлежат классу $C_{(1)}^{1,v}(\bar{D}, \tau)$. Поэтому их композиция $\omega' \in C_{(0)}^{1,v}(\bar{D}, \tau)$ и, значит, $\omega \in C_{(1)}^{2,v}(\bar{D}, \tau)$. Повторяя эту процедуру, после конечного числа шагов получим $\omega \in C_{(1)}^{n,v}(\bar{D}, \tau)$.

Итак, достаточно доказать, что $\sigma \in C_{(1)}^{n,v}(\bar{\mathbb{D}}, 1)$. Пусть $e(t)$ означает единичный касательный вектор к контуру Γ в точке t , направление которого совпадает с обходом контура против часовой стрелки. Тогда в силу леммы 3 функция

$$e(t) \in C_{(0)}^{n,v}(\Gamma, \tau). \quad (25)$$

По отношению к единичной окружности $\mathbb{T}$ аналогичный вектор $e(t) = it$. Известно (см. [1]), что аргумент производной σ' на граничной окружности $|t| = 1$ единичного круга $\mathbb{D}$ дается равенством

$$\arg \sigma'(t) = \arg e[\sigma(t)] - \arg t - \pi/2, \quad |t| = 1. \quad (26)$$

Поскольку σ удовлетворяет двустороннему условию Липшица, в силу леммы 2 из этого равенства следует, что $\arg \sigma' \in C_{(0)}^{0,v}$. В единичном круге функция $-i \ln \sigma'$, реальная часть которой на границе совпадает с $\arg \sigma'$, восстанавливается по формуле Шварца

$$-i \ln \sigma'(z) = \frac{1}{\pi} \int_{\mathbb{T}} \frac{\arg \sigma'(t) dt}{t - z} + ic, \quad c \in \mathbb{R}.$$

Поэтому на основании леммы 4 и теоремы 3 функция $\sigma \in C_{(1)}^{1,v}$. Но тогда при $n \geq 2$ из (26) следует, что $\arg \sigma' \in C_{(0)}^{1,v}$, так что из тех же соображений $\sigma \in C_{(1)}^{1,v}$. Повторяя эту процедуру, после конечного числа шагов получим $\sigma \in C_{(1)}^{n,v}$.

Доказательство теоремы 2. В силу теоремы 1, не ограничивая общности, можно считать, что область D ограничена простым контуром с одной угловой точкой $\tau = 0$, т.е. $(\Gamma, \tau) \in C_{(1)}^{n,v}$. Соответственно в качестве области $\mathbb{D}$ можно выбрать единичный круг. Можно также считать, что функция $\zeta = z^\delta$, где $\delta = \pi/\theta_0$, однолистка в области D и отображает ее на область D_0 , ограниченную гладким контуром Γ_0 . В силу леммы 3 этот контур также принадлежит классу $C_{(1)}^{n,v}$.

Запишем $\omega(z) = \omega_0(z^\delta)$, где ω_0 означает конформное отображение $D_0 \rightarrow \mathbb{D}$. В силу теоремы 4 имеем

$$\omega_0(z) = zA_0(z), \quad A_0 \in C_{(0)}^{n,v}(\mathbb{D}, 0); \quad \omega'_0(z) = B_0(z), \quad B_0 \in C_{(0)}^{n-1,v}(\mathbb{D}, 0).$$

Следовательно, $\omega(z) = z^\delta A_0(z^\delta)$, $\omega'(z) = \delta z^{\delta-1} B_0(z^\delta)$, так что на основании теоремы 3(в) отсюда приходим к справедливости теоремы с

$$A(z) = A_0(z^\delta) \in C_{(0)}^{n,v}(D, 0), \quad B(z) = \delta z^{\delta-1} B_0(z^\delta) \in C_{(0)}^{n-1,v}(D, 0).$$

СПИСОК ЛИТЕРАТУРЫ

1. Голузин Г.М. Геометрическая теория функций комплексного переменного. М.: Наука, 1972.
2. Векуа И.Н. Обобщенные аналитические функции. 2-е изд., М.: Наука, 1988.
3. Солдатов А.П. Сингулярные интегральные операторы и эллиптические краевые задачи // Современ. математика. Фундамент. направления. 2017. Т. 63. С. 1–189.

МАТЕМАТИЧЕСКАЯ
ФИЗИКА

УДК 532.59

**ВНУТРЕННИЕ ГРАВИТАЦИОННЫЕ ВОЛНЫ
ОТ ЛОКАЛИЗОВАННОГО ИСТОЧНИКА В ПОТОКЕ
СТРАТИФИЦИРОВАННОЙ СРЕДЫ С МОДЕЛЬНЫМ
РАСПРЕДЕЛЕНИЕМ ЧАСТОТЫ ПЛАВУЧЕСТИ¹⁾**

© 2023 г. В. В. Булатов^{1,*}, И. Ю. Владимиров^{2,**}

¹ 119526 Москва, пр-т Вернадского, 101–1, ИПМ им. А.Ю. Ишлинского РАН, Россия

² 117997 Москва, Нахимовский пр-т, 36, Институт океанологии им. П.П. Ширшова РАН, Россия

*e-mail: internalwave@mail.ru

**e-mail: iyuvladimirov@rambler.ru

Поступила в редакцию 30.01.2023 г.

Переработанный вариант 02.02.2023 г.

Принята к публикации 30.03.2023 г.

Рассмотрена задача о расчете полей внутренних гравитационных волн, генерируемых локализованным источником возмущений в потоке стратифицированной среды конечной толщины с модельным распределением частоты плавучести. Используя аналитические представления частоты плавучести, получена неявная форма дисперсионного соотношения, аналитическое представление которой зависит от функций Бесселя действительного индекса. Приведены результаты численных расчетов дисперсионных кривых, линий равной фазы и волновых амплитуд для различных волновых мод и скоростей стратифицированного потока. Толщина стратифицированного слоя, вертикальный масштаб частоты плавучести и величина скорости потока являются основными факторами, влияющими на амплитудно-фазовую пространственную трансформацию возбуждаемых вниз по потоку волновых полей. Библ. 24. Фиг. 10.

Ключевые слова: стратифицированная среда, внутренние гравитационные волны, частота плавучести, волновая мода, источник возмущений.

DOI: 10.31857/S0044466923080045, **EDN:** WRZZHG

ВВЕДЕНИЕ

Геофизическая гидродинамика предлагает большой ряд задач по изучению распространения волн в среде, анизотропия которой связана с эффектами стратификации (см. [1–4]). Изменчивость плотности воды в океане, вызванное, в частности, его термохалинной структурой, приводит к генерации внутренних гравитационных волн (ВГВ). В настоящее время растет интерес к исследованию динамики ВГВ для решения фундаментальных задач физической океанологии, а также геофизических, промышленных и экологических приложений (см. [4–6]). Например, для изучения влияния крупномасштабных ВГВ на глобальные энергетические обмены в океане проводятся математическое моделирование, натурные наблюдения и лабораторные эксперименты. Исследования, связанные с приложениями, с одной стороны, в основном сосредоточены на изучении воздействия ВГВ на морские технологические сооружения. С другой стороны, представляют определенный интерес пакеты ВГВ, генерируемые различными источниками возмущений природного и антропогенного характеров (см. [7–10]). Такие пакеты ВГВ могут иметь регулярный характер и сохранять неизменной свою структуру в течение нескольких дней, что дает возможность их обнаружения с помощью радаров с синтезированной апертурой и инфракрасных детекторов (см. [4, 6, 7, 11]). В [8] представлен обзор наблюдений ВГВ больших амплитуд, играющих важную роль в глобальной картине перемешивания вод Мирового океана.

Актуальность исследования возбуждаемых ВГВ заключается также в том, что понимание физических свойств наблюдаемых в океане турбулентных следов остается актуальной и сложной за-

¹⁾ Работа выполнена при финансовой поддержке РНФ (код проекта № 23-21-00194).

дачей физической океанографии. Соответствующие исследования могут предоставить важную информацию об особенностях полей океанических течений и гидродинамических параметрах источников возбуждения, в том числе для разработки эффективных алгоритмов обнаружения. Даже незначительное ускорение или изменение направления движущегося источника возмущений может привести к заметной трансформации возникающих крупномасштабных амплитудно-фазовых структур волнового следа (см. [7, 8, 12–14]).

В современных научных исследованиях при анализе динамики ВГВ в природных стратифицированных средах (океан, атмосфера Земли) широко используются аналитические модели волновой генерации (см. [1, 4, 7, 16–20]). Численные модели не позволяют эффективно рассчитывать конкретные физические задачи волновой динамики океана и атмосферы с учетом их реальной изменчивости, ориентированы на решение достаточно общих задач, требуют большой вычислительной мощности, не всегда учитывают физическую специфику решаемых задач, что существенно ограничивает их практическую применимость. Кроме того, использование мощных численных алгоритмов требует верификации и сравнения с решениями модельных задач. В результате проведения модельных расчетов волновая система может быть приближена к наблюдаемым в натурных условиях волновым картинам, что позволяет оценить физические параметры реальных источников возбуждения ВГВ в природных стратифицированных средах (см. [4–6, 10–12]). В линейном приближении существующие подходы к описанию волновой картины возбуждаемых полей ВГВ основаны на представлении волновых полей интегралами Фурье и анализе получаемых решений (см. [1–4, 7, 17, 18]). Целью настоящей работы является изучение внутренних гравитационных волн, возникающих в потоке стратифицированной среды конечной толщины с модельным распределением частоты плавучести.

1. ПОСТАНОВКА ЗАДАЧИ. ИНТЕГРАЛЬНЫЕ ФОРМЫ РЕШЕНИЙ

Рассматривается невязкая, несжимаемая, стратифицированная по вертикали среда, которая в невозмущенном состоянии имеет плотность $\rho_0(z)$. Стратификация предполагается устойчивой: $d\rho_0(z)/dz < 0$. Линеаризованная система уравнений гидродинамики может быть представлена в виде [1–4, 7]

$$\begin{aligned} \frac{\partial U_1}{\partial t} + \frac{1}{\rho_0(z)} \frac{\partial p}{\partial x} = 0, \quad \frac{\partial U_2}{\partial t} + \frac{1}{\rho_0(z)} \frac{\partial p}{\partial y} = 0, \quad \frac{\partial W}{\partial t} + \frac{1}{\rho_0(z)} \frac{\partial p}{\partial z} + \frac{gp}{\rho_0(z)} = 0, \\ \frac{\partial U_1}{\partial x} + \frac{\partial U_2}{\partial y} + \frac{\partial W}{\partial z} = 0, \quad \frac{\partial p}{\partial t} + \frac{d\rho_0(z)}{dz} W, \end{aligned} \quad (1.1)$$

где g — ускорение свободного падения, (U_1, U_2, W) — вектор скоростей, p и ρ — возмущения давления и плотности от равновесного состояния соответственно. Из системы уравнений (1.1) можно получить одно уравнение, например, для вертикальной компоненты скорости W в виде

$$\begin{aligned} \frac{\partial^2}{\partial t^2} \Delta W + \frac{\partial^4 W}{\partial t^2 \partial z^2} - \frac{N^2(z)}{g} \frac{\partial^3 W}{\partial t^2 \partial z} + N^2(z) \Delta W = 0, \\ \Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}, \end{aligned} \quad (1.2)$$

где $N^2(z) = (g/\rho_0(z))(d\rho_0(z)/dz)$ — квадрат частоты Брента–Вайсяля (частоты плавучести). Оценим третий член уравнения (1.2) по сравнению со вторым. Введем характерный для океана масштаб по вертикали $Z \sim 10^2\text{--}10^3$ м, тогда отношение третьего члена ко второму будет $\delta \sim ZN^2/g$, где $N \sim 10^{-2} \text{ с}^{-1}$ — характерная частота плавучести (см. [1–5]). В реальном океане $\delta \sim 10^{-2}\text{--}10^{-3} \ll 1$, поэтому уравнение (1.2) можно представить в приближении Буссинеска

$$\frac{\partial^2}{\partial t^2} \left(\Delta + \frac{\partial^2}{\partial z^2} \right) W + N^2(z) \Delta W = 0.$$

Данное приближение является общепринятым в большинстве океанологических задач, связанных с изучением динамики линейных ВГВ (см. [1–7]). Интересно отметить, что в случае $N(z) = \text{const}$ и глубины дна H , значение δ можно оценить следующим образом:

$\delta = H^2 N^2 / (gH) = \pi^2 c^2 / C^2$, где $C = \sqrt{gH}$ — скорость длинных поверхностных гравитационных волн на поверхности жидкости глубины H , $c = NH/\pi$ — максимальная групповая скорость ВГВ. В реальном океане эти величины соответственно равны $C \sim 10^2$ м/с, $c \sim 1$ м/с (см. [1–4, 7]).

Рассматривается стационарная картина линейных внутренних гравитационных волн, генерируемых точечным источником мощности q в потоке стратифицированной среды, имеющим скорость V , занимающую область (слой) $0 < z < H$, ось z направлена вниз. Источник расположен в точке $(0, 0, z_0)$, $0 < z_0 < H$. Тогда смещение изопикн $\xi = \xi(x, y, z, z_0)$, где $W = \partial \xi / \partial \xi$, в приближении Буссинеска и “твердой крышки” описывается уравнением (см. [1, 7, 19])

$$\left(V \frac{\partial}{\partial x}\right)^2 \left(\Delta + \frac{\partial^2}{\partial z^2}\right) \xi + N^2(z) \Delta \xi = q V \delta'(x) \delta(y) \delta'(z - z_0), \quad (1.3)$$

$$\xi = 0 \quad \text{при} \quad z = 0, H.$$

Далее будем рассматривать модельное представление частоты плавучести в виде $N(z) = 0$ при $0 \leq z \leq h$, $N(z) = N_0 \exp(-\alpha(z - h))$ при $h \leq z \leq H$, которое применяется в прикладных океанологических расчетах для изучения динамики ВГВ, в частности, при наличии сезонного термоклина толщиной несколько десятков метров (см. [1, 5, 6, 8, 10]). Параметр α характеризует вертикальный масштаб стратификации.

Замечательной особенностью Мирового океана является наличие такого слоя — пространственной области сравнительно быстрого изменения температуры и в то же время области большой устойчивости частоты плавучести. Зависимость частоты Брента–Вайсяля от глубины в модельном представлении может отличаться от эмпирических зависимостей, которые, однако, могут характеризоваться наличием в верхнем слое океана слоя с почти постоянной стратификацией, лежащем на подстилающем слое слабой стратификации и большой глубины. Модельное распределение частоты плавучести позволяет исследовать задачу аналитически, в то время как использование эмпирических зависимостей требует применения только численных методов. Однако, как показывают ряд исследований, основные качественные результаты по описанию динамики ВГВ, как правило, не сильно зависят от конкретной аналитической формы аппроксимации частоты плавучести (см. [5–7, 17, 18, 21]).

Фурье-образ возвышения изопикн

$$\Phi(\mu, \nu, z, z_0) = \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} \exp(i(\nu y + \mu x)) \xi(x, y, z, z_0) dx$$

удовлетворяет уравнению

$$\frac{\partial^2 \Phi}{\partial z^2} + k^2 \left(\frac{N^2(z)}{(\mu V)^2} - 1 \right) \Phi = \frac{i q}{\mu V} \delta'(z - z_0), \quad (1.4)$$

$$\Phi = 0 \quad \text{при} \quad z = 0, H, \quad k^2 = \mu^2 + \nu^2.$$

Решение (1.4) ищется в виде ряда по собственным функциям $\varphi_n(z, k)$ соответствующей однородной спектральной задачи

$$\frac{\partial^2 \varphi_n(z, k)}{\partial z^2} + k^2 \left(\frac{N^2(z)}{\omega_n^2(k)} - 1 \right) \varphi_n(z, k) = 0, \quad (1.5)$$

$$\varphi_n(z, k) = 0 \quad \text{при} \quad z = 0, H.$$

Задача (1.5) имеет полную ортонормированную (с весом $N^2(z)$) систему собственных функций $\varphi_n(z, k)$ и соответствующий набор собственных значений $\omega_n(k)$ (дисперсионных кривых) (см. [1, 4, 7]). Для использованной модельной стратификации собственные функции $\varphi_n(z, k)$ могут быть найдены аналитически (см. [16, 22, 23]):

$$\varphi_n(z, k) = \frac{1}{L_n} \text{sh}(kz) \quad \text{при} \quad 0 \leq z \leq h,$$

$$\varphi_n(z, k) = \frac{1}{L_n} (C_n^1 J_\lambda(\beta_n(z)) + C_n^2 Y_\lambda(\beta_n(z))) \quad \text{при} \quad h \leq z \leq H,$$

$$C_n^{1,2} = \frac{d_n^{1,2}}{D_n}, \quad \beta_n(z) = \lambda \frac{N(z)}{\omega_n(k)}, \quad \lambda = \frac{k}{\alpha},$$

$$d_n^1 = \frac{\omega}{N_0} \operatorname{ch}(kh) Y_\lambda(\beta_n(h)) + \operatorname{sh}(kh) Y'_\lambda(\beta_n(h)),$$

$$d_n^2 = -\frac{\omega}{N_0} \operatorname{ch}(kh) J_\lambda(\beta_n(h)) - \operatorname{sh}(kh) J'_\lambda(\beta_n(h)),$$

$$D_n = J_\lambda(\beta_n(h)) Y'_\lambda(\beta_n(h)) - J'_\lambda(\beta_n(h)) Y_\lambda(\beta_n(h)),$$

$$L_n = \left(\int_h^H N^2(z) (C_n^1 J_\lambda(\beta_n(z)) + C_n^2 Y_\lambda(\beta_n(z)))^2 dz \right)^{1/2},$$

где J_λ, Y_λ — функции Бесселя действительного индекса λ . Тогда из условия обращения в нуль собственной функции $\varphi_n(z, k)$ при $z = H$ можно получить неявную форму дисперсионного соотношения

$$d_n^1 J_\lambda \left(\lambda \frac{N(H)}{\omega_n(k)} \right) + d_n^2 Y_\lambda \left(\lambda \frac{N(H)}{\omega_n(k)} \right) = 0,$$

решение которого представляет определенную вычислительную трудность. При малых значениях волнового числа k аналитические свойства дисперсионных кривых $\omega_n(k)$ (в виде ряда по нечетным степеням k) определяются из задачи (1.5) с помощью метода возмущений (см. [1, 4, 7]). При значениях k , отличных от нуля, можно при достаточно малых значениях параметра α , характеризующего вертикальный масштаб стратификации, построить равномерную (по параметру $\lambda = k/\alpha$) асимптотику функции $J_\lambda(\lambda\tau)$ при $\lambda \rightarrow \infty$ и произвольных значениях τ . Действительно, функция $F = J_\lambda(\lambda\tau)$ удовлетворяет уравнению Бесселя порядка λ (см. [22, 23])

$$\tau^2 F'' + \tau F' + \lambda^2 (\tau^2 - 1) F = 0. \quad (1.6)$$

Перейдем в (1.6) к новой функции $F = u(\tau)/\sqrt{\tau}$. В результате можно получить

$$u'' + \left(\lambda^2 q_1(\tau) + q_2(\tau) \right) u = 0, \quad q_1(\tau) = 1 - \frac{1}{\tau^2}, \quad q_2(\tau) = \frac{1}{4\tau^2}. \quad (1.7)$$

Далее, следуя общим алгоритмам построения асимптотик функций Бесселя, выполним переход к новой независимой переменной s и новой функции $v(s)$: $s = \varphi(\tau)$, $v = \psi(\tau)u(\tau)$, где функции $\varphi(\tau)$, $\psi(\tau)$ подлежат определению (см. [24]). Тогда из (1.7) можно получить

$$\frac{d^2 v}{ds^2} + \left(\varphi'' - \frac{2\psi'\varphi'}{\psi} \right) \frac{dv}{ds} + \left(\lambda^2 \frac{q_1}{(\varphi')^2} + \frac{q_2}{(\varphi')^2} - \frac{\psi''}{\psi(\varphi')^2} + \frac{2(\psi')^2}{\psi^2(\varphi')^2} \right) v = 0. \quad (1.8)$$

Для преобразования уравнения (1.8) потребуем, чтобы коэффициент при dv/ds был равен нулю: $\varphi'' - 2\psi'\varphi'/\psi = 0$. Тогда, интегрируя это уравнение, имеем $\psi = \sqrt{\varphi'}$. Далее, требуя, чтобы $\lambda^2 q_1/(\varphi')^2 = s = \varphi(\tau)$, можно получить явные выражения для функции $\varphi(\tau)$ в виде

$$\begin{aligned} \varphi(\tau) &= \left[\frac{3}{2} \lambda \left(\sqrt{\tau^2 - 1} - \arccos \frac{1}{\tau} \right) \right]^{2/3} \quad \text{при} \quad \tau > 1, \\ \varphi(\tau) &= - \left[\frac{3}{2} \lambda \left(-\sqrt{1 - \tau^2} - \ln \frac{1 + \sqrt{1 - \tau^2}}{\tau} \right) \right]^{2/3} \quad \text{при} \quad \tau < 1. \end{aligned}$$

В результате уравнение (1.8) представимо в форме

$$\frac{d^2 v}{ds^2} + (s + \delta)v = 0, \quad \delta = \frac{q_2}{(\varphi')^2} - \frac{\psi''}{\psi(\varphi')^2} + \frac{2(\psi')^2}{\psi^2(\varphi')^2},$$

где $\delta = O(1/\sqrt{\lambda})$ и $s = \varphi(\tau) = O(\lambda^{2/3})$ при $\lambda \rightarrow \infty$. Поэтому для главного члена асимптотики получается уравнение Эйри: $d^2 v/ds^2 + sv = 0$, ограниченное решение которого $v(s) = C \operatorname{Ai}(-s) = C \operatorname{Ai}(-\varphi(\tau))$ (см. [22, 23]). Тогда асимптотика функции $J_\lambda(\lambda\tau)$ при $\lambda \rightarrow \infty$ имеет вид

$$J_\lambda(\lambda\tau) \sim \frac{u}{\sqrt{\tau}} = \frac{C \operatorname{Ai}(-\varphi(\tau))}{\psi(\tau)\sqrt{\tau}}, \quad \psi(\tau) = \sqrt{\frac{\lambda}{\tau}} \sqrt{\frac{\tau^2 - 1}{\varphi(\tau)}}.$$

Значение постоянной C определяется из условий сшивки соответствующих асимптотических разложений. В результате можно получить равномерную по переменной τ на луче $[0; +\infty)$ асимптотику функции Бесселя $J_\lambda(\lambda\tau)$ при $\lambda \rightarrow \infty$ в виде

$$J_\lambda(\lambda\tau) \sim \sqrt{\frac{2}{\lambda}} \operatorname{Ai}(-\varphi(\tau)) \sqrt{\frac{\varphi(\tau)}{\tau^2 - 1}}.$$

Асимптотический анализ поведения функций Бесселя Y_λ можно провести аналогично [24], что дает возможность эффективно рассчитывать неявную форму дисперсионного соотношения. Далее, в силу полноты и ортонормированности собственных функций $\varphi_n(z, k)$, используя соотношение $\delta'(z - z_0) = -N^2(z) \sum_{n=1}^{\infty} \varphi_n(z, k) \partial \varphi_n(z_0, k) / \partial z_0$, окончательно получаем решение задачи (1.3) в виде суммы волновых мод: $\xi = \sum_{n=1}^{\infty} \xi_n$, где

$$\xi_n = \frac{iqV}{4\pi^2} \int_{-\infty}^{\infty} dv \int_{-\infty}^{\infty} \exp(-i(vy + \mu x)) \frac{\mu \omega_n^2(k)}{(\mu^2 V^2 - \omega_n^2(k))k^2} \varphi_n(z, k) \frac{\partial \varphi_n(z_0, k)}{\partial z_0} d\mu. \quad (1.9)$$

2. ЧИСЛЕННЫЕ РЕЗУЛЬТАТЫ

Чтобы удовлетворить условию излучения, т.е. условию отсутствия волнового возмущения вверх по потоку, контур интегрирования по переменной μ следует сместить в верхнюю полуплоскость, тогда при $x < 0$ интеграл (1.9) экспоненциально мал. При $x > 0$ контур интегрирования в (1.9) замыкается в нижнюю полуплоскость. В этом случае основной вклад будет определяться полюсами подынтегральной функции на действительной оси, которые находятся из уравнения $\mu^2 V^2 = \omega_n^2(\sqrt{\mu^2 + v^2})$. Учитывая вычеты в полюсах $\mu = \pm \mu_n(v)$, для отдельной моды возвышения изопокин можно получить представление в виде

$$\xi_n = \frac{qV}{4\pi^2} \int_{-\infty}^{\infty} A_n [\exp(-i(vy + \mu_n(v)x)) + \exp(i(vy - \mu_n(v)x))] \varphi_n(z, k) \frac{\partial \varphi_n(z_0, k)}{\partial z_0} dv,$$

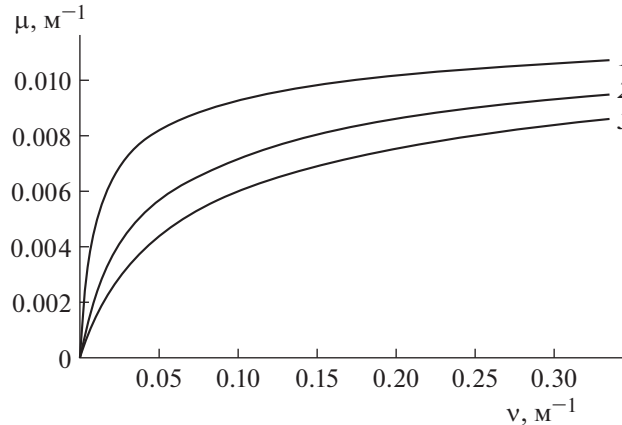
$$A_n = \frac{\omega_n^2(k)}{(V^2 - \omega_n(k)\omega'_n(k)/k)k^2}, \quad k \equiv k(v) = \sqrt{\mu_n^2(v) + v^2}.$$

Используя соотношение

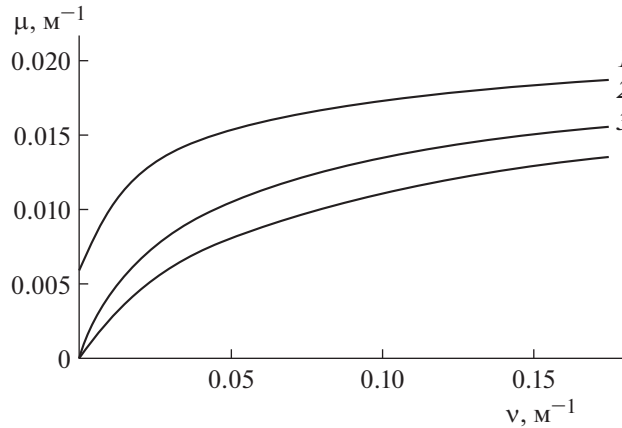
$$\frac{\omega_n(k)\omega'_n(k)}{k} = \frac{\mu_n(v)\mu'_n(v)V^2}{\mu_n(v)\mu'_n(v) + v},$$

окончательно имеем

$$\xi_n = \frac{qV}{\pi} \int_0^{\infty} \frac{\mu_n^2(v)}{\mu_n^2(v) + v^2} \left(\frac{\mu_n(v)\mu'_n(v)}{v} + 1 \right) \cos(\mu_n(v)x) \cos(vy) f_n(z, v) \frac{\partial f_n(z_0, v)}{\partial z_0} dv,$$



Фиг. 1. Дисперсионные кривые $\mu_n(v)$ первых трех мод: 1 – первая мода, 2 – вторая мода, 3 – третья мода. Сверхкритический режим.

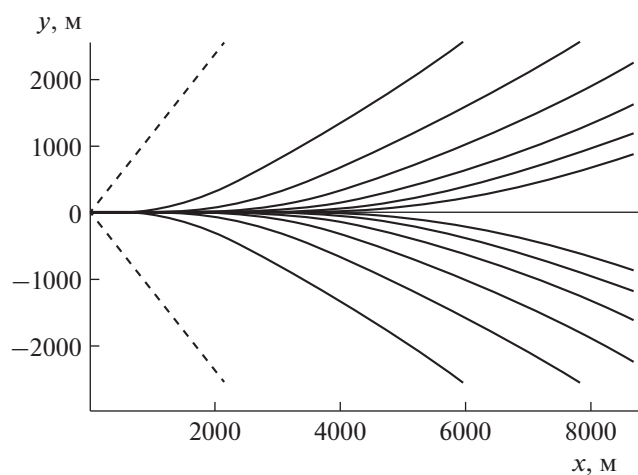


Фиг. 2. Дисперсионные кривые $\mu_n(v)$ первых трех мод: 1 – первая мода, 2 – вторая мода, 3 – третья мода. До-критический режим.

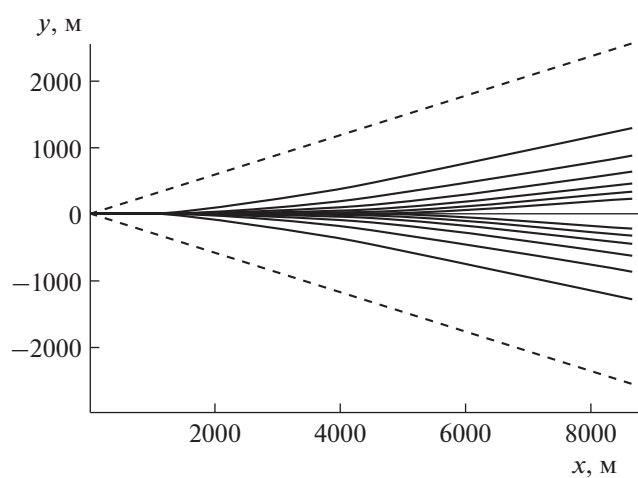
$$f_n(z, v) \equiv \varphi_n \left(z, \sqrt{\mu_n^2(v) + v^2} \right).$$

На иллюстрациях приведены результаты расчетов дисперсионных кривых (фиг. 1, 2), линий равных фаз (фиг. 3–5, 7–9) и волновых полей возвышения изопикн (фиг. 6, 10) первых трех мод для различных значений скорости потока: $M = 1.3$ (фиг. 1, 3–6), $M = 0.7$ (фиг. 2, 7–10), где M – отношение скорости стратифицированного потока V к максимальной групповой скорости первой волновой моды ВГВ. Остальные параметры расчетов, характерные для условий реального океана, были следующие (см. [1–7]): $N_0 = 0.015 \text{ с}^{-1}$, $\alpha_0 = 0.01 \text{ м}^{-1}$, $h = 50 \text{ м}$, $H = 1600 \text{ м}$, $z = 40 \text{ м}$ (горизонт наблюдения), $z_0 = 60 \text{ м}$ (горизонт источника). Сплошные линии на фиг. 3–5, 7–9 – линии равной фазы с шагом 2π , штриховые – волновые фронты.

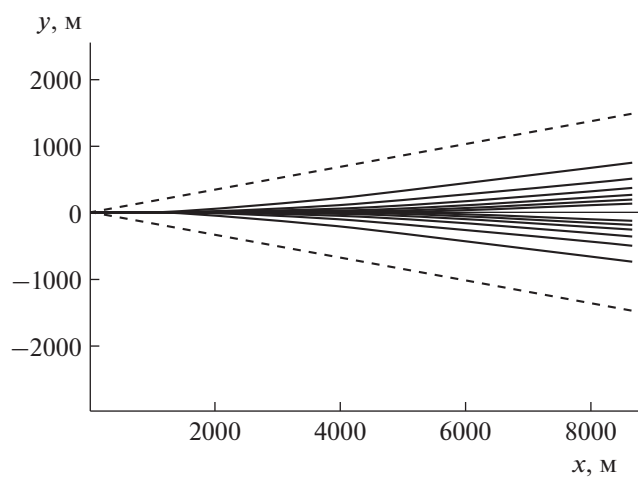
Для заданных параметров численное исследование дисперсионных соотношений показывает, что при $M = 0.7$ только у первой волновой моды значение в нуле положительно ($\mu_1(0) > 0$), у остальных мод $\mu_n(0) = 0$, $n > 1$. Максимальная групповая скорость первой моды больше скорости потока V , максимальные групповые скорости других мод меньше скорости потока. Поэтому полученные результаты можно интерпретировать следующим образом: стратифицированный поток при $M = 0.7$ является докритическим для первой моды (его скорость меньше максимальной



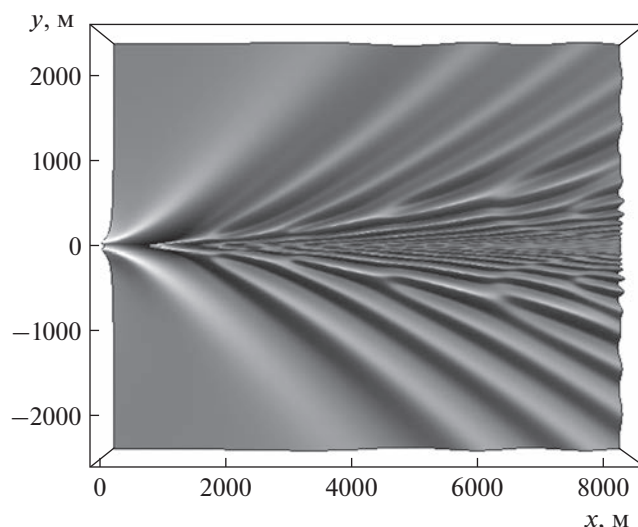
Фиг. 3. Линии равной фазы первой моды.



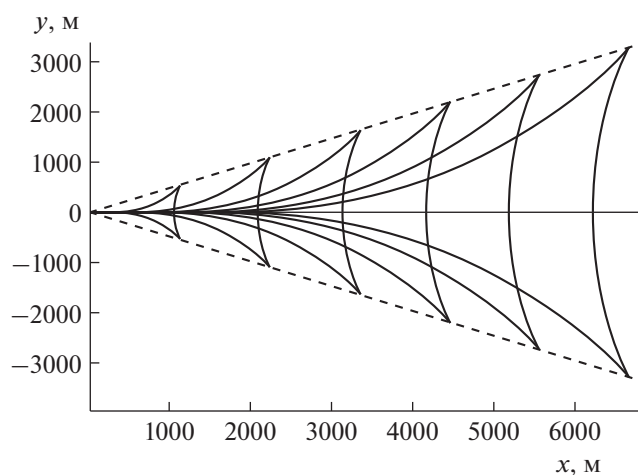
Фиг. 4. Линии равной фазы второй моды.



Фиг. 5. Линии равной фазы третьей моды.

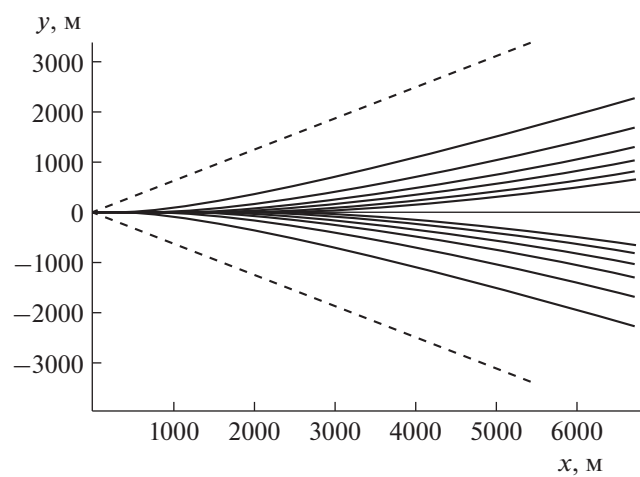


Фиг. 6. Возвышение изопикн: сумма трех первых мод.

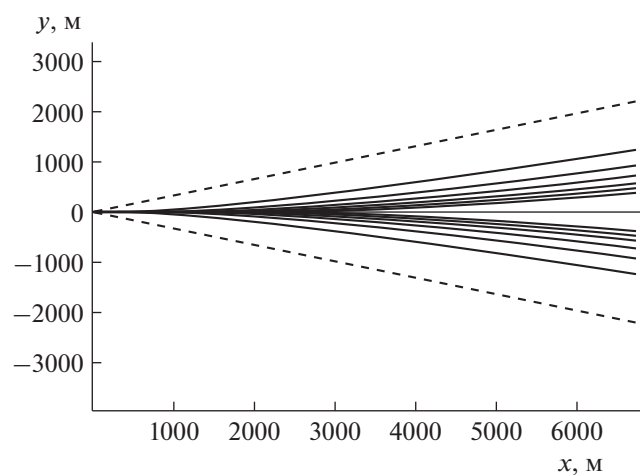


Фиг. 7. Линии равной фазы первой моды.

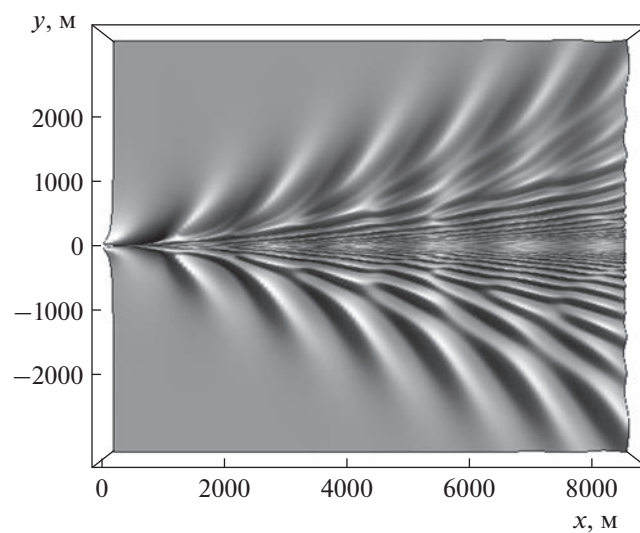
групповой скорости волны первой моды) и сверхкритическим для других мод (максимальная групповая скорость которых меньше скорости потока). Это означает, в частности, что волновая система первой моды (с наибольшим углом полураствора) состоит как из продольных (клиновидных), так и из поперечных (кольцевых) волн (фиг. 7). Остальные моды (с меньшими углами полураствора) представляют собой только продольные (клиновидные) волны (фиг. 8, 9). Известно, что максимальные групповые скорости ВГВ при увеличении номера моды n убывают как n^{-1} , поэтому только конечное число волновых мод могут генерировать поперечные (кольцевые) волны (см. [1, 7, 16, 17]). При $M > 1$ все волновые моды находятся внутри клиновидной области, ограниченной углом полураствора первой моды вниз по потоку. Определяющий вклад в полное поле вносят только первые три моды, суммарный вклад остальных мод не превышает 5%. Такие параметры течения, как толщина стратифицированного слоя h , значение вертикального масштаба стратификации α , величина скорости потока V , являются основными факторами, влияющими на амплитудно-фазовую пространственную трансформацию возбуждаемых вниз по потоку волновых полей.



Фиг. 8. Линии равной фазы второй моды.



Фиг. 9. Линии равной фазы третьей моды.



Фиг. 10. Возвышение изопоикн: сумма трех первых мод.

ЗАКЛЮЧЕНИЕ

Построенные в работе решения позволяют описать амплитудно-фазовые характеристики полей внутренних гравитационных волн от локализованного источника возмущений в потоке стратифицированной среды конечной толщины с модельным распределением частоты плавучести. Получена неявная форма дисперсионного соотношения, аналитическое представление которого выражается через функции Бесселя действительного индекса. Приведены результаты численных расчетов дисперсионных кривых, линий равной фазы и волновых амплитуд для различных скоростей потока и волновых мод. Наибольший выигрыш при использовании данного подхода можно получить при исследовании эволюции волновых пакетов, возбуждаемых распределенными в пространстве возмущениями, так как, используя операцию свертки, построенные в работе решения позволяют рассчитывать волновые поля от нелокальных движущихся источников возмущений различной физической природы. Полученные результаты с различными значениями входящих в них параметров позволяют провести оценку характеристик пакетов внутренних гравитационных волн, наблюдаемых в реальных океанических условиях. Такие волновые картины полей могут наблюдаться при дистанционном зондировании, наблюдении и измерениях внутренних гравитационных волн, возбуждаемых различными источниками возмущений в природных (океан, атмосфера Земли) и искусственных стратифицированных средах.

СПИСОК ЛИТЕРАТУРЫ

1. *Мирополюский Ю.З.* Динамика внутренних гравитационных волн в океане. Ленинград: Гидрометеоздат, 1981. 302 с.
2. *Pedlosky J.* Waves in the ocean and atmosphere: introduction to wave dynamics. Berlin-Heidelberg: Springer, 2010. 260 p.
3. *Sutherland B.R.* Internal gravity waves. Cambridge: Cambridge Univ. Press, 2010. 394 p.
4. *Ozsoy E.* Geophysical fluid dynamics II. Stratified rotating fluid dynamics of the atmosphere-ocean. Springer Textbook in Earth Sciences. Geography and Environment. Switzerland AG Cham: Springer Nature, 2021. 323 p.
5. *Morozov E. G.* Oceanic internal tides. Observations, analysis and modeling. Berlin: Springer, 2018. 317 p.
6. *Velarde M.G., Tarakanov R.Yu., Marchenko A.V. (Eds.)* The ocean in motion. Springer Oceanography. Springer International Publishing AG, 2018. 625 p.
7. *Булатов В.В., Владимиров Ю.В.* Волны в стратифицированных средах. М.: Наука, 2015. 735 с.
8. *Сабинин К.Д., Серебряный А.Н.* “Горячие” точки в поле внутренних волн в океане // Акустич. ж. 2007. Т. 53. № 3. С. 410–436.
9. *Abdilghanie A.M., Diamessis P.J.* The internal gravity wave field emitted by a stably stratified turbulent wake // J. Fluid Mech. 2013. V. 720. P. 104–139.
10. *Voelker G.S., Myers P.G., Walter M., Sutherland B.R.* Generation of oceanic internal gravity waves by a cyclonic surface stress disturbance // Dynamics Atm. Oceans. 2019. V. 86. P. 116–133.
11. *Сидняев Н.И.* Теоретические исследования гидродинамики при подводном взрыве точечного источника // Инженерный журнал: наука и инновации. 2013. № 2. URL: <https://engjournal.ru/catalog/aprmath/hidden/614.html>
<https://doi.org/10.18698/2308-6033-2013-2-614>
12. *Беляев М.Ю., Десинов Л.В., Крикалев С.К., Кумакиев С.А., Секерж-Зенькович С.Я.* Идентификация системы океанских волн по фотоснимкам из космоса // Изв. РАН. Теория и системы управления. 2009. № 1. С. 117–127.
13. *Матюшин П.В.* Процесс формирования внутренних волн, инициированных начальным движением тела в стратифицированной вязкой жидкости // Изв. РАН. МЖГ. 2019. № 3. С. 83–97.
14. *Chai J., Wang Z., Yang Z., Wang Z.* Investigation of internal wave wakes generated by a submerged body in a stratified flow // Ocean Engineer. 2022. V. 266. P. 112840.
15. *Wang J., Wang S., Chen X., Wang W., Xu Y.* Three-dimensional evolution of internal waves rejected from a submarine seamount // Phys. Fluid. 2017. V. 29. P. 106601.
16. *Garrett C., Munk W.* Space-time scales of internal waves // Geophys. Fluid Dyn. 1972. V. 3. P. 225–264.
17. *Свиркунов П.Н., Калашиник М.В.* Фазовые картины диспергирующих волн от движущихся локализованных источников // УФН. 2014. Т. 184. № 1. С. 89–100.
18. *Gnevyshev V., Badulin S.* Wave patterns of gravity–capillary waves from moving localized sources // Fluids. 2020. V. 5. P. 219.

19. *Bulatov V., Vladimirov Yu.* Generation of internal gravity waves far from moving non-local source // *Symmetry*. 2020. V. 12 (11). P. 1899.
20. *Булатов В.В., Владимиров Ю.В.* Аналитические решения уравнения внутренних гравитационных волн в полубесконечном слое стратифицированной среды переменной плавучести // *Ж. вычисл. матем. и матем. физ.* 2019. Т. 59. № 5. С. 792–795.
21. *Рындина В.В.* Собственные частоты внутренних волн в жидкости и частота Брента–Вайсяля. Ростов-на-Дону: Изд-во ЦВВР, 2007. 128 с.
22. *Бейтмен Г., Эрдейи А.* Высшие трансцендентные функции. Т. 2. Функции Бесселя, функции параболического цилиндра, ортогональные многочлены. М.: Наука, 1974. 296 с.
23. *Никифоров А.Ф., Уваров В.В.* Специальные функции математической физики. М.: ИД Интеллект, 2008. 344 с.
24. *Доброхотов С.Ю., Миненков Д.С., Назайкинский В.Е.* Представления функций Бесселя с помощью канонического оператора Маслова // *Теор. и матем. физ.* 2021. Т. 208. № 2. С. 196–217.

МАТЕМАТИЧЕСКАЯ ФИЗИКА

УДК 519.63

МОДЕЛИРОВАНИЕ ЭМИССИОННЫХ ПРОЦЕССОВ В СИЛЬНЫХ ЭЛЕКТРОМАГНИТНЫХ ПОЛЯХ

© 2023 г. Т. А. Кудряшова^{1,*}, С. В. Поляков^{1,**}, Н. И. Тарасов^{1,***}

¹ 125047 Москва, Миусская пл., 4, ИПМ им. М.В. Келдыша РАН, Россия

*e-mail: kudryashova@imamod.ru

**e-mail: polyakov@imamod.ru

***e-mail: nikita_tarasov@imamod.ru

Поступила в редакцию 25.02.2023 г.
Переработанный вариант 14.03.2023 г.
Принята к публикации 28.04.2023 г.

Рассмотрена проблема расчета процессов электронной эмиссии с поверхности металлов при сильных электромагнитных полях с учетом релятивистских эффектов. Одним из методов моделирования в данной области является метод частиц, сочетающийся с сеточным расчетом полей на основе уравнений Максвелла. Подобные методики развиваются с 1960-х годов по настоящее время. При этом существующие подходы все еще имеют определенные ограничения. В настоящей работе для аксиально-симметричной геометрии, генерирующей системы, представлена новая численная методика моделирования процессов эмиссии электронов с поверхности металлических катодов. Методика использует представление крупных сглаженных гауссовых частиц и реализует расчеты электромагнитных полей на декартовых пространственных сетках. Программная реализация ориентирована на параллельные вычисления. Целью численных экспериментов было определение параметров электронной эмиссии. В качестве тестовых задач были выбраны диодные и триодные цилиндрические системы. В численных расчетах получены пространственно-временные характеристики релятивистских электронных пучков, порождаемых эмиссионными процессами, в том числе воспроизведен ток Чайлда–Ленгмюра. Разработанная численная методика подтвердила свою корректность и эффективность. Библ. 42. Фиг. 7.

Ключевые слова: электронная эмиссия, сильные электромагнитные поля, сеточные методы, методы частиц, параллельные вычисления.

DOI: 10.31857/S0044466923080100, **EDN:** WSWCXJ

ВВЕДЕНИЕ

Традиционно большой интерес представляют задачи о строении веществ и композитных материалов на их основе (см. [1, 2]). Один из экспериментальных способов исследования новых перспективных материалов основывается на воздействии пучка заряженных частиц на изучаемое вещество. В рамках этой технологии используется интенсивное импульсное воздействие на материал. В связи с этим разрабатываются генераторы сильноточных пучков и приборы на их основе (см., например, [3–6]). Наиболее востребованными являются коротко-импульсные генераторы релятивистских электронных пучков (РЭП) с длительностью импульса в диапазоне микро- и наносекунд (см. [7, 8]), которые не разрушают структуру материала.

Генерация РЭП часто основывается на явлении взрывной эмиссии электронов (ВЭЭ) (см. [9, 10]), которая существенно превосходит по энергии выхода как термоэлектронную (см. [11]), так и полевую автоэлектронную эмиссию (см. [12, 13]). Экспериментальному и теоретическому анализам этого явления посвящена обширная литература (см. [6]). Однако наряду с уже известными закономерностями остаются еще вопросы реализации конкретных приборов на основе ВЭЭ. Они решаются в том числе с помощью методов математического и компьютерного моделирования.

Для математического моделирования процессов, составляющих основу сильноточной электроники, с 70-х годов прошлого века создавались необходимые вычислительные основы (см., например, [14–18]) и соответствующие пакеты программ COMSOL (см. [19]), Ansys (см. [20]),

КАРАТ (см. [21]), МЕЕР (см. [22]) и др. Общий тренд численного анализа данного класса задач базируется на применении сеточных методов контрольных объемов или конечных элементов в пространстве-времени FDTD (Finite Difference Time Domain) к решению уравнений Максвелла (см. [16, 21–24]). В дополнение применяются методы частиц в различных модификациях (см., например, [25–32]).

Современный технический потенциал, увеличение возможностей натуральных и компьютерных экспериментов, а также широкие перспективы практического использования результатов фундаментальных исследований привели к бурному развитию сильноточного приборостроения. С внедрением этого оборудования в научных лабораториях и на производстве возник интерес к более детальному изучению эмиссионных процессов. В частности, это связано с использованием мощных электронных и ионных пучков в задачах напыления тонких пленок и установках нанолитографии.

Как известно, общей задачей практического применения эффекта эмиссии является создание источников электромагнитного излучения для технологических и хозяйственных целей. Среди типов источников различают электронные, ионные, оптические, смешанные и специальные. Одним из массово используемых механизмов эмиссии является электронная эмиссия, которая чаще всего принимает следующие формы:

- термоэлектронная эмиссия, при которой дополнительная энергия сообщается электронам в результате нагрева катода;
- фотоэлектронная эмиссия, при которой на поверхность катода воздействует электромагнитное излучение;
- полевая эмиссия, при которой сильное электрическое поле у поверхности катода создает силы, способствующие выходу электронов за его пределы;
- взрывная эмиссия — это электронно-ионный вариант в сильных полях, явление вырывания ионов с поверхности твердого тела под действие потоков электронов.

В ряде случаев различные механизмы эмиссии реализуются одновременно. Например, при действии тлеющего разряда на катод одновременно возникают четыре вида эмиссии: ионно-электронная, фотоэлектронная и эмиссия под действием быстрых невозбужденных атомов. Комбинация вторичной электронной эмиссии и эмиссии, возникающей в результате действия сильного электрического поля в тонком слое диэлектрика или полупроводника, называется эффектом Мальтера.

В настоящей работе исследуется взрывная эмиссия, сопровождаемая образованием плазмы и пучков релятивистских электронов (РЭП). Общей целью работы является разработка вычислительной технологии и комплекса параллельных программ для моделирования эмиссионных процессов взрывного типа, а в перспективе — численный анализ взаимодействия пучков частиц с различными средами и материалами. Технология включает разработку комплексной математической модели, множества численных методов, параллельных алгоритмов, комплекса расчетных программ для вычислительных кластеров и суперкомпьютеров, цифровой платформы для проведения детальных численных экспериментов. Для верификации разработанной численной методики использовались данные из [33–35].

1. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ

В математической модели необходимо иметь минимально два уровня описания — сплошная среда для полей (электрических, магнитных, тепловых) и описание в рамках моделей частиц. Мотивацией такого подхода является то, что модели сплошной среды позволяют рассматривать системы произвольного размера, которые актуальны для приложений, а модели частиц позволяют исследовать тонкие эффекты, связанные со свойствами сред и материалов.

Итоговым выбором для математического описания проблемы на первом этапе исследования явились: полная система уравнений Максвелла для электромагнитных полей, уравнение теплопроводности в твердых материалах, уравнения гидродинамики как альтернатива моделям частиц и описание крупных частиц, дополняемое при необходимости моделями молекулярной динамики.

В настоящей работе использованы уравнения Максвелла и специальная модель крупных сглаженных частиц. Последняя имеет тесные связи с методами частиц в ячейках (PiC — Particles in Cells) (см. [25–27]), крупных частиц (LPM — Large-Particle Method) (см. [28, 29]), облаков в ячейках (CIC — Cloud In Cell) (см. [30]), сглаженных частиц в гидродинамике (SPH, SPE —

Smoothed-Particles in Hydrodynamics) и электродинамике (Smoothed-Particles in Electrodynamics) (см. [31, 32]). Поясним эти связи.

Во-первых, рассматриваемая нами проблема эмиссии связана с фактом рождения отдельных заряженных частиц (электронов). Появление этих частиц стимулируется средним полем, отнесенным к ячейке сетки, примыкающей к эмиссионной поверхности. Размер ячейки определяет количество рождаемых частиц. Первоначальная форма частиц также зависит от размера ячейки. В этом смысле применяется метод PiC.

Во-вторых, объединение отдельных заряженных частиц в макрочастицу (облако) и дальнейший анализ ее эволюции описываются уже методом крупных частиц, похожим на LPM. Однако специфика этого метода предполагает чередование эйлеровых и лагранжевых расчетных алгоритмов. Нами же используется несколько иной подход, в рамках которого макрочастицы движутся на фоне неподвижной эйлеровой сетки в условиях сохранения массы и заряда макрочастиц на этой сетке. Также нами предполагается возможность туннелирования одних частиц через другие, что не соответствует LPM, но находится в согласии с волновой природой зарядов.

В-третьих, форма частицы в свободной от границ зоне расчетной области имеет гауссов профиль, выходящий за рамки одной ячейки. Такой подход соответствует методу CIC. Однако при приближении к границам форма макрочастицы может искажаться (в методах CIC она обычно фиксирована). Например, при соприкосновении с металлической поверхностью включается механизм поглощения, и часть макрочастицы исчезает. При соприкосновении с диэлектрической поверхностью макрочастица может растекаться вдоль нее или частично отразиться. Таким образом, здесь присутствует аналогия с методами сглаженных частиц SPH и SPE.

Далее в работе обсуждаются основные компоненты модели, численные подходы для их реализации и результаты модельных расчетов.

1.1. Уравнения Максвелла

Основу модели составляют уравнения Максвелла, которые в СИ вместе с материальными уравнениями имеют вид

$$\begin{aligned} \operatorname{div} \mathbf{D} &= \rho, \quad \operatorname{div} \mathbf{B} = 0, \\ \operatorname{rot} \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t}, \quad \operatorname{rot} \mathbf{H} = \mathbf{j} + \frac{\partial \mathbf{D}}{\partial t}, \\ \mathbf{D} &= \epsilon_a \mathbf{E}, \quad \mathbf{B} = \mu_a \mathbf{H}. \end{aligned} \quad (1)$$

Здесь $\mathbf{D}$ и $\mathbf{B}$ — векторы электрической и магнитной индукций, $\mathbf{E}$ и $\mathbf{H}$ — векторы напряженности электрического и магнитного полей, $\rho = \rho_- + \rho_+$ — объемная плотность зарядов, $\rho_- = q_- n_-$ и $\rho_+ = q_+ n_+$ — плотности отрицательно и положительно заряженных частиц, $q_- < 0$ и $q_+ > 0$ — заряды частиц, n_- и n_+ — объемные концентрации частиц, $\mathbf{j} = \mathbf{j}_- + \mathbf{j}_+$ — суммарная плотность токов частиц, $\mathbf{j}_- = \rho_- \mathbf{v}_-$ и $\mathbf{j}_+ = \rho_+ \mathbf{v}_+$ — плотности токов частиц, $\mathbf{v}_-$ и $\mathbf{v}_+$ — средние скорости частиц, $\epsilon_a = \epsilon \epsilon_0$ и $\mu_a = \mu \mu_0$ — абсолютные диэлектрическая и магнитная проницаемости среды, где μ_0 и ϵ_0 — магнитная и диэлектрическая проницаемости вакуума, c — скорость света.

Уравнения (1) рассматриваются в области цилиндра Ω , не занятой катодом Ω_C и анодом Ω_A , т.е. в $\Omega_D = \Omega / (\Omega_C \cup \Omega_A)$. С учетом материальных уравнений и кусочно-постоянных ϵ и μ из (1) можно получить

$$\begin{aligned} \operatorname{div}(\epsilon_a \mathbf{E}) &= \rho, \quad \operatorname{div} \mathbf{B} = 0, \\ \operatorname{rot} \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t}, \quad \operatorname{rot} \left(\frac{1}{\mu_a} \mathbf{B} \right) = \mathbf{j} + \frac{\partial}{\partial t}(\epsilon_a \mathbf{E}). \end{aligned} \quad (1)'$$

1.2. Релятивистская динамика частиц

Как указано выше, нами используется специальный метод крупных частиц применительно к задачам эмиссионной электродинамики. В его рамках вместо классических уравнений нераз-

рывности и импульса сплошной среды используются уравнения релятивистской динамики, записанные для отдельных заряженных частиц:

$$\begin{aligned} \frac{d\mathbf{r}_{\alpha,k}}{dt} &= \mathbf{v}_{\alpha,k}, \quad \frac{d\mathbf{p}_{\alpha,k}}{dt} = \mathbf{F}_{\alpha,k} = q_{\alpha,k} (\mathbf{E} + [\mathbf{v}_{\alpha,k} \times \mathbf{B}]), \\ \mathbf{p}_{\alpha,k} &= m_{\alpha} \mathbf{v}_{\alpha,k} \gamma_{\alpha,k}, \quad \gamma_{\alpha,k} = 1/\sqrt{1 - (v_{\alpha,k}/c)^2}, \quad k = 1, \dots, N_{\alpha}, \\ n_{\alpha} &= \sum_{k=1}^{N_{\alpha}} \delta(\mathbf{r} - \mathbf{r}_{\alpha,k}), \quad \rho_{\alpha} = \sum_{k=1}^{N_{\alpha}} q_{\alpha,k} \delta(\mathbf{r} - \mathbf{r}_{\alpha,k}), \\ \mathbf{v}_{\alpha} &= \frac{1}{N_{\alpha}} \sum_{k=1}^{N_{\alpha}} \mathbf{v}_{\alpha,k}, \quad \mathbf{j}_{\alpha} = \sum_{k=1}^{N_{\alpha}} q_{\alpha,k} \delta(\mathbf{r} - \mathbf{r}_{\alpha,k}) \mathbf{v}_{\alpha,k}, \quad \mathbf{p}_{\alpha} = \frac{1}{N_{\alpha}} \sum_{k=1}^{N_{\alpha}} \mathbf{p}_{\alpha,k}. \end{aligned} \quad (2)$$

Здесь $\delta(\mathbf{r} - \mathbf{r}_{\alpha,k})$ — дельта-функция Дирака, описывающая плотность заряда частицы сорта α (например, $\alpha = “-”$ — для электронов, $\alpha = “+”$ — для положительно заряженных частиц, например, ионов плазмы), N_{α} — число частиц сорта α , $\mathbf{p}_{\alpha,k}$ — плотность импульса частицы, $\gamma_{\alpha,k}$ — релятивистская поправка к скорости, $v_{\alpha,k} = |\mathbf{v}_{\alpha,k}|$ — модуль скорости частицы.

Уравнения (2) можно модифицировать в целях описания электродинамики крупных частиц. Для этого объединим несколько частиц сорта α в “крупную” частицу и обозначим число таких частиц как $N_{\alpha}^{(p)}$. Тогда вместо (2) получим

$$\begin{aligned} \frac{d\mathbf{r}_{\alpha,l}^{(p)}}{dt} &= \mathbf{v}_{\alpha,l}^{(p)}, \quad \frac{d\mathbf{p}_{\alpha,l}^{(p)}}{dt} = \mathbf{F}_{\alpha,l}^{(p)} = q_{\alpha,l}^{(p)} (\mathbf{E} + [\mathbf{v}_{\alpha,l}^{(p)} \times \mathbf{B}]), \\ \mathbf{p}_{\alpha,l}^{(p)} &= m_{\alpha,l}^{(p)} \mathbf{v}_{\alpha,l}^{(p)} \gamma_{\alpha,l}^{(p)}, \quad \gamma_{\alpha,l}^{(p)} = 1/\sqrt{1 - (v_{\alpha,l}^{(p)}/c)^2}, \quad m_{\alpha,l}^{(p)} = k_{\alpha,l}^{(p)} m_{\alpha}, \quad q_{\alpha,l}^{(p)} = k_{\alpha,l}^{(p)} q_{\alpha}, \quad l = 1, \dots, N_{\alpha}^{(p)}, \\ N_{\alpha} &= \sum_{l=1}^{N_{\alpha}^{(p)}} k_{\alpha,l}^{(p)}, \quad n_{\alpha} \approx n_{\alpha}^{(p)} = \sum_{l=1}^{N_{\alpha}^{(p)}} k_{\alpha,l}^{(p)} \bar{\delta}(\mathbf{r} - \mathbf{r}_{\alpha,l}^{(p)}), \quad \rho_{\alpha} \approx \rho_{\alpha}^{(p)} = \sum_{l=1}^{N_{\alpha}^{(p)}} q_{\alpha,l}^{(p)} \bar{\delta}(\mathbf{r} - \mathbf{r}_{\alpha,l}^{(p)}), \\ q_{\alpha,l}^{(p)} &= \pm e k_{\alpha,l}^{(p)}, \quad \mathbf{v}_{\alpha} \approx \mathbf{v}_{\alpha}^{(p)} = \frac{1}{N_{\alpha}^{(p)}} \sum_{l=1}^{N_{\alpha}^{(p)}} \mathbf{v}_{\alpha,l}^{(p)}, \quad \mathbf{j}_{\alpha} \approx \mathbf{j}_{\alpha}^{(p)} = \sum_{l=1}^{N_{\alpha}^{(p)}} q_{\alpha,l}^{(p)} \bar{\delta}(\mathbf{r} - \mathbf{r}_{\alpha,l}^{(p)}) \mathbf{v}_{\alpha,l}^{(p)}, \\ \mathbf{p}_{\alpha} &\approx \mathbf{p}_{\alpha}^{(p)} = \frac{1}{N_{\alpha}^{(p)}} \sum_{l=1}^{N_{\alpha}^{(p)}} \mathbf{p}_{\alpha,l}^{(p)}, \quad \rho^{(p)} = \rho_{-}^{(p)} + \rho_{+}^{(p)}, \quad \mathbf{j}^{(p)} = \mathbf{j}_{-}^{(p)} + \mathbf{j}_{+}^{(p)}. \end{aligned} \quad (2)'$$

Здесь $\bar{\delta}(\mathbf{r} - \mathbf{r}_{\alpha,l}^{(p)})$ — модифицированная дельта-функция, описывающая плотность заряда, создаваемую одной крупной частицей во всем пространстве, e — заряд электрона, знак “+” или “-” в выражении для $q_{\alpha,l}^{(p)}$ определяется типом частиц.

Здесь и ниже будем считать, что частицы сорта $\alpha = “-”$ представляют собой облака электронов, частицы сорта $\alpha = “+”$ представляют собой скопления положительно заряженных ионов.

1.3. Модификация уравнений Максвелла

Уравнения Максвелла в исходном виде не очень удобны для последующего численного решения сеточными методами. В связи с этим многие исследователи используют принцип суперпозиции полей для выделения отдельных их компонент. Наш подход не является исключением. В его рамках векторы электрического и магнитного полей представим в виде сумм квазистатических и нестационарных компонент:

$$\mathbf{E} = \bar{\mathbf{E}} + \tilde{\mathbf{E}}, \quad \mathbf{B} = \bar{\mathbf{B}} + \tilde{\mathbf{B}}, \quad (3)$$

где $\bar{\mathbf{E}}$ и $\bar{\mathbf{B}}$ удовлетворяют уравнениям Лапласа

$$\operatorname{div}(\epsilon_a \nabla u) = 0, \quad \bar{\mathbf{E}} = -\nabla u; \quad \operatorname{div}\left(\frac{1}{\mu_a} \nabla \phi\right) = 0, \quad \bar{\mathbf{B}} = -\nabla \phi. \quad (4)$$

Здесь u и ϕ — электростатический и магнитостатический скалярные потенциалы. Уравнения (1)' с учетом (3), (4) примут вид

$$\begin{aligned} \operatorname{div}(\epsilon_a \tilde{\mathbf{E}}) &= \rho^{(p)}, \quad \operatorname{div} \tilde{\mathbf{B}} = 0, \\ \frac{\partial \tilde{\mathbf{B}}}{\partial t} &= -\operatorname{rot} \tilde{\mathbf{E}}, \quad \operatorname{rot} \left(\frac{1}{\mu_a} \tilde{\mathbf{B}} \right) = \mathbf{j}^{(p)} + \frac{\partial}{\partial t} (\epsilon_a \tilde{\mathbf{E}}). \end{aligned} \quad (5)$$

Поле $\tilde{\mathbf{E}}$ представим в виде поля зарядов и волн:

$$\tilde{\mathbf{E}} = \tilde{\mathbf{E}}^{(p)} + \tilde{\mathbf{E}}^{(w)}. \quad (6)$$

Тогда уравнения (5) для слагаемых $\tilde{\mathbf{E}}$ примут вид

$$\begin{aligned} \operatorname{div}(\epsilon_a \tilde{\mathbf{E}}^{(p)}) &= \rho^{(p)}, \quad \operatorname{div}(\epsilon_a \tilde{\mathbf{E}}^{(w)}) = 0, \\ \frac{\partial}{\partial t} (\epsilon_a \tilde{\mathbf{E}}^{(w)}) &= \operatorname{rot} \left(\frac{1}{\mu_a} \tilde{\mathbf{B}} \right) - \mathbf{j}^{(p)} - \frac{\partial}{\partial t} (\epsilon_a \tilde{\mathbf{E}}^{(p)}). \end{aligned} \quad (7)$$

Зарядовую часть динамического поля определяем через квазистатический потенциал:

$$\tilde{\mathbf{E}}^{(p)} = -\nabla u^{(p)}, \quad -\operatorname{div}(\epsilon_a \nabla u^{(p)}) = \rho^{(p)}. \quad (8)$$

1.4. Граничные и начальные условия

Системы уравнений (3)–(8) и (2)' дополняются граничными и начальными условиями, которые зависят от конкретной конструкции исследуемой технической системы. Будем считать, что в качестве таких систем выступают множества проводников в вакууме при наличии или отсутствии компактных слоев низкотемпературной квазинейтральной плазмы. При необходимости плазма задается системой зарядов, также описываемой в рамках вышеуказанного метода крупных частиц.

1.4.1. Граничные условия для поля. Граничные условия для потенциала электростатического поля задаются следующим образом:

$$u = u_A = 0, \quad \mathbf{r} \in \partial\Omega_A; \quad u = u_C, \quad \mathbf{r} \in \partial\Omega_C; \quad (\bar{\mathbf{E}}, \mathbf{n}) = 0, \quad \mathbf{r} \in \partial\Omega_F. \quad (9_1)$$

Здесь $\partial\Omega_A$ — внутренняя поверхность анодов исследуемой технической системы, $\partial\Omega_C$ — внешняя поверхность ее катодов, $\partial\Omega_F$ — свободные поверхности.

Магнитостатическое поле не вычисляется, а задается некоторым распределением компонент во всей расчетной области:

$$\bar{\mathbf{B}} = (B_r, B_\phi, B_z)^T, \quad \mathbf{r} \in \Omega. \quad (9_2)$$

Граничные условия для зарядовой части динамического электрического поля имеют следующий вид:

$$u^{(p)} = 0, \quad \mathbf{r} \in \partial\Omega_A \cup \partial\Omega_C; \quad (\tilde{\mathbf{E}}^{(p)}, \mathbf{n}) = 0, \quad \mathbf{r} \in \partial\Omega_F. \quad (9_3)$$

Граничные условия для волновой части нестационарных полей являются следствием из общих граничных условий. Так, в приближении идеальной проводимости на металлических поверхностях последние задаются в виде

$$[\mathbf{E} \times \mathbf{n}] = 0, \quad (\mathbf{B} \cdot \mathbf{n}) = 0, \quad \mathbf{r} \in \partial\Omega_A \cup \partial\Omega_C. \quad (9_4)$$

Граничные условия на свободных границах имеют вид

$$(\mathbf{E} \cdot \mathbf{n}) = 0, \quad (\mathbf{B} \cdot \mathbf{n}) = 0, \quad \mathbf{r} \in \partial\Omega_F. \quad (9_5)$$

Граничные условия на оси симметрии ($r = 0$) задаются следующим образом:

$$[\mathbf{E} \times \mathbf{l}_z] = 0, \quad (\mathbf{B} \cdot \mathbf{l}_z) = 0, \quad r = \sqrt{x^2 + y^2} = 0. \quad (9_6)$$

Здесь $\mathbf{l}_z = (0, 0, 1)^T$.

1.4.2. Граничные условия для частиц. Граничные условия на эмиссионной поверхности для вновь появляющихся в результате эмиссии отрицательно заряженных частиц при использовании уравнений (2) имеют вид

$$k_{-,l}^{(p)} = \left[\frac{|\epsilon_a \mathbf{E} \cdot \mathbf{n}|}{|q_-|} \Delta S_C \right], \quad \mathbf{v}_{\alpha,l}^{(p)} = \bar{\mathbf{v}}_{\alpha,l}^{(p)}, \quad \mathbf{p}_{\alpha,l}^{(p)} = \bar{\mathbf{p}}_{\alpha,l}^{(p)}, \quad l = N_{\alpha}^{(p)} + 1, \dots, N_{\alpha}^{(p)} + \bar{N}_{\alpha}^{(p)}, \quad \mathbf{r} \in \partial\Omega_C. \quad (10)$$

Здесь $k_{\alpha,l}^{(p)}$ — число электронов, составляющих новую отрицательно заряженную крупную частицу, связанную с элементом эмиссионной поверхности (эмиссионным центром) с площадью ΔS_C ; $\bar{\mathbf{v}}_{\alpha,l}^{(p)}$ и $\bar{\mathbf{p}}_{\alpha,l}^{(p)}$ — начальная скорость и импульс новой крупной частицы соответственно; $\bar{N}_{\alpha}^{(p)}$ — число образовавшихся крупных частиц на всей эмиссионной поверхности $\partial\Omega_C$.

Относительно выбранного граничного условия сделаем несколько замечаний.

Во-первых, в настоящей работе рассматривается автоэлектронная эмиссия при сильном самосогласованном электрическом поле в предположении идеальности эмиссионной поверхности и без учета тепловых и магнитных эффектов. Такое допущение справедливо в диапазоне электрических полей меньших, чем поле пробоя некоторых материалов, и при умеренных магнитных полях. В качестве примера реальных технических решений можно привести стальные катоды, покрытые слоем графита (см. [12, 13]).

Во-вторых, в этой части модели реализуется именно метод PiC, поскольку площадь ΔS_C определяется размерами ячейки используемой расчетной сетки. Эта площадь определяет число эмитированных электронов на каждом шаге по времени. Параметром модели является количество электронов, объединяемых в крупную частицу. В нашем подходе задается количество таких частиц, слетающих с поверхности одного сеточного элемента. Для уменьшения вычислительных затрат оно варьируется в диапазоне 1–10.

Граничные условия свободного выхода частиц из расчетной области реализуются с помощью геометрического анализа их положения и соответствующей процедуры удаления. В результате свободного выхода количество крупных частиц $N_{\alpha}^{(p)}$ в рассматриваемом объеме уменьшается.

1.4.3. Начальные условия. Будем считать, что в начальный момент времени выполнены следующие условия:

$$\begin{aligned} \tilde{\mathbf{E}}^{(p)} = 0, \quad \tilde{\mathbf{E}}^{(w)} = 0, \quad \tilde{\mathbf{B}} = 0, \\ \mathbf{v}_{\alpha,l}^{(p)} = \bar{\mathbf{v}}_{\alpha,l}^{(p)}, \quad l = 1, \dots, N_{\alpha}^{(p)}, \quad \rho_{\alpha}^{(p)} = \bar{\rho}_{\alpha}^{(p)} \equiv \sum_{l=1}^{N_{\alpha}^{(p)}} q_{\alpha,l}^{(p)} \bar{\mathbf{v}}_{\alpha,l}^{(p)}. \end{aligned} \quad (11)$$

Здесь $N_{\alpha}^{(p)}$, $\bar{\mathbf{v}}_{\alpha,l}^{(p)}$ и $\bar{\rho}_{\alpha}^{(p)}$ — начальные количества крупных заряженных частиц, их скорости и плотности заряда, соответствующие количеству и распределению частиц по расчетной области.

Распределения частиц по пространству в начальный момент времени могут задаваться различными способами: равномерно или в соответствии с каким-либо физическим процессом. В рассматриваемом нами случае либо частицы отсутствуют (случай неразвитой эмиссии; $N_{\alpha}^{(p)} = 0$, $\rho_{\alpha}^{(p)} = 0$), либо в некоторой подобласти сформирована низкотемпературная квазинейтральная плазма, занимающая часть или весь расчетный объем. В этом случае $N_{\alpha}^{(p)} > 0$, а пространственные распределения отрицательно и положительно заряженных частиц обеспечивают квазинейтральность плазмы ($\rho_{-}^{(p)} + \rho_{+}^{(p)} \approx 0$) в каждой точке расчетного объема.

Распределения скоростей частиц во втором случае выбираются в соответствии с температурной моделью плазмы. В частности, в рамках однотемпературной модели все частицы имеют модули скорости, соответствующие распределению Максвелла для заданной температуры, а направления скоростей обеспечивают равенство нулю суммарного импульса частиц:

$$\sum_{\alpha=-,+} \sum_{l=1}^{N_{\alpha}^{(p)}} \bar{\mathbf{v}}_{\alpha,l}^{(p)} = 0.$$

В рамках двухтемпературной модели вышесказанное справедливо для каждой из подсистем частиц, и, следовательно,

$$\sum_{l=1}^{N_{\alpha}^{(p)}} \bar{\mathbf{v}}_{\alpha,l}^{(p)} = 0, \quad \alpha = -, +.$$

2. ЧИСЛЕННАЯ И ПРОГРАММНАЯ РЕАЛИЗАЦИИ МОДЕЛИ

2.1. Численная реализация

Численная реализация вышеуказанной модели проводилась для случая аксиальной симметрии в координатах (r, z) . При этом предполагалось, что выполняется условие независимости полей от азимутальной переменной φ .

Сеточная модель заряженных частиц сорта α представляет собой аппроксимацию дельта-функции с модифицированным гауссовым профилем:

$$\begin{aligned} \rho_{\alpha}^{(p)} &= A_{p,\alpha} \exp \left[-R^2(r, z; r_{p,\alpha}, z_{p,\alpha}) / R_p^2 \right], \\ R(r, z; r_{p,\alpha}, z_{p,\alpha}) &= \sqrt{(r - r_{p,\alpha})^2 + (z - z_{p,\alpha})^2}, \end{aligned} \quad (12)$$

где $(r_{p,\alpha} = r_{p,\alpha}(t), z_{p,\alpha} = z_{p,\alpha}(t))$ – текущие координаты частицы, R_p – эффективный радиус частиц, составляющий несколько шагов сетки ($R_p \sim n \cdot h$, $h = \sqrt{h_r^2 + h_z^2}$, $n = 1, 2, 3, \dots$). Коэффициент нормировки $A_{p,\alpha}$ зависит неявно от времени и выбирается из условия

$$\iint_{\Omega} \rho_{\alpha}^{(p)} r dr dz = 1. \quad (13)$$

Численная реализация итоговых уравнений и дополнительных условий представляет собой совокупность четырех блоков.

Первый блок связан с построением расчетной сетки во всей области и численным решением уравнения Пуассона для электростатического потенциала u (см. (4)). В качестве сетки используется либо ортогональная декартова сетка, либо криволинейная сетка, топологически эквивалентная декартовой. Аппроксимация уравнения Пуассона основана на известной схеме “крест” (см. [36]), модифицированной для случая (r, z) -геометрии и выбранного типа сеток. Решение дискретного аналога (4) проводится с помощью итераций. Ввиду возможной сингулярности расчетной области в качестве итерационного процесса применяется алгоритм Якоби из [36] с диагональной матрицей перехода.

Второй блок включает решение динамических уравнений Максвелла для расчета магнитного и электрического полей $\tilde{\mathbf{B}}$ и $\tilde{\mathbf{E}}^{(w)}$. Численная схема в этом случае реализована в рамках метода FDTD (см. [23, 24]). Фактически она объединяет методы расщепления по времени и контрольных объемов на случай согласованных векторных полей. Особенностью представляемой реализации также является использование криволинейных сеток, топологически эквивалентных декартовым. Также подчеркнем, что используемая аппроксимация уравнений Максвелла для магнитной индукции $\tilde{\mathbf{B}}$ обеспечивает точное выполнение условия соленоидальности поля $\text{div } \tilde{\mathbf{B}} = 0$ на криволинейной сетке.

Третий блок включает процедуру интегрирования уравнений динамики частиц. Для этого используется известная симметричная схема по времени из [36]. Фактически она представляет собой неявную схему Адамса второго порядка точности, записанную в векторной форме.

Четвертый блок связан с решением квазистатического уравнения (5) для компоненты электрического поля $\tilde{\mathbf{E}}^{(p)}$ (также с помощью итераций на основе метода Якоби) и позволяет найти суммарные поля $\mathbf{E}$ и $\mathbf{B}$.

2.2. Программная реализация

На основе разработанной численной модели была создана параллельная программа. При ее реализации использовались языки программирования ANSI C и C++, а также стандарты параллельных вычислений MPI (см. [37]) и OpenMP (см. [38]). Методика распараллеливания базируется на методе разделения областей (DDM – Domain Decomposition Method) (см. [39, 40]) и ал-

горитмах динамической балансировки загрузки вычислителей (Dynamic Load Balancing) (см. [41, 42]).

Методика распараллеливания на основе DDT связана с разбиением декартовой цилиндрической сетки на компактные домены одинаковой мощности. Во время вычислений каждый домен прикрепляется к конкретному MPI-процессу, который последовательно выполняет все вычисления по четырем блокам динамического алгоритма.

В каждый фиксированный момент времени не все домены могут быть заняты частицами. Вследствие этого распределение сеточных доменов проводится только между MPI-процессами. Когда в конкретный домен попадает некоторое количество частиц, они обрабатываются группой параллельных потоков CPU, организованных с помощью стандарта OpenMP. При этом, если частиц немного, используется один поток. В противном случае добавляется такое число потоков, которое обеспечивает примерное равенство загрузки между MPI-процессами. Поскольку частицы динамически переходят из одного домена в другой (т.е. переходят в другой MPI-процесс), то возникает необходимость корректировки вычислительной нагрузки, которую реализует алгоритм динамической балансировки загрузки.

3. РЕЗУЛЬТАТЫ РАСЧЕТОВ

3.1. Эмиссия с боковой поверхности цилиндрического катода

Первая серия расчетов была посвящена численному анализу эмиссии с поверхности цилиндрического диода. Данная задача в условиях бесконечной протяженности поверхности диода имеет в стационарном варианте аналитическое решение, характеризующееся тем, что радиальная компонента плотности тока убывает пропорционально радиусу ($j_r \cdot r = \text{Const}$) (см. [33, 34]). В проведенных численных экспериментах удалось воспроизвести это решение путем подбора настроечных параметров, а именно, шагов сеток по пространству и времени и константы эмиссии.

При локализации поверхности эмиссии до конечных размеров аналитическое решение не выражается в квадратурах. Однако в центре эмиссионной поверхности реализуется решение, близкое к аналитическому. В качестве примера приведем распределение частиц в зоне развития РЭП (фиг. 1), распределение потенциала электрического поля (фиг. 2) в момент времени 0.1325 нс, а также зависимость суммарного тока эмиссии от времени (фиг. 3). В расчетах шаги пространственной сетки были равны 0.125 см, шаг по времени — 0.125 пс. Напряжение на катоде

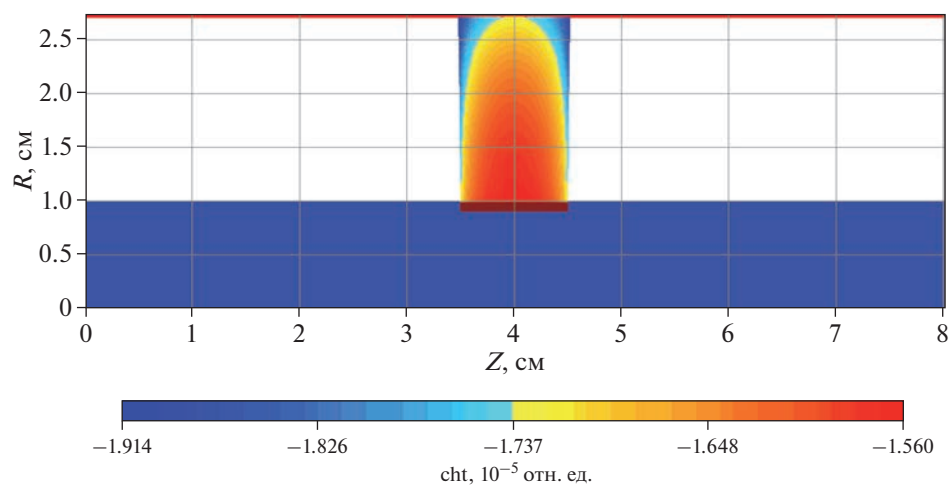
было равно 511 кВ. Константа эмиссии c_{ks}/c_{ks}^0 равна 0.001, где $c_{ks}^0 = \frac{\epsilon_0 E_n r_n^2}{e} \approx 2.824 \times 10^{11}$ — константа, получаемая при обезразмеривании граничного условия (10) ($\epsilon_0 \approx 8.85 \times 10^{-14}$ Ф/см — диэлектрическая проницаемость вакуума, $E_n = 511$ кВ/см, $r_n = 1$ см — нормировочные параметры, $e \approx 1.602 \times 10^{-19}$ Кл — заряд электрона).

На фиг. 1 и 2 видна характерная структура РЭП и потенциала, соответствующая в центральном сечении $z = 4$ см аналитическому решению. Зависимость эмиссионного тока от времени на фиг. 3 иллюстрирует эффект начального скачка (поскольку потенциал был включен мгновенно) и последующий выход тока на более низкий стационарный уровень.

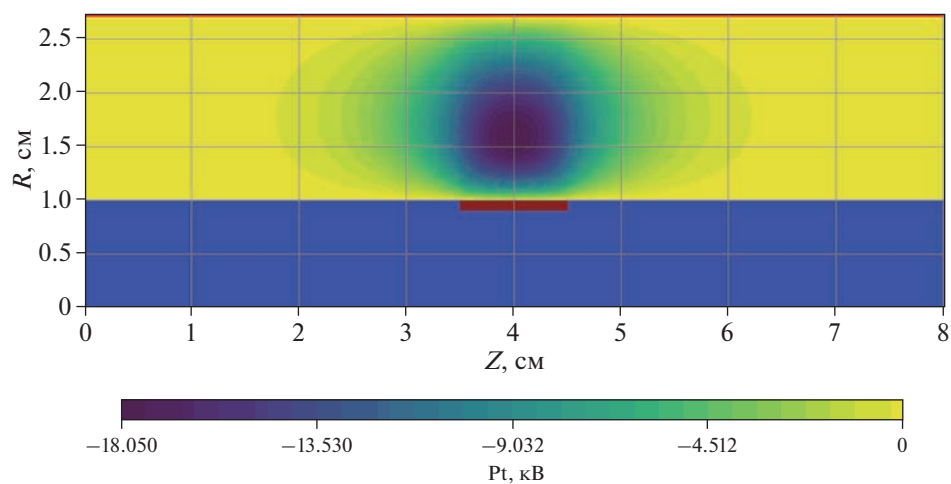
3.2. Эмиссия с торцевой поверхности цилиндрического катода

Вторая серия расчетов была посвящена подбору константы эмиссии в случае торцевого цилиндрического катода (см. фиг. 4). В качестве примера была рассмотрена ситуация, когда эмиссия идет с малого участка торца в целях создания трубчатого РЭП в условиях сильного однородного продольного магнитного поля. Физические параметры задачи: величина B_z равна 1 Тл, напряжение на катоде 511 кВ; радиус катода равен 1 см, длина — 10 см; радиус анода — 2.72 см, длина — 30 см. Эмиссионный процесс запускается с помощью ТЕМ-волны, набегающей слева, с напряженностью поля 511 кВ/см. Расчеты проводились на сетке с шагами 0.05 см, шаг по времени составлял 0.5 пс.

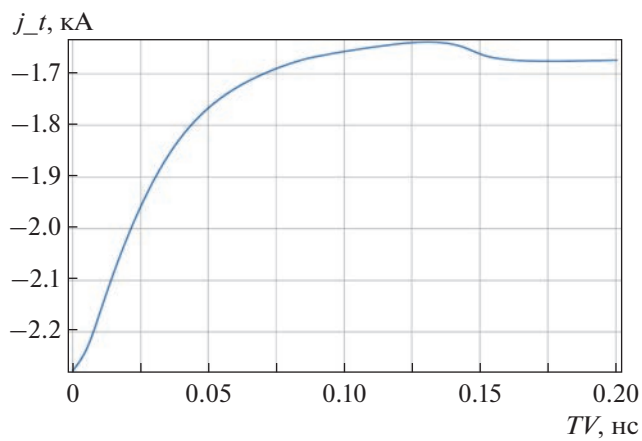
В численных экспериментах варьировалась величина безразмерной константы эмиссии c_{ks} относительно значения c_{ks}^0 , составлявшей соответственно 0.004, 0.01, 0.02, 0.04, 0.06, 0.08. Результаты расчетов динамики полного тока эмиссии показаны на фиг. 5, кривые 1–6 соответству-



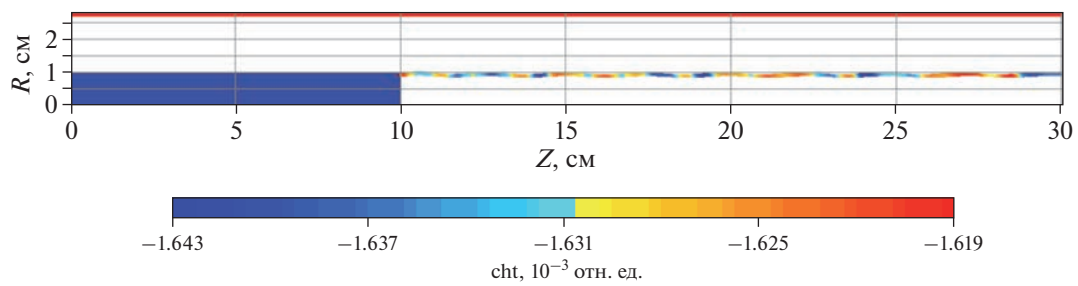
Фиг. 1. Распределение частиц при эмиссии с поверхности цилиндрического диода в момент времени 0.1325 нс.



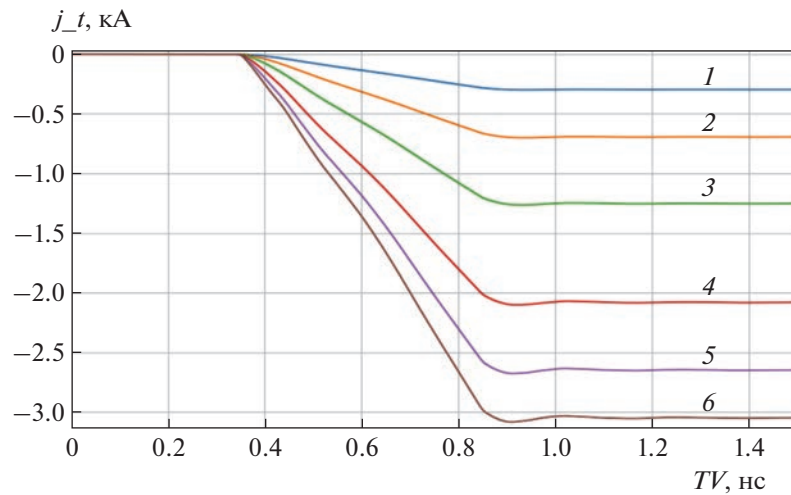
Фиг. 2. Распределение потенциала электрического поля в момент времени 0.1325 нс.



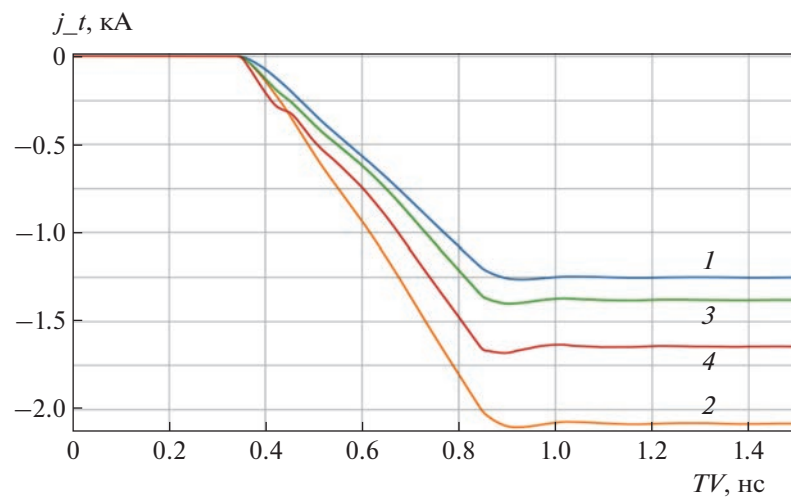
Фиг. 3. Зависимость суммарного тока эмиссии от времени.



Фиг. 4. Расчетная геометрия и конфигурация стационарного распределения трубчатого РЭП.



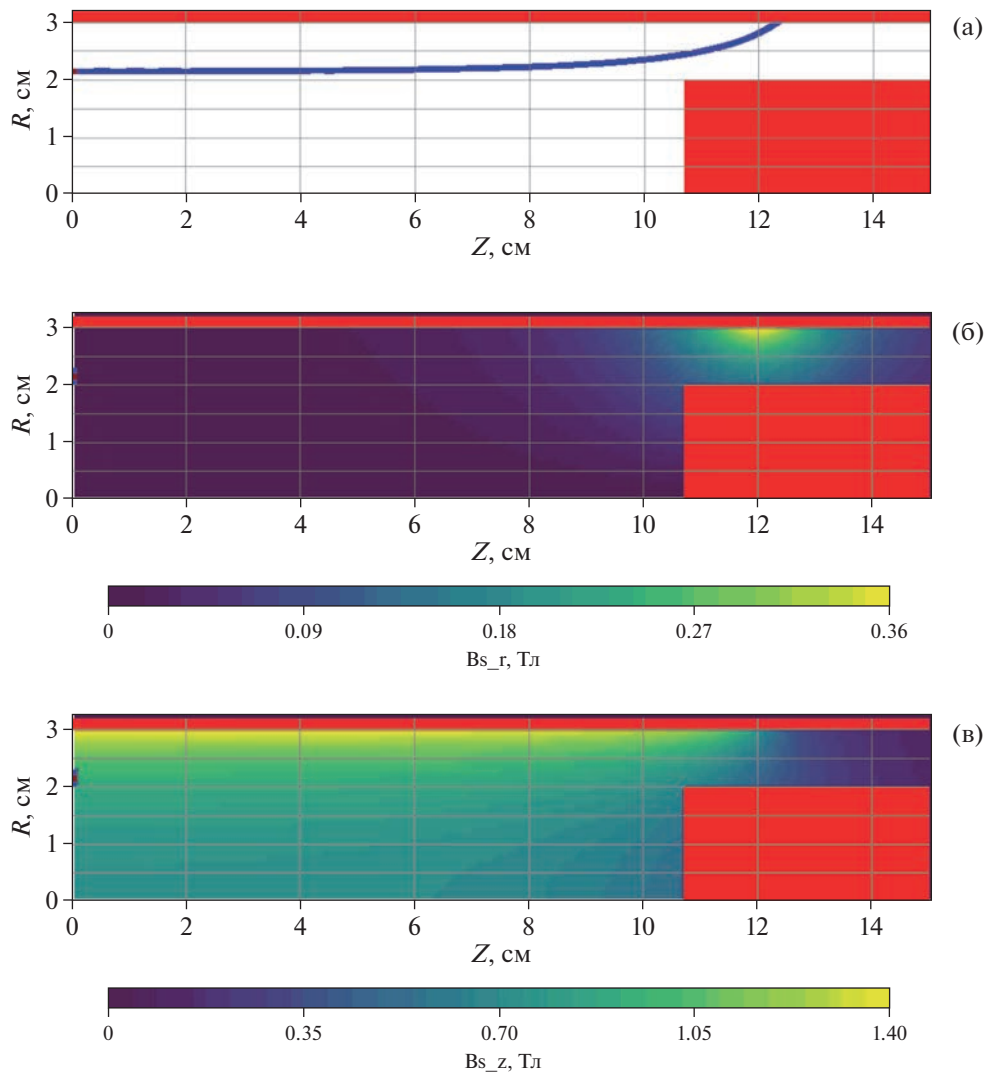
Фиг. 5. Зависимости эмиссионного тока от времени для различных значений константы эмиссии.



Фиг. 6. Зависимости эмиссионного тока от времени для различных значений амплитуды волны и константы эмиссии.

ют вышеуказанным значениям параметра c_{ks} . Они показывают, что полный ток примерно за 1 нс выходит на насыщение, величина тока пропорциональна c_{ks} .

Также величина тока зависит от амплитуды ТЕМ-волны. На фиг. 6 показаны кривые эмиссионного тока для амплитуды 511 кВ/см (кривые 1, 2) и 255.5 кВ/см (кривые 3, 4) для значений



Фиг. 7. Расчетная геометрия и конфигурация РЭП (а), распределения радиальной (б) и продольной (в) компонент стационарного магнитного поля.

$c_{ks} = 0.02, 0.04, 0.08, 0.16$ (кривые 1–4). При этом коэффициент эмиссии выбирается уже из другого диапазона.

Заметим, что в натурном эксперименте амплитуда волны является контролируемым параметром. С другой стороны, величина c_{ks} определяется материалом катода. Вносит свой вклад и геометрия эмиттеров. В связи с этим подбор константы эмиссии в численных расчетах подчиняется обычно каким-либо известным экспериментальным данным, например, данным о величине тока Чайлда–Ленгмюра (см. [33–35]).

3.3. Тестирование кода в условиях неоднородного магнитного поля

Третья серия расчетов была посвящена моделированию эмиссии в условиях неоднородного магнитного поля. Данная задача актуальна для управления направлением РЭП. При этом часто требуется обеспечить генерацию постоянного тока РЭП. В качестве примера была выбрана геометрия генерирующей системы, показанная на фиг. 7а. Там же показана итоговая квазистационарная конфигурация РЭП. На фиг. 7б,в приведены распределения радиальной и продольной компонент вектора магнитной индукции. Рассмотрен случай, когда генерирующая система частично помещена внутрь соленоида. Параметры соленоида: радиус 3.5 см, полудлина 12 см, амплитуда магнитного поля в точке генерации РЭП ($z = 0, r = 2.3$ см) составляла 1 Тл. Пучок гене-

рировался ТЕМ-волной с амплитудой 511 кВ/см, проникающей слева через сечение над катодом. Целью численного эксперимента был подбор константы эмиссии, соответствующей току, наблюдаемому в натурном эксперименте.

Численные расчеты проводились для различных значений константы эмиссии. В результате для реализации постоянного тока РЭП величиной 2 кА была определена константа $c_{ks} = 0.01$. При этом значении была подтверждена численно генерация постоянного тока РЭП в течение всего времени действия постоянного входного электромагнитного импульса.

ЗАКЛЮЧЕНИЕ

Рассмотрена проблема расчета процессов электронной эмиссии с поверхности металлов при сильных электромагнитных полях с учетом релятивистских эффектов. Для аксиально-симметричной геометрии генерирующей системы представлена новая численная методика решения задач эмиссии, сочетающая метод частиц и метод сеток. Методика использует представление крупных сглаженных гауссовых частиц и реализует расчеты электромагнитных полей на декартовых пространственных сетках. Программная реализация выполнена в параллельном варианте с использованием стандартов MPI и OpenMP. В численных экспериментах изучены характеристики электронной эмиссии. В качестве модельных геометрий были выбраны коаксиальный диод и цилиндрические системы типа анод–катод–коллектор. В численных расчетах были получены пространственно-временные характеристики релятивистских электронных пучков, порождаемых эмиссионными процессами. Предложенная численная методика подтвердила свою корректность и эффективность. Дальнейшее направление работы будет связано с рассмотрением технических систем реальной геометрии и уточнением механизма эмиссии на микроскопическом уровне.

СПИСОК ЛИТЕРАТУРЫ

1. *Васильев В.В.* Механика конструкций из композиционных материалов. М.: Машиностроение, 1988. С. 272.
2. *Кербер М.Л.* Полимерные композиционные материалы. Структура. Свойства. Технологии. СПб.: Профессия, 2008. С. 560.
3. *Рухадзе А.А., Богданкевич Л.С., Росинский С.Е.* Физика сильноточных релятивистских электронных пучков. М.: Атомиздат, 1980. С. 167.
4. *Миллер Р.* Введение в физику сильноточных пучков заряженных частиц. М.: Мир, 1984. С. 432.
5. *Бойко В.И., Евстигнеев В.В.* Введение в физику взаимодействия сильноточных пучков заряженных частиц с веществом. М.: Энергоатомиздат, 1988. С. 136.
6. *Месяц Г.А.* Эктоны. Екатеринбург: УИФ “Наука”, 1993. С. 183.
7. *Диденко А.Н., Юшков Ю.Г.* Мощные СВЧ-импульсы наносекундной длительности. М.: Энергоатомиздат, 1984. С. 112.
8. *Воронков С.Н., Лоза О.Т., Стрелков П.С.* Ограничение длительности импульса излучения СВЧ генераторов на микросекундных РЭП // Физика плазмы. 1991. Т. 17. Вып. 6. С. 751–760.
9. *Бугаев С.П., Литвинов Е.А., Месяц Г.А., Проскуровский Д.И.* Взрывная эмиссия электронов // УФН 1975. Т. 115. С. 101–120.
10. *Месяц Г.А.* Взрывная электронная эмиссия. М.: Физматлит, 2011. С. 280.
11. *Херинг К., Никольс М.* Термоэлектронная эмиссия. М.: Изд-во иностр. лит., 1950. С. 196.
12. *Шешин Е.П.* Структура поверхности и автоэмиссионные свойства углеродных материалов. М.: Изд-во МФТИ, 2001. С. 288.
13. *Иванов О.А., Лобаев М.А., Чернов В.В. и др.* Экспериментальное исследование сильноточных катодов на основе алмазных пленок в составе мощного компрессора сверхвысокочастотных импульсов // Изв. вузов. Радиофизика. 2014. Т. LVII. № 10. С. 797–806.
14. *Mumtaz P.* (ред.) Вычислительные методы в электродинамике. М.: Мир, 1977. С. 485.
15. *Birdsall C.K., Langdon A.B.* Plasma Physics via Computer Simulation. New-York, McGraw-Hill book, 1985. P. 479.
16. *Taflove Allen, Hagness Susan C.* Computational Electrodynamics. The Finite-Difference Time-Domain Method. Third Ed. Artech House. 2005. P. 1038.
17. *Inan U.S., Marshall R.A.* Numerical Electromagnetics. The FDTD Method. Edinburgh, Cambridge (UK), Cambridge Univ. Press, 2011. P. 406.
18. *Григорьев А.Д.* Методы вычислительной электродинамики. М.: Физматлит, 2012. С. 432.

19. Программное обеспечение COMSOL Multiphysics URL: <https://www.comsol.ru/comsol-multiphysics>
20. Официальный сайт компании ANSYS Inc. URL: <http://www.ansys.com/>
21. *Tarakanov V.P.* User's Manual for Code KARAT. Springfield, VA: Berkeley Res. VA, 1992. С. 262.
22. Программное обеспечение MEEP. URL: <https://mEEP.readthedocs.io/en/latest/Materials/>
23. *Kane Yee.* Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media // IEEE Transact. Anten. Propagat. 1966. V. 14. №. 3. P. 302–307.
24. *Benford J., Swegle J., Schamiloglu E.* High Power Microwaves. Taylor & Francis, New York, 2nd ed. 2007. P. 1–12.
25. *Харлоу Ф.Х.* Численный метод частиц в ячейках для задач гидродинамики. Вычислительные методы в гидродинамике. М.: Мир, 1967. С. 316–342.
26. *Дьяченко В.Ф.* О расчетах задач бесстолкновительной плазмы // Ж. вычисл. матем. и матем. физ. 1985. Т. 25. № 4. С. 622–627.
27. *Григорьев Ю.Н., Вишников В.А., Федорук М.П.* Численное моделирование методами частиц в ячейках. Новосибирск: Изд-во СО РАН, 2004. С. 358.
28. *Белоцерковский О.М., Давыдов Ю.М.* Метод крупных частиц в газовой динамике. М.: Наука, 1982. С. 370.
29. *Садин Д.В.* Эффективная реализация гибридного метода крупных частиц // Матем. моделирование. 2022. Т. 34. № 4. С. 113–127.
30. *Birsdall C.K., Fuss D.* Clouds-in-clouds, clouds-in-cells physics for many-body plasma simulation // J. Comp. Phys. 1969. V. 3. Iss. 4. P. 494–511.
31. *Jianguo Wang, Dianhui Zhang, Chunliang Liu, Yongdong Li, Yue Wang, Hongguang Wang, Hailiang Qiao, Xiaozhe Li.* UNIPIC code for simulations of high power microwave devices // Phys. Plasm. 2009. № 16. P. 1–11.
32. *Monaghan J.J.* An introduction to SPH. // Comp. Phys. Comm. 1988. V. 48. P. 88–96.
33. *Добрецов Л.Н.* Электронная и ионная эмиссия. М.-Л., 1952. С. 312.
34. *Райзер Ю.П.* Физика газового разряда. М.: Наука, 1992. С. 536.
35. *Lisovskiy V., Yegorenkov V.* Validating the collision-dominated Child-Langmuir law for a dc discharge cathode sheath in an undergraduate laboratory // Europ. J. Phys. 2009. V. 30. № 6. P. 1345–1351.
36. *Самарский А.А., Гулин А.В.* Численные методы. М.: Наука, 1989. С. 432.
37. Official documentation and manuals on MPI. [Online]. Available from: <http://mpi-forum.org/>
38. Official documentation and manuals on OpenMP. [Online]. Available from: <http://www.openmp.org, http://www.llnl.gov/computing/tutorials/openMP>
39. *Smith B.F.* Domain Decomposition Methods for Partial Differential Equations / In: *Keyes, D.E., Sameh, A., Venkatakrishnan, V.* (eds) Parallel Numerical Algorithms. ICASE/LaRC Interdisciplinary Series in Science and Engineering, V. 4. Springer, Dordrecht, 1997. P. 225–243.
40. *Dolean V., Jolivet P., Nataf F.* An Introduction to Domain Decomposition Methods: algorithms, theory and parallel implementation. Master. France. 2015. P. 289. <https://hal.science/cel-01100932v6>
41. *Alakeel A.A.* Guide to Dynamic Load Balancing in Distributed Computer Systems // Inter. J. Comp. Sci. Network Security (IJCSNS). 2009. V. 10. № 6. P. 153–160.
42. *Sanders P., Mehlhorn K., Dietzfelbinger M., Dementiev R.* Sequential and parallel algorithms and data structures: the basic toolbox. Springer Nature, Cham (Switzerland), 2019. P. 516.

МАТЕМАТИЧЕСКАЯ ФИЗИКА

УДК 519.245

ОЦЕНКА ПОГРЕШНОСТИ И ОПТИМИЗАЦИЯ МЕТОДА ПРЯМОГО СТАТИСТИЧЕСКОГО МОДЕЛИРОВАНИЯ С УЧЕТОМ ПРОСТРАНСТВЕННОЙ РЕГУЛЯРИЗАЦИИ¹⁾

© 2023 г. М. Ю. Плотников^{1,*}, Е. В. Шкарупа^{2,**}

¹ 630090 Новосибирск, пр-т Лаврентьева, 1, Ин-т теплофизики СО РАН, Россия

² 630090 Новосибирск, пр-т Лаврентьева, 6, Ин-т вычисл. матем. и матем. геофиз. СО РАН, Россия

*e-mail: plotnikov@itp.nsc.ru

**e-mail: sev@osmf.sccc.ru

Поступила в редакцию 26.12.2022 г.
Переработанный вариант 28.02.2023 г.
Принята к публикации 30.03.2023 г.

Метод прямого статистического моделирования широко применяется для решения задач динамики разреженного газа. Представленная работа направлена на исследование погрешности, вносимой пространственной регуляризацией взаимодействия двух частиц. Рассматриваются два подхода к пространственной регуляризации и три алгоритма метода прямого статистического моделирования, реализующих эти подходы. Для этих алгоритмов построена верхняя граница погрешности в метрике пространства непрерывных функций и получены условно-оптимальные параметры, гарантирующие по вероятности заданный уровень погрешности. На примере классической задачи Фурье проведено численное исследование погрешности, вносимой регуляризацией, и тестирование построенных условно-оптимальных параметров. Библ. 28. Фиг. 4. Табл. 4.

Ключевые слова: прямое статистическое моделирование, погрешность, оптимизация, регуляризация.

DOI: 10.31857/S0044466923070128, EDN: ZXWLAB

ВВЕДЕНИЕ

Метод прямого статистического моделирования (ПСМ) был разработан в начале 60-х годов прошлого столетия Г. Бердом для моделирования течений разреженного газа. Впоследствии он получил дальнейшее развитие, в частности были разработаны эффективные процедуры моделирования столкновений частиц (схема No Time Counter (NTC) [1] и метод мажорантной частоты [2]). В настоящее время, несмотря на развитие альтернативных подходов (см., например, [3–6]), метод ПСМ [7], [8] сохраняет свои лидирующие позиции среди численных методов решения задач динамики разреженного газа.

В основе “классического” алгоритма метода ПСМ лежит расщепление непрерывного движения частиц и их столкновений на два последовательных этапа, определяемых временным интервалом Δt .

Этап 1. Все частицы перемещаются на расстояние, определяемое их скоростями и временем Δt . Проводятся определенные действия, если частицы пересекают граничные поверхности.

Этап 2. Моделируются столкновения между частицами, соответствующие интервалу Δt . Скорости частиц до столкновения заменяются скоростями, приобретенными ими после столкновения.

При этом обычно моделируемый объем физического пространства разбивается на ячейки и в каждой ячейке моделируется некоторое количество частиц. Изменение времени проводится дискретными шагами Δt , малыми по сравнению со средним временем между столкновениями частиц. На каждом временном шаге производится суммирование информации, необходимой для вычисления макропараметров течения газа.

¹⁾ Работа выполнена при финансовой поддержке бюджетных грантов (¹121031800218-5 и ²0251-2021-0002).

Несмотря на продолжительное и активное использование метода ПСМ, работа по изучению влияния на точность вычисления макропараметров течения отдельных параметров алгоритма (шага по времени, размера ячейки, количества шагов по времени, количества частиц в ячейке, соотношения числа реальных и моделируемых частиц) ввиду очевидной сложности метода далека от завершения. Различными исследователями проделан большой объем работы по изучению погрешности алгоритма и получен ряд важных результатов (см., например, [9–18]). Развитие компьютерной техники и появление пакетов программ для расчетов методом ПСМ стимулируют интерес к разработке новых подходов к моделированию течений газа, дальнейшему анализу погрешности метода и выбору оптимальных параметров [19–22].

Для дальнейшего анализа удобно разделить погрешность метода на внутреннюю и внешнюю [23]. Внутренняя погрешность метода обусловлена тем, что непрерывный процесс движения и столкновений частиц заменяется дискретным по времени и пространству процессом, а также тем, что каждая моделируемая частица представляет собой тысячи реальных частиц. Кроме того, для ускорения расчета при реализации столкновений частиц используется вероятностный процесс выбора партнера. Поэтому внутренняя погрешность зависит от шага по времени, количества моделируемых частиц и способа реализации столкновений частиц, в частности выбора партнера для столкновения (пространственной регуляризации). Внешняя погрешность связана со сбором информации для оценки макропараметров течения. Как правило, требуется пространственное распределение макропараметров течения, поэтому в интересующей области строится сетка, в узлах которой оцениваются макропараметры, а затем производится интерполяция на всю область. Поэтому внешняя погрешность зависит от объема выборки, выбора сетки и способа выполнения.

Одной из составляющих внутренней погрешности метода ПСМ является погрешность, обусловленная регуляризацией столкновения двух частиц по пространственным переменным. В “классическом” алгоритме метода ПСМ (см., например [7]) выбор партнеров для столкновения на каждом временном шаге осуществляется среди частиц, находящихся в одной ячейке. При этом не рассматривается возможность столкновения частиц, находящихся в разных ячейках, даже если они расположены близко друг к другу. Естественно, что такая реализация алгоритма приводит к возникновению дополнительной погрешности. В [15] проведено численное исследование модификации алгоритма ПСМ, основанной на использовании ячеек различного размера для моделирования столкновений и сбора информации. В частности показано, что введение ячеек уменьшенного размера для моделирования столкновений позволяет существенно улучшить точность вычислений. В работе [24] проведен теоретический анализ погрешности метода ПСМ, обусловленной пространственной регуляризацией. В работе [19] предложен имитационный метод с расщеплением (ИМР), в котором этап столкновений разбивается на два этапа. При этом на втором этапе производится поправка на столкновения частиц, принадлежащих на первом этапе к различным группам. В [25] на примере задачи о продольном обтекании пластины показано различие в макропараметрах течения, рассчитанных “классическим” алгоритмом метода ПСМ и методом ИМР.

Представленная работа направлена на исследование влияния погрешности, обусловленной пространственной регуляризацией, на точность вычислений методом ПСМ. Рассматривается решение стационарных задач, в которых для нахождения решения суммируются данные о частицах вдоль одной траектории. Построена верхняя граница погрешности (с учетом регуляризации) метода ПСМ в метрике пространства непрерывных функций и получены условно-оптимальные параметры алгоритма, гарантирующие по вероятности заданный уровень погрешности.

1. СТАТИСТИЧЕСКИЕ ОЦЕНКИ МЕТОДА ПСМ

Макропараметры течения разреженного газа представимы в виде математического ожидания $E\xi(X)$ некоторой функции $\xi(X)$ от состояния N частичной системы X или в виде рациональной функции от таких математических ожиданий. Так, числовая плотность n в некоторой точке g представима в виде

$$n(g) \approx C_n E\xi(X), \quad \xi(X) = \frac{1}{\Delta g} \sum_{l=1}^N h(r_l), \quad (1.1)$$

где C_n — нормирующая константа, Δg — мера ячейки, $h(r)$ — индикатор ячейки, r_l — точка пространственного расположения частицы с номером l . Отметим, что в представленной работе размерность физического пространства может принимать значения $d = 1, 2, 3$, а пространство ско-

ростей частиц трехмерно. Полная температура течения в некоторой точке области складывается из трех покомпонентных компонент $T = (T_1 + T_2 + T_3)/3$. Для скорости течения $\mathbf{v} = (v_1, v_2, v_3)$ и температуры T используются следующие представления:

$$v_i(\mathbf{r}) \approx \frac{\mathbf{E}\eta_i(X)}{\mathbf{E}\xi(X)}, \quad \eta_i(X) = \frac{1}{\Delta r} \sum_{l=1}^N h(\mathbf{r}_l)(v_l)_i, \quad i = 1, 2, 3, \quad (1.2)$$

$$RT_i(\mathbf{r}) \approx \frac{\mathbf{E}\zeta_i(X)}{\mathbf{E}\xi(X)} - \left(\frac{\mathbf{E}\eta_i(X)}{\mathbf{E}\xi(X)} \right)^2, \quad \zeta_i(X) = \frac{1}{\Delta r} \sum_{l=1}^N h(\mathbf{r}_l)(v_l)_i^2, \quad i = 1, 2, 3. \quad (1.3)$$

Здесь $(v_l)_i$ есть i -я компонента скорости частицы с номером l , R — газовая постоянная.

При решении стационарных задач основу метода ПСМ составляет предположение об эргодичности рассматриваемого стационарного процесса. В этом случае математические ожидания оцениваются осреднением значений функций от состояний N -частичной системы, взятых через равные интервалы времени Δt . Для определенности зафиксируем три параметра метода, определяющих непосредственно процесс моделирования: полное число частиц N , размер ячейки Δr и шаг по времени Δt . Пусть X_k обозначает состояние N -частичной системы в момент времени $t_k = k\Delta t$, $k = 1, 2, \dots, K$. Тогда осредненная по времени оценка математического ожидания, например, для вычисления числовой плотности (1.1) имеет вид

$$E\xi(X^{N, \Delta x, \Delta t}) \approx \frac{1}{K} \sum_{k=1}^K \xi(X_k^{N, \Delta x, \Delta t}).$$

Аналогичные выражения получаются для вычисления скорости и температуры (1.2), (1.3).

2. ПРОСТРАНСТВЕННАЯ РЕГУЛЯРИЗАЦИЯ И ВНОСИМАЯ ЕЮ ПОГРЕШНОСТЬ

Пусть расчетная область $D \subset \mathbf{R}^d$ ограниченная, с кусочно-гладкой границей. Обозначим через ΔJ разность между оценками функционала от решения, полученными с использованием регуляризации и без нее. Рассмотрим два подхода к регуляризации парного взаимодействия частиц по пространственным переменным.

1. Расчетная область D разбивается на непересекающиеся ячейки, каждая из которых представляет собой выпуклое множество. Обозначим через h максимальный диаметр ячеек. Выбор партнеров для столкновения на каждом временном шаге осуществляется среди частиц, находящихся в одной ячейке. Такая регуляризация соответствует “классической” реализации метода ПСМ [7], [8].

2. Для произвольной частицы с координатами $\mathbf{r}_1 \in D$ партнер для столкновения выбирается среди частиц, попадающих в шар заданного радиуса h с центром в точке $\mathbf{r}_1$.

В работе [24] для трехмерного случая расчетной области получены следующие оценки погрешности регуляризации:

$$|\Delta J_1| < C_1 h, \quad |\Delta J_2| < C_2 h^4 \quad (2.1)$$

для первого и второго подхода соответственно. Здесь C_1 и C_2 — некоторые константы.

Для случая использования расчетной области размерности $d = 1$ и 2 по аналогии с рассуждениями [24] можно получить следующие оценки:

$$|\Delta J_1| < C_1 h, \quad (2.2)$$

$$|\Delta J_2| < C_2 h^{d+1}. \quad (2.3)$$

Отметим, что аналитические выражения для констант C_1 и C_2 , позволяющие вычислить их значения для конкретных задач, в [24] не получены, поэтому далее будем использовать оценки этих констант, вычисленные в ходе предварительных расчетов.

3. ТРИ АЛГОРИТМА МЕТОДА ПСМ

Для численного исследования влияния регуляризации на точность вычислений в представленной работе рассмотрим три алгоритма метода ПСМ на основе схемы столкновений НТС [1], [7]. Кратко сформулируем эти алгоритмы, уделяя основное внимание реализации регуляризации.

Алгоритм 1 (первый подход к регуляризации) [7].

Область разбивается на непересекающиеся ячейки. На каждом временном шаге частицы сортируются по ячейкам. Для каждой ячейки номера m вычисляется число столкновений $N_{col,m}$ за временной шаг Δt ; далее из числа частиц в ячейке $N_{col,m}$ раз равномерно выбираются пары партнеров для столкновения, для каждой пары с некоторой вероятностью происходит реальное столкновение, в противном случае — фиктивное.

Алгоритм 2 (второй подход к регуляризации) [26].

Для числа частиц N во всей расчетной области вычисляется число столкновений N_{col} за временной шаг Δt . Далее из общего числа частиц N_{col} раз равномерно выбираются партнеры для столкновения. Если расстояние между выбранными частицами больше заданного h , то происходит фиктивное столкновение. Если расстояние между ними меньше заданного, то с некоторой вероятностью реализуется реальное столкновение, в противном случае — фиктивное.

Второй алгоритм был реализован на основе алгоритма 6 из [26], сформулированного для схемы столкновений на основе мажорантной частоты. Отметим, что использование ячеек для регуляризации позволяет получить более эффективные алгоритмы (см., например [26]). Проведенные нами расчеты (см. ниже разд. 5) также подтвердили этот вывод. Поэтому нами был сформулирован приближенный комбинированный алгоритм метода ПСМ для второго подхода к регуляризации.

Алгоритм 3 (второй подход к регуляризации).

Область разбивается на непересекающиеся ячейки. На каждом временном шаге частицы сортируются по ячейкам. Для каждой ячейки m вычисляется число столкновений $N_{col,m}$ за временной шаг Δt ; из числа частиц в ячейке равномерно выбирается частица для столкновения и далее равномерно выбирается партнер для столкновения из ее окрестности. Для выбранной таким образом пары с некоторой вероятностью происходит реальное столкновение, в противном случае — фиктивное. Вклад в число столкновений по ячейкам вносится согласно принадлежности выбранных для столкновения частиц ячейкам.

4. ВЕРХНИЕ ГРАНИЦЫ ПОГРЕШНОСТИ И УСЛОВНО-ОПТИМАЛЬНЫЕ ПАРАМЕТРЫ АЛГОРИТМОВ

Для построения верхней границы погрешности и выбора оптимальных параметров метода ПСМ будем использовать теорию функциональных алгоритмов статистического моделирования [27]. Методы статистического моделирования (Монте-Карло) позволяют вычислять отдельные функционалы от решений интегральных и дифференциальных уравнений, в том числе значения решений в выбранных точках. Для построения решения уравнения на всей области необходимо сначала построить сетку, оценить решения в узлах сетки методом Монте-Карло, а затем восполнить решения по полученным приближенным значениям в узлах. Такие алгоритмы будем называть *функциональными алгоритмами статистического моделирования*. Особенностью таких алгоритмов является наличие детерминированных и стохастических компонент погрешности, что приводит к некоторым сложностям в исследовании сходимости и оптимизации. Важной задачей является выбор оптимальных параметров функциональных алгоритмов статистического моделирования, например, числа узлов сетки и объема выборки, позволяющих достичь заданного уровня погрешности. Здесь мы используем подход к оптимизации из [27], т.е. предполагаем, что верхняя граница погрешности достаточно точно воспроизводит зависимость погрешности от параметров. В этом случае мы приравниваем верхнюю границу к заданному уровню погрешности и минимизируем функцию трудоемкости, зависящую от параметров.

С точки зрения внешней погрешности метод ПСМ можно рассматривать и исследовать как функциональный алгоритм статистического моделирования. Пусть $u(r)$ один из макропараметров течения, который должен быть аппроксимирован на некоторой ограниченной области

$D \subset \mathbf{R}^d$. Тогда в общем виде функциональный алгоритм метода Монте-Карло строится следующим образом.

1. В области D строится сетка $\{\mathbf{r}^{(i)}\}_{m=1}^M$.
2. Значения функции $u(\mathbf{r})$ в узлах сетки оцениваются методом Монте-Карло:

$$u(\mathbf{r}^{(m)}) \approx \tilde{u}(\mathbf{r}^{(m)}) = \frac{1}{K} \sum_{k=1}^K \xi_k^{(m)}.$$

3. Для приближения функции $u(\mathbf{r})$ на области D применяется процедура интерполяции:

$$u(\mathbf{r}) \approx L_{(M)} \tilde{u}(\mathbf{r}) = \sum_{m=1}^M \tilde{u}(\mathbf{r}^{(m)}) \chi_m(\mathbf{r}), \quad (4.1)$$

где $\chi_m(\mathbf{r})$ — базисные функции интерполяции. Здесь мы используем мультилинейную интерполяцию [27], т.е. $\chi_m(\mathbf{r}) = \chi(\mathbf{r}/h - \mathbf{m})$ в случае $d = 1$, где $\chi(s)$ — кусочно-линейная финитная образующая функция:

$$\chi(s) = \begin{cases} 1+s & \text{при } -1 \leq s \leq 0, \\ 1-s & \text{при } 0 \leq s \leq 1, \\ 0 & \text{иначе.} \end{cases}$$

В случае $d = 2, 3$ базисные функции $\chi_m(\mathbf{r})$ являются покоординатными произведениями соответствующих функций типа $\chi(s)$.

Для выбора оптимальных параметров функционального алгоритма, т.е. объема выборки K и числа узлов M , позволяющих достичь заданного уровня погрешности, нужно выполнить два шага.

1. Построить верхнюю границу погрешности аппроксимации (4.1). Здесь мы рассматриваем верхнюю границу погрешности в метрике пространства непрерывных функций со сходимостью по вероятности, т.е. строим отношения вида

$$\mathbf{P}\left\{\delta = \sup_{\mathbf{r} \in D} |u(\mathbf{r}) - L_{(M)} \tilde{u}(\mathbf{r})| \leq T(M, K)\right\} > 1 - \varepsilon,$$

где $T(M, K) \rightarrow 0$ при $M, K \rightarrow \infty$, а $\varepsilon > 0$ — малая величина.

2. Решить задачу оптимизации, т.е. найти условно-оптимальные значения параметров функционального алгоритма, минимизирующих функцию трудоемкости при заданном уровне погрешности:

$$\min_{M, N} S(M, K) \quad \text{при условии} \quad T(M, K) = \alpha,$$

где $S(M, K)$ — функция трудоемкости, $T(M, K)$ — верхняя граница погрешности, а $\alpha > 0$ — заданный уровень погрешности.

При решении стационарных задач методом ПСМ сбор информации на каждом временном шаге приводит к возникновению корреляции данных, собранных с близких временных шагов [28]. В работе [23] при построении верхней границы внешней погрешности метода ПСМ была показана важность использования разреженного сбора информации для уменьшения статистической погрешности оцениваемых макропараметров. Разреженный сбор информации подразумевает суммирование информации с каждого J -го шага по времени, что уменьшает корреляции и соответственно статистическую погрешность. В [23] были построены верхние границы внешней погрешности функционального алгоритма ПСМ для плотности, скорости и температуры:

$$\mathbf{P}\left\{\delta \leq \frac{H_I}{M^{2/d}} + \sqrt{\frac{H_V M}{K} \frac{1 + \beta_\xi^{JM^{1/d}}}{1 - \beta_\xi^{JM^{1/d}}} \theta_M(\varepsilon)}\right\} \geq 1 - \varepsilon.$$

Здесь первое слагаемое в правой части неравенства соответствует погрешностям интерполяции и смещения, а второе — статистической погрешности. Константы H_I , H_V и β_ξ оцениваются на ос-

нове предварительных вычислений. Константа H_I выражается через частные производные решения

$$H_I = \frac{(\text{mes } D)^{1/2}}{6} \sum_{i=1}^d \sup_{r \in D} \left| \frac{\partial^2}{\partial r_i^2} u(r) \right|.$$

Константа H_V оценивается через максимум дисперсии статистической оценки ξ по формуле

$$V\xi^{(m)} \leq H_{V,\xi} M \quad \text{для любых } m = 1, \dots, M.$$

Константа β_ξ оценивается через максимум корреляций статистической оценки ξ по формуле

$$|\rho_{\xi,k}| \leq \beta_\xi^{kM^{1/d}},$$

где $\rho_{\xi,k}$ — корреляции стационарного стохастического процесса $\xi(X_k)$:

$$\rho_{\xi,k} = \frac{1}{V\xi} E(\xi(X_1) - E(\xi))(\xi(X_{1+k}) - E(\xi)).$$

Величина $\theta_M(\varepsilon)$ определяется по формуле

$$\theta_M(\varepsilon) = \Phi_{0,1}^{-1} \left(1 - \frac{\varepsilon}{2M} \right),$$

где $\Phi_{0,1}$ — стандартная нормальная функция распределения.

По аналогии с выкладками из [23] учет погрешности регуляризации приводит к формулам

$$\mathbf{P} \left\{ \delta \leq \frac{H_I}{M^{2/d}} + \frac{H_R}{M_R^{1/d}} + \sqrt{\frac{H_V M}{K} \frac{1 + \beta_\xi^{JM^{1/d}}}{1 - \beta_\xi^{JM^{1/d}}} \theta_M(\varepsilon)} \right\} \geq 1 - \varepsilon \quad (4.2)$$

для первого подхода к регуляризации и

$$\mathbf{P} \left\{ \delta \leq \frac{H_I}{M^{2/d}} + \frac{H_R}{M_R^{1+1/d}} + \sqrt{\frac{H_V M}{K} \frac{1 + \beta_\xi^{JM^{1/d}}}{1 - \beta_\xi^{JM^{1/d}}} \theta_M(\varepsilon)} \right\} \geq 1 - \varepsilon \quad (4.3)$$

для второго подхода.

Здесь M_R — число ячеек для моделирования столкновений частиц для первого подхода к регуляризации. Для второго подхода M_R — целое число, определяемое диаметром окрестности h : $M_R = \text{mes } D/h^d$. Для нахождения оптимальных параметров алгоритма метода ПСМ нужно решить оптимизационную задачу вида

$$\min_{J, M_S, M, K} S(J, M, M_R, K) \quad \text{при условии} \quad T(J, M, M_R, K) = \alpha \quad (4.4)$$

для верхней границы погрешности из формул (4.2) и (4.3):

$$T(J, M, M_R, K) = \frac{H_I}{M^{2/d}} + \frac{H_R}{M_R^{1/d}} + \sqrt{\frac{H_V M}{K} \frac{1 + \beta_\xi^{JM^{1/d}}}{1 - \beta_\xi^{JM^{1/d}}} \theta_M(\varepsilon)}$$

и

$$T(J, M, M_R, K) = \frac{H_I}{M^{2/d}} + \frac{H_R}{M_R^{1+1/d}} + \sqrt{\frac{H_V M}{K} \frac{1 + \beta_\xi^{JM^{1/d}}}{1 - \beta_\xi^{JM^{1/d}}} \theta_M(\varepsilon)}$$

соответственно.

Для рассматриваемых алгоритмов функция трудоемкости имеет вид:

$$S(J, M, M_R, K) = (t_1 J + t_2 + t_3 M + t_4(M_R) J) K. \quad (4.5)$$

Здесь t_1 — время, затрачиваемое на моделирование процесса переноса частиц за один шаг Δt , $(t_2 + t_3 M)$ — время, затрачиваемое на сбор информации за один шаг выборки, $t_4(M_R)$ — время, за-

трачиваемое на моделирование процесса столкновений частиц за один шаг Δt . Предварительные расчеты показали, что величина t_4 зависит от выбранного алгоритма: для первого и второго алгоритмов она прямо пропорциональна M_R , $t_4(M_R) = t_5 M_R$, а для третьего алгоритма — обратно пропорциональна M_R , $t_4(M_R) = t_5 / M_R$. Для первого алгоритма рост t_4 с увеличением M_R связан с необходимостью затрат на сортировку по ячейкам для столкновений, для второго алгоритма — с большим количеством фиктивных столкновений. Для третьего алгоритма рост t_4 с уменьшением M_R связан с необходимостью затрат на выбор партнера для столкновения из окрестности.

Чтобы решить задачу оптимизации, мы выражаем K из условия (4.4) и подставляем его в функцию затрат (4.5). Таким образом, мы получаем следующие функции в зависимости от J , M и M_R :

$$\tilde{S}(J, M, M_R) = (t_1 J + t_2 + t_3 M + t_4(M_R) J) \frac{H_V M \theta_M^2(\epsilon)}{\left(\alpha - \frac{H_I}{M^{2/d}} - \frac{H_R}{M_R^{1/d}} \right)^2} \frac{1 + \beta_\xi^{JM^{1/d}}}{1 - \beta_\xi^{JM^{1/d}}} \quad (4.6)$$

для первого подхода к регуляризации и

$$\tilde{S}(J, M, M_R) = (t_1 J + t_2 + t_3 M + t_4(M_R) J) \frac{H_V M \theta_M^2(\epsilon)}{\left(\alpha - \frac{H_I}{M^{2/d}} - \frac{H_R}{M_R^{1+1/d}} \right)^2} \frac{1 + \beta_\xi^{JM^{1/d}}}{1 - \beta_\xi^{JM^{1/d}}} \quad (4.7)$$

для второго подхода.

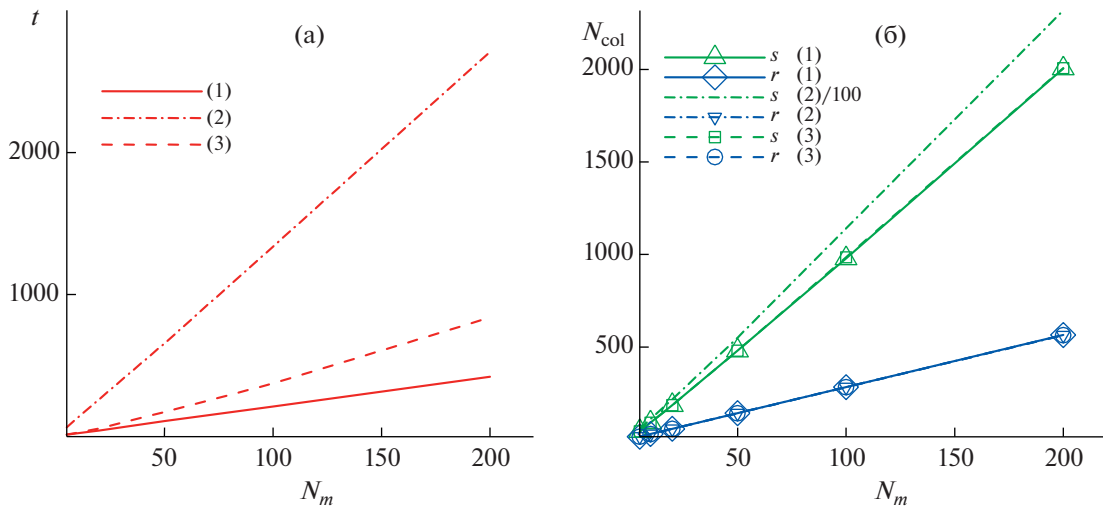
Значения M_{opt} , $M_{R\text{opt}}$, K_{opt} , J_{opt} , минимизирующие функции (4.6), (4.7), можно найти численно путем перебора всех возможностей. Конечно, для каждого макропараметра течения газа эти значения могут быть разными.

5. РЕЗУЛЬТАТЫ ЧИСЛЕННЫХ ЭКСПЕРИМЕНТОВ

Тестирование разработанных подходов проводилось на примере классической задачи Фурье о теплопередаче между двумя бесконечными параллельными пластинами. Введем декартову систему координат XYZ . Пусть газ со средней числовой плотностью n_0 заключен между двумя бесконечными пластинами, расположенными в плоскостях $z = 0$ и $z = H$, температуры пластин T_0 и T_H постоянны. Предполагается, что отражение частиц от пластин является диффузным с полной аккомодацией энергии и импульса. Для моделирования межмолекулярных столкновений применяется модель твердых сфер [7]. Для приведения задачи к безразмерному виду использовались следующие величины: числовая плотность n_0 , средняя длина свободного пробега L частиц в газе при плотности n_0 , температура T_0 и наиболее вероятная тепловая скорость газа $c_m = \sqrt{2kT_0/m}$ (здесь k — постоянная Больцмана, m — масса частицы). После перехода к безразмерным переменным задача определяется только отношением температур T_H/T_0 и числом Кнудсена $\text{Kn} = L/H$. Требуется определить распределение макропараметров газа (плотность, скорость и температура) между пластинами при заданном отношении температур пластин T_H/T_0 и степени разреженности газа. Расчеты проводились для следующего набора параметров: $\text{Kn} = 0.025$; три отношения температур: $T_H/T_0 = 1$, $T_H/T_0 = 4$ и $T_H/T_0 = 9$; $\Delta t = 0.05$. При $T_H/T_0 = 1$ известно точное решение: $n = 1$, $T_x = T_y = T_z = 1$, $u_x = u_y = u_z = 0$.

Численные эксперименты проводились также для двумерной задачи Фурье. Для двумерного случая ($d = 2$) отражающие пластины были расположены в плоскостях $z = 0$, $z = H$, $y = 0$, $y = H$. Температура пластин $z = H$ и $y = H$ равнялась T_H , а пластин $z = 0$ и $y = 0$ — T_0 . Расчеты проводились для следующего набора параметров: $\text{Kn} = 0.2$; $T_H/T_0 = 2$; $\Delta t = 0.05$; полное число частиц в расчетной области $N = 32000$.

На первом этапе был выполнен ряд расчетов для тестирования рассматриваемых алгоритмов. На фиг. 1 для одномерной задачи приведены зависимости полного времени расчета, количества реальных столкновений и полного количества столкновений (суммы реальных и фиктивных) за один временной шаг от числа частиц в ячейке N_m . Моделировалась задача для $T_H/T_0 = 1$, число



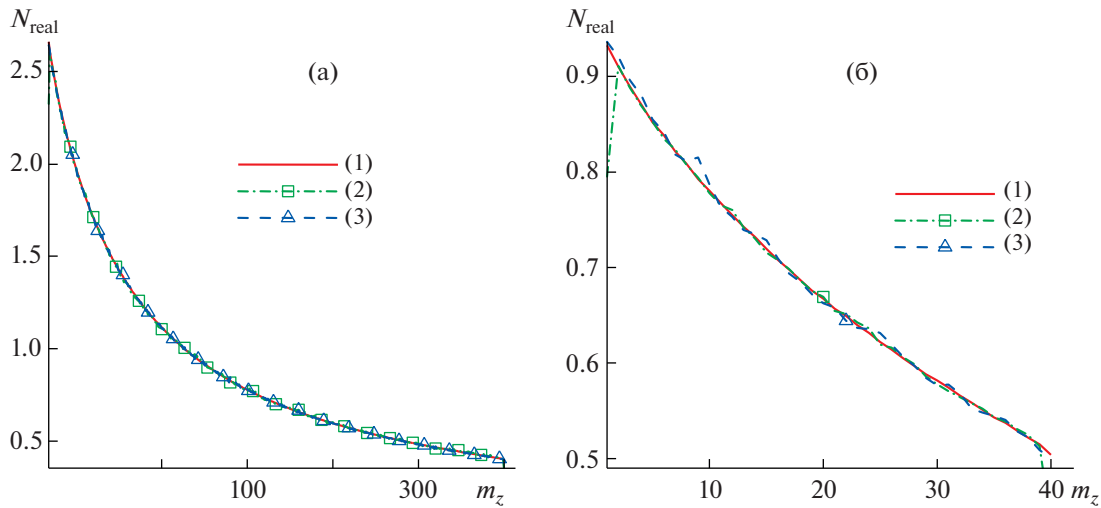
Фиг. 1. Зависимость полного времени расчета t (а) и количества столкновений за один временной шаг N_{col} (б) от числа частиц в ячейке для алгоритмов (1), (2) и (3). $T_H/T_0 = 1$. Здесь линии s относятся к полному числу столкновений, а r — к числу реальных столкновений.

ячеек для сбора информации равнялось 100; размер ячейки для моделирования столкновений (размер окрестности) $h = 0.4$, полное число частиц в расчетной области $N = 100 \cdot N_m$, число временных шагов $K = 200000$. Второй алгоритм существенно проигрывает по вычислительным затратам первому из-за большого количества фиктивных столкновений, поэтому и был разработан приближенный комбинированный алгоритм 3. Отметим, что количество реальных столкновений фактически совпадает для всех трех алгоритмов, что подтверждает корректность их реализации.

На фиг. 2а для одномерной задачи приведено пространственное распределение количества реальных столкновений за временной шаг. На фиг. 2б для двумерной задачи приведено пространственное распределение (вдоль диагонали области от самой холодной точки ($y = z = 0$) до самой горячей ($y = z = H$)) количества реальных столкновений за временной шаг. Отметим хорошее совпадение количества реальных столкновений для всех алгоритмов, за исключением области у границ для второго алгоритма. Это отклонение связано с влиянием границы (уменьшением размера области для выбора партнера для столкновения).

На втором этапе проводилось исследование погрешности, обусловленной пространственной регуляризацией. Для этого вместо неизвестного точного решения задачи использовалось численное решение, насчитанное с заведомо большей точностью. Отметим, что в настоящее время развита методика, позволяющая оценить уровень внешней погрешности (детерминированной и стохастической) [23]. Использование этой методики позволило вычислить “точное” решение с контролируемым малым уровнем внешней погрешности. В табл. 1 приведены параметры алгоритмов, использованные для расчета “точных” решений и полученные при этом уровни внешней погрешности. Здесь δ_i и δ_s — погрешность интерполяции и стохастическая погрешность соответственно. Выбранные временной шаг, размер ячейки и количество частиц в ячейке позволяют рассчитывать на минимальный уровень внутренней погрешности, вносимый этими параметрами алгоритма ПСМ. Отметим, что для каждого алгоритма было получено свое “точное” решение для минимизации возможных других погрешностей.

На фиг. 3а для одномерной задачи представлена зависимость максимума погрешности вычисления плотности n и температуры T_x от размера ячейки (окрестности) регуляризации. Под максимумом погрешности мы подразумеваем максимум модуля разности между “точным” и полученным решением в узлах (т.е. в центрах ячеек для сбора информации). Отметим, что здесь при проведении расчетов использовался тот же набор параметров процесса моделирования (число частиц N , шаг по времени Δt и размер ячейки для сбора информации), что и при вычислении “точного” решения. Это было сделано для того, чтобы исключить из анализа погрешность, обусловленную этими параметрами (см. более подробно в [23]).



Фиг. 2. Пространственное распределение количества реальных столкновений за временной шаг для алгоритмов (1), (2) и (3) для одномерного случая ($T_H/T_0 = 4$) (а) и двумерного случая (вдоль диагонали области от самой холодной точки до самой горячей) ($T_H/T_0 = 2$) (б).

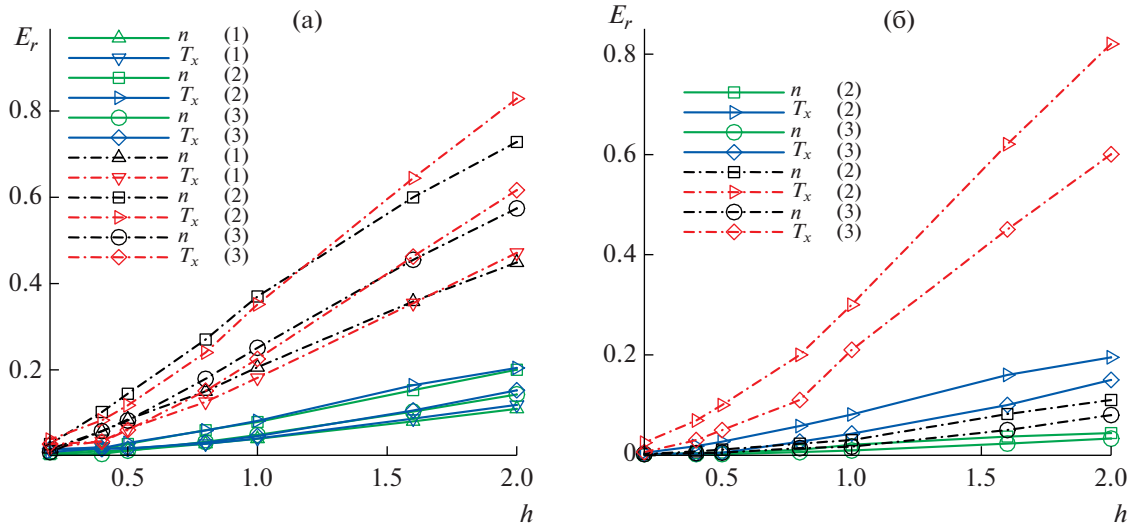
Видим, что максимум погрешности убывает с уменьшением размера ячейки (окрестности) (фиг. 3а). Причем характер убывания близок к линейному. Для первого подхода к регуляризации (алгоритм 1) это соответствует формуле (2.2). Для второго подхода к регуляризации (алгоритмы 2 и 3) ожидалось уменьшение погрешности согласно квадратичному закону в соответствии с формулой (2.3). Возможно, что такое поведение максимума погрешности для второго подхода к регуляризации в проведенных расчетах связано с влиянием граничных условий. Характер изменения погрешности в центральной области близок к квадратичному закону (фиг. 3б).

Из аналитических выкладок при построении оценок (2.1) в [24] можно сделать вывод о том, что константы C_1 и C_2 зависят от градиента решения. Увеличение отношения температур пластин приводит к росту градиента плотности и температуры. Этим можно объяснить рост величины максимума погрешности с увеличением отношения температур пластин (фиг. 3).

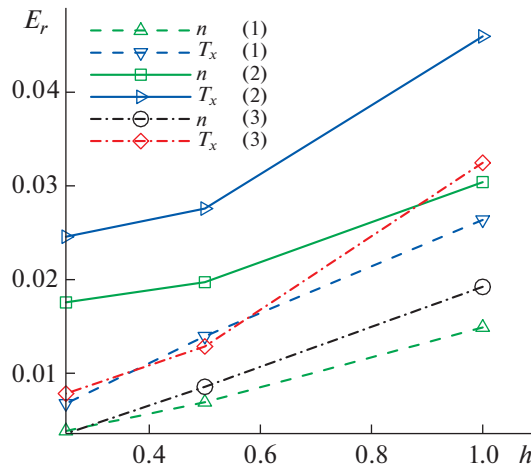
На фиг. 4 представлена зависимость максимума погрешности от размера ячейки (окрестности) регуляризации для двумерной задачи. Погрешность для всех трех алгоритмов убывает с

Таблица 1. Параметры алгоритма для вычисления “точного” решения и оценка получаемой внешней погрешности

Номер алгоритма	M	M_R	K	δ_s	δ_i
$T_H/T_0 = 4, d = 1, N = 8000$					
1	400	400	2400000	0.00158	0.00097
2	400	400	2400000	0.00158	0.00097
3	400	400	2400000	0.00158	0.00097
$T_H/T_0 = 9, d = 1, N = 8000$					
1	400	400	9000000	0.00112	0.00406
2	400	400	1000000	0.00336	0.00406
3	400	400	9000000	0.00112	0.00406
$T_H/T_0 = 2, d = 2, N = 32000$					
1	40×40	40×40	2000000	0.00133	0.00115
2	40×40	40×40	150000	0.00495	0.00115
3	40×40	40×40	2000000	0.00133	0.00115



Фиг. 3. Зависимость максимума погрешности вычисления плотности n и температуры T_x от размера ячейки (окрестности) регуляризации h : (а) для трех алгоритмов (1), (2) и (3) на всей области; (б) для алгоритмов (2) и (3) в центральной области. Одномерная задача Фурье для отношения температур $T_H/T_0 = 4$ (сплошная линия) и $T_H/T_0 = 9$ (штрихпунктир).



Фиг. 4. Зависимость максимума погрешности вычисления плотности n и температуры T_x от размера ячейки (окрестности) регуляризации h для трех алгоритмов (1), (2) и (3) на всей области. Двумерная задача Фурье для отношения температур $T_H/T_0 = 2$.

уменьшением размера ячейки (окрестности), причем характер убывания близок к линейному. Анализ изменения погрешности в центральной точке расчетной области показал похожий характер изменения, что возможно связано с более существенным влиянием граничных условий в силу малых размеров области ($Kn = 0.2$).

На третьем этапе было проведено тестирование полученных в разд. 4 условно-оптимальных параметров. Исходя из близкого к линейному характеру изменения погрешности регуляризации условно-оптимальные параметры определялись по формуле (4.6). Константа H_R была оценена из данных, представленных на фиг. 3а и 4. В табл. 2–4 представлены результаты расчетов числовой плотности газа с использованием условно-оптимальных параметров. Здесь M_{opt} и $M_{R\text{opt}}$ — количество оптимальных ячеек для сбора информации и столкновений соответственно, K_{opt} — оптимальное количество шагов по времени, δ_i и δ_s — погрешность интерполяции и стохастическая по-

Таблица 2. Условно-оптимальные параметры и погрешность вычисления числовой плотности. $T_H/T_0 = 4$

α	M_{opt}	$M_{R\text{opt}}$	J_{opt}	K_{opt}	δ_i	δ_s	δ	t
1 алгоритм								
0.1	189	123	6	367	0.0053	0.0601	0.0654	1.5
0.05	261	171	5	1864	0.0024	0.0278	0.0302	7.0
2 алгоритм								
0.1	205	120	2	1086	0.0036	0.0512	0.0548	19.4
0.05	299	168	1	9171	0.0013	0.0263	0.0276	115.3
3 алгоритм								
0.1	187	400	6	362	0.0052	0.0463	0.0515	2.5
0.05	259	400	5	1841	0.0023	0.0319	0.0352	10.9

Таблица 3. Условно-оптимальные параметры и погрешность вычисления числовой плотности. $T_H/T_0 = 9$. $d = 1$

α	M_{opt}	$M_{R\text{opt}}$	J_{opt}	K_{opt}	δ_i	δ_s	δ	t
1 алгоритм								
0.2	261	177	6	188	0.0097	0.0696	0.0793	0.9
0.1	359	246	5	956	0.0035	0.0312	0.0347	4.1
2 алгоритм								
0.2	291	165	2	548	0.0073	0.0809	0.0882	18.8
0.1	419	234	1	4621	0.0014	0.0441	0.0455	126.6
3 алгоритм								
0.2	291	400	4	276	0.0085	0.0745	0.0830	1.5
0.1	383	400	4	1180	0.0016	0.0343	0.0359	6.4

Таблица 4. Условно-оптимальные параметры и погрешность вычисления числовой плотности. $T_H/T_0 = 2$. $d = 2$

α	M_{opt}	$M_{R\text{opt}}$	J_{opt}	K_{opt}	δ_i	δ_s	δ	t
1 алгоритм								
0.1	9×9	38×38	9	26	0.0144	0.0345	0.0489	0.5
0.05	12×12	49×49	7	206	0.0111	0.0160	0.0271	3.0
0.02	19×19	67×67	6	2777	0.0046	0.0081	0.0127	36.0
2 алгоритм								
0.1	9×9	7×7	2	120	0.019	0.037	0.056	2.5
0.05	13×13	10×10	2	775	0.014	0.024	0.038	27.9
0.02	20×20	15×15	2	9242	0.0061	0.0082	0.0143	729
3 алгоритм								
0.1	8×8	40×40	7	30	0.017	0.032	0.049	0.9
0.05	12×12	40×40	6	224	0.011	0.020	0.031	5.9
0.02	19×19	40×40	4	3890	0.0053	0.0086	0.0139	74.2

грешность, $\delta = \delta_i + \delta_s$ – полная погрешность, t – время расчета в секундах. Сбор информации для оценки макропараметров газа осуществлялся через J_{opt} временных шагов.

Отметим, что в формуле (4.6) для алгоритма 3 имеем $t_4(M_R) = t_5/M_R$. Из этого следует, что чем больше M_R , тем меньше трудоемкость. Однако в представленном подходе к оценке погрешности не учитывается часть внутренней погрешности, обусловленная числом частиц в ячейке. Умень-

шение среднего по времени числа частиц в ячейке до величины менее единицы начинает вносить вклад во внутреннюю погрешность и, соответственно, в точность расчетов [15]. Поэтому в расчетах для третьего алгоритма было использовано значение M_R , соответствующее расчету “точного” решения.

Видим, что использование условно-оптимальных параметров позволяет оценивать числовую плотность с погрешностью δ , не превышающей заданный уровень α . Также отметим, что самым эффективным с точки зрения затрат компьютерного времени для данной задачи является первый алгоритм, далее идет третий алгоритм. Второй алгоритм требует существенно большего времени.

ЗАКЛЮЧЕНИЕ

Метод ПСМ широко применяется для решения задач динамики разреженного газа. Несмотря на продолжительное и широкое использование метода ПСМ, работа по изучению влияния отдельных параметров алгоритма (шаг по времени, размер ячейки, количество шагов по времени, количество частиц в ячейке, соотношение числа реальных частиц и моделируемых частиц) на точность вычислений макропараметров течения газа ввиду очевидной сложности метода далека от завершения.

В представленной работе проведено исследование влияния погрешности, вносимой регуляризацией взаимодействия двух частиц по пространственным переменным, на точность вычислений методом ПСМ. Рассматриваются два подхода к регуляризации: столкновение частиц, находящихся в одной ячейке, и выбор второго партнера для столкновения из окрестности первой частицы. Для численного исследования реализованы три алгоритма метода ПСМ, использующих эти регуляризации. Проведено численное исследование влияния погрешности, вносимой пространственной регуляризацией, на результаты расчетов на примере классической задачи Фурье (одномерной и двумерной).

Построена верхняя граница погрешности метода ПСМ, учитывающая погрешность регуляризации, в метрике пространства непрерывных функций. Получены условно-оптимальные параметры алгоритма, гарантирующие по вероятности уровень заданной погрешности. Проведено тестирование полученных условно-оптимальных параметров на примере задачи Фурье.

СПИСОК ЛИТЕРАТУРЫ

1. Bird G.A. Perception of numerical methods in rarefied gas dynamics // in: *Proc. of 16-th Intern. Symp. on Rarefied Gas Dynamics*, Eds. by E.P. Muntz, D.P. Weaver, D.H. Campbell, V. 118. (Progress in Astro. and Aero., 1989), P. 211.
2. Ivanov M.S., Rogasinsky S.V. Analysis of the numerical techniques of the direct simulation Monte Carlo method in the rarefied gas dynamics // *Soviet J. Numer. Anal. Math. Modelling*. 1988. V. 3. № 6. P. 453.
3. Черемусин Ф.Г. Решение кинетического уравнения Больцмана для высокоскоростных течений // *Ж. вычисл. матем. и матем. физ.* 2006. Т. 46. № 2. С. 329.
4. Титарев В.А., Шахов Е.М. Гибридный метод расчета струи разреженного газа при истечении через очень длинный канал в вакуум // *Ж. вычисл. матем. и матем. физ.* 2020. Т. 60. № 11. С. 1998.
5. Titarev V.A., Frolova A.A., Rykov V.A., Vashchenkov P.V., Shevyrin A.A., Bondar Ye.A. Comparison of the Shakhov kinetic equation and DSMC method as applied to space vehicle aerothermodynamics // *J. Comput. Appl. Math.* 2020. V. 364. 112354.
6. Shi Yangyang, Wu Lei and Shan Xiaowen. Accuracy of high-order lattice Boltzmann method for non-equilibrium gas flow // *J. Fluid Mech.* 2021. V. 907, A25.
7. Bird G.A. *Molecular gas dynamics and the direct simulation of gas flows*. Oxford: Clarendon Press, 1994.
8. Иванов М.С., Рогазинский С.В. Метод прямого статистического моделирования в динамике разреженного газа. Новосибирск: ВЦ СО РАН, 1988.
9. Wagner W. A convergence proof for Bird's direct simulation Monte Carlo method for the Boltzmann equation // *J. Stat. Phys.* 1992. V. 66. P. 101.
10. Rogasinsky S.V. On the pair correlations of particle evolution in the direct statistical simulation // *Monte Carlo methods and applications*. 1996. V. 2. № 1. P. 25.
11. Alexander F.J., Garcia A.L., Alder B.J. Cell size dependence of transport coefficients in stochastic particle algorithms // *Phys. Fluids*. 1998. V. 10. № 6. P. 1540.
<https://doi.org/10.1063/1.869674>
12. Garcia A.L., Wagner W. Time step truncation error in direct simulation Monte Carlo // *Phys. Fluids*. 2000. V. 12. P. 2621.

13. *Hadjiconstantinou N.G.* Analysis of discretization in the direct simulation Monte Carlo // *Phys. Fluids*. 2000. V. 12. P. 2634.
14. *Bobylev A.V., Ohwada T.* The error of the splitting scheme for solving evolutionary equations // *Appl. Math. Lett.* 2001. V. 14. P. 45.
15. *Gallis M.A., Torczynski J., Rader D., and Bird G.A.* Convergence behavior of a new DSMC algorithm // *J. Comput. Phys.* 2009. V. 228. P. 4532.
16. *Rogasinsky S.V., Levin D.A., Ivanov M.S.* Statistical errors of DSMC results for rarefied gas flow // In: *Proc. of 25-th Intern. Symp. on Rarefied Gas Dynamics*, Eds. by A.K. Rebrov, M.S. Ivanov, (Publish House of the Siberian Branch of Russian Academy of Sciences, Novosibirsk; 2007), P. 391.
17. *Plotnikov M.Yu., Shkarupa E.V.* Theoretical and numerical analysis of approaches to evaluation of statistical error of the DSMC method // *Comput. Fluids*. 2014. V. 105. P. 251.
18. *Плотников М.Ю., Шкарупа Е.В.* Комбинированный подход к оцениванию статистической погрешности метода прямого статистического моделирования // *Ж. вычисл. матем. и матем. физ.* 2015. Т. 55. № 11. С. 138.
19. *Khisamutdinov A., Velker N.* Algorithms and numerical implementation of imitation Monte Carlo methods with splitting for problems of the Boltzmann equation // *Journal of Computational and Theoretical Transport*. 2016. V. 45. № 3. P.230.
20. *Rogasinsky Sergey V.* Two variants of Monte Carlo projection method for numerical solution of nonlinear Boltzmann equation // *Russ. J. Numer. Anal. Math. Modelling*. 2019. V. 34. № 3. P. 143.
21. *Myong R.S., Karchani A., Ejtehadi O.* A review and perspective on a convergence analysis of the direct simulation Monte Carlo and solution verification // *Phys. Fluids*. 2019. V. 31. 066101.
22. *Stefanov Stefan, Roohi Ehsan, and Shoja-Sani Ahmad.* A novel transient-adaptive subcell algorithm with a hybrid application of different collision techniques in direct simulation Monte Carlo (DSMC) // *Phys. Fluids*. 2022. V. 34. 092003.
23. *Plotnikov M.Yu., Shkarupa E.V.* Selection of sampling numerical parameters for the DSMC method // *Comput. Fluids*. 2012. V. 58. P. 102.
24. *Rogasinsky S.V.* Statistical modelling of the solution of the nonlinear Boltzmann equation in the spatially inhomogeneous case // *Russ. J. Numer. Anal. Math. Modelling*. 2009. V. 24. № 5. P. 495.
25. *Хисамутдинов А.И.* Влияние области взаимодействий пар частиц на результаты статистического моделирования течений разреженных газов. Препринт ИНГГ СО РАН, Новосибирск. 2021. С. 1–9.
26. *Иванов М.С., Рогазинский С.В.* Экономичные схемы прямого статистического моделирования течений разреженного газа // *Матем. моделирование*. 1988. Т. 1. № 7. С.130.
27. *Shkarupa E.V., Voytishchek A.V.* Optimization of discretely stochastic procedures for globally estimating the solution of an integral equation of the second kind // *Russ. J. Numer. Anal. Math. Modelling*. 1997. V.12. № 6. P. 525.
28. *Плотников М.Ю., Шкарупа Е.В.* Оценка статистической погрешности метода прямого статистического моделирования // *Ж. вычисл. матем. и матем. физ.* 2010. Т. 50. № 2. С. 1.

МАТЕМАТИЧЕСКАЯ
ФИЗИКА

УДК 51-76

NUMERICAL SIMULATION OF PHASE TRANSFER
DURING CRYOSURGERY FOR AN IRREGULAR TUMOR
USING HYBRID APPROACH¹⁾

© 2023 г. Surabhi Nishada^{1,*}, Rama Bhargava^{2,**}

¹ Department of Mathematics, Indian Institute of Technology, Roorkee, India

² D.G. Roorkee College of Engineering, Roorkee, India

*e-mail: snishad@ma.iitr.ac.in,

**e-mail: rama.bhargava@ma.iitr.ac.in

Поступила в редакцию 29.01.2023 г.

Переработанный вариант 23.03.2023 г.

Принята к публикации 28.04.2023 г.

Численное моделирование фазового перехода при криохирургии опухоли неправильной формы с использованием гибридного подхода. Проведено численное решение уравнения теплопроводности в двумерной нестационарной модели Пеннеса, описывающей пораженные опухолевые клетки. Эллиптический кризонд различных размеров берется в центре вычислительной области таким образом, что местоположение зонда фиксируется на протяжении всего вычисления. Фазовый переход происходит за счет воздействия зонда с введением различных наночастиц золота Au, окиси алюминия Al_2O_3 и окиси железа Fe_3O_4 . Эффективность охлаждения этих наночастиц, вводимых при очень низкой температуре, была изучена с помощью гибридного метода конечных элементов, в котором вся область разбивается на две подобласти. Результаты показаны в виде температурного профиля внутри расчетной области. Скорость охлаждения получена для различных наночастиц, и замечено, что введение наночастиц золота очень эффективно увеличивает скорость нагревания в области пораженных опухолевых клеток.

Ключевые слова: криохирургия, гибридный метод конечных элементов, наночастицы.

DOI: 10.31857/S0044466923080112, **EDN:** WTEQNM

¹⁾ Полный текст статьи печатается в английской версии журнала.

УДК 517.977

МЕТОДЫ ПОСТРОЕНИЯ ОЦЕНОК МНОЖЕСТВ ДОСТИЖИМОСТИ В ЗАДАЧЕ МОДЕЛИРОВАНИЯ ПОТОКОВ ЛЮДЕЙ¹⁾

© 2023 г. М. В. Зайцева^{1,*}, П. А. Точилин^{1,**}

¹ 119991 Москва, Ленинские горы, МГУ им. М.В. Ломоносова, ВМК, Россия

* e-mail: zaimarko@gmail.com

** e-mail: tochilin@cs.msu.ru

Поступила в редакцию 23.01.2023 г.
Переработанный вариант 23.01.2023 г.
Принята к публикации 30.03.2023 г.

Работа посвящена математическому моделированию потоков людей в помещении. За основу взята модификация дискретной макромоделей СТМ, построенная на гарантированных оценках. Для описанной модели предложено два способа приближенного вычисления множества достижимости — количества людей в каждой комнате в последующий момент времени. Строятся интервальные оценки и оценки в форме совокупностей двумерных проекций. Предложенные алгоритмы проиллюстрированы численными примерами. Библ. 14. Фиг. 3.

Ключевые слова: моделирование потоков людей, линейное программирование, множество достижимости, нелинейная динамика.

DOI: 10.31857/S0044466923070190, **EDN:** ZXXNGL

1. ВВЕДЕНИЕ

В математическом моделировании потоков людей можно выделить два основных подхода: микромоделирование и макромоделирование. В первом случае необходимо учитывать положение и скорость каждого элемента системы, во втором используются их усредненные характеристики.

Настоящая работа посвящена математическому моделированию движения групп людей в помещении на основе известной макромоделей СТМ (Cell Transmission Model) (см. [1], [2]), которая часто используется при математическом моделировании транспортных потоков и является дискретным аналогом гидродинамической модели LWR (Lighthill, Whitham, Richards) (см. [3]). Подробное описание и возможный метод калибровки для моделирования движения групп людей были предложены в [4]. Существуют и другие ее модификации, например, PedCTM (Pedestrian Cell Transmission Model) (см. [5]). Наряду с дискретными моделями рассматривают также непрерывные модели движения групп людей, построенные как обобщение одномерных транспортных моделей (см. [6]).

Недостатками указанных методов являются предположения об известном поведении людей или известной плотности их распределения в помещении. Поэтому предлагается использовать модификацию из [4] модели СТМ, основанную на гарантированных оценках количества людей в комнатах. Однако построение этих оценок также является непростой задачей. Данная работа посвящена двум методам аппроксимации множества достижимости в последующий момент времени на основе интервальных оценок и оценок двумерных проекций. В дальнейшем такие краткосрочные численные прогнозы могут быть полезны при выборе стратегии управления и при оценке ее качества. Особенно актуальны задачи оптимизации эвакуации людей из помещений (см. [7]), а также перераспределения потоков людей в общественном транспорте (см. [5], [8]). Построение оценок распределений людей в будущие моменты времени является важным элементом решения указанных задач.

¹⁾ Работа выполнена при финансовой поддержке Минобрнауки РФ в рамках реализации программы Московского центра фундаментальной и прикладной математики по соглашению № 075-15-2022-284.

В настоящей статье предложены два метода численного прогноза в модели движения групп людей, описаны алгоритмы построения оценок множества достижимости, а также представлены численные результаты двух подходов.

2. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ

2.1. Схема помещения

План помещения или его части представим в виде ориентированного графа $\Gamma = \{\mathcal{R}, \mathcal{T}\}$, где множество вершин $\mathcal{R} = \{R_1, \dots, R_N\}$ обозначает N комнат (это могут быть не только отдельные комнаты, но и части больших холлов и коридоров), множество ребер $\mathcal{T}$ — переходы между ними. Множество $\mathcal{T}$ состоит из элементов T_{ij} , $i, j \in \{1, \dots, N\}$, T_{ij} обозначает переход между комнатами R_i и R_j , если он существует.

Каждой комнате R_i сопоставлены следующие характеристики:

- S_i — площадь i -й комнаты;
- C_i — максимальная вместимость (наибольшее количество человек, которое может одновременно находиться в i -й комнате);
- $n_i(t)$ — количество людей в комнате i в момент времени t .

Замечание 1. В рассматриваемой модели время t — дискретное.

Замечание 2. Величина $n_i(t)$ является усредненной по небольшому интервалу времени, поэтому функция $n_i(t)$ не обязательно является целочисленной, однако условие $n_i(t) \in [0, C_i]$ выполняется всегда.

Для комнаты R_i , связанной переходами с комнатами $R_{j_1}, \dots, R_{j_k}$, дополнительно вводится $\alpha^{(i)}(t) = (\alpha_0^{(i)}(t), \alpha_{j_1}^{(i)}(t), \dots, \alpha_{j_k}^{(i)}(t))$ — набор коэффициентов расщепления, где k — количество соседних комнат, в которые можно перейти из i -й. При этом $\alpha_s^{(i)}(t) \geq 0$, $s = 0, j_1, \dots, j_k$, и $\alpha_0^{(i)}(t) + \alpha_{j_1}^{(i)}(t) + \dots + \alpha_{j_k}^{(i)}(t) = 1 \ \forall t$. Коэффициент расщепления $\alpha_j^{(i)}(t)$ определяет долю людей, находящихся в комнате R_i в момент времени t , которые собираются перейти в комнату R_j . Коэффициент $\alpha_0^{(i)}(t)$ задает долю людей, которые собираются остаться в комнате R_i .

Каждому переходу T_{ij} сопоставлена тройка (v_{ij}, F_{ij}, w_{ij}) , где $v_{ij} \geq 0$, $F_{ij} \geq 0$, $w_{ij} \geq 0$ — коэффициенты, подобные коэффициентам фундаментальной диаграммы (см. [3]), определяющие зависимость потока людей $f_{ij}(t)$ от плотности $\rho_{ij}(t)$: $f_{ij}(t) = f_{ij}(\rho_{ij}(t))$, v_{ij} характеризует скорость свободного движения, w_{ij} — скорость распространения затора, F_{ij} — максимальную пропускную способность перехода между комнатами.

Коэффициенты v_{ij} , F_{ij} , w_{ij} в рассматриваемой задаче будем считать известными и постоянными. Оценку на их численные значения для каждого случая можно получить, например, с помощью способа, предложенного в [4], опирающегося на молекулярную модель взаимодействия из [9].

2.2. Пересчет числа людей в комнатах

Рассмотрим моменты времени $t = t_0, t_0 + \Delta t, \dots$ и уравнения пересчета количеств людей n_i в комнатах:

$$n_i(t + \Delta t) = n_i(t) + \Delta t \left(\sum_{T_{ji} \in \mathcal{T}} f_{ji}(t) - \sum_{T_{ij} \in \mathcal{T}} f_{ij}(t) \right) \quad \forall i = 1, \dots, N, \quad (1)$$

где $f_{ij}(t)$ — поток людей на соединении T_{ij} в момент времени t , где шаг $\Delta t > 0$ фиксирован, является решением задачи оптимизации

$$\sum_{T_{ij} \in \mathcal{T}} f_{ij}(t) \rightarrow \max_{\{f_{ij}\}} \quad (2)$$

$$0 \leq f_{ij}(t) \leq F_{ij} - f_{ji}(t) \quad \forall i, j : T_{ij} \in \mathcal{T}, \quad (3)$$

$$f_{ij}(t) \leq \alpha_j^{(i)}(t) v_{ij} \frac{n_i(t)}{S_i} \quad \forall i, j : T_{ij} \in \mathcal{T}, \quad (4)$$

$$\sum_{T_{ij} \in \mathcal{T}} \frac{f_{ij}(t)}{w_{ij}} \leq \left(\frac{C_j - n_j(t)}{S_j} \right) \quad \forall j. \quad (5)$$

Неравенство (3) характеризует влияние противонаправленных потоков f_{ij} и f_{ji} друг на друга, (4) определяет количество желающих перейти из комнаты i в комнату j , неравенство (5) является ограничением на суммарный входящий поток в j -ю комнату. Максимизация потоков (2) соответствует тому принципу, что люди всегда будут занимать с максимально возможной скоростью все доступное им свободное место, если это соответствует желаемому направлению их движения.

Замечание 3. В случае соединения T_{ij} с возможным двусторонним движением будем считать, что $F_{ij} = F_{ji}$.

2.3. Решение задачи оптимизации

С учетом ограничений на противонаправленный поток (3) и суммарный входящий поток (5) задача линейного программирования (2)–(5) должна быть рассмотрена сразу для всех комнат одновременно, т.е. в общем случае ее нельзя разбить на несколько подзадач, относящихся к отдельным комнатам. Это значительно усложняет вычисления, а также существенно отличает рассматриваемую математическую модель от аналогичной транспортной модели, где потоки на перекрестках можно рассчитывать независимо.

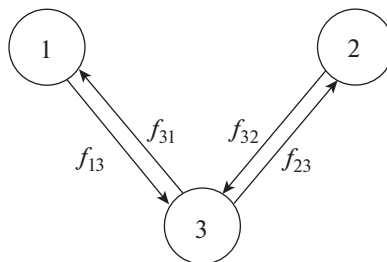
Кроме того, решение может не быть единственным — это соответствует неопределенности в распределении потоков. Если в какой-либо комнате имеет место давка (неравенство (5) превращается в равенство), то, вообще говоря, существует бесконечное количество вариантов того, как при этом распределятся входящие в эту комнату потоки. Одним из выходов в этой ситуации является предположение о том, что входящие потоки должны быть пропорциональны спросу на свободное место в комнате со стороны соседних комнат (т.е. поток пропорционален размеру толпы на входе) или максимальным пропускным способностям (аналогично подходу из [10] для транспортной модели). Это приводит к необходимости вводить дополнительные параметры — коэффициенты приоритетов — и усложнять задачу оптимизации. В настоящей работе мы не будем использовать такой подход, а будем интерпретировать неоднозначность в определении потоков как неопределенность.

Пример. Рассмотрим три комнаты, причем комнаты с номерами 1 и 2 соединены с комнатой 3, но не соединены между собой (фиг. 1).

Пусть $S_i = S$, $C_i = C$, $i = 1, 2, 3$, для каждого перехода T_{ij} : $v_{ij} = v$, $F_{ij} = F$, $w_{ij} = w$, и в некоторый момент t : $\alpha_3^{(1)}(t) = \alpha_3^{(2)}(t) = 1$, $\alpha_1^{(3)}(t) = \alpha_2^{(3)}(t) = 1/2$, $n_1(t) = n_2(t) = n_3(t) = C/2$. Пусть также $v > 2w$. Требуется определить потоки $f_{13}(t)$, $f_{23}(t)$, $f_{31}(t)$, $f_{32}(t)$.

Решая задачу оптимизации (2)–(5), получим, что $f_{13} + f_{23} + f_{31} + f_{32} \rightarrow \max$ при условиях

$$f_{13} + f_{31} \leq F, \quad f_{23} + f_{32} \leq F, \quad f_{13} + f_{23} \leq \frac{wC}{2S}, \quad f_{31} \leq \frac{wC}{2S}, \quad f_{32} \leq \frac{wC}{2S}. \quad (6)$$



Фиг. 1. Схема помещения из примера.

Пусть $2F \leq wC/(2S)$. Тогда последние три неравенства в (6) являются следствием первых двух, а значит, максимальный суммарный поток равен $2F$, и максимум достигается в любой точке $(f_{13}, f_{31}, f_{23}, f_{32})$, для которой $f_{13} \in [0, F]$, $f_{23} \in [0, F]$, $f_{13} + f_{31} = F$, $f_{23} + f_{32} = F$.

Пусть теперь $wC/S \leq F$. Тогда, наоборот, первые два неравенства (6) с учетом неотрицательности потоков являются следствием из последних трех, а потому максимальный суммарный поток равен $3wC/(2S)$, и максимум достигается в любой точке $(f_{13}, f_{31}, f_{23}, f_{32})$, для которой $f_{13} \geq 0$, $f_{23} \geq 0$, $f_{13} + f_{23} = wC/(2S)$, $f_{31} = wC/(2S)$, $f_{32} = wC/(2S)$. Видим, что в двух разобранных случаях решение задачи оптимизации является неединственным.

Вне зависимости от предположений о распределении потоков в конфликтных ситуациях будем считать, что в каждый момент времени количество людей в той или иной комнате точно не известно. Можно построить оценку, используя особенности рассматриваемой математической модели, а также поступающие результаты измерений — неточную информацию о текущем количестве людей в некоторых комнатах, например, в результате обработки изображений с камер видеонаблюдения.

3. ОЦЕНИВАНИЕ МНОЖЕСТВА ДОСТИЖИМОСТИ

Предполагается, что задано некоторое начальное множество $\mathcal{X}(t_0)$ возможных значений переменных $n_i(t_0)$, $i = 1, \dots, N$. Множеством достижимости $\tilde{\mathcal{X}}(t)$ (см. [11]) в некоторый момент времени $t \geq t_0$ называется множество, содержащее все векторы $\bar{n}(t) \in \mathbb{R}^N$, которые могут быть получены в качестве решений задач (1)–(5) для некоторого начального условия $\bar{n}(t_0) \in \mathcal{X}(t_0)$. Поиск точного множества достижимости $\tilde{\mathcal{X}}(t)$ для нелинейной системы является сложной задачей, поэтому будем искать $\mathcal{X}(t)$ — внешнюю оценку множества достижимости в некоторый момент времени $t \geq t_0$. На основании этой оценки можно будет судить о сложившейся ситуации, предсказывать места возможного возникновения давки и, в частности, анализировать качество предлагаемой стратегии управления. Далее будут описаны два разных вида таких множеств, задающих внешние оценки.

Основной идеей, использованной в указанных ниже алгоритмах, является разделение большой задачи линейного программирования (2)–(5) на множество более простых подзадач, относящихся к парам связанных между собой комнат, что позволяет легко интерпретировать получающиеся при решении соотношения геометрически.

3.1. Интервальные оценки

Предположим, что задано начальное множество $\mathcal{X}(t_0)$ возможных значений $n_i(t_0)$ во всех комнатах, являющееся декартовым произведением интервалов:

$$\mathcal{X}(t_0) = \{[n_{i,-}(t_0), n_{i,+}(t_0)], i = 1, \dots, N\},$$

где $0 \leq n_{i,-}(t_0) \leq n_{i,+}(t_0) \leq C_i$ — заданные величины. Необходимо найти внешнюю оценку множества $\mathcal{X}(t)$ в некоторый момент времени $t > t_0$ также в виде декартового произведения интервалов:

$$\mathcal{X}(t) = \{[n_{i,-}(t), n_{i,+}(t)], i = 1, \dots, N\}.$$

Для любой траектории системы (1), для которой $n_i(t_0) \in [n_{i,-}(t_0), n_{i,+}(t_0)]$, $i = 1, \dots, N$, обязательно должно быть выполнено условие $n_i(t) \in [n_{i,-}(t), n_{i,+}(t)]$, $i = 1, \dots, N$ (подробнее о применении интервальных оценок для множества достижимости см. [12], [13]).

Фиксируем произвольный момент времени $t \geq t_0$, и будем считать, что множество $\mathcal{X}(t)$ уже найдено. Найдем множество $\mathcal{X}(t + \Delta t)$ в следующий момент времени. Для этого сперва необходимо вычислить потоки $f_{ij}(t)$ или же множество их значений $\{f_{ij}(t)\}$, учитывая неединственность решения задачи и неопределенность в состоянии системы. Далее опишем алгоритм построения множеств возможных значений $f_{ij}(t)$.

1. Будем использовать вспомогательные переменные σ_{ij} для всех $i, j: T_{ij} \in \mathcal{T}$. Изначально положим $\sigma_{ij} = 0 \forall i, j$. Переменная σ_{ij} имеет смысл минимального суммарного входящего потока в комнату j из комнат $k \neq i, T_{kj} \in \mathcal{T}$.

Для каждой пары $i, j, T_{ij} \in \mathcal{T}$, необходимо определить множество $\mathcal{F}_{ij} \subset \mathbb{R}_+^2$, содержащее все возможные пары величин f_{ij} и f_{ji} . Далее всюду предполагается, что для любой пары индексов i, j множества $\mathcal{F}_{ij}$ и $\mathcal{F}_{ji}$ отличаются друг от друга лишь перестановкой компонент входящих в них векторов.

Изначально положим

$$\mathcal{F}_{ij} = \{(f_{ij}, f_{ji}) : 0 \leq f_{ij} \leq F_{ij} - f_{ji}, 0 \leq f_{ji} \leq F_{ji} - f_{ij}\}. \quad (7)$$

Зафиксируем некоторое значение параметра $\Theta \in \mathbb{N}$, а также положим $\theta = 1$. Параметр θ будет соответствовать номеру итерации основного цикла алгоритма, а Θ — максимальное количество таких итераций.

2. Необходимо уточнить множества $\mathcal{F}_{ij}$ так, чтобы были выполнены неравенства² (4), (5) с учетом вспомогательной переменной σ_{ij} :

$$f_{ij} \leq \alpha_j^{(i)} v_{ij} \frac{n_i}{S_i}, \quad f_{ji} \leq \alpha_i^{(j)} v_{ji} \frac{n_j}{S_j}, \quad (8)$$

$$f_{ij} \leq w_{ij} \left(\frac{C_j - n_j}{S_j} \right) - \sigma_{ij}, \quad f_{ji} \leq w_{ji} \left(\frac{C_i - n_i}{S_i} \right) - \sigma_{ji}, \quad (9)$$

для каких-либо значений $(n_i, n_j) \in [n_{i,-}, n_{i,+}] \times [n_{j,-}, n_{j,+}]$.

Сначала отдельно рассмотрим пару неравенств

$$f_{ij} \leq \alpha_j^{(i)} v_{ij} \frac{n_i}{S_i}, \quad f_{ji} \leq w_{ji} \left(\frac{C_i - n_i}{S_i} \right) - \sigma_{ji}.$$

При каждом фиксированном значении n_i они определяют на плоскости (f_{ij}, f_{ji}) , где $f_{ij} \geq 0$, $f_{ji} \geq 0$ — некоторый прямоугольник. Так как точное значение n_i неизвестно, то построим объединение всех таких прямоугольников при $n_i \in [n_{i,-}, n_{i,+}]$. Верхняя правая вершина таких прямоугольников будет находиться на прямой

$$\frac{f_{ij} S_i}{\alpha_j^{(i)} v_{ij}} = C_i - \frac{(f_{ji} + \sigma_{ji}) S_i}{w_{ji}} \Leftrightarrow f_{ji} + \frac{w_{ji}}{\alpha_j^{(i)} v_{ij}} f_{ij} = \frac{w_{ji} C_i}{S_i} - \sigma_{ji}.$$

Таким образом, искомое объединение прямоугольников $\hat{\mathcal{F}}_{ij}$ задается неравенствами

$$\left\{ (f_{ij}, f_{ji}) : f_{ij} \leq \alpha_j^{(i)} v_{ij} \frac{n_{i,+}}{S_i}, f_{ji} \leq w_{ji} \left(\frac{C_i - n_{i,-}}{S_i} \right) - \sigma_{ji}, f_{ji} + \frac{w_{ji}}{\alpha_j^{(i)} v_{ij}} f_{ij} \leq \frac{w_{ji} C_i}{S_i} - \sigma_{ji} \right\}.$$

Тогда уточненное множество $\mathcal{F}_{ij}$ можно представить в виде пересечения трех множеств

$$\mathcal{F}_{ij} := \mathcal{F}_{ij} \cap \hat{\mathcal{F}}_{ij} \cap \hat{\mathcal{F}}_{ji}$$

(формула для множества $\hat{\mathcal{F}}_{ij}$ отличается от $\hat{\mathcal{F}}_{ji}$ перестановкой индексов).

3. Предположим, что вычисления из предыдущего пункта проделаны для фиксированного j и всех k таких, что $T_{kj} \in \mathcal{T}$, т.е. известны множества $\mathcal{F}_{kj}$.

Множество $\mathcal{F}_{ij}$ может быть избыточным, так как содержит малые величины f_{ij} , f_{ji} , которые заведомо не являются решением задачи максимизации (2). Поэтому определим нижние границы

² Всюду в формулах для краткости опущена зависимость функций f_{ij} , f_{ji} , $\alpha_j^{(i)}$ от переменной t .

на эти величины, учитывая, что движущиеся люди всегда по максимуму заполняют все доступное им свободное место. Допустимые значения потоков f_{ij} и f_{ji} задаются неравенствами

$$0 \leq f_{ij} \leq w_{ij} \left(\frac{C_j - n_j}{S_j} - \sum_{\substack{k \neq i, \\ T_{kj} \in \mathcal{T}}} \frac{f_{kj}}{w_{kj}} \right), \quad 0 \leq f_{ji} \leq w_{ji} \left(\frac{C_i - n_i}{S_i} - \sum_{\substack{m \neq j, \\ T_{mi} \in \mathcal{T}}} \frac{f_{mi}}{w_{mi}} \right), \quad (10)$$

$$f_{ij} + f_{ji} \leq F_{ij}, \quad f_{ij} \leq \alpha_j^{(i)} v_{ij} \frac{n_i}{S_i}, \quad f_{ji} \leq \alpha_i^{(j)} v_{ji} \frac{n_j}{S_j}. \quad (11)$$

(а) Пусть в переходе между комнатами i и j допустимо движение в противоположных направлениях. Предположим, что нам известны величины n_i , n_j , $\bar{f}_{kj}$, $\bar{f}_{mi}$, где

$$\bar{f}_{kj} = \sum_{\substack{k \neq i, \\ T_{kj} \in \mathcal{T}}} \frac{f_{kj}}{w_{kj}} \quad \text{и} \quad \bar{f}_{mi} = \sum_{\substack{m \neq j, \\ T_{mi} \in \mathcal{T}}} \frac{f_{mi}}{w_{mi}}.$$

Рассмотрим задачу линейного программирования $f_{ij} + f_{ji} \rightarrow \max$ при условиях (10), (11). Если неравенство

$$\min \left\{ \alpha_j^{(i)} v_{ij} \frac{n_i}{S_i}, w_{ij} \left(\frac{C_j - n_j}{S_j} - \bar{f}_{kj} \right) \right\} + \min \left\{ \alpha_i^{(j)} v_{ji} \frac{n_j}{S_j}, w_{ji} \left(\frac{C_i - n_i}{S_i} - \bar{f}_{mi} \right) \right\} \leq F_{ij} \quad (12)$$

выполнено, то решение этой задачи единственно. Найдем первую компоненту максимизатора

$$f_{ij}^*(n_i, n_j, \bar{f}_{kj}, \bar{f}_{mi}) = \min \left\{ \alpha_j^{(i)} v_{ij} \frac{n_i}{S_i}, w_{ij} \left(\frac{C_j - n_j}{S_j} - \bar{f}_{kj} \right) \right\}.$$

Если же указанное выше неравенство не выполнено, то решения задачи образуют отрезок прямой $f_{ij} + f_{ji} = F_{ij}$. В качестве компоненты максимизатора $f_{ij}^*(n_i, n_j, \bar{f}_{kj}, \bar{f}_{mi})$ возьмем ту точку этого отрезка, которая соответствует наименьшему значению f_{ij} :

$$f_{ij}^*(n_i, n_j, \bar{f}_{kj}, \bar{f}_{mi}) = F_{ij} - \min \left\{ \alpha_i^{(j)} v_{ji} \frac{n_j}{S_j}, w_{ji} \left(\frac{C_i - n_i}{S_i} - \bar{f}_{mi} \right) \right\}.$$

Теперь учтем, что параметры n_i , n_j , $\bar{f}_{kj}$, $\bar{f}_{mi}$ являются компонентами неопределенности. Следовательно, для потока f_{ij} справедлива оценка

$$f_{ij} \geq r_{ij} = \max \left\{ 0, \min \left\{ f_{ij}^*(n_i, n_j, \bar{f}_{kj}, \bar{f}_{mi}) : n_i \in [n_{i,-}, n_{i,+}], n_j \in [n_{j,-}, n_{j,+}], \right. \right. \\ \left. \left. f_{kj} \in \text{Pr}_1 \mathcal{F}_{kj}, f_{mi} \in \text{Pr}_1 \mathcal{F}_{mi}, \forall k \neq i, m \neq j \right\} \right\},$$

где $\text{Pr}_1 \mathcal{F}_{kj} = \{f_{kj}\}$. Используя геометрическую интерпретацию указанной выше задачи линейного программирования, легко установить, что минимум f_{ij}^* будет достигаться при $n_j = n_{j,+}$, $n_i = n_{i,-}$, $f_{kj} = f_{kj,\max}$, $f_{mi} = f_{mi,\min}$. Здесь $f_{kj,\max} = \max \{f_{kj} : (f_{kj}, f_{jk}) \in \mathcal{F}_{kj}\}$, $f_{mi,\min} = \min \{f_{mi} : (f_{mi}, f_{im}) \in \mathcal{F}_{mi}\}$.

Аналогичную оценку можно выписать и для потока f_{ji} . Величина r_{ij} имеет смысл гарантированного ресурса свободного места в ячейке j , который может быть использован людьми, идущими из ячейки i .

(б) Рассмотрим частный случай, когда в переходе между комнатами i и j движение является односторонним (из i -й комнаты в j -ю). Тогда $r_{ji} = 0$, а для r_{ij} можно получить более простую формулу:

$$r_{ij} = 0 \vee \min \left\{ \alpha_j^{(i)} v_{ij} \frac{n_{i,-}}{S_i}, F_{ij}, w_{ij} \left(\left(\frac{C_j - n_{j,+}}{S_j} \right) - \sum_{\substack{k \neq i, \\ T_{kj} \in \mathcal{T}}} \frac{f_{kj, \max}}{w_{kj}} \right) \right\}.$$

Если $r_{ij} > 0$, то модифицируем множество $\mathcal{F}_{ij} := \mathcal{F}_{ij} \cap \{(f_{ij}, f_{ji}) : f_{ij} \geq r_{ij}\}$.

Указанные модификации множеств $\mathcal{F}_{ij}$ проведем для всех возможных пар индексов i, j .

4. Если в п. 3 было изменено множество $\mathcal{F}_{kj}$ для некоторых значений индексов k, j (т.е. увеличена нижняя граница возможного потока f_{kj}), то пересчитаем значение σ_{ij} для фиксированного j и разных $i \neq k$. А именно,

$$\sigma_{ij} = w_{ij} \sum_{\substack{k \neq i, \\ T_{kj} \in \mathcal{T}}} \frac{f_{kj, \min}}{w_{kj}} > 0. \quad (13)$$

5. Если в п. 4 были изменены значения σ_{ij} хотя бы для одной пары индексов i, j , и при этом $\theta < \Theta$, то значение θ увеличивается на 1, и алгоритм снова переходит к п. 2. Иначе алгоритм завершает работу, на выходе — совокупность множеств $\mathcal{F}_{ij}$.

Для полученных множеств справедливы следующие утверждения.

Теорема 1. Пусть f_{ij}^* — какое-либо решение задачи оптимизации (2)–(5) в момент времени t при условии, что истинное состояние системы находится в некоторой точке множества $\mathcal{X}(t)$. Тогда $f_{ij}^* \in \mathcal{F}_{ij} \forall i, j$.

Теорема 2. Пусть комнаты с номерами i, j являются частью системы связанных комнат с односторонним движением без разветвлений, т.е.

$$\alpha_j^{(i)} = 1, \quad \alpha_i^{(j)} = 0, \quad \alpha_j^{(k)} = 0 \quad \forall k \neq i.$$

Тогда множество $\mathcal{F}_{ij}$ совпадет со множеством допустимых потоков f_{ij} , рассчитанным на основании интервальных оценок при моделировании транспортного потока (см. [14]):

$$f_{ij} \in \left[\max \left\{ 0, \min \left\{ v_{ij} \frac{n_{i,-}}{S_i}, F_{ij}, w_{ij} \frac{C_j - n_{j,+}}{S_j} \right\} \right\}, \min \left\{ v_{ij} \frac{n_{i,+}}{S_i}, F_{ij}, w_{ij} \frac{C_j - n_{j,-}}{S_j} \right\} \right]. \quad (14)$$

Доказательство. Найдем допустимые значения f_{ij}, f_{ji} с помощью приведенного алгоритма. В пунктах 1 и 2 получаем

$$f_{ji} = 0, \quad 0 \leq f_{ij} \leq \min \left\{ F_{ij}, \alpha_j^{(i)} v_{ij} \frac{n_{i,+}}{S_i}, w_{ij} \frac{C_j - n_{j,-}}{S_j} \right\}, \quad \sigma_{ij} = 0.$$

Далее, в п. 3 алгоритма

$$r_{ij} = \max \left\{ 0, \min \left\{ \alpha_j^{(i)} v_{ij} \frac{n_{i,-}}{S_i}, F_{ij}, w_{ij} \left(\frac{C_j - n_{j,+}}{S_j} \right) \right\} \right\}, \quad f_{ij} \geq r_{ij}.$$

В п. 4 получим $\sigma_{ij} = 0$, а потому на дальнейших итерациях работы алгоритма (при $\theta > 1$) уточнение верхней оценки для f_{ij} не проводится. Объединяя полученные неравенства и учитывая $\alpha_j^{(i)} = 1$, получим (14). Теорема доказана.

В итоге на основании оценки потоков можно аппроксимировать множества достижимости

$$\mathcal{X}(t + \Delta t) = \{[n_{i,-}(t + \Delta t), n_{i,+}(t + \Delta t)], i = 1, \dots, N\},$$

$$n_{i,-}(t + \Delta t) = \max \left\{ 0, n_{i,-}(t) + \Delta t \sum_{T_{ij} \in \mathcal{T}} \min \{ f_{ji} - f_{ij} : (f_{ij}, f_{ji}) \in \mathcal{F}_{ij} \} \right\},$$

$$n_{i,+}(t + \Delta t) = \min \left\{ C_i, n_{i,+}(t) + \Delta t \sum_{T_{ij} \in \mathcal{T}} \max \{ f_{ji} - f_{ij} : (f_{ij}, f_{ji}) \in \mathcal{F}_{ij} \} \right\}.$$

Замечание 4. Рассмотрим частный случай, когда $w_{ij} = w \ \forall i, j$, и для некоторого номера i выполнено следующее неравенство:

$$\sum_{j: T_{ji} \in \mathcal{T}} \min \left\{ \alpha_i^{(j)} v_{ji} \frac{n_{j,-}}{S_j}, f_{ji} - f_{ji, \max} \right\} \geq w \frac{C_i - n_{i,-}}{S_i}.$$

Тогда для реальных потоков неравенство (5) обязательно должно обратиться в равенство (это условие затора на входах в комнату с номером i). Значит, суммарный входящий поток будет не меньше, чем $w(C_i - n_{i,-})/S_i$, т.е.

$$n_{i,-}(t + \Delta t) := \max \left\{ n_{i,-}(t + \Delta t), n_{i,-}(t) + \Delta t \left(w \frac{C_i - n_{i,-}}{S_i} - \sum_{T_{ij} \in \mathcal{T}} f_{ij, \max} \right) \right\},$$

где в правой части $n_{i,-}(t + \Delta t)$ — ранее полученная оценка. Эта формула в отдельных случаях позволяет улучшить оценку снизу на величину n_i .

3.2. Оценки двумерных проекций

Из приведенных выше уравнений видно, что потоки f_{ij} и f_{ji} одновременно зависят от двух фазовых переменных: n_i и n_j . В случае интервальных оценок эти переменные аппроксимируются с помощью двух независимых отрезков — одномерных проекций множества достижимости. Однако, если их заменить на двумерные, можно получить более качественные оценки.

Теперь рассмотрим оценки множества достижимости вида

$$\mathcal{X}(t) = \{ \mathcal{N}_{ij}(t), i, j = 1, \dots, N : T_{ij} \in \mathcal{T} \},$$

где $\mathcal{N}_{ij}(t) \subset [0, C_i] \times [0, C_j]$ — некоторое компактное множество, содержащее возможные значения пары переменных $(n_i(t), n_j(t))$. Будем далее считать, что множество $\mathcal{N}_{ij}(t)$ является выпуклым многоугольником, задаваемым совокупностью своих вершин $g^{(ij,1)}(t), \dots, g^{(ij,m)}(t)$, где $m = m(t, i, j) \in \mathbb{N}$.

Далее приведем алгоритм построения оценки возможных значений потоков $f_{ij}(t)$ при условии, что в некоторый произвольный момент времени $t \geq t_0$ заданы множества $\mathcal{N}_{ij}(t)$ указанного выше вида.

1. Как и в случае интервальных оценок, будем использовать вспомогательные переменные σ_{ij} для всех $i, j, T_{ij} \in \mathcal{T}$. Изначально положим $\sigma_{ij} = 0 \ \forall i, j$.

Для каждой пары $i, j, T_{ij} \in \mathcal{T}$, необходимо определить множество $\mathcal{F}_{ij} \subset \mathbb{R}_+^2$, содержащее все возможные пары величин f_{ij} и f_{ji} . Множества $\mathcal{F}_{ij}$ и $\mathcal{F}_{ji}$ отличаются друг от друга лишь перестановкой компонент входящих в них векторов.

Изначально определим множества $\mathcal{F}_{ij}$ аналогично формуле (7).

Зафиксируем некоторое значение параметра $\Theta \in \mathbb{N}$, а также положим $\theta = 1$.

2. Необходимо уточнить множества $\mathcal{F}_{ij}$ так, чтобы были выполнены неравенства (8), (9) для какой-либо пары значений $(n_i, n_j) \in \mathcal{N}_{ij}(t)$. Пусть

$$\hat{\mathcal{F}}_{ij} = \bigcup_{(n_i, n_j) \in \mathcal{N}_{ij}(t)} \left\{ (f_{ij}, f_{ji})^T : 0 \leq f_{ij} \leq \min \left\{ \alpha_j^{(i)} v_{ij} \frac{n_i}{S_i}, w_{ij} \left(\frac{C_j - n_j}{S_j} \right) - \sigma_{ij} \right\}, \right. \\ \left. 0 \leq f_{ji} \leq \min \left\{ \alpha_i^{(j)} v_{ji} \frac{n_j}{S_j}, w_{ji} \left(\frac{C_i - n_i}{S_i} \right) - \sigma_{ji} \right\} \right\}. \quad (15)$$

Тогда уточненное множество $\mathcal{F}_{ij}$ можно представить в виде $\mathcal{F}_{ij} := \mathcal{F}_{ij} \cap \hat{\mathcal{F}}_{ij} \cap \hat{\mathcal{F}}_{ji}$. Покажем, как можно построить выпуклые многоугольники, оценивающие сверху множества $\hat{\mathcal{F}}_{ij}$. Введем обозначения для двух вспомогательных прямых на плоскости:

$$H_1 = \left\{ (n_i, n_j)^T : \alpha_j^{(i)} v_{ij} \frac{n_i}{S_i} = w_{ij} \left(\frac{C_j - n_j}{S_j} \right) - \sigma_{ij} \right\}, \\ H_2 = \left\{ (n_i, n_j)^T : \alpha_i^{(j)} v_{ji} \frac{n_j}{S_j} = w_{ji} \left(\frac{C_i - n_i}{S_i} \right) - \sigma_{ji} \right\}.$$

Эти прямые разбивают $\mathcal{N}_{ij}(t)$ на несколько многоугольных частей (не более четырех), в каждой из которых оба минимума из (15) раскрываются однозначно. Вершины k -й такой части обозначим через $z^{(k),1}, \dots, z^{(k),s_k}$. Рассмотрим опорную функцию³ ко множеству $\mathcal{F}_{ij}$ в направлении $l \in \mathbb{R}^2$:

$$\rho(l | \hat{\mathcal{F}}_{ij}) = \sup_{(n_i, n_j) \in \mathcal{N}_{ij}(t)} \left\{ \max \left\{ 0, l_1 \min \left\{ \alpha_j^{(i)} v_{ij} \frac{n_i}{S_i}, w_{ij} \frac{C_j - n_j}{S_j} - \sigma_{ij} \right\} \right\} + \right. \\ \left. + \max \left\{ 0, l_2 \min \left\{ \alpha_i^{(j)} v_{ji} \frac{n_j}{S_j}, w_{ji} \frac{C_i - n_i}{S_i} - \sigma_{ji} \right\} \right\} \right\}. \quad (16)$$

Здесь нужно учесть, что если какая-то из упомянутых выше четырех частей множества $\mathcal{N}_{ij}(t)$ пуста, то соответствующий максимум по ее вершинам отсутствует. Таким образом, $\hat{\mathcal{F}}_{ij}$ представляет собой многоугольник, являющийся выпуклой оболочкой точек

$$(0, 0)^T, \quad \left(\alpha_j^{(i)} v_{ij} \frac{z_1^{(1),s}}{S_i}, \alpha_i^{(j)} v_{ji} \frac{z_2^{(1),s}}{S_j} \right)^T, \quad \left(\alpha_j^{(i)} v_{ij} \frac{z_1^{(2),s}}{S_i}, w_{ji} \frac{C_i - z_1^{(2),s}}{S_i} - \sigma_{ji} \right)^T, \\ \left(w_{ij} \frac{C_j - z_2^{(3),s}}{S_j} - \sigma_{ij}, \alpha_i^{(j)} v_{ji} \frac{z_2^{(3),s}}{S_j} \right)^T, \quad \left(w_{ij} \frac{C_j - z_2^{(4),s}}{S_j} - \sigma_{ij}, w_{ji} \frac{C_i - z_1^{(4),s}}{S_i} - \sigma_{ji} \right)^T, \\ \left(0, \max_s \left\{ \alpha_i^{(j)} v_{ji} \frac{z_2^{(1),s}}{S_j}, w_{ji} \frac{C_i - z_1^{(2),s}}{S_i} - \sigma_{ji}, \alpha_i^{(j)} v_{ji} \frac{z_2^{(3),s}}{S_j}, w_{ji} \frac{C_i - z_1^{(4),s}}{S_i} - \sigma_{ji} \right\} \right)^T, \\ \left(\max_s \left\{ \alpha_j^{(i)} v_{ij} \frac{z_1^{(1),s}}{S_i}, \alpha_j^{(i)} v_{ij} \frac{z_1^{(2),s}}{S_i}, w_{ij} \frac{C_j - z_2^{(3),s}}{S_j} - \sigma_{ij}, w_{ij} \frac{C_j - z_2^{(4),s}}{S_j} - \sigma_{ij} \right\}, 0 \right)^T,$$

где s — вершины соответствующего многоугольника с номером $k = 1, 2, 3, 4$.

3. Предположим, что известны множества $\mathcal{F}_{kj}$. Уточним множество $\mathcal{F}_{ij}$ аналогично тому, как это было сделано в случае с интервальными оценками.

(а) Пусть в переходе между комнатами i и j допустимо движение в двух противоположных направлениях. Определим величины

$$f_{kj,\max} = \max \{ f_{kj} : (f_{kj}, f_{jk})^T \in \mathcal{F}_{kj} \}, \quad f_{mi,\min} = \min \{ f_{mi} : (f_{mi}, f_{im})^T \in \mathcal{F}_{mi} \}.$$

³ Опорная функция для некоторого компакта X в направлении l определяется соотношением $\rho(l | X) = \sup_{x \in X} \langle l, x \rangle$.

Рассмотрим разные варианты раскрытия минимумов в левой части (12) и найдем минимальные значения величины f_{ij}^* . Пусть N_1 — это та часть множества $\mathcal{N}_{ij}(t)$, для которой выполнено неравенство

$$\min \left\{ w_{ij} \left(\frac{C_j - n_j}{S_j} - \sum_{\substack{k \neq i, \\ T_{kj} \in \mathcal{T}}} \frac{f_{kj, \max}}{w_{kj}} \right), \alpha_j^{(i)} v_{ij} \frac{n_i}{S_i} \right\} + \min \left\{ w_{ji} \left(\frac{C_i - n_i}{S_i} - \sum_{\substack{m \neq j, \\ T_{mi} \in \mathcal{T}}} \frac{f_{mi, \min}}{w_{mi}} \right), \alpha_i^{(j)} v_{ji} \frac{n_j}{S_j} \right\} \leq F_{ij},$$

и $N_2 = \mathcal{N}_{ij}(t) \setminus N_1$. Подсчитаем следующие величины:

$$F_1 = \min_{(n_i, n_j) \in N_1} \left\{ \min \left\{ w_{ij} \left(\frac{C_j - n_j}{S_j} - \sum_{\substack{k \neq i, \\ T_{kj} \in \mathcal{T}}} \frac{f_{kj, \max}}{w_{kj}} \right), \alpha_j^{(i)} v_{ij} \frac{n_i}{S_i} \right\} \right\},$$

$$F_2 = F_{ij} - \max_{(n_i, n_j) \in N_2} \left\{ \min \left\{ w_{ji} \left(\frac{C_i - n_i}{S_i} - \sum_{\substack{m \neq j, \\ T_{mi} \in \mathcal{T}}} \frac{f_{mi, \min}}{w_{mi}} \right), \alpha_i^{(j)} v_{ji} \frac{n_j}{S_j} \right\} \right\}.$$

Теперь определим величину

$$r_{ij} = 0 \vee \min \{F_1, F_2\}.$$

Как и в предыдущем разделе, эта величина имеет смысл максимального гарантированного ресурса свободного места в комнате j , который будет использован людьми из комнаты i .

(б) Рассмотрим частный случай, когда в переходе между комнатами i и j движение является односторонним, из i -й в j -ю. Тогда $r_{ji} = 0$,

$$r_{ij} = 0 \vee \min_{(n_i, n_j) \in \mathcal{N}_{ij}(t)} \left\{ \min \left\{ \alpha_j^{(i)} v_{ij} \frac{n_i}{S_i}, F_{ij}, w_{ij} \left(\frac{C_j - n_j}{S_j} - \sum_{\substack{k \neq i, \\ T_{kj} \in \mathcal{T}}} \frac{f_{kj, \max}}{w_{kj}} \right) \right\} \right\}.$$

Если $r_{ij} > 0$, то модифицируем множество $\mathcal{F}_{ij} := \mathcal{F}_{ij} \cap \{(f_{ij}, f_{ji}) : f_{ij} \geq r_{ij}\}$. Аналогичные модификации $\mathcal{F}_{ij}$ проведем для всех возможных пар i, j .

4. Если в п. 3 было изменено множество $\mathcal{F}_{kj}$ для некоторых значений индексов k, j , то пересчитаем значение σ_{ij} для фиксированного j и разных $i \neq k$ по формуле (13).

5. Если в п. 4 были изменены значения σ_{ij} хотя бы для одной пары индексов i, j , и при этом $\theta < \Theta$, то значение θ увеличивается на 1, и алгоритм снова переходит к п. 2. Иначе алгоритм завершает работу, на выходе — совокупность множеств $\mathcal{F}_{ij}$.

На основании оценки потоков можно аппроксимировать множества достижимости в момент времени $t + \Delta t$. При этом для потоков между двумя соседними комнатами i и j будем использовать оценку в виде многоугольника $\mathcal{F}_{ij}$, а для внешних (относительно i -й и j -й комнат) — интервальную оценку. В итоге получаем

$$\mathcal{X}(t + \Delta t) = \{\mathcal{N}_{ij}(t + \Delta t), i, j = 1, \dots, N : T_{ij} \in \mathcal{T}\},$$

$$\mathcal{N}_{ij}(t + \Delta t) = (\mathcal{N}_{ij}(t) + \Delta t (C \mathcal{F}_{ij} + f_{ij, \text{ext}}(t))) \cap ([0, C_i] \times [0, C_j]), \quad C = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix},$$

$$f_{ij, \text{ext}}(t) = [f_{ji, \text{ext}, -}, f_{ji, \text{ext}, +}] \times [f_{ij, \text{ext}, -}, f_{ij, \text{ext}, +}],$$

$$f_{ji, \text{ext}, -} = \sum_{k \neq j: T_{ki} \in \mathcal{T}} \min_{(f_{ki}, f_{ik})^T \in \mathcal{F}_{ki}} \{f_{ki} - f_{ik}\}, \quad f_{ji, \text{ext}, +} = \sum_{k \neq j: T_{ki} \in \mathcal{T}} \max_{(f_{ki}, f_{ik})^T \in \mathcal{F}_{ki}} \{f_{ki} - f_{ik}\}.$$

Так как линейная комбинация выпуклых многоугольников является выпуклым многоугольником, то множество $\mathcal{N}_{ij}(t + \Delta t)$ — выпуклый многоугольник. Его вершины можно найти из приведенных выше формул, зная вершины $\mathcal{N}_{ij}(t)$ и $\mathcal{F}_{ij}$.

Замечание 5. Полученные оценки можно улучшить за счет сопоставления между собой полученных множеств, соответствующих соседним парам комнат (т.е. множеств вида $\mathcal{N}_{ij}(t + \Delta t)$ и $\mathcal{N}_{ik}(t + \Delta t)$). Для каждой пары i, j положим

$$\begin{aligned}\mathcal{N}_{ij}(t + \Delta t) &:= \mathcal{N}_{ij}(t + \Delta t) \cap ([n_{ij,\text{ext},-}, n_{ij,\text{ext},+}] \times [n_{ji,\text{ext},-}, n_{ji,\text{ext},+}]), \\ n_{ij,\text{ext},-} &= \max_{k \neq j, T_{ik} \in \mathcal{T}} \min \{n_i : (n_i, n_k) \in \mathcal{N}_{ik}(t + \Delta t)\}, \\ n_{ij,\text{ext},+} &= \min_{k \neq j, T_{ik} \in \mathcal{T}} \max \{n_i : (n_i, n_k) \in \mathcal{N}_{ik}(t + \Delta t)\}.\end{aligned}$$

Данную процедуру можно проводить до тех пор, пока на каждой итерации хотя бы одна из оценок $\mathcal{N}_{ij}(t + \Delta t)$ изменяется (уточняется).

3.3. Количество вершин в многоугольных оценках $\mathcal{N}_{ij}(t)$

В описанном алгоритме построения оценок множества достижимости в форме выпуклых многоугольников может иметь место эффект увеличения с течением времени количеств вершин, задающих многоугольники $\mathcal{N}_{ij}(t)$. Это может привести к значительному росту вычислительной сложности при больших значениях $t > t_0$. Для решения этой проблемы можно на отдельных шагах алгоритма после подсчета очередного набора оценок $\mathcal{N}_{ij}(t + \Delta t)$ дополнительно огрублять их, уменьшая количество используемых вершин.

Пусть $v \in \mathbb{N}$ — фиксированное число, задающее максимально допустимое количество вершин многоугольника, $g^{(ij,1)}, \dots, g^{(ij,m)}$ — вершины многоугольника $\mathcal{N}_{ij}(t + \Delta t)$. Пусть $m > v$, опишем алгоритм удаления вершины (индексы ij опущены для краткости).

1. Для каждой стороны $[g^{(k)}, g^{(k+1)}]$ (индексацию следует брать по модулю m) вычисляется $\tilde{g}^k$ — точка пересечения лучей $[g^{(k-1)}, g^{(k)}]$ и $[g^{(k+2)}, g^{(k+1)}]$.
2. Для всех треугольников с вершинами вида $\tilde{g}^k, g^{(k)}, g^{(k+1)}$ вычисляются площади.
3. Находим сторону $[\bar{g}^{(k)}, \bar{g}^{(k+1)}]$, на которой располагается треугольник наименьшей площади.
4. Пару вершин $\bar{g}^{(k)}, \bar{g}^{(k+1)}$ в многоугольной оценке заменяем на найденную точку $\tilde{g}^k$.

Указанный алгоритм применяется до тех пор, пока количество вершин не станет равным заданному числу v . На выходе получаем многоугольник $\tilde{\mathcal{N}}_{ij}(t + \Delta t)$ с $m = v$ вершинами. Легко видеть, что $\mathcal{N}_{ij}(t + \Delta t) \subseteq \tilde{\mathcal{N}}_{ij}(t + \Delta t)$.

4. УЧЕТ РЕЗУЛЬТАТОВ ИЗМЕРЕНИЙ

Рассмотрим теперь ситуацию, когда в некоторый момент времени t появляется неточная информация (оценка) о количестве людей в комнате i . Например, это могут быть результаты обработки изображений, полученных с видеокамер. С точки зрения рассматриваемой математической модели можно использовать уравнение измерения

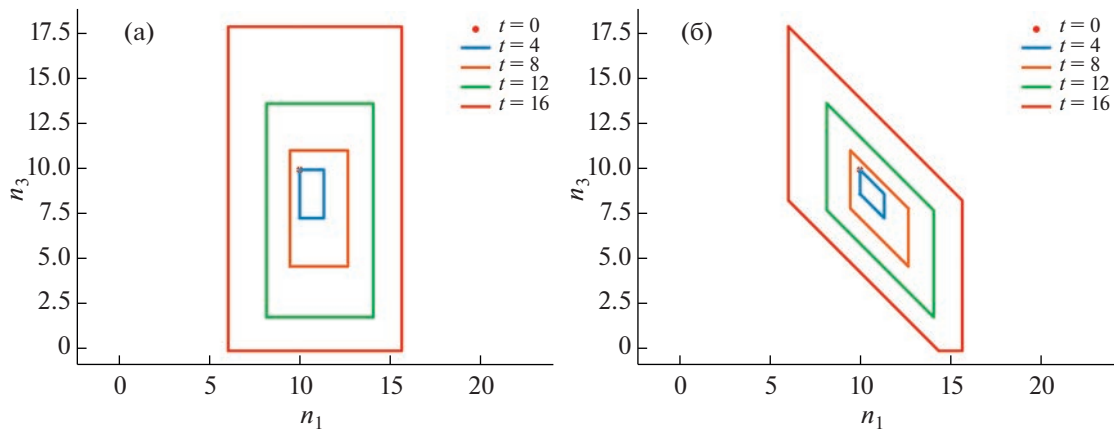
$$y_i(t) = n_i(t) + \xi_i(t). \quad (17)$$

Здесь $\xi_i(t)$ — помеха измерения, относительно которой известны ограничения $|\xi_i(t)| \leq \Xi$, где константа Ξ считается заданной и не зависит от номера комнаты или момента времени. Таким образом, $n_i(t) \in [y_i(t) - \Xi, y_i(t) + \Xi]$.

Поступившие результаты измерений могут быть использованы для уточнения текущей оценки величины $n_i(t)$. В случае с интервальными оценками такое уточнение выглядит следующим образом:

$$\tilde{n}_i(t) \in [\tilde{n}_{i,-}(t), \tilde{n}_{i,+}(t)] = [n_{i,-}(t), n_{i,+}(t)] \cap [y_i(t) - \Xi, y_i(t) + \Xi].$$

Будем считать, что результаты измерений не противоречат математической модели, и потому полученное пересечение отрезков не пусто.



Фиг. 2. Эволюция проекции оценки множества достижимости на плоскость (n_1, n_3) : (a) — интервальные оценки, (б) — двумерные проекции.

В случае с оценками двумерных проекций множества достижимости результаты измерений могут быть использованы следующим образом:

$$\tilde{\mathcal{N}}_{ij}(t) = \mathcal{N}_{ij}(t) \cap ([y_i(t) - \Xi, y_i(t) + \Xi] \times \mathbb{R}) \quad \forall i, j : T_{ij} \in \mathcal{T}.$$

Заметим, что в отдельных случаях результаты измерения количества людей в комнате с номером i могут быть также использованы для уточнения оценок величин $n_j(t)$, соответствующих соседним комнатам. Например, если используются интервальные оценки и известны оценки $\mathcal{F}_{ij}$ потоков, то

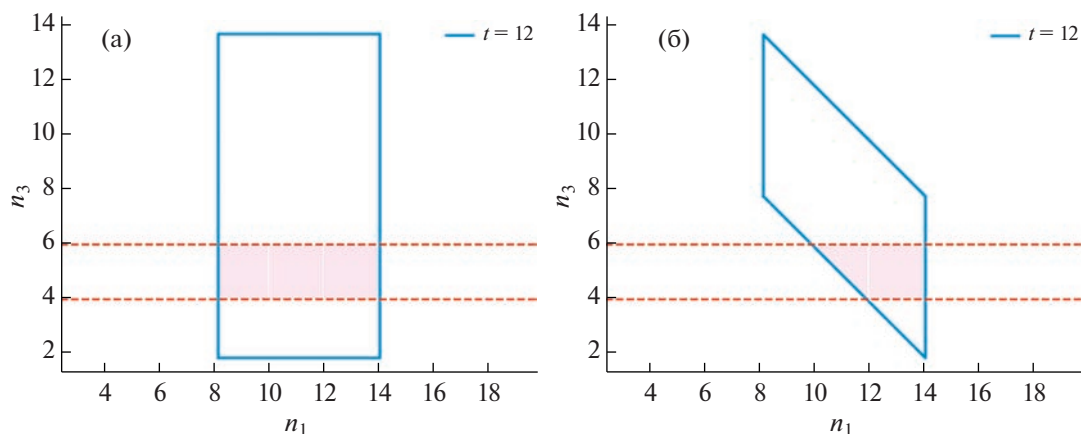
$$\begin{aligned} \tilde{n}_{j,-}(t) &= \max \left\{ n_{j,-}(t), n_{j,-}(t - \Delta t) + \Delta t \sum_{T_{jk} \in \mathcal{T}, k \neq i} \min \{ f_{kj} - f_{jk} : (f_{jk}, f_{kj}) \in \mathcal{F}_{jk} \} + s_{ij,-} \right\}, \\ \tilde{n}_{j,+}(t) &= \min \left\{ n_{j,+}(t), n_{j,+}(t - \Delta t) + \Delta t \sum_{T_{jk} \in \mathcal{T}, k \neq i} \max \{ f_{kj} - f_{jk} : (f_{jk}, f_{kj}) \in \mathcal{F}_{jk} \} + s_{ij,+} \right\}, \\ s_{ij,-} &= \tilde{n}_{i,-}(t) - n_{i,+}(t - \Delta t) + \Delta t \sum_{T_{ik} \in \mathcal{T}, k \neq j} \min \{ f_{ki} - f_{ik} : (f_{ik}, f_{ki}) \in \mathcal{F}_{ik} \}, \\ s_{ij,+} &= \tilde{n}_{i,+}(t) - n_{i,-}(t - \Delta t) + \Delta t \sum_{T_{ik} \in \mathcal{T}, k \neq j} \max \{ f_{ki} - f_{ik} : (f_{ik}, f_{ki}) \in \mathcal{F}_{ik} \}. \end{aligned}$$

5. РЕЗУЛЬТАТЫ РАБОТЫ ПРОГРАММЫ

Для численной иллюстрации двух алгоритмов возьмем ранее приведенный пример, граф которого изображен на фиг. 1. Пусть $n = 10$, $S = 15$, $C = 20$, $v = 1.2$, $w = 0.5$, $F = 3$, параметры $\alpha_3^{(1)} = \alpha_3^{(2)} = 1$, $\alpha_1^{(3)} = \alpha_2^{(3)} = 1/2$ считаем не зависящими от времени.

Построим оценку множества достижимости $\mathcal{X}(t)$ двумя способами с шагом $\Delta t = 4$ из начального состояния в момент времени $t_0 = 0$, при котором точно известно начальное количество людей в каждой комнате $n_{i,-} = n_{i,+} = 10$, $i = 1, 2, 3$. Далее достаточно рассмотреть ситуацию в переходе T_{13} (T_{31}), в соседнем переходе T_{23} ситуация будет в точности повторяться из-за симметричности схемы помещения и начальных условий.

Пересчитывая количества людей в комнатах $n_i(t)$ с помощью интервальных оценок потоков, получаем набор отрезков $\{[n_{i,-}(t), n_{i,+}(t)]\}$, $t > 0$. Для удобного сравнения двух подходов изобразим декартово произведение отрезков $\{[n_{i,-}(t), n_{i,+}(t)]\}$ и $\{[n_{3,-}(t), n_{3,+}(t)]\}$ на плоскости (n_1, n_3) (фиг. 2а).



Фиг. 3. Учет результата измерения $y_3(t)$ в момент времени $t = 12$: (а) — интервальные оценки, (б) — двумерные проекции.

Уточним множества возможных потоков до пересчета $n_{i,-}$, $n_{i,+}$ в момент времени $t = 4$. Так как выполняется $wC/S \leq F$, в результате работы первого алгоритма получаем оценку $\mathcal{F}_{13} = \{(f_{13}, f_{31}) : 0 \leq f_{13} \leq 1/3, f_{31} = 1/3\}$ (множество $\mathcal{F}_{31}$ отличается от $\mathcal{F}_{13}$ перестановкой компонент). Далее, так как $\min\{f_{31} - f_{13}\} = \max\{f_{13} - f_{31}\} = 0$, $n_{1,-}$ не уменьшается, а $n_{3,+}$ не увеличивается. Это соответствует тому, что при рассматриваемых параметрах модели из третьей комнаты в первую переходит больше человек, чем в противоположном направлении. В следующие моменты времени такая ситуация не наблюдается, так как значение n_i известно неточно, поэтому нижняя граница $n_{1,-}$ начнет уменьшаться, а верхняя $n_{3,+}$ увеличиваться.

Применение двумерных проекций в данном случае позволяет одновременно учитывать увеличение и уменьшение значения потоков в двух комнатах. Это можно интерпретировать следующим образом: при пересчете значений n_i мы сразу же учитываем то, что если, например, увеличилось n_1 , то при этом обязательно должно уменьшиться n_3 , оно в данном случае никак не может остаться неизменным, так как в комнату 1 нет других входящих переходов, кроме T_{31} . Таким образом, получается более сложная оценка в виде выпуклого многоугольника $\mathcal{N}_{13}$ (фиг. 2б). Заметим, что одномерные проекции $[n_{1,-}, n_{1,+}]$ и $[n_{3,-}, n_{3,+}]$ множества $\mathcal{N}_{13}$ совпадают с интервальной оценкой, построенной с помощью первого метода.

Наиболее осязаемое различие между двумя подходами можно заметить в случае учета результатов измерений. Пусть в момент времени $t = 12$ поступило измерение $y_3(t) = 5$, $\Xi = 1$. Очевидно, что в случае интервальных оценок такая ситуация уточняет только значения $[\tilde{n}_{3,-}, \tilde{n}_{3,+}] = [4, 6]$, на текущую оценку количества людей в соседних комнатах она никак не повлияет (фиг. 3а): $[\tilde{n}_{1,-}, \tilde{n}_{1,+}] = [8.17, 14.07]$.

Напротив, в случае двумерных оценок уточнение может затрагивать значения в соседней комнате. Для данного примера проекция множества $\tilde{\mathcal{N}}_{13}$ улучшается по сравнению с интервальными оценками, построенными с помощью первого метода (фиг. 3б): $[\tilde{n}_{1,-}, \tilde{n}_{1,+}] = [9.93, 14.07]$.

6. ЗАКЛЮЧЕНИЕ

Рассмотрена задача оценивания числа людей в каждой комнате для модифицированной модели СТМ, описывающей перемещение потоков людей внутри помещения. Предложено два способа оценки множеств достижимости задачи: с помощью одномерных интервальных оценок и двумерных проекций. Работа двух алгоритмов проиллюстрирована численными примерами, показано преимущество второго способа оценивания.

СПИСОК ЛИТЕРАТУРЫ

1. *Daganzo C.F.* The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory // *Transp. Res. B.* 1994. V. 28B. № 4. P. 269–287.
2. *Daganzo C.F.* The cell transmission model, part II: network traffic // *Transp. Res. B.* 1995. V. 29B. № 2. P. 79–93.
3. *Piccoli B., Garavello M.* Traffic flow on networks. American institute of mathematical sciences. Springfield, 2006.
4. *Зайцева М.В., Точилин П.А.* Управление потоками людей в здании во время эвакуации // *Вестник Московского ун-та. Сер. 15. Вычисл. матем. и кибернетика.* 2020. № 4. С. 3–17.
5. *Hänseler F.S., Bierlaire M., Farooq B., Mühlematter T.* A macroscopic loading model for time-varying pedestrian flows in public walking areas // *Transp. Res. B.* 2014. V. 69. P. 60–80.
6. *Kachroo P., Al-nasur S.J., Wadoo S.A., Shende A.* Pedestrian dynamics. Feedback control of crowd evacuation. Springer, 2008.
7. *Акопов А.С., Бекларян Л.А.* Агентная модель поведения толпы при чрезвычайных ситуациях // *Автомат. и телемех.* 2015. Вып. 10. С. 131–143.
8. *Samson B.P.V., Aldanese IV C.R., Chan D.M.C., San Pascual J.J.S., Sido M.V.A.P.* Crowd dynamics and control in high-volume metro rail stations // *Proced. Comput. Sci.* 2007. V. 108. P. 195–204.
9. *Helbing D., Farcas I., Vicsek T.* Simulating dynamical features of escape panic // *Nature.* 2000. V. 407. P. 487–490.
10. *Tampere C.M.J., Corthout R., Catrysse D., Immers L.H.* A generic class of first order node models for dynamic macroscopic simulation of traffic flows // *Trans. Rep. B.* 2011. № 45. P. 289–309.
11. *Kurzhanski A.B., Varaiya P.* Dynamics and control of trajectory tubes. Birkhäuser, 2014.
12. *Корнушенко Е.К.* Интервальные по координатным оценкам для множества достижимых состояний линейной стационарной системы // *Автомат. и телемех.* 1980. Вып. 5. С. 12–22.
13. *Tang W., Wang Z., Wang Y., Raissi T., Shen Y.* Interval estimation methods for discrete-time linear time-invariant systems // *IEEE Trans. Automat. Control.* 2019. V. 64. № 11. P. 4717–4724.
14. *Куржанский А.Б., Куржанский А.А., Варайя П.* Роль макро моделирования в активном управлении транспортной сетью // *Тр. МФТИ.* 2010. Т. 2. № 4. С. 100–118.

О МНОЖЕСТВЕ СТАБИЛЬНЫХ МАТЧИНГОВ ДВУДОЛЬНОГО ГРАФА

© 2023 г. А. В. Карзанов^{1,*}

¹ 117418 Москва, Нахимовский пр-т, 47, ЦЭМИ РАН, Россия

*e-mail: akarzanov7@gmail.com

Поступила в редакцию 16.02.2023 г.

Переработанный вариант 16.02.2023 г.

Принята к публикации 28.04.2023 г.

Тема стабильных матчингов (марьяжей) в двудольных графах приобрела широкую популярность, начиная с выхода классической работы Гейла и Шепли. В настоящей работе дается развернутый обзор избранных и взаимосвязанных утверждений в этой области, описывающих структурные, полиэдральные и алгоритмические свойства таких объектов и их множеств, снабжая эти утверждения короткими доказательствами. Библ. 24. Фиг. 3.

Ключевые слова: стабильный матчинг, посет ротаций, стабильный матчинг минимальной стоимости, медианный стабильный матчинг.

DOI: 10.31857/S0044466923080082, **EDN:** WSLYAY

1. ВВЕДЕНИЕ

Начиная с классической работы Гейла и Шепли [1], область задач о стабильных марьяжах и их обобщений привлекала внимание многочисленных исследователей, и в этой области был собран богатый урожай интересных результатов, как теоретического, так и прикладного характера.

В задаче о стабильном марьяже (или, более определенно, о стабильном матчинге в двудольном графе) рассматривается двудольный граф $G = (V, E)$, в котором для каждой вершины v задан линейный порядок $<_v$, указывающий предпочтения на множестве ребер $\delta_G(v)$, инцидентных v . Пусть I и J — вершинные доли в G . В [1] и других работах вершины в I и J понимались как лица разных полов (“мужчины” и “женщины” соответственно), а ребра как возможные союзы (или “браки”) между ними: если для вершины $m \in I$ (“man”) и инцидентных ей ребер $e = mw$ и $e' = mw'$ выполняется $e <_m e'$, это означает, что m предпочитает союз с w союзу с w' . И аналогично для $w \in J$ (“woman”) относительно порядка $<_w$.

Матчинг в G — это подмножество ребер $M \subseteq E$, где никакие два ребра не имеют общей вершины; таким образом, можно считать, что M описывает множество “моногамных браков”. Матчинг M называется **стабильным**, если для любого ребра $e = mw \in E - M$ по крайней мере одна из вершин m и w имеет инцидентное ребро в M (определяющее выбранного “партнера”), и это ребро является более предпочтительным для данной вершины, чем e . Стабильный матчинг существует для любого двудольного $G = (V, E)$ с произвольными линейными порядками $<_v$ на $\delta_G(v)$, $v \in V$, что первоначально было доказано Гейлом и Шепли [1] для полных двудольных графов с долями одного размера, и обобщено последующими авторами на произвольные двудольные графы.

Впоследствии развитие в области стабильности, касающейся графов, шло по нескольким направлениям. В одном из них понятие стабильности переносилось на матчинги в произвольных графах (здесь следует особо выделить основополагающие работы Ирвинга [2] и Тана [3] о “стабильных расселениях по комнатам” (stable roommates)). В другом направлении изучались вопросы стабильности при рассмотрении вариантов весовых функций на ребрах и вершинах графа (отметим, как одну из наиболее общих, задачу о “стабильном распределении” (stable allocation problem), которая была введена и изучена Бейу и Балински [4]).

В то же время целый ряд глубоких результатов был получен в пределах собственной теории стабильных матчингов для двудольных графов, и цель данной работы — сделать обзор избранных, наиболее значительных на наш взгляд, известных результатов в этой теории. Они в боль-

шой степени связаны между собой и охватывают комбинаторные, полиэдральные, алгоритмические и др. аспекты. Ряд задач, освещаемых в обзоре, состоит в описании структурных свойств множества $\mathcal{M}(G)$ стабильных матчингов в $(G, <)$.

Наше изложение организовано следующим образом.

Раздел 2 содержит описание базовых утверждений и конструкций. Он начинается с напоминания классических результатов, восходящих к работе Гейла и Шепли о существовании стабильного матчинга и о методе “отложенных принятий” для его построения. Затем объясняется, что в множестве $\mathcal{M}(G)$ стабильных матчингов произвольного двудольного графа G все матчинги покрывают одно и то же множество вершин (предложение 2), и описываются операции на парах стабильных матчингов, позволяющих определить частичный порядок $<$ на множестве $\mathcal{M}(G)$, представляющий его в виде дистрибутивной решетки (предложение 3).

Раздел 3 посвящен обсуждению ключевого понятия *ротации* в стабильном матчинге. Оно было введено Ирвингом [2] в связи с задачей на произвольном графе (stable roommates problem), но оказалось очень полезным и для исследования структуры множества стабильных матчингов двудольного графа G . В графе G со стабильным матчингом M мы понимаем под ротацией определенный цикл в G , чередующийся относительно M . К числу основных из известных утверждений, обсуждаемых в этом разделе, относятся такие: (а) трансформация вдоль ротации преобразует стабильный матчинг в стабильный (предложение 4); (б) ротационные трансформации соответствуют отношениям немедленного предшествования в решетке $(\mathcal{M}(G), <)$ (предложение 6); (в) во всех максимальных цепях решетки $(\mathcal{M}(G), <)$ множество ротаций одно и то же, обозначаемое $\mathcal{R}_G$ (предложение 7).

В разделе 4 продолжают обсуждения, связанные с ротациями. Здесь вводится частичный порядок $\leq$ на множестве $\mathcal{R}_G$ и приводится альтернативное представление для множества стабильных матчингов $\mathcal{M}(G)$, установленное Ирвингом и Лейтером [5], а именно: стабильные матчинги в G взаимно-однозначно соответствуют идеалам (или анти-цепям) посета $(\mathcal{R}_G, \leq)$ (предложение 8).

В разделе 5 рассматривается усиленный, “стоимостной”, вариант задачи о стабильном матчинге. Здесь множество ребер двудольного графа $G = (V, E, <)$ снабжаются вещественной весовой функцией $c : E \rightarrow \mathbb{R}$, и требуется найти стабильный матчинг M , минимизирующий общий вес $\sum (c(e) : e \in E)$. В частном случае, когда для ребра $e = tw$ вес $c(e)$ равняется сумме его рангов в порядках $<_m$ и $<_w$, получаем задачу об эгалитарном стабильном матчинге (в котором стороны I и J “уравниваются”), поставленную в [6]. Для произвольных весов c задача минимизации (или максимизации) решается эффективным комбинаторным алгоритмом, разработанным в [7]. Это является следствием того, что: во-первых, стабильные матчинги представляются идеалами посета ротаций $(\mathcal{R}_G, \leq)$, в котором число элементов $|\mathcal{R}_G|$ не превосходит числа ребер $|E|$; во-вторых, вес стабильного матчинга выражается, с точностью до константы, как вес соответствующего идеала (при назначении подходящих весов ротаций); и, в-третьих, задача о нахождении идеала (или “замкнутого множества”) минимального веса в произвольном конечном посете сводится к задаче о минимальном разрезе ориентированного графа, как показано Пикаром [8].

Раздел 6 посвящен полиэдральным аспектам. Здесь описывается многогранник стабильных матчингов $\mathcal{P}_{st}(G)$ через систему линейных неравенств (близкую к той, что указана Ванде Вейтом [9]). Также, следуя полиэдральной конструкции Тео и Сетурамана [10], показывается, что для произвольного набора стабильных матчингов $M_1, \dots, M_\ell$ и любого $k \leq \ell$, выбирая для каждой открытой вершины в доле I ребро, имеющее порядок k среди этих матчингов, мы снова получаем стабильный матчинг. В частности, когда ℓ нечетно и $k = (\ell + 1)/2$, определяется т.н. *медианный* стабильный матчинг для заданных $M_1, \dots, M_\ell$.

В заключительном разд. 7 формулируется результат, полученный в [7], о труднорешаемости ($\#P$ -полноте) задачи определения числа стабильных матчингов в двудольном графе. Это отвечает на вопрос Кнута [11] об алгоритмической сложности вычисления такого числа.

Для полноты изложения мы, как правило, сопровождаем представленные утверждения доказательствами, часто альтернативными и более короткими по сравнению с оригинальными доказательствами в работах авторов. Отметим, что везде, где мы даем ссылки на работы [5, 7] и некоторые другие, соответствующие результаты в этих работах были получены для случая полного

двудольного графа $K_{n,n}$ (с произвольными порядками $<$), но эти результаты распространяются и на случай произвольного двудольного графа G .

2. ОПРЕДЕЛЕНИЯ И БАЗОВЫЕ ФАКТЫ

На протяжении всего изложения мы рассматриваем двудольный граф $G = (V, E)$ с разбиением множества вершин V на подмножества I и J (где каждое ребро соединяет вершины из разных подмножеств). Иногда мы будем считать граф G ориентированным, с ребрами, направленными от I к J . Ребро, соединяющее вершины $m \in I$ (“man”) и $w \in J$ (“woman”), обозначается через mw . Множество ребер, инцидентных вершине $v \in V$, обозначается $\delta(v) = \delta_G(v)$. Каждая вершина v снабжена линейным порядком $<_v$, задающим предпочтения на множестве $\delta(v)$. А именно для ребер $e, e' \in \delta(v)$, если $e <_v e'$, то ребро e считается более предпочтительным для v , чем ребро e' ; иногда нам также будет удобно говорить, что e расположено раньше или левее, чем e' (а e' позднее, или правее, чем e). Если вершина v ясна из контекста, можно писать $e < e'$. Обычно мы включаем порядки $<$, а также разбиение V на доли I, J в описание графа, применяя обозначение $G = (V, E, <)$ или $G = (I \sqcup J, E, <)$.

Матчинг M в G называется *стабильным*, если он не допускает блокирующих ребер. Здесь ребро $e = mw \in E - M$ считается блокирующим для M , если вершина m либо не покрывается M , либо ребро e' в M , инцидентное m , менее предпочтительно: $e <_m e'$, и аналогичное верно для вершины w .

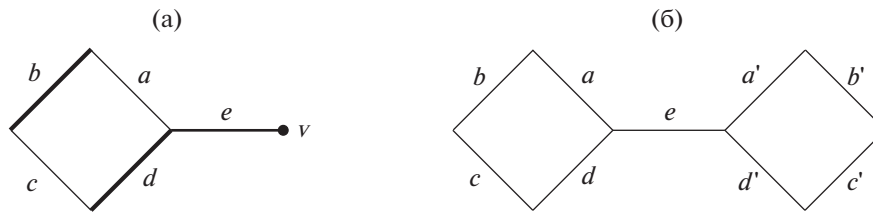
Ниже в этом разделе мы приводим две группы базовых свойств стабильных матчингов в G .

I. В основании теории стабильных матчингов лежит эффективный алгоритм построения стабильного матчинга Гейла и Шепли [1] (в изложении для полного двудольного графа с долями I, J одинакового размера n : $G \simeq K_{n,n}$). Рекурсивное изложение этого алгоритма, который часто называют “алгоритмом отложенных принятий” (АОП), дано МакВайти и Вилсоном в работе [6] (где также описан алгоритм последовательного конструирования множества всех стабильных матчингов для $G \simeq K_{n,n}$). Короткое неконструктивное доказательство существования стабильного матчинга в двудольном случае можно найти в [12, Sec. 18.5g].

В АОП роли сторон I и J неравнозначны: одна сторона “предлагает”, а другая “принимает или отвергает” предложения. (Напомним, что на каждом шаге АОП произвольно выбирается не включенный в текущий марьяж “man” $m \in I$, который делает предложение “woman” $w \in J$ согласно наилучшему ребру mw из еще не использованных им ребер в $\delta(m)$; это предложение сразу отвергается, если w уже имеет лучшее предложение от $m' \in I$ (т.е. $m'w <_w mw$), и принимается в ином случае, отвергая предыдущее предложение, если таковое имеется.) Результаты естественно расширяются и для произвольного двудольного графа G . Это приводит к следующему важному свойству.

Предложение 1 (см. [1, 6]). *В случае, когда I предлагает, а J принимает/отвергает, АОП находит (за линейное время $O(|V| + |E|)$) канонический стабильный матчинг, в котором выбор ребер для вершин в I наилучший, а для вершин в J наихудший по всем стабильным матчингам.*

Мы обозначаем этот матчинг $M^{\min} = M^{\min}(G)$. По симметрии, если АОП применяется в варианте, когда J предлагает, то алгоритм строит стабильный матчинг, для которого выборы ребер для J наилучшие, а для I наихудшие; обозначим его $M^{\max} = M^{\max}(G)$. Уточним смысл терминов “наилучший-наихудший”, рассматривая любой другой стабильный матчинг L для G . А именно, если ребра $e \in M^{\min}$ и $e' \in L$ имеют общую вершину $m \in I$, то $e \leq_m e'$, а если общую вершину $w \in J$, то $e' \leq_w e$. Здесь мы опираемся на инвариантность множества вершин, покрытых стабильным матчингом (что тривиально в случае полного двудольного графа $K_{n,n}$, где стабильный матчинг является совершенным, т.е. покрывающим все вершины). А именно, справедливо следующее свойство, установленное в [13].



Фиг. 1. Два примера.

Предложение 2. Для двудольного $G = (V, E, <)$ все стабильные матчинги покрывают одно и то же множество вершин.

Чтобы показать это свойство, примем следующие определения и обозначения (которые понадобятся и в дальнейшем).

Путь в ориентированном графе — это последовательность $P = (v_0, e_1, v_1, \dots, e_k, v_k)$, где e_i — ребро, соединяющее вершины v_{i-1} и v_i . Для пути P мы можем также применять сокращенную запись через вершины: $v_0 v_1 \dots v_k$, или ребра: $e_1 e_2 \dots e_k$. Ребро e_i в P называется *прямым* (forward), если $e_i = v_{i-1} v_i$, и *обратным* (backward), если $e_i = v_i v_{i-1}$. (Мы обозначаем ребро, выходящее из u и входящее в v , как uv , вместо (u, v) .) Путь называется *ориентированным*, если все его ребра прямые, и называется *простым*, если все его вершины различные. Путь из вершины u в вершину v может быть назван $u-v$ путем. Для подмножества ребер $A \subseteq E$ множество вершин, покрытых A , обозначается V_A . Для подмножеств X, Y множества S мы пишем $X - Y$ вместо $X \setminus Y$ ($= \{s \in X : s \notin Y\}$) и обозначаем $X \Delta Y$ симметрическую разность $(X - Y) \cup (Y - X)$.

Пусть теперь M, L — два стабильных матчинга в двудольном G . Рассмотрим подграф $\Delta_{M,L}$ в G , индуцированный множеством ребер $M \Delta L$. Заметим, что

$$\begin{aligned} &\text{если } e, e', e'' \text{ — различные ребра в } \Delta_{M,L} \text{ такие, что } e, e' \text{ имеют общую} \\ &\text{вершину } u, \text{ а } e', e'' \text{ имеют общую вершину } v, \text{ и если } e >_u e', \text{ то } e' >_v e''. \end{aligned} \quad (2.1)$$

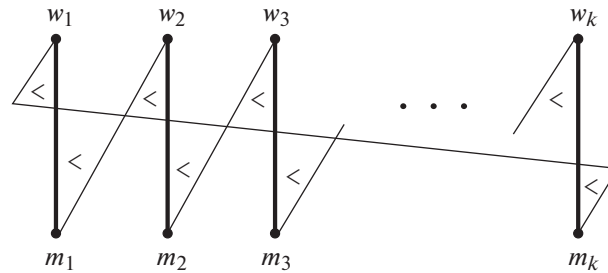
Действительно, пусть для определенности $e' \in M$. Тогда $e, e'' \in L$. Если бы было $e' <_v e''$, то ребро e' было бы блокирующим для L .

Перейдем теперь к доказательству предложения 2. Предположим, что $V_M \neq V_L$. Тогда, по крайней мере, одна из компонент в $\Delta_{M,L}$ является простым путем $P = (v_0, e_1, v_1, \dots, e_k, v_k)$. В этом пути ребра из M и L чередуются, и из стабильности M, L следует $k \geq 2$. Без ограничения общности можно считать, что $e_1 \in M$. Тогда $v_0 \notin V_L$, и так как ребро e_1 не является блокирующим для L , то должно выполняться $e_1 >_{v_1} e_2$.

Ввиду этого, последовательно применяя (2.1) к тройкам соседних ребер в P , получаем, что $e_{k-1} >_{v_{k-1}} e_k$. Но тогда при нечетном k ребро e_k является блокирующим для L (так как $e_{k-1} \in L$, $e_k \notin L$ и $v_k \notin V_L$), а при четном k ребро e_k является блокирующим для M (так как $e_{k-1} \in M$, $e_k \notin M$ и $v_k \notin V_M$); противоречие.

Будем обозначать множество вершин в G , покрытых стабильным матчингом, через $\tilde{V}$. Очевидно, при удалении вершин в $V - \tilde{V}$ каждый стабильный матчинг для G остается стабильным и для результирующего графа $\tilde{G}$. Однако в $\tilde{G}$ могут появиться новые стабильные матчинги (и, следовательно, включение $\mathcal{M}(G) \subseteq \mathcal{M}(\tilde{G})$ может быть строгим), как показывает простой пример в левом фрагменте на фиг. 1. Здесь отношения предпочтений устроены так: $a < b$, $b < c$, $c < d$, $d < e < a$, и можно проверить, что имеется единственный стабильный матчинг, а именно, $M = \{b, d\}$. Тогда вершина v не покрыта M , однако при ее удалении (вместе с ребром e) появляется новый стабильный матчинг $\{a, c\}$.

Таким образом, при исследовании множества стабильных матчингов для G мы, вообще говоря, не можем выкидывать из рассмотрения непокрытые вершины, оставляя только подграф $\tilde{G} = (\tilde{V} = (\tilde{I} \sqcup \tilde{J}), \tilde{E})$ (где доли $\tilde{I} := I \cap \tilde{V}$ и $\tilde{J} := J \cap \tilde{V}$ имеют одинаковый размер, и все стабиль-



Фиг. 2. Повышающий цикл для (M, L) . Матчинг M изображен жирными линиями, а L — светлыми линиями. Вершины m_i принадлежат доле I .

ные матчинги совершенные). По похожим причинам в случае отсутствия непокрытых вершин удаление ребра, не входящего ни в один стабильный матчинг, может повлечь появление нового стабильного матчинга. На фиг. 1а изображено расширение предыдущего графа с аналогичными предпочтениями для новых ребер, а именно, $a' < b'$, $b' < c'$, $c' < d'$ и $d' < e < a'$. Здесь имеются три стабильных матчинга, а именно: $\{a, c, b', d'\}$, $\{b, d, a', c'\}$ и $\{b, d, b', d'\}$, однако, при удалении непокрытого ребра e появляется четвертый стабильный матчинг $\{a, c, a', c'\}$, для которого ранее имелось блокирующее ребро e .

II. Далее рассмотрим упорядоченную пару (M, L) стабильных матчингов в G . Из предложения 2 следует, что граф $\Delta_{M,L}$, индуцированный $M \Delta L$, распадается на некоторое множество $\mathcal{C} = \mathcal{C}(M, L)$ непересекающихся чередующихся циклов. Считая G ориентированным от I к J , будем полагать каждый цикл $C = (v_0, e_1, v_1, \dots, e_k, v_k = v_0) \in \mathcal{C}$ направленным согласно ориентации ребер в L , т.е. прямые ребра в C принадлежат L , а обратные принадлежат M . Ввиду (2.1), все предпочтения вдоль C “направлены в одну сторону”, а именно:

$$\begin{aligned} &\text{для цикла } C = e_1 e_2 \dots e_k \in \mathcal{C}(M, L) \text{ (применяя обозначение } C \text{ через ребра),} \\ &\text{если } e_i < e_{i+1} \text{ выполнено для некоторого } i, \text{ то это выполняется} \\ &\text{для всех } i = 1, \dots, k \text{ (полагая } e_{k+1} := e_1). \end{aligned} \quad (2.2)$$

В этом случае скажем, что цикл C повышающий, или правый, относительно M . Иначе C считается понижающим, или левым. Обозначим через $\mathcal{C}^+(M, L)$ и $\mathcal{C}^-(M, L)$ множества правых и левых циклов для (M, L) соответственно. Если матчинг M' получается из M заменой ребер вдоль цикла C , будем писать $M \xrightarrow{C} M'$ или $M' = \text{Repl}(M, C)$, и аналогично, будем писать $L \xrightarrow{C} L'$ или $L' = \text{Repl}(L, C)$ для замены вдоль C в L . Заметим, что если цикл C правый относительно M , то по сравнению с M матчинг M' является менее предпочтительным для всех вершин в $I \cap C$ (“men”) и более предпочтительным для всех вершин в $J \cap C$ (“women”), а для L и L' поведение противоположное. (Иначе говоря, при замене вдоль повышающего цикла мы как бы удаляемся от $M^{\min}$ и приближаемся к $M^{\max}$.) Повышающий цикл показан на фиг. 2.

Можно было бы ожидать, что матчинг $M' = \text{Repl}(M, C)$ должен быть стабильным. Однако это может быть неверным. Например, рассмотрим граф как в правом фрагменте на фиг. 1 (с указанными выше предпочтениями) и возьмем стабильные матчинги $M = \{b, d, a', c'\}$ и $L = \{a, c, b', d'\}$. Тогда ребра a, b, c, d порождают цикл C в $\Delta_{M,L}$, но матчинг $M' = \{a, c, a', c'\}$, получающийся в результате замены ребер в M вдоль C , не является стабильным.

Тем не менее стабильность сохраняется при замене сразу во всех циклах в $\mathcal{C}^+(M, L)$ или в $\mathcal{C}^-(M, L)$. В этих случаях будем писать $M' = \text{Repl}(M, \mathcal{C}^+(M, L))$ и $M' = \text{Repl}(M, \mathcal{C}^-(M, L))$ соответственно.

Лемма 1 (см. [11]). Если M, L — стабильные матчинги, то $M' = \text{Repl}(M, \mathcal{C}^+(M, L))$ — тоже стабильный матчинг. Аналогично для $\text{Repl}(M, \mathcal{C}^-(M, L))$.

Доказательство. Очевидно, M' является матчингом. Предположим, M' имеет блокирующее ребро $e = tw$ (где $t \in I$). Легко видеть, что это в принципе возможно, только если одна из вершин t, w принадлежит правому циклу, а другая — левому циклу в $\mathcal{C}(M, L)$. Без ограничения общности можно считать, что t принадлежит циклу $C \in \mathcal{C}^+(M, L)$, а w — циклу $C' \in \mathcal{C}^-(M, L)$. Пусть вершина t инцидентна ребрам $a \in M$ и $b \in L$ в C , а w инцидентна ребрам $c \in M$ и $d \in L$ в C' . Тогда $a <_m b$ и $c <_w d$ (учитывая $t \in I$ и $w \in J$). При преобразовании $M \mapsto M'$ в матчинг M' попадают ребра b и c , и должно выполняться $e <_m b$ и $e <_w c$ (так как по предположению e блокирует M'). Но тогда получаем $e < d$ (ввиду $c < d$), и, следовательно, e блокирует L ; противоречие.

Это позволяет определить решетку на стабильных матчингах в G . Заметим, что для $M, L \in \mathcal{M}(G)$ заменить в M ребра вдоль всех повышающих циклов для M — это все равно, что сделать в L замены вдоль всех повышающих циклов для L , т.е. $\text{Repl}(M, \mathcal{C}^+(M, L)) = \text{Repl}(L, \mathcal{C}^+(L, M))$. Аналогично $\text{Repl}(M, \mathcal{C}^-(M, L)) = \text{Repl}(L, \mathcal{C}^-(L, M))$. Ввиду Предложения 1, в $\mathcal{C}(M, L)$ нет понижающих циклов при $M = M^{\min}$, и нет повышающих циклов при $M = M^{\max}$.

Предложение 3 (см. [11]). Для различных $M, L \in \mathcal{M}(G)$ положим $M \prec L$, если для каждой пары ребер $e \in M$ и $e' \in L$, инцидентных вершине в $\tilde{I} (= I \cap \tilde{V})$ выполняется $e \leq e'$ (давая противоположные соотношения для вершин в $\tilde{J}$). Тогда $\prec$ определяет дистрибутивную решетку на $\mathcal{M}(G)$ с минимальным элементом $M^{\min}(G)$ и максимальным элементом $M^{\max}(G)$, в которой для $M, L \in \mathcal{M}(G)$ точной нижней гранью $M \wedge L$ является $\text{Repl}(M, \mathcal{C}^-(M, L)) = \text{Repl}(L, \mathcal{C}^-(L, M))$, а точной верхней гранью $M \vee L = \text{Repl}(M, \mathcal{C}(M, L)) = \text{Repl}(L, \mathcal{C}^+(L, M))$.

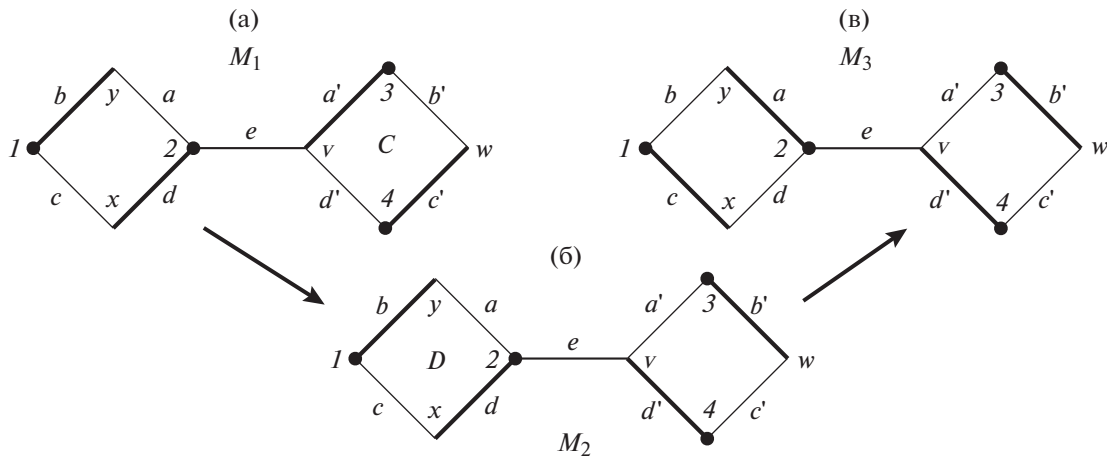
3. РОТАЦИИ

Благодаря решеточности множества $\mathcal{M}(G)$, можно, имея два или более стабильных матчингов, конструировать новые, делая переназначения в соответствующих циклах для пар матчингов, как указано в лемме 1. Другой метод, связанный с понятием ротации, позволяет порождать новые стабильные матчинги, исходя из одиночных элементов в $\mathcal{M}(G)$. Это понятие было введено Ирвингом и Лейтером [5] для задачи о стабильном марьяже, основываясь на концепции “цикла типа все или ничего” (all-or-nothing cycle) из работы Ирвинга [2], посвященной задаче о стабильных матчингах в общем графе (“stable roommates problem”). Впоследствии понятие ротации было распространено и на другие задачи о стабильности, такие как задачи о стабильных b-матчингах, распределениях (allocations) и др. (см., например, [14]). Отметим также недавнюю работу [15], где устанавливается связь между выпуклой оболочкой стабильных распределений при отсутствии ограничений на ребрах и порядковым многогранником посета ротаций.) Заметим, что хотя в [5] рассматривался случай полного двудольного графа $G \simeq K_{n,n}$, конструкции и результаты без особого труда переносятся на произвольный случай, и как выше, мы далее будем рассматривать произвольный двудольный граф $G = (V = I \sqcup J, E, <)$.

Пусть M — стабильный матчинг в G .

Определение 1. Скажем, что ребро $a = tw \in E - M$ (где $t \in I$) допустимое, если в M есть ребро e , инцидентное t , и это ребро более предпочтительное, чем a (т.е. $e <_m a$), и при этом либо вершина w непокрытая, либо M имеет ребро e' , инцидентное w , и это ребро менее предпочтительное, чем a (т.е. $a <_w e'$). Если для вершины $t \in I$ множество допустимых ребер в $\delta(t)$ непустое, то самое первое (наиболее предпочтительное) из них назовем *активным*. Обозначим через A множество активных ребер для M . Подграф в G , индуцированный множеством ребер $M \cup A$, обозначим $\Gamma = \Gamma(M)$ и назовем *активным графом*.

Учитывая то, что из каждой вершины в I исходит не более одного активного ребра (в то время как в вершину в J может входить несколько активных ребер), всякий цикл в Γ должен быть чередующимся относительно M и A . Кроме того, каждая компонента K в Γ либо является деревом, либо содержит ровно один цикл. (Действительно, в противном случае в K имелась бы пара циклов и соединяющий их простой $u-v$ путь P , все вершины которого, кроме u, v , не принадлежат циклам.) Концевые ребра этого пути должны принадлежать A , и, по крайней мере, одна из вер-



Фиг. 3. Граф с тремя стабильными матчингами, выделенными жирными линиями: $M_1 = M^{\min}$ (а), M_2 (б) и $M_3 = M^{\max}$ (в).

шин u, v должна находиться в I . Эта вершина имеет два инцидентных ребра в A (одно на P , другое на цикле); противоречие. При наличии цикла мы называем компоненту K *цикло-содержащей*, а сам цикл обозначаем через C_K и называем *ротацией* для M . (Заметим, что в [5] ротацией называется не сам цикл, а его пересечение с M .) Ребра в C_K , принадлежащие M , назовем *матчинговыми*, а остальные — *активными*. При замене ребер вдоль ротации C возникает матчинг, который по сравнению с M менее предпочтителен для вершин в $I \cap C$ и более предпочтительный для вершин в $J \cap C$. (Для операции замены вдоль C в [5] применяется термин “rotation elimination”.) Когда C — цикл в $\Gamma(M)$, мы также можем говорить, что матчинг M допускает ротацию C .

Пример. На фиг. 3 изображены три стабильных матчинга M_1, M_2, M_3 (выделенных жирной линией) в графе $G = (V, E, <)$ как на фиг. 1б. Здесь доля I составлена вершинами 1, 2, 3, 4, а доля J — вершинами x, y, v, w , и предпочтения заданы как и раньше, а именно:

$$\begin{aligned} a <_y b, \quad b <_1 c, \quad c <_x d, \quad d <_2 e <_2 a, \\ a' <_3 b', \quad b' <_w c', \quad c' <_4 d', \quad d' <_v e <_v a'. \end{aligned}$$

Для матчинга M_1 все нематчинговые ребра a, c, e, b', d' являются допустимыми, но из них активными оказываются только c, e, b', d' (так как $e <_2 a$). Поэтому подграф $\Gamma(M_1)$ порожден $M_1 \cup \{c, e, b', d'\}$, имеет ровно одну компоненту, и она является цикло-содержащей с циклом $C = a'b'c'd'$. При замене вдоль ротации C получаем матчинг M_2 (являющийся стабильным согласно предложению 4 ниже). Для M_2 допустимыми ребрами в $E - M_2$ являются только a и c (а e, a', c' — не допустимые, ввиду $d' <_v e, a'$ и $b' <_w c'$, учитывая $b', d' \in M_2$ и $v, w \in J$). Поэтому активными ребрами являются a и c , и граф $\Gamma(M_2)$ состоит из трех компонент, порожденных $\{a, b, c, d\}, \{b'\}$ и $\{d'\}$. Первая из них образует цикл $D = abcd$. Наконец, при замене вдоль ротации D получаем стабильный матчинг M_3 . Для него ни одно из ребер в $E - M_3$ не является допустимым, и $\Gamma(M_3)$ составлено просто ребрами в M_3 . Можно видеть, что $M_1 = M^{\min}$ и $M_3 = M^{\max}$.

Следующие два утверждения демонстрируют важные свойства $\Gamma(M)$.

Предложение 4 (см. [5]). Для любого цикла (ротации) C в $\Gamma(M)$ матчинг $M' = \text{Repl}(M, C)$ является стабильным.

Доказательство. Оба M и M' покрывают одно и то же множество вершин $\tilde{V}$. Предположим, что M' допускает блокирующее ребро $e = tw \in E - M'$. Заметим, что $e \notin M$ (учитывая то, что ребра в $C \cap M$ не могут быть блокирующими для M'). Возможны два случая.

Случай 1: $w \in \tilde{V}$. Тогда вершина w принадлежит ребрам $a \in M'$ и $b \in M$, и мы имеем $e <_w a$ (ввиду блокирования) и $a \leq_w b$ (так как в случае $a \neq b$ ребро a лежит в C и является допустимым). Следовательно, $e <_w b$. Тогда $m \in \tilde{V}$ (иначе e было бы блокирующим для M).

Пусть вершина m принадлежит ребрам $c \in M'$ и $d \in M$. Тогда $e <_m c$, и должно выполняться $d <_m e$ (иначе e было бы блокирующим для M , учитывая $e <_w b$). Следовательно, e является допустимым для M и более предпочтительным, чем ребро c , которое принадлежит A ; противоречие.

Случай 2: $w \notin \tilde{V}$. Тогда $m \in \tilde{V}$. Пусть m принадлежит ребрам $c \in M'$ и $d \in M$. Так как e является блокирующим для M' , но не для M , имеем $d <_m e <_m c$. Но тогда ребро e допустимое и более предпочтительное, чем активное ребро c ; противоречие.

Говоря далее о чередующихся (относительно M и A) путях и циклах в $\Gamma(M)$, мы считаем, что в них выбрано такое направление, чтобы ребра из A были прямыми, а ребра из M обратными.

Предложение 5. Пусть e — ребро в M , принадлежащее древесной компоненте K в $\Gamma(M)$. Тогда для любого стабильного матчинга $L \in \mathcal{M}(G)$ либо $e \in L$, либо e содержится в понижающем цикле в $\mathcal{C}(M, L)$.

Доказательство. Предположим, ребро $e = mw \in M$ из древесной компоненты K находится в повышающем цикле $C \in \mathcal{C}^+(M, L)$ для некоторого $L \in \mathcal{M}(G)$. Пусть C имеет вид $e_1 e_2 \dots e_k$ (применяя обозначение через ребра и учитывая направленность C), и пусть $e := e_1 \in M$. Возьмем максимальный чередующийся путь $P = p_1 p_2 \dots p_r$ (через ребра) в $\Gamma(M)$, начиная с $p_1 = e$. Можно считать, что никакое ребро в $P \cap M$, кроме p_1 , не лежит в повышающем цикле в $\mathcal{C}(M, L)$ (в противном случае возьмем его в качестве e). Ребра e_1 и $e_2 = mw'$ инцидентны вершине $m \in I$, и поскольку C повышающий, $e_1 <_m e_2$ и $e_2 <_{w'} e_3$. Кроме того, $e_1, e_3 \in M$ и $e_2 \in L$. Следовательно, e_2 — допустимое ребро для M , и в $\delta(m)$ есть активное ребро $a = mv$, для которого $e_1 <_m a \leq_m e_2$. Это a совпадает с ребром p_2 в P . Кроме того, $a \neq e_2$ (иначе $e_3 \in P$ и $e_3 \in C \cap M$, вопреки условию на $p_1 = e$).

Мы утверждаем, что ребро $a = p_2$ — блокирующее для L . Это следует из $a <_m e_2 \in L$, если $r = 2$ (в этом случае концевая вершина v в P является непокрытой), а также если $r \geq 3$ и $p_3 \in L$ (так как $p_3 \in M$ и $a <_v p_3$). Наконец, пусть $p_3 \notin L$. Тогда p_3 принадлежит понижающему циклу $C' \in \mathcal{C}^-(M, L)$. Для ребра b в C' , инцидентного v и отличного от p_3 , выполняется $b \in L$ и $p_3 <_v b$ (учитывая $v \in J$ и то, что C' понижающий). Тогда из $a <_v p_3 <_v b$ и $a <_m e_2 \in L$ следует, что a блокирует L ; противоречие.

Ясно, что для матчинга $M' = \text{Repl}(M, C)$, полученного заменой вдоль ротации C в $\Gamma(M)$, выполняется $M \prec M'$, где $\prec$ — частичный порядок в решетке на $\mathcal{M}(G)$ (см. предложение 3). Как было установлено в [5] (при $G \simeq K_{n,n}$), начиная с $M^{\min}$ и используя ротации, можно исчерпать все $\mathcal{M}(G)$.

Определение 2. Назовем *трассой* последовательность стабильных матчингов $\tau = (M_0, M_1, \dots, M_N)$ в G такую, что $M_0 = M^{\min}$, $M_N = M^{\max}$, и каждое M_i ($1 \leq i \leq N$) получается из M_{i-1} заменой вдоль одной ротации C_i в $\Gamma(M_{i-1})$. Последовательность ротаций $C_1, \dots, C_N$ обозначим через $\mathcal{R}(\tau)$.

Предложение 6. Трассы покрывают всю решетку $(\mathcal{M}(G), \prec)$.

Доказательство. Достаточно показать, что если матчинги $M, L \in \mathcal{M}$ удовлетворяют $M \prec L$, то найдется ротация C для M такая, что для $M' = \text{Repl}(M, C)$ справедливо $M' \preceq L$. Чтобы построить такую ротацию, возьмем вершину $m \in \tilde{I}$, для которой инцидентные ребра $e = mw \in M$ и $f = mv \in L$ различны. Тогда $e <_m f$ (ввиду $M \prec L$), и ребро f допустимое для M (так как имеет ребро $b \in M$, инцидентное v , и для него выполняется $f <_v b$, ввиду $M \prec L$ и $w \in \tilde{J}$). Поэтому в $\delta(m)$ есть активное ребро $a = mw'$, и выполняется $a \leq_m f$. Заметим, что $w' \in \tilde{V}$ (иначе должно быть $a \notin L$; тогда $a \neq f$, и a блокирует L).

Следовательно, имеется ребро $e' = m'w' \in M$, а также ребро $f' \in L$, инцидентное m' . Мы утверждаем, что $f' \neq e'$. Действительно, в случае $f' = e'$, мы имели бы $a <_{w'} f'$ и $a <_m f$, и, следовательно, a блокировало бы L .

Теперь из $f' \neq e'$ получаем $e' <_{m'} f'$, аналогично $e <_m f$ в начальной ситуации. Продолжая построение, мы получаем “неограниченный” чередующийся путь P в $\Gamma(M)$ с последовательностью ребер $e = e_1 = m_1w_1$, $a = a_1 = m_2w_1$, $e' = e_2 = m_2w_2$, $a_2 = m_2w_3, \dots$, и для каждого i имеется ребро $f_i \in L$, инцидентное m_i и отличное от e_i . Тогда $e_i <_{m_i} a_i \leq_{m_i} f_i$, и P содержит искомую ротацию C .

Назовем *рангом* матчинга M сумму $\rho(M)$ порядковых номеров его ребер mw в множествах $\delta(m)$ (упорядоченных согласно $<_m$), где $m \in I$. Очевидно, $\rho(M) \leq |E|$, и если $M < L$, то $\rho(M) < \rho(L)$. Поэтому

$$\begin{aligned} \text{любая трасса } \tau \text{ в } G = (V, E, <) \text{ содержит не более } |E| \text{ матчингов;} \\ \text{следовательно, } |\mathcal{R}(\tau)| < |E|. \end{aligned} \quad (3.1)$$

Из предложений 4–6 мы также получаем

Следствие 1. (i) Граф $\Gamma(M^{\max}(G))$ является лесом (и ротации для $M^{\max}(G)$ отсутствуют).

(ii) Ребра минимального стабильного матчинга $M = M^{\min}(G)$, находящиеся в древесных компонентах активного графа $\Gamma(M)$, принадлежат всем стабильным матчингам в $\mathcal{M}(G)$.

(iii) G имеет единственный стабильный матчинг тогда и только тогда, когда $\Gamma(M^{\min})$ является лесом.

Следующее утверждение, установленное в [5], несмотря на относительную простоту доказательства, является наиболее существенным в области ротаций.

Предложение 7. Множество ротаций $\mathcal{R}(\tau)$ для всех трасс τ в $(\mathcal{M}(G), <)$ одно и то же.

Доказательство. Для $M \in \mathcal{M}(G)$ определим $\mathcal{T}(M)$ как множество отрезков трасс (“подтрасс”), начинающихся с M и оканчивающихся в $M^{\max}$. Для такой подтрассы $\tau \in \mathcal{T}(M)$ соответствующее множество ротаций обозначим через $\mathcal{R}(\tau)$ и скажем, что матчинг M *особый*, если в $\mathcal{T}(M)$ имеются подтрассы τ и τ' с $\mathcal{R}(\tau) \neq \mathcal{R}(\tau')$. Надо доказать, что $M^{\min}$ не является особым.

Предположим, что это не так, и рассмотрим особый матчинг M максимальной высоты, т.е. такой, что матчинг L не является особым, если $M < L$. Пусть $\mathcal{C}(M)$ – множество ротаций в $\Gamma(M)$; тогда для любой подтрассы $\tau = (M = M_0, M_1, M_2, \dots, M_k = M^{\max})$ матчинг M_1 получается из M заменой вдоль некоторой ротации $C \in \mathcal{C}(M)$. Ввиду максимальной M , найдутся ротации $C', C'' \in \mathcal{C}(M)$ такие, что матчинги $M' = \text{Repl}(M, C')$ и $M'' = \text{Repl}(M, C'')$ неособые, и множества ротаций $\mathcal{R}(\tau')$ и $\mathcal{R}(\tau'')$ различны для подтрасс $\tau', \tau'' \in \mathcal{T}(M)$ таких, что τ' проходит через M' , а τ'' проходит через M'' .

Теперь заметим, что поскольку ротации C' и C'' не пересекаются, то C' является ротацией в $\Gamma(M'')$, а C'' – ротацией в $\Gamma(M')$. Поэтому в $\mathcal{T}(M)$ есть подтрассы τ', τ'' такие, что τ' начинается с $M, M', \text{Repl}(M', C'')$, и τ'' начинается с $M, M'', \text{Repl}(M'', C')$, а после матчинга $\text{Repl}(M', C'') = \text{Repl}(M'', C')$ подтрассы τ' и τ'' совпадают. Но тогда $\mathcal{R}(\tau') = \mathcal{R}(\tau'')$; противоречие.

Это множество ротаций $\mathcal{R}(\tau)$, не зависящее от трасс $\tau \in \mathcal{T}(M^{\min})$, обозначим через $\mathcal{R}_G$. Ввиду (3.1), получаем

Следствие 2. Число ротаций в $\mathcal{R}_G$ не превосходит $|E|$.

Приведем еще одно полезное свойство:

$$\begin{aligned} \text{множества матчинговых ребер двух различных ротаций не пересекаются} \\ \text{(и поэтому общее число ребер в ротациях не превосходит } 2|E| \text{)}. \end{aligned} \quad (3.2)$$

Действительно, если $e = mw$ – матчинговое ребро в ротации C , где $m \in I$, то при трансформации в C новым матчинковым ребром в $\delta(m)$ становится активное ребро $a = mv$ в C ; при этом выполняется $e <_m a$. При дальнейшем движении вдоль трассы матчинговые ребра в $\delta(m)$ могут перемещаться только “вправо” (по аналогичным причинам), и возврата к ребру e не происходит.

В заключение этого раздела скажем о компонентах графа объединения стабильных матчингов Σ_G . Для этого обозначим через M^0 множество ребер минимального матчинга $M^{\min}$, принадлежащих древесным компонентам в $\Gamma(M^{\min})$, и обозначим через U_G подграф в G , являющийся объединением всех ротаций в $\mathcal{R}_G$. Как указано в следствии 1(ii), M^0 принадлежит всем матчингам в $\mathcal{M}(G)$ (поскольку активный граф $\Gamma(M^{\min})$ не имеет понижающих циклов, и учитывая предложение 5); отсюда получаем, что каждое ребро в M^0 образует компоненту в Σ_G . В то же время очевидно, что каждая ротация целиком лежит в Σ_G , и поэтому $U_G \subseteq \Sigma_G$. Более того, так как каждый стабильный матчинг получается из $M^{\min}$ в результате последовательности ротационных трансформаций, то справедливо

Следствие 3. Σ_G является объединением $U_G = \cup\{C \in \mathcal{R}_G\}$ и $M^{\min}$. Каждое ребро в M^0 и каждая компонента в U_G является компонентой в Σ_G .

Таким образом, Σ_G строится эффективно (за время $O(|V||E|)$, ввиду сказанного в следующем разделе). А также каждое ребро в M^0 и каждая компонента в U_G является компонентой в Σ_G . В связи с этим можно спросить, есть ли в Σ_G другие компоненты? Если да, то ими могут быть только ребра в $M^{\min}$, которые находятся в цикло-содержащих компонентах K графа $\Gamma(M^{\min})$, но при этом не входят в объединение ротаций U_K . Вопрос о существовании таких компонент открытый для нас.

4. ИДЕАЛЫ ПОСЕТА РОТАЦИЙ И СТАБИЛЬНЫЕ МАТЧИНГИ

Как было сказано выше, естественный частичный порядок $<$ на стабильных матчингах превращает множество $\mathcal{M}(G)$ в дистрибутивную решетку (см. предложение 3). Известно, что любая дистрибутивная решетка изоморфна решетке идеалов некоторого посета (и наоборот). Для решетки $(\mathcal{M}(G), <)$ Ирвинг и Лейтер [5] указали явную конструкцию. (Хотя [5] и последующая работа [7] имеют дело с полным двудольным графом $K_{n,n}$, полученные там результаты без большого труда распространяются на произвольный двудольный граф G .) А именно, опираясь на ключевой факт о постоянстве множества ротаций, ассоциированных с трассами (см. предложение 7), а также на определенный способ задания структуры посета на множестве ротаций $\mathcal{R}_G$, было показано соответствие между стабильными матчингами и идеалами в таком посете. Это представление для $\mathcal{M}(G)$ имеет важные применения, в частности, позволяет эффективно решать задачи линейной оптимизации для $\mathcal{M}(G)$ и установить труднорешаемость задачи вычисления числа $|\mathcal{M}(G)|$, о чем будет сказано в разд. 6 и 7.

Начнем с построения упомянутого посета на $\mathcal{R}_G$. Напомним, что для трассы $\tau = (M_0, M_1, \dots, M_N)$ мы обозначаем через $\mathcal{R}(\tau)$ последовательность ротаций $C_1, \dots, C_N$, где M_i получается из M_{i-1} заменой вдоль C_i .

Определение 3. Для ротаций $C, D \in \mathcal{R}_G$ скажем, что C предшествует D , и обозначим это как $C \leq D$, если для каждой трассы τ в $(\mathcal{M}(G), <)$ ротация C встречается в последовательности $\mathcal{R}(\tau)$ раньше, чем D . Иными словами, трансформация матчинга в D не может произойти ранее, чем трансформация в C .

Это бинарное отношение антисимметричное и транзитивное и дает частичный порядок на $\mathcal{R}_G$.

Чтобы объяснить связь с матчингами, рассмотрим $M \in \mathcal{M}(G)$. Пусть $\tau = (M_0, M_1, \dots, M_N)$ — трасса, содержащая M , скажем, $M = M_j$, и пусть $\mathcal{R}(\tau) = (C_1, \dots, C_N)$. Тогда M получается из $M_0 = M^{\min}$ последовательностью трансформаций относительно ротаций $C_1, \dots, C_i$. Из предложения 7 легко следует, что

$$\begin{aligned} & \text{(неупорядоченное) множество } \mathcal{R}_M := \{C_1, \dots, C_i\} \text{ не зависит от трассы } \tau, \\ & \text{проходящей через } M; \text{ аналогично для множества } \mathcal{R}_M^+ := \{C_{i+1}, \dots, C_N\}, \\ & \text{и для множества } \mathcal{R}_{M, M'} := \{C_{i+1}, \dots, C_j\}, \text{ где } M' = M_j \text{ для } j \geq i. \end{aligned} \quad (4.1)$$

В соответствии с определением $\leq$ отношение $C_j \leq C_k$ может выполняться, только если $j < k$. Более того, $\mathcal{R}_M$ образует замкнутое множество, или идеал, в посете $(\mathcal{R}_G, \leq)$ (т.е. из $C \leq D$ и $D \in \mathcal{R}_M$ следует $C \in \mathcal{R}_M$). Это дает инъективное отображение β множества $\mathcal{M}(G)$ в множество идеалов в $(\mathcal{R}_G, \leq)$. В действительности, β является биекцией, как показано в [5, Th. 4.1].

Предложение 8. Отображение $M \mapsto \mathcal{R}_M$ устанавливает изоморфизм между решеткой стабильных матчингов $(\mathcal{M}(G), \prec)$ и решеткой $\mathcal{I}(G)$ идеалов в $(\mathcal{R}_G, \leq)$.

Доказательство. Рассмотрим стабильные матчинги M и L , и пусть E' и E'' — множества ребер, принадлежащих циклам в $\mathcal{C}^+(M, L)$ и $\mathcal{C}^-(M, L)$ соответственно. Очевидно, подграфы, порожденные этими множествами, не пересекаются. Из доказательства предложения 6 следует, что E' является объединением множества R' ротаций, требуемых для получения $M \vee L$ из M (или, эквивалентно, L из $M \wedge L$), а E'' — объединением множества R'' ротаций, требуемых для получения $M \vee L$ из L (эквивалентно, M из $M \wedge L$). Следовательно, ротации R' и R'' “перестановочны”: чтобы получить матчинг $M \vee L$ из $M \wedge L$, можно применять сначала R' , а затем R'' , или наоборот. Это дает $\beta(M) \cap \beta(L) = \beta(M \wedge L)$ и $\beta(M) \cup \beta(L) = \beta(M \vee L)$ (т.е. β определяет гомоморфизм решеток). Теперь утверждение легко следует из того факта, что для ротаций $C, D \in \mathcal{R}_G$ не выполняется $C \leq D$ тогда и только тогда, когда найдется стабильный матчинг M такой, что $D \in \mathcal{R}_M \not\leq C$.

Как следствие, $\mathcal{M}(G)$ биективно множеству $\mathcal{A}(G)$ анти-цепей в $(\mathcal{R}_G, \leq)$. (Напомним, что анти-цепь в посете $(P, \leq)$ — это множество A попарно несравнимых элементов. Оно определяет идеал $\{p \in P : p \leq a \exists a \in A\}$, и определяется идеалом, будучи множеством его максимальных элементов.)

Мы завершаем этот раздел, объясняя, как эффективно проверять отношение $\leq$. Заметим, что C, D не связаны этим отношением, если они являются ротациями для одного и того же стабильного матчинга M (и, следовательно, трансформации в C и D могут производиться в произвольном порядке); в этом случае C и D лежат в разных компонентах в $\Gamma(M)$, и поэтому $V_C \cap V_D = \emptyset$. С другой стороны, если C и D имеют общую вершину v , то они сравнимы по $\leq$.

Действительно, для C и D матчинговые ребра, инцидентные v , различные, и порядок следования C и D для любой трассы один и тот же и определяется порядком $<_v$ для этих ребер (см. объяснение свойства (3.2)). Вообще, множество ротаций $\mathcal{R}^v$, содержащих фиксированную вершину v , упорядочено по $\leq$, скажем, $\mathcal{R}^v = (C_1 \leq C_2 \leq \dots \leq C_k)$, и этот порядок легко определяется, так как он согласуется с порядком в $<_v$, если $v \in I$, и обратным порядком, если $v \in J$.

Есть еще одна локальная причина того, что две ротации C, D должны быть сравнимы по $\leq$. А именно, предположим, что

$$\begin{aligned} &C \text{ содержит матчинговое ребро } e = mw \text{ и активное ребро } vw, \\ &\text{и } D \text{ содержит матчинговое ребро } e' = m'w', \text{ для которых: } b = m'w - \text{ребро в } G, \\ &\text{не принадлежащее никаким ротациям и являющимся первым после } e' \\ &\text{ребром в } \delta(m'), \text{ и при этом выполняется: } a <_w b <_w e. \end{aligned} \quad (4.2)$$

Тогда $C \leq D$. Действительно, ввиду $a <_w b <_w e$, после трансформации в C ребро b становится недопустимым. С другой стороны, если бы D предшествовало C в какой-то трассе, то именно b было бы активным ребром в $\delta(m')$ и вошло бы в ротацию.

Пример в разд. 3 именно таков: здесь после замены вдоль ротации C ребро $e = 2v$ становится недопустимым, ввиду $d' <_v e <_v a'$ и $d, d' \in M_2$ (при этом e — первое после d ребро в $\delta(2)$). В этом примере посет состоит из двух ротаций C, D при соотношении $C \leq D$, и есть ровно три идеала: $\emptyset, \{C\}$ и $\{C, D\}$, отвечающие стабильным матчингам $M_1 = M^{\min}$, M_2 и $M_3 = M^{\max}$ (а $\{D\}$ — не идеал).

Заметим, что для нахождения множества ротаций $\mathcal{R}_G$ достаточно определить $M^{\min}$ и затем построить произвольную трассу, делая трансформации на ротациях в текущих матчингах. (Это вы-

полняется естественным алгоритмом за время $O(|V||E|)$. В [16] предложен алгоритм формирования $\mathcal{R}_G$ за время $O(|V|^2)$.

В свою очередь в [7, Сес. 5] был установлен важный факт, что для эффективного описания порядка $\leq$, действительно достаточно рассмотреть пересечения ротаций и учесть (4.2). А именно, образуем ориентированный граф $H = (\mathcal{R}_G, \mathcal{E})$, ребрами которого являются пары ротаций (C, D) такие, что либо $V_C \cap V_D \neq \emptyset$ и $C \leq D$, либо выполняется (4.2) (этот граф легко строится за время $O(|V||E|)$; с некоторой изобретательностью время можно уменьшить до $O(|E|)$). Можно показать, что

граф H содержит ребро (C, D) тогда и только тогда, когда найдется стабильный матчинг M такой, что M допускает ротацию C , но не D , а после трансформации M относительно C полученный стабильный матчинг M' уже допускает D . (4.3)

Граф H ациклический, и если какая-либо вершина D в H достижима ориентированным путем из C , то выполняется $C \leq D$. Верно и обратное.

Предложение 9. Если $C \leq D$, то в H имеется ориентированный путь из C в D . Следовательно, отношение достижимости для вершин в H совпадает с отношением $\leq$.

Доказательство. Пусть $\mathcal{F}$ — множество идеалов в посете для H ; тогда $\mathcal{F} \supseteq \mathcal{I}(G)$. Надо доказать, что $\mathcal{F} \supseteq \mathcal{I}(G)$. Предположим, это не так, т.е. имеется $X \in \mathcal{F}$, не являющийся идеалом в $(\mathcal{R}_G, \leq)$. Пусть при этом X выбрано так, чтобы число элементов $|X|$ было максимальным. Также выберем идеал $R \in \mathcal{I}(G)$ такой, что $R \subset X$, и при этом $|R|$ максимально. Согласно предложению 8, $R = \mathcal{R}_M$ для некоторого стабильного матчинга M . Возьмем допустимую ротацию C для M . Ввиду максимальной R , элемент C должен быть вне X . Тогда $R' := R \cup \{C\}$ — идеал в $(\mathcal{R}_G, \leq)$ (соответствующий матчингу M' , получаемому из M заменой вдоль C). Также $X' := X \cup \{C\} \in \mathcal{F}$.

Ввиду максимальной X , имеем $X' \in \mathcal{I}(G)$. Поскольку $X' \supset R'$, и оба X', R' — идеалы в $(\mathcal{R}_G, \leq)$, множество $X' - R'$ содержит элемент (ротацию) D такой, что $R'' := R' \cup \{D\} \in \mathcal{I}(G)$. В то же время $X'' := R \cup \{D\}$ не является идеалом в $(\mathcal{R}_G, \leq)$; иначе, ввиду $R \subset X'' \subseteq X$, мы получили бы противоречие с максимальной R (если $X'' \neq X$) или с условием на X (если $X'' = X$).

Итак, мы пришли к ситуации, когда $R, R', R'' \in \mathcal{I}(G)$, $R' = R \cup \{C\}$, $R'' = R' \cup \{D\}$, но $R \cup \{D\} \notin \mathcal{I}(G)$. Тогда ротация D недопустима для матчинга M , но становится допустимой сразу после того, как в M производится замена вдоль ротации C . Это в точности означает, что граф H содержит ребро (C, D) ; ср. (4.3). Но такое ребро идет из $\mathcal{R}_G - X$ в X , вопреки тому, что X принадлежит $\mathcal{F}$.

Можно показать, что ребра (C, D) графа H как в (4.3) определяют отношения непосредственного предшествования в посете $(\mathcal{R}_G, \leq)$, а именно, выполнение $C \leq D$ и отсутствие ротации C' между C и D (т.е. $C \leq C' \leq D$).

Замечание 1. Пусть нам даны стабильные матчинги M и M' в двудольном G такие, что $M \prec M'$, и пусть $[M, M']$ обозначает множество (“интервал”) стабильных матчингов L , для которых $M \preceq L \preceq M'$. (Как специальные случаи можно рассматривать $M = M^{\min}$ или $M' = M^{\max}$.) Нас интересует возможность “очистки” графа G путем удаления части его ребер с тем, чтобы для полученного графа G' множество стабильных матчингов в точности совпадало с интервалом $[M, M']$ для G . Чтобы попытаться ответить на данный вопрос, возьмем произвольную трассу τ , проходящую через оба M, M' , и рассмотрим множество ротаций, используемых на участке трассы от M до M' , т.е. $\mathcal{R}_{M, M'}$, см. (4.1). (Трасса τ и множество $\mathcal{R}_{M, M'}$ строятся эффективно; ср. доказательство предложения 6.) Тогда интервал $[M, M']$ состоит ровно из тех $L \in \mathcal{M}(G)$, которые получаются из M применением последовательности ротаций из множества $\mathcal{R}_{M, M'}$. Поэтому искомым граф G' должен включать M и объединение всех ротаций из $\mathcal{R}_{M, M'}$. Однако он также должен учитывать структуру графа H (см. (4.3)), чтобы избежать появления “лишних” матчингов, использующих ротации из $\mathcal{R}_{M, M'}$. Следовательно, нужно добавить ребра как в (4.2) (вида $m'w$), чтобы гарантировать сохранение указанного там отношения $C \leq D$ для соответствующих ротаций

C, D в $\mathcal{R}_{M, M'}$. Гипотеза: подграф G' существует и получается из G удалением всех остальных ребер.

5. ОПТИМАЛЬНЫЕ СТАБИЛЬНЫЕ МАТЧИНГИ

В этом разделе мы рассматриваем ситуацию, когда двудольный граф $G = (V, E, <)$ дополнительно снабжен вещественной функцией *весов* (или *стоимостей*) ребер $c : E \rightarrow \mathbb{R}$. Нас интересует задача о стабильном матчинге минимального веса:

найти стабильный матчинг $M \in \mathcal{M}(G)$, минимизирующий общий вес $(M) := \sum_{e \in M} c(e)$. (5.1)

(Ввиду того, что все стабильные матчинги в G имеют одинаковый размер, добавление константы к функции c фактически не меняет задачи; поэтому можно считать функцию c неотрицательной. При замене c на $-c$ получаем задачу максимизации веса $c(M)$.)

В важном частном случае этой задачи, известном как задача об *оптимальном стабильном матчинге* (см. [6]), вес $c(e)$ ребра $e = mw$ определяется как

$$c(e) := r_I(e) + r_J(e), \quad (5.2)$$

где $r_I(e)$ — порядковый номер ребра e в упорядоченном (согласно $<_m$) списке $\delta(m)$, и, аналогично, $r_J(e)$ — номер в списке $\delta(w)$. (Такая постановка широко обсуждалась в литературе, в частности, в [11, 17]. Здесь интересуются матчингом, который “наиболее благоприятен по суммарным (или средним) предпочтениям для всех персон” (в то время как “алгоритм отложенных принятий” (АОП) дает наилучшее решение только для одного множества из I и J).) В [7] задача с весовой функцией как в (5.2) также называется задачей об *эгалитарном стабильном матчинге*.

Задача (5.1) может быть сформулирована как задача линейного программирования с $0, \pm 1$ -матрицей размера $(|V| + |E|) \times |E|$ (о чем будет сказано в следующем разделе), поэтому она может быть решена универсальным сильно полиномиальным алгоритмом (усиленной версией метода эллипсоидов в [18]). Однако изложенное в предыдущем разделе представление $\mathcal{M}(G)$ в виде множества идеалов в посете ротаций $(\mathcal{R}_G, <)$ дает возможность решать (5.1) намного более экономным и наглядным методом. Это было предложено Ирвингом—Лейтером—Гасфилдом в [7], и ниже мы описываем этот метод (применительно к рассматриваемому нами произвольному двудольному графу G).

Для заданной функции весов c и произвольной ротации $R = e_1 a_1 e_2 a_2 \cdots e_k a_k$ в $\mathcal{R}_G$, где ребра e_i матчинговые, а ребра a_i активные, определим вес R как

$$c^R := \sum (c(a_i) - c(e_i) : i = 1, \dots, k). \quad (5.3)$$

При трансформации относительно ротации R к весу текущего матчинга прибавляются веса активных ребер и из него вычитаются веса матчинговых ребер в R . Поэтому вес любого стабильного матчинга M выражается как

$$c(M) = c(M^{\min}) + \sum (c^R : R \in \mathcal{R}_M),$$

где $\mathcal{R}_M$ — идеал в $(\mathcal{R}_G, <)$, соответствующий M (т.е. множество ротаций, возникающих на участке (любой) трассы от $M^{\min}$ до M).

Таким образом, (5.1) эквивалентно задаче нахождения идеала минимального веса. В общем случае задача такого рода выглядит следующим образом:

для ориентированного графа $Q = (V_Q, E_Q)$ и функции $\zeta : V_Q \rightarrow \mathbb{R}$ найти замкнутое множество $X \subseteq V_Q$ минимального веса $\zeta(X) := \sum (\zeta(v) : v \in X)$. (5.4)

(Напомним, что множество X считается *замкнутым*, если нет ребер, идущих из $V_Q - X$ в X . В частности, замкнутыми множествами являются $\emptyset$ и V_Q . Без ограничения общности можно считать граф Q ациклическим, так как любой ориентированный цикл не может разделяться замкнутым множеством, и его можно стянуть в одну вершину суммарного веса.)

Решение задачи (5.4), предложенное Пикаром [8], состоит в ее сведении к задаче о минимальном разрезе в ориентированном графе $\hat{Q} = (\hat{V}, \hat{E})$ с пропускными способностями ребер $h(e)$, $e \in \hat{E}$, определяемыми следующим образом.

Положим $V^+ := \{v \in V_Q : \zeta(v) > 0\}$ и $V^- := \{v \in V_Q : \zeta(v) < 0\}$. Граф $\hat{Q}$ получается из Q добавлением двух вершин: “источника” s и “стока” t , а также множества ребер E^+ из s в v для всех $v \in V^+$ и множества ребер E^- из u в t для всех $u \in V^-$. Пропускные способности ребер $e \in \hat{E}$ задаются так:

$$h(e) := \begin{cases} \zeta(v), & \text{если } e = sv \in E^+, \\ |\zeta(u)|, & \text{если } e = ut \in E^-, \\ \infty, & \text{если } e \in E_Q. \end{cases}$$

Для подмножества $S \subseteq \hat{V}$ такого, что $s \in S \not\equiv t$, обозначим через $\delta(S)$ множество ребер в $\hat{Q}$, идущих из S в $\hat{V} - S$, называемое $s-t$ разрезом; величина $h(\delta(S)) := \sum (h(e) : e \in \delta(S))$ считается пропускной способностью этого разреза.

Лемма 2 (см. [8]). *Подмножество $X \subseteq V_Q$ является замкнутым множеством минимального веса в (Q, ζ) тогда и только тогда, когда $\delta((V_Q - X) \cup \{s\})$ является $s-t$ разрезом минимальной пропускной способности в $(\hat{Q}, h)$.*

Доказательство. Заметим, что $X \subseteq V_Q$ — замкнутое множество тогда и только тогда, когда разрез $E' = \delta((V_Q - X) \cup \{s\})$ не содержит ребер бесконечной пропускной способности; эквивалентно, E' содержится в $E^+ \cup E^-$. Для такого разреза, состоящего из ребер вида sv для $v \in X$ и ребер вида ut для $u \in V_Q - X$, пропускная способность выглядит так:

$$\begin{aligned} h(E') &= \zeta(X \cap V^+) + \sum (|\zeta(u)| : u \in (V_Q - X) \cap V^-) = \\ &= \zeta(X \cap V^+) + \zeta(X \cap V^-) - \zeta(V^-) = \zeta(X) - \zeta(V^-). \end{aligned}$$

Следовательно, вес замкнутого множества отличается от пропускной способности соответствующего разреза на постоянную величину $-\zeta(V^-)$, откуда получаем требуемое утверждение.

Таким образом, задача (5.1) сводится к задаче о минимальном двухполюсном разрезе в сети с $N = O(|E|)$ вершинами и $A = O(|E|)$ ребрами. (Учитывая то, что вместо всего посета $(\mathcal{R}_G, <)$ достаточно рассматривать его порождающий граф H , имеющий $O(|E|)$ ребер, см. предложение 9. Применяя быстрые алгоритмы для задачи о максимальном потоке и минимальном разрезе (см., например, обзор в [12, Сек. 10.8]), можно получить временную оценку $O(NA \log N) \simeq O(|V|^4 \log |V|)$. В [7] приведена имплементация, решающая (5.1) за время $O(|V|^4)$.)

Замечание 2. В [19] доказывалась труднорешаемость некоторых вариантов задачи о замкнутых множествах. Используя один из них, а также тот факт, что любой транзитивно замкнутый граф реализуется как посет ротаций (о чем будет сказано в разд. 7), можно показать NP-трудность следующего усиления задачи (5.1): для двудольного $G = (V, E, <)$, функций $c, g : E \rightarrow \mathbb{R}_+$ и числа $K \in \mathbb{R}_+$ найти стабильный матчинг $M \in \mathcal{M}(G)$, минимизирующий $c(M)$ при условии $g(M) \geq K$. (Здесь $g(M)$ можно интерпретировать как прибыль, а $c(M)$ как затраты при организации союзов (или контрактов) в M .)

6. ПОЛИЭДРАЛЬНЫЕ АСПЕКТЫ И МЕДИАННЫЕ СТАБИЛЬНЫЕ МАТЧИНГИ

Как мы упоминали ранее, для двудольного графа $G = (V = I \sqcup J, E, <)$ имеется полиэдральная характеристика многогранника стабильных матчингов $\mathcal{P}_{\text{st}}(G)$, задаваемая линейным от $|V|, |E|$ числом неравенств. Здесь $\mathcal{P}_{\text{st}}(G)$ — выпуклая оболочка множества характеристических векторов χ^M стабильных матчингов M в пространстве $\mathbb{R}^E$. Первоначальное описание $\mathcal{P}_{\text{st}}(G)$ (в случае $G \simeq K_{n,n}$) было дано в работе Ванде Вейта [9] (и повторено в работе [20]). Ниже мы при-

водим описание (несколько отличающееся по виду, но близкое к тому, что в [9]) и доказательство, основываясь на изложении в [12], Сек. 18.5g.

Для ребер $e, f \in E$ будем писать $f <_v e$, если они имеют общую вершину v , и выполняется $f <_v e$ (т.е. f предпочтительнее, чем e). Для $e \in E$ обозначим $\gamma(e)$ множество ребер f таких, что $f \preceq e$ (в частности, $e \in \gamma(e)$).

Напомним, что многогранник матчингов в двудольном случае описывается системой линейных неравенств

$$x(e) \geq 0, \quad e \in E; \quad (6.1)$$

$$x(\delta(v)) \leq 1, \quad v \in V. \quad (6.2)$$

В случае стабильных матчингов добавляется еще один тип неравенств:

$$x(\gamma(e)) \geq 1, \quad e \in E. \quad (6.3)$$

Предложение 10. Система (6.1)–(6.3) описывает в точности множество векторов $x \in \mathbb{R}^E$, принадлежащих $\mathcal{P}_{\text{st}}(G)$.

Доказательство. Легко видеть, что для любого стабильного матчинга M вектор $x = \chi^M$ удовлетворяет (6.1)–(6.3). Поэтому достаточно доказать, что если x — вершина многогранника $\mathcal{P}'$, определяемого (6.1)–(6.3), то вектор x целочисленный.

Положим $E^+ := \{e \in E : x(e) > 0\}$ и обозначим V^+ множество вершин в G , покрытых E^+ . Для $v \in V^+$ обозначим через e_v наилучшее ребро в $\delta(v) \cap E^+$ относительно порядка $<_v$. Справедливо следующее свойство:

$$\begin{aligned} \text{для } v \in V^+ \text{ и } e_v = \{v, u\} \text{ ребро } e_v \text{ является наихудшим в } (\delta(u) \cap E^+, <_u); \\ \text{кроме того, выполняется } x(\delta(u)) = 1. \end{aligned} \quad (6.4)$$

Действительно, полагая $e := e_v$, имеем

$$\begin{aligned} 1 \leq \sum (x(f) : f \preceq e) \text{ (ввиду (6.3))} &= \sum (x(f) : f \leq_u e) \text{ (ввиду определения } e_v) = \\ &= x(\delta(u)) - \sum (x(f) : f >_u e) \leq 1 - \sum (x(f) : f >_u e) \text{ (применяя (6.2) к } u). \end{aligned}$$

Здесь все неравенства должны обращаться в равенства. Это дает $\sum (x(f) : f >_u e) = 0$ и $x(\delta(u)) = 1$, откуда следует (6.4).

Образует множества $M := \{e \in E^+ : e = e_v \text{ для } v \in I\}$ и $L := \{e \in E^+ : e = e_v \text{ для } v \in J\}$. Для любой вершины $v \in I \cap V^+$ наилучшее ребро в $\delta(v) \cap E^+$ принадлежит M , а наихудшее ребро, и только оно, принадлежит L (ввиду (6.4)); для вершин в $J \cap V^+$ поведение обратное. Отсюда можно заключить, что оба M и L являются матчингами. Каждое ребро $e \in M \cap L$ образует компоненту в подграфе (V^+, E^+) ; это дает $x(e) = 1$ (ввиду (6.2), (6.3)). В частности, x целочисленный, если $M = L$.

Пусть теперь $M \neq L$. Очевидно, для любого ребра e в $M' := M - L$ или в $L' := L - M$ выполняется $0 < x(e) < 1$. Поэтому можно выбрать достаточно малое число $\varepsilon > 0$ так, чтобы вектора $x' := x + \varepsilon \chi^{M'} - \varepsilon \chi^{L'}$ удовлетворяли (6.1) и (6.2). Проверим, что оба x' и x'' удовлетворяют также и (6.3).

Для этого рассмотрим ребро $e = mw$ с $x(e) < 1$ и предположим, что $x'(\gamma(e)) < x(\gamma(e))$. Это возможно только если $a \leq_w e <_w b$, где $\{a\} = L \cap \delta(w)$ и $\{b\} = M \cap \delta(w)$. При этом должно выполняться либо (i) $m \notin V^+$, либо (ii) $m \in V^+$, и e более предпочтительное, чем ребро в $M \cap \delta(m)$ (иначе такое ребро давало бы положительный вклад в $x'(\gamma(e))$, откуда $x'(\gamma(e)) = x(\gamma(e))$). Но в этих случаях из равенства $x(\delta(w)) = 1$ (ввиду (6.4)) и неравенства $x(b) > 0$ следует

$$x(\gamma(e)) = \sum (x(f) : f \leq_w e) \leq x(\delta(w)) - x(b) < 1,$$

что невозможно. Аналогично, (6.3) верно для x'' .

Таким образом, $x', x'' \in \mathcal{P}'$. Но $x' \neq x''$ и $(x' + x'')/2 = x$. Это противоречит тому, что x — вершина в $\mathcal{P}'$.

Имеется еще одно полезное свойство, показанное в [21]; мы приводим его в несколько иной, но эквивалентной, форме.

Лемма 3. Пусть $x \in \mathcal{P}_{\text{st}}(G)$, и пусть $e = tw$ — ребро в G , для которого $x(e) > 0$. Тогда $x(\delta(m)) = x(\delta(w)) = x(\gamma(e)) = 1$.

Доказательство. Представим x как $\alpha_1 \chi^{M_1} + \dots + \alpha_k \chi^{M_k}$, где M_i — стабильный матчинг, $\alpha_i > 0$, и $\alpha_1 + \dots + \alpha_k = 1$. Обозначим $x_i := \chi^{M_i}$. Ввиду $x(e) > 0$, ребро e принадлежит некоторому M_i . Тогда $x_i(\delta(m)) = x_i(\delta(w)) = x_i(\gamma(e)) = 1$. Аналогичные равенства имеют место и для матчинга M_j , не содержащего e . Действительно, так как M_i и M_j покрывают одно и то же множество вершин (ввиду предложения 2), в M_j есть ребро e' , инцидентное m , и ребро e'' , инцидентное w . Более того, рассматривая пару M_i, M_j и применяя (2.1), получаем либо $e' <_m e <_w e''$, либо $e' >_m e >_w e''$. В обоих случаях в $\gamma(e)$ попадает ровно одно ребро из e', e'' ; поэтому $x_j(\gamma(e)) = 1$. Теперь утверждение следует из $\alpha_1 + \dots + \alpha_k = 1$.

Эта лемма помогает получить интересный результат Тео и Сетурамана.

Предложение 11 (см. [10]). Пусть $M_1, \dots, M_\ell$ — стабильные матчинги в G . Для каждого $m \in \tilde{I}$ обозначим E_m список (с возможными повторениями) ребер в $\delta(m)$, принадлежащих этим матчингам и упорядоченных согласно $<_m$, и пусть $e_m(i)$ обозначает i -й элемент в E_m , $i = 1, \dots, \ell$. Аналогично, для каждого $w \in \tilde{J}$ пусть $e_w(j)$ обозначает j -й элемент в списке E_w ребер в $\delta(w)$, принадлежащих данным матчингам и упорядоченных согласно $<_w$, $j = 1, \dots, \ell$. Тогда для любого $k \in \{1, \dots, \ell\}$ множество ребер $A(k) := \{e_m(k) : m \in \tilde{I}\}$ совпадает с множеством $B(\ell - k + 1) := \{e_w(\ell - k + 1) : w \in \tilde{J}\}$ и образует стабильный матчинг в G .

Доказательство. Положим $x_i := \frac{1}{\ell} \chi^{M_i}$. Тогда $x = x_1 + \dots + x_\ell$ лежит в многограннике $\mathcal{P}_{\text{st}}(G)$ (является т.н. “дробным стабильным матчингом”).

Вначале предположим для простоты, что все ребра в $M_1, \dots, M_\ell$ различны. Тогда для $m \in \tilde{I}$ список E_m состоит из ℓ различных ребер, и ребро $e_m(k)$ является k -м элементом в E_m . Из Леммы 3, примененной к x и $e = e_m(k)$ следует, что множество $\gamma(e)$ состоит из ℓ элементов. Тогда e является в точности $(\ell - k + 1)$ -м элементом в списке E_w (по предположению состоящем из ℓ различных ребер), и тем самым $e_m(k) = e_w(\ell - k + 1)$.

Таким образом, $A(k) = B(\ell - k + 1)$, откуда следует, что $A(k)$ является матчингом в G . Чтобы показать стабильность $A(k)$ рассмотрим произвольное ребро $e = tw \notin A(k)$ в G . Надо проверить, что e не блокирует $A(k)$, или, эквивалентно, что $A(k) \cap \gamma(e) \neq \emptyset$.

По крайней мере один из концов e , скажем, m , принадлежит $\tilde{V}$ (и покрывается $A(k)$); иначе для x и e мы имели бы $x(\gamma(e)) = 0$ вопреки (6.3). Если $w \notin \tilde{V}$, то $x(\delta(w)) = 0$, и из неравенства $x(\gamma(e)) \geq 1$ следует, что для всех ребер $f \in E_m$ (включая $f = e_m(k)$) выполняется $f <_m e$. Это дает требуемое $A(k) \cap \gamma(e) \neq \emptyset$.

Пусть теперь $w \in \tilde{V}$. Ввиду $x(\gamma(e)) \geq 1$, число ребер f в $E_m \cup E_w$, для которых $f \preceq e$, не меньше ℓ . Тогда должно выполняться по крайней мере одно из $e_m(k) \leq_m e$ и $e_w(\ell - k + 1) \leq_w e$, откуда опять следует $A(k) \cap \gamma(e) \neq \emptyset$.

В том случае, когда в $M_1, \dots, M_\ell$ есть общие ребра, можно рассмотреть мультиграф, получаемый из G заменой каждого ребра $e = tw$ с $x(e) > 0$ на $\ell x(e) =: r$ параллельных ребер $e^1, \dots, e^r$. При этом продолжение порядка $<_m$ на эти ребра назначается противоположным продолжению порядка $<_w$, скажем, $e^1 <_m \dots <_m e^r$ и $e^1 >_w \dots >_w e^r$. Требуемое утверждение для этого общего случая получается повторением (с незначительными уточнениями) рассуждений для случая непересекающихся матчингов выше.

Следствие 4. Если ℓ нечетно, то для $k := (\ell + 1)/2$ множество, состоящее из k -х по порядку элементов в списках ребер E_v , инцидентных v и принадлежащих $M_1, \dots, M_\ell$, для всех вершин $v \in \tilde{V}$, является стабильным матчингом.

Такой матчинг называют *медианным* для $M_1, \dots, M_\ell$. В [10] был поставлен вопрос о возможности эффективного нахождения медианного стабильного матчинга среди всех стабильных матчингов в G (или “почти медианного”, когда число стабильных матчингов четное). Нам это представляется маловероятным в свете того, что задача вычисления числа $|\mathcal{M}(G)|$ является труднорешаемой, о чем будет сказано ниже.

7. О ПОДСЧЕТЕ ЧИСЛА СТАБИЛЬНЫХ МАТЧИНГОВ

Кнут [11] указал примеры, когда число стабильных матчингов двудольного графа экспоненциально велико по сравнению с размером графа, и задал вопрос о сложности точного вычисления этого числа. Ответ был дан Ирвингом и Лейтером [5], которые показали труднорешаемость такой задачи (рассматривая графы вида $K_{n,n}$). Ниже мы кратко поясним идею их доказательства.

Напомним некоторые понятия (отсылая за точными определениями к [22] или [23, разд. 7.3]). Не вдаваясь в полную логическую строгость, можно понимать, что описание той или иной перечислительной задачи (enumeration problem) $\mathcal{P}$ состоит из бесконечного семейства $\mathcal{S}$ конечных множеств, и для каждого $S \in \mathcal{S}$ имеется семейство $\mathcal{F}(S)$ подмножеств в S (“объектов”). В задаче $\mathcal{P}$ требуется для заданного $S \in \mathcal{S}$ определить число $|\mathcal{F}(S)|$. Говорят, что $\mathcal{P}$ является *#P-задачей* (или *KP-задачей*, в терминологии некоторых авторов), если распознавание объекта проводится за полиномиальное время, т.е. имеется алгоритм, который для любых $S \in \mathcal{S}$ и $X \subseteq S$ за время, полиномиальное от $|S|$, определяет, принадлежит или нет данное множество X семейству $\mathcal{F}(S)$. Говорят, что *#P-задача* $\mathcal{P} = \{\mathcal{F}(S), S \in \mathcal{S}\}$ является *#P-полной* (или универсальной в классе *#P*), если любая другая *#P-задача* $\mathcal{P}' = \{\mathcal{F}'(S'), S' \in \mathcal{S}'\}$ сводится к ней за полиномиальное время (т.е. есть отображение $\omega: \mathcal{S}' \rightarrow \mathcal{S}$ такое, что для каждого $S' \in \mathcal{S}'$ число $|\mathcal{F}'(S')|$ определяется из $|\mathcal{F}(\omega(S'))|$ за время, полиномиальное от $|S'|$).

В рассматриваемой нами задаче роль семейства $\mathcal{S}$ играет совокупность реберных множеств E двудольных графов $G = (V, E, <)$, а роль семейства $\mathcal{F}(S)$, $S \in \mathcal{S}$, — соответствующее множество стабильных матчингов в G . Задача определения $|\mathcal{M}(G)|$ — это действительно *#P-задача*, так как для любого подмножества $M \subseteq E$ можно определить, является ли M стабильным матчингом, за время $O(|E|)$.

Заметим, что *#P-аналог* любой *NP-полной* задачи является “труднорешаемым” (поскольку в последней задаче требуется “всего лишь” определить, является ли непустым соответствующее семейство объектов $\mathcal{F}(S)$ (например, содержит ли заданный граф хотя бы один гамильтонов цикл), а в первой нужно найти количество объектов). Однако есть *P-задачи*, чьи перечислительные аналоги являются *#P-полными*. Именно таковой и является задача определения $|\mathcal{M}(G)|$. Это вытекает из следующих двух результатов.

Предложение 12. Задача определения числа анти-цепей в конечном посете является *#P-полной*.

Предложение 13. Пусть $(P, <')$ — посет на n элементах. Существует и может быть построен за время полиномиальное от n двудольный граф $G = (V, E, <)$ такой, что его ротационный посет $(\mathcal{R}_G, <)$ изоморфен $(P, <')$. Следовательно (ввиду предложения 8), число $|\mathcal{M}(G)|$ стабильных матчингов в G равно числу анти-цепей (или числу идеалов) в посете $(P, <')$.

Предложение 12 было установлено Прованом и Боллом [24]. Предложение 13 было доказано в [5, Сек. 5] путем явной конструкции требуемого графа G для заданного посета $(P, <')$.

Автор благодарит рецензента за анализ начальной версии статьи и сделанные замечания.

СПИСОК ЛИТЕРАТУРЫ

1. Gale D., Shapley L.S. College admissions and the stability of marriage // Amer. Math. Monthly. 1962. V. 69. № 1. P. 9–15.
2. Irving R.W. An efficient algorithm for the “stable roommates” problem // J. Algorithms. 1985. V. 6. P. 577–595.

3. *Tan J.* A necessary and sufficient condition for the existence of a complete stable matching // *J. Algorithms*. 1991. V. 12. P. 154–178.
4. *Baiou M., Balinski M.* Erratum: the stable allocation (or ordinal transportation) problem // *Math. Oper. Res.* 2002. V. 27. № 4. P. 662–680.
5. *Irving R.W., Leather P.* The complexity of counting stable marriages // *SIAM J. Comput.* 1986. V. 15. P. 655–667.
6. *McVitie D., Wilson L.B.* The stable marriage problem // *Commun. ACM*. 1971. V. 14. P. 486–492.
7. *Irving R.W., Leather P., Gusfield D.* An efficient algorithm for the optimal stable marriage problem // *J. ACM*. 1987. V. 34. P. 532–543.
8. *Picard J.* Maximum closure of a graph and applications to combinatorial problems // *Manage. Sci.* 1976. V. 22. P. 1268–1272.
9. *Vande Vate J.H.* Linear programming brings marital bliss // *Oper. Res. Lett.* 1989. V. 8. P. 147–153.
10. *Teo C.P., Sethuraman J.* The geometry of fractional stable matchings and its applications // *Math. Oper. Res.* 1998. V. 23. № 4. P. 874–891.
11. *Knuth D.E.* *Mariages stables*. Les Presses de l'Université de Montreal. Montreal, 1976.
12. *Schrijver A.* *Combinatorial Optimization*. Vol. A. Springer, 2003.
13. *McVitie D., Wilson L.B.* Stable marriage assignments for unequal sets // *BIT*. 1970. V. 10. P. 295–309.
14. *Dean B.C., Munshi S.* Faster algorithms for stable allocation problems // *Algorithmica*. 2010. V. 58. № 1. P. 59–81.
15. *Mourtos I., Samaris M.* Stable allocations and partially ordered sets // *Discr. Optimiz.* 2022. V. 46. P. 100731.
16. *Gusfield D.* Three fast algorithms for four problems in stable marriage // *SIAM J. Comput.* 1987. V. 16. P. 111–128.
17. *Polya G., Tarjan R.E., Woods D.R.* *Notes on Introductory Combinatorics*. Birkbauser Verlag, Boston, Mass., 1983.
18. *Tardos É.* A strongly polynomial algorithm to solve combinatorial linear problems // *Oper. Research*. 1986. V. 34. P. 250–256.
19. *Карзанов А.В.* О замкнутых множествах ориентированного графа // *Ж. вычисл. матем. и матем. физ.* 1984. Т. 24. С. 1903–1906.
20. *Rothblum U.G.* Characterization of stable matchings as extreme points of a polytope // *Math. Programming*. 1992. V. 54. P. 57–67.
21. *Roth A.E., Rothblum U.G., Vande Vate J.H.* Stable matching, optimal assignments and linear programming // *Math. Oper. Res.* 1993. V. 18. P. 808–828.
22. *Valiant L.G.* The complexity of computing the permanent // *Theoret. Comp. Sci.* 1979. V. 8. P. 189–201.
23. *Гэри М., Джонсон Д.* *Вычислительные машины и труднорешаемые задачи*. М.: Мир, 1982.
24. *Provan J.S., Ball M.O.* The complexity of counting cuts and of computing the probability that a graph is connected // *SIAM J. Comput.* 1983. V. 12. P. 777–788.